
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
MASTER THESIS



ALCIE: Active Learning for Continual Image Captioning Enhancement

submitted by
AKASH KUMAR
Saarbrücken
August 1, 2025

Advisor:

Aliko Anagnostopoulou
German Research Center for Artificial Intelligence
Marie-Curie-Str. 1
Oldenburg, Germany

Reviewers

Prof. Dr. Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Saarbrücken Germany

Prof. Dr. Antonio Krüger
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Saarbrücken Germany

Submitted

August 1, 2025

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Declarations

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, _____
(Datum/Date)

(Unterschrift/Signature)

Erklärung

Ich erkläre hiermit, dass die vorliegende Arbeit mit der elektronischen Version übereinstimmt.

Statement

I hereby confirm the congruence of the contents of the printed data and the electronic version of the thesis.

Saarbrücken, _____
(Datum / Date)

(Unterschrift / Signature)

AI Tools Declaration

I declare that I used the following AI tools during the thesis process:

| Tool | Description of use |
|--------------------|--|
| Claude (Anthropic) | Grammar checking and LaTeX formatting assistance. All research content, analysis, and conclusions are original work by the author. |
| GitHub Copilot | Code completion suggestions during implementation of the AL-CIE framework. All algorithms, experimental design, and technical contributions are original work by the author. |
| Mermaid AI | Assistance in creating flow diagrams and visualizations for thesis figures. All conceptual design and content of diagrams are original work by the author. |

I confirm that:

- All AI-generated content has been clearly identified and properly attributed
- The core research contributions, experimental design, and analysis are my original work
- AI tools were used as assistants only and did not replace my own critical thinking and research
- All AI assistance has been used in compliance with Saarland University guidelines and academic integrity policies

Saarbrücken, _____
(Datum/Date)

(Unterschrift / Signature)

Acknowledgements

I would like to express my heartfelt gratitude to my advisor, Alikí Anagnostopoulou (M.Sc.), for her dedicated guidance and continuous support throughout every stage of my thesis work. Her insightful advice, patience, and encouragement were instrumental from the early obstacles to the successful completion of this thesis. I am truly grateful for her mentorship and commitment.

My sincere thanks also go to Prof. Dr.-Ing. Daniel Sonntag and Prof. Dr. Antonio Krüger for granting me the opportunity to pursue this thesis project. Their support and the collaborative environment at the Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) greatly contributed to the progress and completion of my research.

I am deeply appreciative of the participants who took part in my user study, whose valuable time and honest feedback were essential to this work. I would also like to thank Sara-Jane Bittner for conducting the pilot study and for her thoughtful guidance in the design and implementation of the user study, which significantly strengthened the quality of my research. I am grateful to Hannes Kath for his valuable advice during our meeting.

Finally, I wish to express my deepest appreciation to my wife, as well as my beloved parents and supportive friends. Their unwavering encouragement, understanding, and unconditional support have been the foundation of my academic journey. Their presence has turned challenges into growth and made every achievement possible.

Saarbrücken, 01 August 2025, Akash Kumar

Abstract

Continual learning in vision-language models faces the fundamental challenge of catastrophic forgetting, where adaptation to new domains results in loss of previously learned knowledge. While current episodic memory approaches rely on random sampling, a critical question emerges: can strategic sample selection significantly improve continual learning performance beyond computationally efficient random approaches?

This thesis introduces Active Learning for Continual Image Captioning Enhancement (ALCIE), systematically comparing uncertainty-based sampling, diversity-based sampling, hybrid strategies, and random sampling baselines across Bootstrapping Language-Image Pre-training (BLIP-2) and One For All (OFA) architectures. Experiments on the Fashion Captioning Dataset (FACAD) dataset simulate continual learning across six sequential fashion domains, with evaluation using both standard captioning metrics and human assessment to ensure comprehensive analysis.

Our investigation reveals counterintuitive findings that challenge fundamental assumptions about memory strategy optimization. Although diversity sampling initially achieves superior early transition stability (56% vs 30-36% retention), all strategies exhibit universal early transition vulnerability with severe performance drops during initial domain shifts. Most significantly, we discover a systematic lexical-semantic forgetting disconnect that reframes understanding of catastrophic forgetting: while lexical generation suffers severe degradation (0-35% retention), semantic understanding remains remarkably stable (86-93% retention) across all approaches. This pattern holds consistently regardless of memory strategy sophistication, suggesting that catastrophic forgetting primarily affects surface-level capabilities rather than fundamental knowledge preservation.

Human evaluation validates these technical findings, with 15 participants showing practical equivalence among strategies despite substantial metric differences. The lexical-semantic disconnect explains why computationally efficient random sampling achieves competitive performance: sophisticated memory strategies primarily optimize surface-level generation while essential semantic capabilities remain stable regardless of algorithmic complexity. These paradigm-shifting findings establish that computational efficiency should be prioritized over sophistication, fundamentally redirecting continual learning research toward addressing forgetting mechanisms rather than optimizing memory selection complexity.

Keywords: Continual Learning, Image Captioning, Episodic Memory, Active Learning, Catastrophic Forgetting, Vision-Language Models, Human Evaluation

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation and Problem Statement | 1 |
| 1.1.1 | Current Limitations in Image Captioning Systems | 1 |
| 1.1.2 | Continual Learning Paradigm | 2 |
| 1.1.3 | The Challenge of Catastrophic Forgetting | 2 |
| 1.1.4 | Memory Management and Strategic Selection Challenges | 3 |
| 1.1.5 | Active Learning Integration Opportunity | 3 |
| 1.1.6 | Research Synthesis and Core Challenge | 4 |
| 1.1.7 | Research Questions | 4 |
| 1.2 | Research Framework and Contributions | 5 |
| 1.2.1 | The ALCIE Framework | 5 |
| 1.3 | Thesis Outline | 6 |
| 2 | Related Work | 8 |
| 2.1 | Image Captioning Foundations | 8 |
| 2.1.1 | Early Template-Based Approaches | 8 |
| 2.1.2 | Neural Encoder-Decoder Architectures | 9 |
| 2.1.3 | Transformer-Based Architectures | 9 |
| 2.2 | Interactive Image Captioning | 12 |
| 2.3 | Continual Learning for Vision-Language Tasks | 13 |
| 2.3.1 | Fundamental Approaches and Categorization | 13 |
| 2.3.2 | Task-Incremental Learning Strategies | 14 |
| 2.3.3 | Memory-Efficient Learning Strategies | 15 |
| 2.3.4 | Parameter-Efficient Adaptation Techniques | 15 |
| 2.3.5 | Multimodal Continual Learning | 16 |
| 2.4 | Active Learning for Sample Selection | 16 |
| 2.4.1 | Uncertainty-Based Selection Strategies | 16 |
| 2.4.2 | Diversity-Based Selection Strategies | 17 |
| 2.4.3 | Hybrid Active Learning Strategies | 18 |
| 2.5 | Evaluation and Datasets | 18 |
| 2.5.1 | Generic Image Captioning Datasets | 18 |
| 2.5.2 | Domain-Specific Image Captioning Datasets | 20 |

| | | |
|----------|---|-----------|
| 2.5.3 | Evaluation Metrics | 20 |
| 3 | Technical Background | 22 |
| 3.1 | Vision-Language Model Architectures | 22 |
| 3.1.1 | Evolution from Attention-Based Captioning to Transformers | 22 |
| 3.1.2 | Transformer Foundations for Multimodal Learning | 25 |
| 3.1.3 | Cross-Modal Attention Mechanisms | 29 |
| 3.1.4 | BLIP-2 Modular Architecture | 29 |
| 3.1.5 | OFA Unified Transformer Framework | 31 |
| 3.2 | Continual Learning Fundamentals | 32 |
| 3.2.1 | Catastrophic Forgetting in Neural Networks | 32 |
| 3.2.2 | Episodic Memory Mechanisms | 34 |
| 3.3 | Active Learning Principles | 36 |
| 3.3.1 | Query Strategy Framework | 36 |
| 3.3.2 | Uncertainty Sampling Strategies | 36 |
| 3.3.3 | Diversity Sampling Methodologies | 37 |
| 3.3.4 | Hybrid Sampling Strategies | 39 |
| 3.4 | Evaluation Frameworks | 40 |
| 3.4.1 | Traditional Image Captioning Metrics | 41 |
| 3.4.2 | Continual Learning Evaluation Metrics | 43 |
| 3.4.3 | Human Evaluation Statistical Methods | 44 |
| 4 | Methodology | 46 |
| 4.1 | ALCIE Framework Overview | 46 |
| 4.1.1 | Active Learning for Continual Image Captioning | 47 |
| 4.1.2 | Theoretical Motivation and Design Principles | 47 |
| 4.1.3 | Framework Architecture and Pipeline | 48 |
| 4.1.4 | Main Contributions and Novel Aspects | 50 |
| 4.2 | Benchmark Architecture Design | 51 |
| 4.2.1 | Benchmark Design Philosophy and Model Selection | 51 |
| 4.2.2 | BLIP-2 Architecture Implementation | 52 |
| 4.2.3 | OFA Unified Transformer Framework | 52 |
| 4.3 | Active Learning Integration | 52 |
| 4.3.1 | Random Sampling Baseline | 52 |
| 4.3.2 | Uncertainty Sampling | 53 |
| 4.3.3 | Diversity Sampling | 53 |
| 4.3.4 | Hybrid Sampling | 54 |
| 4.4 | Memory Management Strategies | 54 |

| | | |
|----------|--|-----------|
| 4.4.1 | Episodic Memory Buffer Architecture | 55 |
| 4.4.2 | Memory Replay Mechanisms | 56 |
| 4.4.3 | Memory Deletion Policies | 57 |
| 4.5 | Training Infrastructure and Implementation | 58 |
| 4.5.1 | System Architecture and Modular Design | 58 |
| 4.5.2 | Continual Learning Protocol Implementation | 59 |
| 4.5.3 | Hyperparameter Configuration and Training Settings | 60 |
| 5 | Experiments and Results | 62 |
| 5.1 | Experimental Setup | 62 |
| 5.1.1 | Dataset and Domain Configuration | 62 |
| 5.2 | Evaluation Metrics | 65 |
| 5.3 | Experiments | 66 |
| 5.3.1 | Experimental Protocol | 66 |
| 5.3.2 | Experiment 1: No Memory Baseline | 67 |
| 5.3.3 | Experiments 2-3: Random Sampling Memory Management | 68 |
| 5.3.4 | Experiment 4: Uncertainty Sampling Memory Management | 71 |
| 5.3.5 | Experiment 5: Diversity Sampling Memory Management | 73 |
| 5.3.6 | Experiment 6: Hybrid Memory Management | 74 |
| 5.3.7 | Continual Learning Metrics Analysis | 76 |
| 5.4 | Discussion | 78 |
| 5.4.1 | Key Experimental Findings | 78 |
| 5.4.2 | Memory Strategy Performance | 79 |
| 5.4.3 | Continual Learning Performance Metrics | 81 |
| 5.4.4 | Cross-Domain Interference Patterns | 81 |
| 5.4.5 | Sequential Knowledge Evolution | 82 |
| 5.4.6 | Universal Early Transition Vulnerability | 83 |
| 5.4.7 | Architectural Performance Differences | 84 |
| 5.4.8 | Practical Implications | 84 |
| 6 | User Study | 85 |
| 6.1 | Introduction | 85 |
| 6.2 | Research Questions | 85 |
| 6.3 | Methodology | 86 |
| 6.3.1 | Experimental Design | 86 |
| 6.3.2 | Participants | 86 |
| 6.3.3 | Stimuli and Conditions | 86 |
| 6.3.4 | Evaluation Metrics | 87 |

| | | |
|----------|---|------------|
| 6.3.5 | Analysis Framework | 87 |
| 6.4 | Results | 88 |
| 6.4.1 | Participant Characteristics and Data Quality | 88 |
| 6.4.2 | RQ1: Strategy Performance Analysis | 89 |
| 6.4.3 | RQ2: Catastrophic Forgetting Detection | 90 |
| 6.4.4 | RQ3: User Preference Analysis | 92 |
| 6.5 | Discussion | 93 |
| 6.5.1 | Strategy Equivalence: Confirming Technical-Human Disconnect | 93 |
| 6.5.2 | Catastrophic Forgetting: A Detectable User Experience Issue | 93 |
| 6.5.3 | No Clear User Preferences | 93 |
| 6.5.4 | Practical Implications | 93 |
| 7 | Conclusion and Future Work | 95 |
| 7.1 | Summary of Contributions and Key Insights | 95 |
| 7.1.1 | Research Question Answers | 95 |
| 7.1.2 | Paradigm-Shifting Insights | 96 |
| 7.2 | Practical Implications | 97 |
| 7.2.1 | Strategy Selection Guidelines | 97 |
| 7.3 | Limitations | 97 |
| 7.4 | Future Work | 98 |
| 7.4.1 | Addressing Fundamental Continual Learning Limitations | 98 |
| 7.4.2 | Domain and Task Expansion | 98 |
| 7.4.3 | Architectural Innovation and Scalability | 98 |
| 7.4.4 | Human-Centered Continual Learning and Methodological Innova- tions | 99 |
| | Bibliography | 100 |
| | Appendix: Supplementary Experiments Results | 110 |

Acronyms

- A-GEM** Averaged Gradient Episodic Memory. 3
- ACC** Average Accuracy. 43, 65, 66, 76
- AF** Average Forgetting. 44, 66, 76
- AL** Active Learning. xiv, xvi, 1, 3, 4, 5, 8, 16, 17, 22, 31, 34, 36, 46, 47, 49, 50, 51, 53, 55, 56, 58, 59, 60, 61, 95
- ALCIE** Active Learning for Continual Image Captioning Enhancement. vi, xiv, 5, 8, 16, 22, 44, 46, 47, 48, 49, 50, 51, 57, 59, 62, 85, 95, 97, 98
- BERTScore** Bidirectional Encoder Representations from Transformers Score. xv, xvi, 21, 42, 65, 66, 68, 70, 71, 72, 74, 75, 76, 77, 79, 80, 85, 110, 117, 118
- BLEU** Bilingual Evaluation Understudy. xvi, 20, 21, 41, 65, 66, 67, 68, 69, 71, 72, 73, 75, 85, 93, 110, 117, 118
- BLIP-2** Bootstrapping Language-Image Pre-training. vi, xiv, xv, xvi, xvii, 4, 5, 12, 13, 22, 29, 30, 31, 46, 49, 50, 51, 52, 56, 59, 60, 61, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 82, 83, 84, 86, 95, 96, 97, 98, 111, 112, 113, 114, 115, 116, 117
- BWT** Backward Transfer. 43, 65, 66
- CF** Catastrophic Forgetting. xiv, 1, 2, 3, 4, 12, 13, 14, 16, 18, 30, 32, 33, 34, 35, 43, 47, 48, 49, 56, 91, 96, 97, 98
- CI** Confidence Interval. 45, 87, 90, 91
- CIDeR** Consensus-based Image Description Evaluation. 21, 41, 42
- CL** Continual Learning. xiii, xiv, 1, 2, 3, 4, 5, 13, 14, 15, 16, 22, 29, 32, 33, 35, 36, 37, 40, 46, 47, 48, 50, 51, 52, 55, 57, 58, 59, 60, 61, 65, 95, 96, 97
- CLIP** Contrastive Language-Image Pretraining. xiv, 5, 17, 18, 38, 39, 47, 48, 52, 53, 54, 59, 61, 66, 73, 87
- CNN** Convolutional Neural Network. 9, 23
- CV** Computer Vision. 8
- EWC** Elastic Weight Consolidation. 3, 13, 34
- FACAD** Fashion Captioning Dataset. vi, xiv, 4, 7, 20, 49, 51, 52, 62, 63, 86, 97
- GEM** Gradient Episodic Memory. 3, 13

HUDS Hybrid Uncertainty and Diversity Sampling. 5, 18, 39, 40, 48, 50, 54, 66, 74

IC Image Captioning. xiv, 1, 2, 3, 4, 5, 8, 12, 13, 14, 16, 17, 18, 22, 23, 31, 33, 37, 46, 47, 49, 50, 51, 65, 95

IML Interactive Machine Learning. 3, 4, 12

LLMs Large Language Models. 13

LSTM Long Short Term Memory. 9, 12, 13, 23, 24

M² Transformer Meshed-Memory Transformer. 10

METEOR Metric for Evaluation of Translation with Explicit ORdering. xvii, xviii, 21, 66, 110, 111, 113, 114, 115, 116, 117, 118

MS-COCO Microsoft Common Objects in Context. 10, 14, 19, 20

MTE Mean Token Entropy. 5, 17, 18, 37, 48, 52, 53, 54, 66, 71, 87

NLP Natural Language Processing. 8, 13

NoCaps Novel Object Captioning at Scale. 20

OFA One For All. vi, xiv, xv, xvi, xvii, 4, 5, 10, 16, 22, 31, 46, 49, 50, 51, 52, 56, 59, 60, 61, 66, 67, 68, 69, 70, 71, 72, 73, 74, 76, 77, 78, 79, 80, 81, 84, 95, 96, 97, 98, 111, 112, 113, 114, 115, 116, 117

ROUGE Recall-Oriented Understudy for Gisting Evaluation. xvii, xviii, 21, 41, 66, 110, 111, 112, 114, 115, 116, 117, 118

SPICE Semantic Propositional Image Caption Evaluation. 21

TextCaps Text in Images Caption Dataset. 20

ViT Vision Transformer. xiv, 29, 30, 39

VizWiz Visual Question Answering from Blind People. 20

VQA Visual Question Answering. 16

List of Figures

| | | |
|-----|---|----|
| 1.1 | Catastrophic forgetting phenomenon in neural networks. Each task performs well during learning (solid lines) but experiences severe degradation when subsequent tasks are learned (dashed lines), demonstrating the fundamental challenge in Continual Learning (CL) scenarios. | 2 |
| 2.1 | Overview of the <i>Show, Attend and Tell</i> architecture [1]. The input image is first processed by a convolutional neural network (CNN) to extract a spatial feature map. An attention mechanism allows the recurrent neural network (RNN) with LSTM units to focus on different regions of the image when generating each word of the caption. Colored boxes illustrate how specific regions of the image are attended to for predicting corresponding words in the caption. | 9 |
| 2.2 | OFA unified architecture [2] handling multiple vision-language tasks through shared parameters and instruction-based prompting. The model processes diverse inputs (images, text, task instructions) through the same encoder-decoder framework, demonstrating the versatility of unified transformer designs for multimodal learning. | 11 |
| 2.3 | BLIP-2 architecture [3] showing the three-component design: frozen vision encoder, trainable Q-Former with learnable queries, and frozen language model. The Q-Former serves as an information bottleneck, extracting relevant visual information for text generation. | 11 |
| 2.4 | ContCap architecture showing the integration of freezing, knowledge distillation, and pseudo-labeling mechanisms for continual image captioning. The framework demonstrates how task-specific knowledge can be preserved while adapting to new captioning domains without catastrophic forgetting. | 14 |
| 2.5 | Core Tokensets [4] approach for memory-efficient continual learning. The method selects and stores the most informative token fragments from image-caption pairs based on attribution scores, enabling effective knowledge retention with minimal memory overhead. | 15 |
| 2.6 | Generic vs domain-specific dataset comparison: Flickr30k [5] focuses on human activities with descriptive captions, while TextCaps [6] specializes in text-rich images requiring OCR-based descriptions. | 19 |
| 3.1 | Comparison of soft and hard attention in the <i>Show, Attend and Tell</i> model [1]. Top: Soft attention assigns weights to all spatial locations (heat maps), blending features smoothly. Bottom: Hard attention selects one location at each step (points), yielding focused, discrete selections. Generated captions illustrate the influence of each mechanism. | 24 |
| 3.2 | Complete Transformer architecture [7] showing encoder-decoder structure with self-attention, cross-attention, and feed-forward components. The parallel processing capability and attention mechanisms form the foundation for modern vision-language models. | 26 |

| | | |
|-----|--|----|
| 3.3 | Multi-head attention visualization showing different attention patterns across heads. Each head captures different types of relationships, enabling comprehensive sequence understanding crucial for multimodal tasks [7]. . | 28 |
| 3.4 | Vision Transformer (ViT) [8] architecture used as the frozen vision encoder in BLIP-2. Input images are divided into non-overlapping patches (typically 16×16), flattened, and linearly projected into embeddings. A [CLS] token and positional embeddings are added before processing through the Transformer encoder layers. | 30 |
| 3.5 | Parameter evolution timeline demonstrating the mathematical basis of Catastrophic Forgetting (CF). Starting from random initialization (θ_0), each phase shows parameter updates through gradient descent optimization for new domains. While current domain performance remains high, parameter drift causes systematic degradation of previous domain performance, with Domain 1 accuracy dropping from 90% to 25% as parameters evolve from θ_1 to θ_3 . This visualization captures the core challenge in CL: how parameter updates beneficial for new tasks can be detrimental to previously learned capabilities. | 33 |
| 3.6 | Class-incremental CL with episodic memory replay (taken from [9]). The model sequentially learns from a stream of data containing new classes while maintaining a memory buffer. At each time step t , the model receives new data and updates using both current samples and replayed samples from memory. The memory buffer stores representative samples (data, features, and logits) from previous tasks to mitigate CF through strategic replay. | 35 |
| 3.7 | Contrastive Language–Image Pretraining (CLIP) architecture and training process [10]. (1) Contrastive pre-training: Image and text encoders learn joint representations by maximizing similarity between correct image-text pairs and minimizing similarity between incorrect pairs in the batch. (2) Create dataset classifier: Class labels are converted to text prompts (e.g., "A photo of a object") and processed through the text encoder. (3) Zero-shot prediction: For inference, the image encoder processes the input image, and classification is performed by comparing the image embedding with all text embeddings to find the highest similarity. | 38 |
| 4.1 | ALCIE Framework Architecture: The framework integrates four Active Learning (AL) strategies (Random, Uncertainty, Diversity, Hybrid) with episodic memory management for continual Image Captioning (IC). Pre-trained vision-language models (BLIP-2/OFA) process sequential fashion domain clusters from the FACAD dataset. The systematic comparison evaluates whether strategic AL strategies provide meaningful advantages over random sampling baselines while managing bounded memory constraints and preventing CF through balanced rehearsal mechanisms. . . . | 49 |
| 4.2 | Progressive evolution of episodic memory buffer composition during continual learning across six fashion categories. Each stage (1-6) corresponds to the buffer state after training on a new domain: (1) Accessories, (2) Bottoms, (3) Dresses, (4) Outerwear, (5) Shoes, and (6) Tops. The proportional allocation demonstrates the cluster-aware deletion strategy defined in Equation 4.17, showing how different sampling strategies prioritize sample retention while maintaining balanced cross-domain representation within the fixed capacity constraint $C_{\max} = 10,000$ | 55 |

| | | |
|-----|---|----|
| 5.1 | Representative FACAD dataset samples from six fashion categories used in experiments. | 64 |
| 5.2 | Average Accuracy across memory management strategies for Bidirectional Encoder Representations from Transformers Score (BERTScore)-F1 metrics. BLIP-2 consistently achieves higher average accuracy than OFA across all strategies, with multiple strategies achieving optimal performance of 0.883 for BLIP-2. | 76 |
| 5.3 | Average Forgetting across memory management strategies for BERTScore-F1 metrics. OFA demonstrates consistently lower forgetting rates than BLIP-2, with Random DEL- achieving the lowest forgetting of 0.034 for OFA. | 77 |
| 5.4 | Catastrophic forgetting analysis revealing systematic lexical-semantic disconnect. While lexical generation (BLEU-4, left panels) suffers severe degradation with universal early transition vulnerability, semantic understanding (BERTScore-F1, right panels) remains remarkably stable across all memory strategies. This 60-85 percentage point gap explains why random sampling achieves competitive performance: sophisticated approaches primarily optimize surface-level generation while essential semantic capabilities are preserved regardless of algorithmic complexity. | 78 |
| 5.5 | Memory investment efficiency across architectures. BLIP-2 demonstrates diminishing returns (6× memory for 1% performance gain), while OFA shows clear benefits from additional memory capacity (6× memory for 7% performance gain), indicating architecture-dependent memory utilization patterns critical for deployment planning. | 80 |
| 6.1 | Overall performance by strategy showing practically equivalent results. Error bars represent 95% confidence intervals. The 0.07-point range falls well below the practical significance threshold. | 89 |
| 6.2 | Performance by evaluation dimension showing strategy equivalence. Error bars represent 95% confidence intervals. | 90 |
| 6.3 | Catastrophic forgetting pattern across learning phases. The non-monotonic trajectory shows initial low performance (Accessories), recovery (Bottoms), mid-sequence vulnerability (Dresses), and late-phase stabilization. Error bars represent 95% confidence intervals. | 90 |
| 6.4 | Early versus late learning phase comparison showing detectable forgetting effects. The 0.14-point difference represents meaningful quality degradation. Error bars represent 95% confidence intervals. | 91 |
| 6.5 | User preference distribution in forced-choice evaluation showing no clear preference pattern. The 2.5 percentage point range indicates practically equivalent user satisfaction across all memory management strategies. . . | 92 |

List of Tables

| | | |
|------|---|----|
| 4.1 | Comprehensive hyperparameter configuration ensuring reproducible training and fair comparison across model architectures and AL strategies. | 60 |
| 5.1 | FACAD dataset statistics for continual learning experiments. | 63 |
| 5.2 | Bilingual Evaluation Understudy (BLEU) scores showing complete catastrophic forgetting without episodic memory. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 67 |
| 5.3 | BERTScore-F1 scores demonstrating severe semantic degradation despite better retention than lexical metrics. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 68 |
| 5.4 | BLEU-4 performance comparison for OFA with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 69 |
| 5.5 | BLEU-4 performance comparison for BLIP-2 with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 69 |
| 5.6 | BERTScore-F1 performance comparison for OFA with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 70 |
| 5.7 | BERTScore-F1 performance comparison for BLIP-2 with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 70 |
| 5.8 | BLEU-4 performance with uncertainty-based memory selection. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 72 |
| 5.9 | BERTScore-F1 performance with uncertainty-based memory selection showing robust semantic retention. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 72 |
| 5.10 | BLEU-4 performance with diversity-based memory selection demonstrating superior transition stability. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 73 |
| 5.11 | BERTScore-F1 performance with diversity-based memory selection showing exceptional semantic stability. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 74 |
| 5.12 | BLEU-4 performance with hybrid memory management demonstrating balanced multi-criteria selection. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 75 |
| 5.13 | BERTScore-F1 performance with hybrid memory management showing balanced semantic retention. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 75 |

| | | |
|------|---|-----|
| 5.14 | Cross-domain interference examples for OFA with constrained (DEL+) and unconstrained (DEL-) memory strategies. The arrow notation ($X \rightarrow Y$) indicates models trained on domain X and tested on domain Y images. Red highlights indicate domain-inappropriate terminology, blue highlights show semantic confusion, and purple highlights demonstrate persistent domain overfitting. | 81 |
| 5.15 | Cross-domain interference examples for BLIP-2 with constrained (DEL+) and unconstrained (DEL-) memory strategies. The arrow notation ($X \rightarrow Y$) indicates models trained on domain X and tested on domain Y images. Red highlights show abrupt truncation in DEL+, orange highlights show more complete DEL- descriptions, blue highlights show domain confusion, and green highlights show appropriate domain descriptions. | 82 |
| 5.16 | Sequential knowledge retention across learning phases for BLIP-2. The image is shown from cluster Accessories, the first domain in the training sequence, which faces maximum catastrophic forgetting risk as it is evaluated after every subsequent domain. Green highlights show diversity sampling's terminological stability, orange highlights show random sampling's hibernation-recovery pattern, purple highlights show uncertainty sampling's vocabulary inconsistency, blue highlights show hybrid's conservative elaboration, and red highlights show systematic domain replacement without memory. | 83 |
| 6.1 | Range analysis across evaluation dimensions for significance assessment. | 89 |
| 6.2 | Summary of catastrophic forgetting effects across learning phases. | 91 |
| 6.3 | Practical significance assessment of user preference patterns. | 92 |
| 1 | Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L scores showing severe structural degradation without episodic memory. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 111 |
| 2 | Metric for Evaluation of Translation with Explicit ORdering (METEOR) scores demonstrating semantic degradation without memory mechanisms. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 111 |
| 3 | ROUGE-L performance comparison for OFA with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. . . | 112 |
| 4 | ROUGE-L performance comparison for BLIP-2 with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. . . | 112 |
| 5 | METEOR performance comparison for OFA with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 113 |
| 6 | METEOR performance comparison for BLIP-2 with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with bold indicating higher values between strategies. | 113 |

| | | |
|----|--|-----|
| 7 | ROUGE-L performance with uncertainty-based memory selection. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 114 |
| 8 | METEOR performance with uncertainty-based memory selection showing robust semantic retention. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 114 |
| 9 | ROUGE-L performance with diversity-based memory selection demonstrating superior transition stability. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 115 |
| 10 | METEOR performance with diversity-based memory selection demonstrating exceptional structural preservation. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 115 |
| 11 | ROUGE-L performance with hybrid memory management demonstrating balanced multi-criteria selection. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 116 |
| 12 | METEOR performance with hybrid memory management showing balanced semantic retention. Subscript percentages show retention levels, with bold indicating higher values between architectures. | 116 |

Chapter 1

Introduction

This chapter establishes the foundation for investigating *strategic memory management* in *continual image captioning systems*. We begin with an examination of the fundamental challenges facing current image captioning approaches, particularly the limitations imposed by *Catastrophic Forgetting (CF)* and *resource constraints* in real-world implementation scenarios (section 1.1). Building upon this analysis, we introduce our proposed research framework and outline the key contributions of this research, demonstrating how *Active Learning (AL)* principles can be systematically integrated with *episodic memory selection* to enhance *Continual Learning (CL)* performance (section 1.2).

1.1 Motivation and Problem Statement

1.1.1 Current Limitations in Image Captioning Systems

Image Captioning (IC) aims to automatically generate natural language descriptions that accurately capture visual content, semantic relationships, and contextual information within images. The field has evolved from template-based systems through retrieval methods to modern encoder-decoder architectures enhanced by attention mechanisms and transformer designs [11, 1, 3]. This progression has established strong performance in controlled settings, yet significant challenges remain for real-world deployment.

The transition from laboratory settings to real-world implementation reveals critical limitations. Unlike static training environments where models are trained once on large, curated datasets, practical applications require systems that can incrementally adapt to new domains, user preferences, and evolving visual content [12]. For example, a fashion e-commerce platform may need to generate product descriptions for new clothing styles, seasonal collections, and regional preferences, requiring the system to learn terminology and concepts while retaining knowledge of previously learned categories.

1.1.2 Continual Learning Paradigm

CL represents a machine learning paradigm that enables artificial intelligence systems to acquire knowledge incrementally from a succession of tasks while preserving previously learned capabilities [13]. Unlike conventional machine learning approaches that assume simultaneous access to complete datasets during training, CL addresses scenarios where data becomes available progressively across distinct task domains, requiring models to adapt continuously without experiencing performance degradation on prior tasks. This paradigm proves particularly relevant for IC systems deployed in dynamic environments where new visual domains, linguistic patterns, or user requirements emerge over time, necessitating adaptive learning mechanisms that maintain historical knowledge while incorporating novel information.

1.1.3 The Challenge of Catastrophic Forgetting

CF represents the most significant obstacle to effective CL across machine learning domains. This phenomenon, first formally characterized by McCloskey and Cohen in 1989 [14], arises when neural networks acquire new tasks or adapt to novel domains, resulting in marked performance degradation on previously learned tasks, as illustrated in Figure 1.1. The challenge is particularly pronounced in gradient-based optimization, where iterative parameter updates to minimize loss on new data systematically overwrite previously acquired knowledge [15].

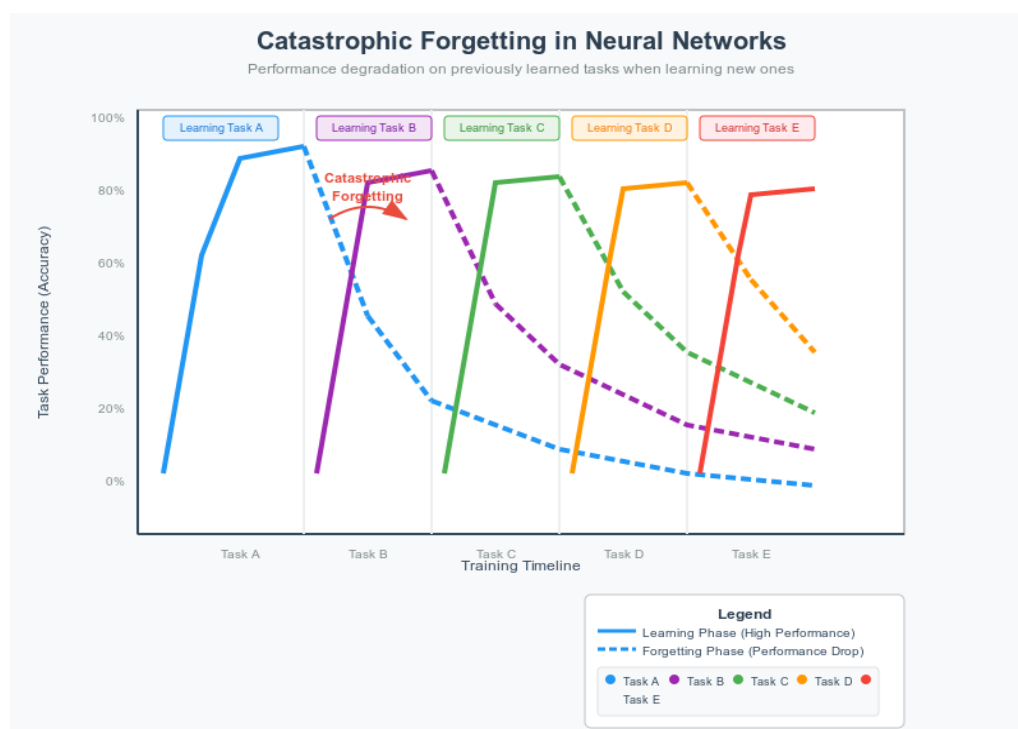


Figure 1.1: Catastrophic forgetting phenomenon in neural networks. Each task performs well during learning (solid lines) but experiences severe degradation when subsequent tasks are learned (dashed lines), demonstrating the fundamental challenge in CL scenarios.

The fundamental challenge in CL lies in achieving optimal balance between *plasticity*: the capacity to learn new information, and *stability*: the retention of previously acquired knowledge. Traditional neural network training typically overwrites existing parameters when learning new tasks, leading to performance degradation on previously mastered domains. This plasticity-stability trade-off becomes increasingly difficult to manage as the number of CL tasks increases [13].

Within CL, several mitigation strategies have been proposed, including *Elastic Weight Consolidation (EWC)* [15], *Gradient Episodic Memory (GEM)* [16], and knowledge distillation techniques [17]. These approaches, however, often require careful hyperparameter tuning and may not scale effectively to the complex, multimodal nature of IC tasks, where both visual and linguistic representations must be preserved simultaneously.

1.1.4 Memory Management and Strategic Selection Challenges

While CF presents a fundamental challenge, strategic memory management is equally critical. Practical implementation of CL systems involves constraints related to memory usage and computational efficiency [13]. Storing all previously encountered training examples, known as *rehearsal* [18], is often impractical due to storage limitations and computational overhead during training [19, 16]. Consequently, *selective episodic memory mechanisms* that can identify and retain the most informative examples while operating within limited memory constraints are essential.

This challenge is particularly pronounced in Interactive Machine Learning (IML) scenarios, where systems must continually incorporate user feedback while operating under strict memory constraints. These limitations create a fundamental trade-off between memory efficiency and the retention of previously acquired knowledge. As the size of episodic memory buffers increases, research has shown that computational and storage demands rise significantly [16]. In response, alternatives such as Averaged Gradient Episodic Memory (A-GEM) [20] have been developed to maintain competitive performance while substantially reducing computational requirements.

Many CL approaches rely on *random sampling* or *simple heuristics* to select which examples are retained in episodic memory buffers [12, 18, 21]. However, such strategies often fail to optimize memory utilization, potentially retaining redundant or less informative instances while discarding more valuable examples. This limitation underscores the need for principled and adaptive memory selection strategies to maximize learning outcomes in resource-constrained environments [19].

1.1.5 Active Learning Integration Opportunity

The integration of AL principles with episodic memory management represents an underexplored opportunity for improving CL performance. AL focuses on identifying the most informative examples for annotation [22], demonstrating significant potential for reducing annotation costs while enhancing model performance across various tasks.

Strategic sample selection approaches include: *uncertainty sampling* (selecting examples where model confidence is lowest), *diversity sampling* (ensuring comprehensive input space coverage), and *hybrid sampling* that combine both criteria [23, 24, 25]. Recent work in continual AL [26] has shown that AL can be effectively combined with replay-based CL algorithms, highlighting the potential for this integration in vision-language tasks.

The emergence of continual AL further highlights the value of integrating AL principles with CL frameworks. Recent studies have shown that AL can be effectively combined with replay-based CL algorithms, accelerating training and sustaining model performance over sequential tasks [26]. These advances underscore that the convergence of AL and CL constitutes a promising research direction with substantial practical implications for adaptive vision-language systems.

1.1.6 Research Synthesis and Core Challenge

The strategic selection of examples for episodic memory storage represents a critical but underexplored research direction in CL for IC. Current episodic memory approaches for IC systems primarily rely on *random sampling* strategies, which may not effectively capture the most informative user interactions [12]. This limitation becomes particularly pronounced in resource-constrained environments where memory capacity is limited and every stored example must contribute maximally to the model’s learning process.

The convergence of AL, CL, and IML presents a unique opportunity to address fundamental challenges in adaptive IC systems. While AL has been extensively studied in traditional supervised learning contexts [22], its integration with *episodic memory mechanisms* for CL remains largely unexplored [13, 27], with even fewer studies addressing vision-language tasks.

Central Research Challenge: This thesis systematically investigates whether *strategic sample selection* can significantly improve CL performance in IC tasks beyond random sampling baselines. We address the fundamental question of whether replacing random episodic memory selection with principled AL strategies: uncertainty sampling, diversity sampling, and hybrid sampling approaches can effectively mitigate CF while maintaining computational efficiency.

Through controlled experiments across multiple fashion categories of Fashion Captioning Dataset (FACAD) [28] using both *Bootstrapping Language-Image Pre-training (BLIP-2)* [3] and *One For All (OFA)* [2] architectures, we evaluate the comparative effectiveness of memory management strategies in sequential learning scenarios. This investigation directly challenges the assumption that algorithmic sophistication necessarily outperforms simple baselines, while establishing whether technical improvements translate to meaningful user experience benefits.

1.1.7 Research Questions

Building upon this central challenge, this thesis addresses the following *fundamental research questions*:

RQ1: Memory Strategy Effectiveness and Architecture Dependency

How do different *episodic memory management strategies* affect *catastrophic forgetting mitigation* in continual learning for vision-language models, and how does *model architecture* influence strategy effectiveness?

RQ2: Lexical versus Semantic Forgetting Patterns

How does catastrophic forgetting differentially affect *lexical generation capabilities* versus *semantic understanding preservation*, and what implications does this distinction have for memory strategy selection and evaluation frameworks?

RQ3: Memory Capacity Optimization and Resource Efficiency

What is the relationship between *memory buffer capacity* and *continual learning performance*, and under what conditions does *constrained memory* outperform unlimited memory?

1.2 Research Framework and Contributions

1.2.1 The ALCIE Framework

This research introduces *Active Learning for Continual Image Captioning Enhancement (ALCIE)*, a framework that systematically advances beyond our previous random sampling approach [12] by integrating *AL* principles with *episodic memory selection* in continual IC systems.

Our approach centers on three key methodological innovations:

Intelligent Sample Selection: Developing multiple selection strategies tailored for episodic memory retention in continual multimodal learning, including uncertainty sampling using our proposed *Mean Token Entropy (MTE)* [29], diversity sampling employing K-means clustering on *Contrastive Language–Image Pretraining (CLIP)* [10] feature representations, and hybrid sampling that adaptively combines both criteria using the *Hybrid Uncertainty and Diversity Sampling (HUDS)* framework [25].

Multi-Architecture Framework: Evaluation across different state-of-the-art vision-language models, including *BLIP-2* [3] and *OFA* [2], revealing architecture-dependent strategy effectiveness patterns.

Comprehensive Evaluation Protocol: Establishing a rigorous evaluation methodology that combines automated metrics with human assessment, enabling thorough validation across multiple dimensions of caption quality and learning effectiveness.

1. **Strategic Memory Management Framework:** We systematically compare AL-based memory selection strategies against random sampling baselines, revealing architecture-dependent strategy effectiveness patterns and challenging assumptions about the necessity of sophisticated memory management in CL.
2. **Methodological Innovation:** We develop multimodal uncertainty sampling and diversity sampling techniques specifically for vision-language CL scenarios, including adaptation of *MTE* for uncertainty sampling, *CLIP*-based clustering for diversity sampling, and *HUDS* integration for hybrid sampling approaches.
3. **Lexical-Semantic Forgetting Paradigm:** We demonstrate that catastrophic forgetting operates differentially across knowledge types, with severe lexical generation degradation occurring alongside robust semantic understanding preservation. This discovery explains the competitive effectiveness of random sampling approaches and fundamentally challenges community assumptions about memory strategy sophistication requirements in continual learning systems.
4. **Human-Centered Evaluation:** We conduct comprehensive human evaluation studies involving fashion-interested participants to validate that improvements in automated metrics translate to meaningful improvements in human-perceived caption quality across dimensions of *relevance*, *fluency*, *descriptiveness*, and *novelty*.
5. **Comprehensive Baseline Evaluation Framework:** We establish rigorous comparative analysis protocols that systematically evaluate AL-based memory selection

strategies against established baselines. This framework enables quantitative assessment of strategy effectiveness through controlled experimental conditions and statistical significance testing across multiple performance dimensions.

In summary, this thesis compares multiple memory management strategies: random, uncertainty-based, diversity-based, and hybrid sampling using BLIP-2 and OFA vision-language models for continual image captioning. Our experiments reveal that simple random sampling achieves comparable retention to sophisticated strategies, with diversity-based approaches offering only temporary early advantages. We discover a systematic lexical-semantic forgetting disconnect, where catastrophic forgetting primarily affects surface-level generation while preserving essential semantic understanding. These findings demonstrate that computational efficiency can be prioritized over algorithmic sophistication, fundamentally realigning continual learning research toward addressing forgetting mechanisms rather than optimizing memory selection complexity.

1.3 Thesis Outline

This thesis is structured to provide a comprehensive investigation of active learning strategies for continual image captioning, progressing from theoretical foundations through empirical evaluation to practical applications. The organization follows a logical progression that establishes the research context, develops the proposed methodology, evaluates its effectiveness, and validates findings through human-centered evaluation.

Chapter 2: Related Work establishes the theoretical foundation by examining four key research areas that intersect in this work. The chapter begins with image captioning foundations, tracing the evolution from early template-based approaches through neural encoder-decoder architectures to modern transformer-based systems. Interactive image captioning research is then reviewed, followed by comprehensive coverage of continual learning approaches for vision-language tasks, including fundamental categorization, task-incremental strategies, and memory-efficient techniques. The chapter concludes with active learning methodologies for sample selection and evaluation frameworks, providing the necessary background for understanding the proposed approach.

Chapter 3: Technical Background provides detailed technical exposition of the core components underlying the research. Vision-language model architectures are examined, covering the evolution from attention-based captioning to transformers, with specific focus on *BLIP-2* and *OFA* architectures used in the experimental evaluation. Continual learning fundamentals are presented, including catastrophic forgetting mechanisms and episodic memory approaches. Active learning principles are detailed, covering uncertainty sampling, diversity sampling, and hybrid strategies. The chapter concludes with evaluation frameworks for both traditional image captioning and continual learning scenarios.

Chapter 4: Methodology introduces the *ALCIE* framework, providing comprehensive coverage of the proposed approach. The chapter begins with a framework overview and theoretical motivation, followed by detailed benchmark architecture design covering both *BLIP-2* and *OFA* implementations. Active learning integration is thoroughly described, including random sampling baselines, uncertainty sampling, diversity sampling, and hybrid approaches. Memory management strategies are presented, covering episodic memory buffer architecture, replay mechanisms, and deletion policies. The chapter concludes with training infrastructure and implementation details.

Chapter 5: Experiments and Results presents a comprehensive empirical evaluation of the proposed approach. The chapter begins with the experimental setup, including dataset configuration using the FACAD fashion dataset organized into six domain clusters. Detailed evaluation metrics are presented, followed by systematic experimental protocols. Six major experiments are conducted: no-memory baseline, random sampling memory management (constrained and unconstrained), uncertainty sampling, diversity sampling, and hybrid approaches. Results are analyzed through both traditional captioning metrics and continual learning performance measures, with a comprehensive discussion of key findings, memory strategy performance, cross-domain interference patterns, and architectural differences.

Chapter 6: User Study validates the technical findings through human-centered evaluation, addressing the critical question of whether technical improvements in continual learning translate to perceivable user benefits. The chapter covers experimental design, participant recruitment, and evaluation methodology for assessing human perception of caption quality across different memory strategies. Results demonstrate the practical significance of the research by examining user preferences and the detectability of catastrophic forgetting effects in real-world scenarios.

Chapter 7: Conclusion and Future Work synthesizes the research contributions, discusses limitations, and outlines directions for future work. The chapter consolidates findings from both technical experiments and human evaluation, providing recommendations for practical deployment and identifying opportunities for further research in continual learning for vision-language tasks.

This structure ensures a thorough examination of active learning strategies for continual image captioning, providing both theoretical insights and practical guidance for researchers and practitioners working in multimodal machine learning and continual learning systems.

Chapter 2

Related Work

This chapter reviews the research landscape that informs our ALCIE framework for *strategic memory management* in *continual image captioning* systems. We examine four interconnected research areas: image captioning foundations (section 2.1), continual learning for vision-language tasks (section 2.3), active learning for sample selection (section 2.4), and evaluation methodologies with datasets (section 2.5). This review establishes the theoretical foundations for integrating AL principles with episodic memory selection, highlighting the research gaps our work addresses.

2.1 Image Captioning Foundations

IC connects Computer Vision (CV) and Natural Language Processing (NLP) by creating written descriptions for images [30, 31]. Over the past decade, the field has evolved from simple rule-based systems to state-of-the-art deep learning models, driven by advances in neural architectures, improved training strategies, and a growing emphasis on generalization across domains [32, 33].

This section traces the evolution of IC from early template-based systems to state-of-the-art *transformer architectures*. We examine early *template-based approaches* (subsection 2.1.1), the development of neural *encoder-decoder architectures* with revolutionary *attention mechanisms* (subsection 2.1.2), and the advancement to *transformer-based architectures* including unified vision-language models and large-scale pretraining approaches (subsection 2.1.3). This progression establishes the foundation for our continual learning approach.

2.1.1 Early Template-Based Approaches

Early *image captioning* approaches relied on template-based systems that detected objects and actions to populate fixed sentence structures [34], or retrieval methods that matched new images with existing captioned examples [35]. While interpretable and grammatically correct, these methods performed poorly on novel visual content due to rigid templates and limited vocabulary coverage [30].

2.1.2 Neural Encoder-Decoder Architectures

Deep learning transformed *image captioning* through encoder-decoder architectures combining Convolutional Neural Network (CNN) for visual feature extraction and Long Short Term Memory (LSTM) networks for sequential text generation [31, 36, 37]. Early neural systems used CNN encoders to compress images into fixed vectors that initialized LSTM decoders for word-by-word generation [11]. However, this static encoding limited spatial understanding and multi-object reasoning [32].

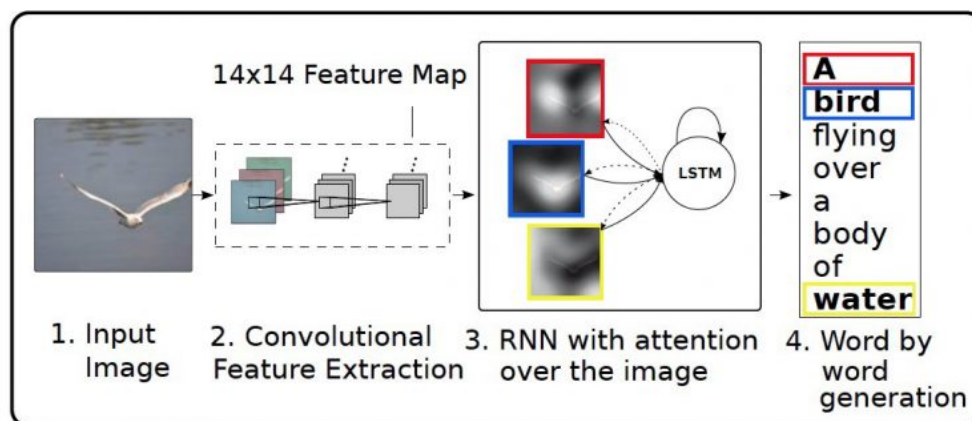


Figure 2.1: Overview of the *Show, Attend and Tell* architecture [1]. The input image is first processed by a convolutional neural network (CNN) to extract a spatial feature map. An attention mechanism allows the recurrent neural network (RNN) with LSTM units to focus on different regions of the image when generating each word of the caption. Colored boxes illustrate how specific regions of the image are attended to for predicting corresponding words in the caption.

Show, Attend and Tell [1] introduced spatial attention mechanisms that dynamically focus on relevant image regions during caption generation, as illustrated in Figure 2.1. *Soft attention* computes differentiable weighted sums of spatial features, while *hard attention* stochastically selects discrete regions. Soft attention became standard due to its end-to-end trainability and interpretable attention visualizations. *Bottom-up attention* [38] leveraged object detection to focus on detected objects rather than spatial grids, providing more semantically meaningful attention targets that enhanced scene understanding and caption accuracy.

2.1.3 Transformer-Based Architectures

The *Transformer* architecture [7] addressed RNN limitations by replacing recurrence with self-attention, enabling parallel processing and superior long-range dependency modeling that significantly enhanced training efficiency and performance. When adapted to *image captioning*, *Transformer-based models* offered significant advantages in capturing global context and producing more coherent captions. Self-attention allowed the decoder to consider relationships between all previously generated words simultaneously, unlike sequential LSTM networks [39, 40]. This evolution benefited directly from attention mechanisms first demonstrated in *Show, Attend and Tell*.

Building on the strengths of Transformer-based models in image captioning, specialized architectures such as the Meshed-Memory Transformer (M^2 Transformer) [39] have further advanced the field. The M^2 Transformer introduces a meshed memory mechanism, forming dense skip connections between encoder and decoder layers, which enables the model to access multi-level visual features. This design enhances visual grounding and allows for more accurate modeling of complex scenes with multiple objects. As a result, the M^2 Transformer achieves state-of-the-art performance on benchmarks like Microsoft Common Objects in Context (MS-COCO), producing captions that are both descriptive and semantically aligned with the images.

The *X-Linear Attention* model [40] further improves the alignment between visual and linguistic information by employing a bilinear pooling attention module. This mechanism captures higher-order interactions between image regions and words, enabling more precise and contextually relevant captions. X-Linear Attention consistently demonstrates strong results on challenging datasets such as MS-COCO, highlighting the value of advanced attention mechanisms in cross-modal understanding.

Unified Vision-Language Models

The field evolved toward *unified vision-language models* capable of joint reasoning across modalities [2, 3]. This development encompasses three approaches: multimodal pretraining, instruction-tuned architectures, and modular frozen-component systems.

Multimodal Pretraining Foundations Early unified models like ViLBERT [41] employed dual-stream architectures with separate visual and textual processing pathways connected through co-attentional layers, while UNITER [42] adopted single-stream designs processing concatenated visual-textual sequences through unified self-attention mechanisms.

Instruction-Tuned Unified Architectures OFA [2] frames vision-language tasks as sequence-to-sequence generation guided by natural language instructions. Its unified Transformer encoder-decoder architecture processes multimodal inputs and generates task-specific outputs based on instructional prompts, as shown in Figure 2.2. Joint training with task-specific prompts enables consistent performance across diverse scenarios. OFA exemplifies this approach through its comprehensive Transformer encoder-decoder architecture that processes multimodal inputs and generates task-specific outputs based on instructional prompts [2]. The model’s architecture, depicted in Figure 2.2, demonstrates how natural language instructions systematically guide task execution across image captioning, visual question answering, and text generation. Joint training on diverse multimodal datasets with task-specific prompts enables consistent performance across both high-resource and low-resource scenarios, establishing instruction tuning as a systematic approach for unified vision-language modeling.

mPLUG [43] extends the instruction-tuned approach by incorporating architectural innovations, including cross-modal skip connections and hierarchical feature fusion mechanisms. The model’s asymmetric dual-stream design processes visual and textual modalities through separate encoders before integration via cross-modal Transformer layers equipped with skip connections. This architectural configuration enables efficient processing of extended visual sequences while preserving cross-modal information flow. The hierarchical feature fusion mechanism captures both coarse scene-level semantics

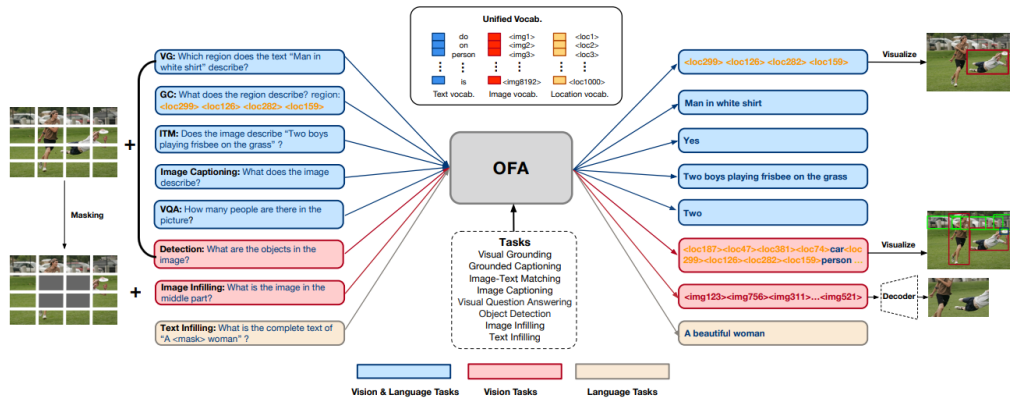


Figure 2.2: OFA unified architecture [2] handling multiple vision-language tasks through shared parameters and instruction-based prompting. The model processes diverse inputs (images, text, task instructions) through the same encoder-decoder framework, demonstrating the versatility of unified transformer designs for multimodal learning.

and fine-grained object-level details, demonstrating particular effectiveness for tasks requiring detailed visual analysis.

Modular Architectures with Frozen Components Modular architectures utilize frozen pretrained components to minimize adaptation costs. ClipCap [44] projects CLIP visual features into GPT-2 through lightweight networks, enabling efficient zero-shot captioning without joint training. BLIP [45] employs flexible encoder-decoder configurations supporting contrastive learning, matching, and generation tasks, with CapFit bootstrapping to enhance noisy web-scraped training data.

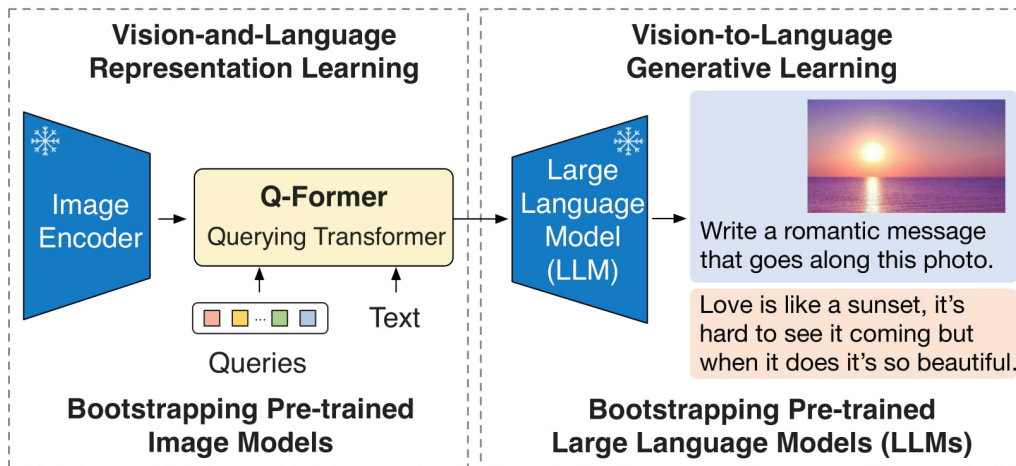


Figure 2.3: BLIP-2 architecture [3] showing the three-component design: frozen vision encoder, trainable Q-Former with learnable queries, and frozen language model. The Q-Former serves as an information bottleneck, extracting relevant visual information for text generation.

BLIP-2 [3] introduces the Q-Former, a lightweight Transformer bridge between frozen vision encoders and language models (Figure 2.3). The Q-Former uses learned query tokens to extract visual features and align them with language model inputs through two-stage bootstrapping: vision-language representation learning and vision-to-language generative learning. This modular design achieves state-of-the-art performance while maintaining parameter efficiency.

2.2 Interactive Image Captioning

This section examines interactive and adaptive image captioning systems that incorporate human feedback and continuous learning, providing the conceptual foundation for human-AI collaborative captioning systems informing our continual learning framework.

As IC systems continue to advance, models trained in static settings prove insufficient for practical applications involving dynamic user needs, evolving visual domains, or sparse labeled data. This has led to *interactive image captioning*, where models continuously learn from user feedback and adapt to changing contexts. Rooted in IML, these systems treat users as active participants in the learning loop, transforming image captioning into a collaborative, adaptive process supporting personalized outputs and efficient learning from limited supervision.

A foundational contribution to interactive captioning introduces a dual-level attention mechanism that combines top-down global scene features with bottom-up object-level representations, enabling flexible attention across visual hierarchies according to semantic context [46, 47]. Building on this foundation, the LSTM-based decoder incorporates beam search and re-ranking mechanisms to generate multiple candidate captions, which can be ranked using content similarity heuristics or refined interactively by users. This approach has proven particularly effective in multilingual settings, where cross-lingual pretraining results in substantial improvements in caption fluency for underrepresented languages such as German [47]. The process of interactive captioning has also been systematically formalized into three core IML components: feedback collection, data augmentation, and model update. This structure enables flexible adaptation across domains where annotation is costly and user input varies in form and frequency [48].

Advanced interactive systems address cognitive load by employing implicit feedback mechanisms that detect user disagreement through *gaze behavior* and *facial expressions* [49]. By continuously monitoring eye movements and emotional cues, these systems can infer *user dissatisfaction* and selectively request feedback only when necessary. This targeted approach not only improves user experience by minimizing interruptions but also ensures that model updates are focused where they are most impactful. Empirical results show that such systems achieve high accuracy in predicting disagreement using multimodal signals, demonstrating that implicit feedback can effectively guide caption correction and adaptation without constant explicit user input.

Domain-specific interactive approaches address unique challenges in specialized contexts by leveraging contextualized systems that incorporate *episodic memory* buffers for incremental learning [12]. Research shows that traditional data augmentation can degrade performance on user-generated images, whereas integrating episodic memory is highly effective in mitigating *CF* during incremental updates. In news captioning, *context-aware strategies* that provide controlled information, such as extracted named entities, substantially enhance model performance across different architectures. This highlights the importance of interactive methods for addressing complex contextual requirements [50].

These advances are demonstrated through the evaluation of large multimodal models like BLIP-2 [3] and two-stage pipelines that combine dedicated captioning systems with post-hoc contextualization using Large Language Models (LLMs).

Comprehensive interactive platforms such as *No-IDLE* (Interactive Deep Learning Enterprise) bridge the gap between research prototypes and practical implementation by incorporating multimodal interaction capabilities, particularly human gaze and pointing gestures, to enhance human-machine collaboration [51]. No-IDLE formalizes these capabilities into a unified infrastructure that supports a range of applications, including image captioning and medical diagnostics such as melanoma detection. By enabling adaptive models to incrementally learn from human collaborators, these platforms address the limitations of pre-trained models in accommodating dynamic circumstances and user-specific variations. Through the integration of continuous feedback and collaboration mechanisms, No-IDLE exemplifies how interactive machine learning can be operationalized for real-world, user-centered AI systems.

2.3 Continual Learning for Vision-Language Tasks

These IML advances directly inform our ALCIE framework by establishing the importance of strategic sample selection in human-AI collaborative scenarios, where episodic memory management becomes crucial for maintaining performance while incorporating user feedback across sequential learning episodes.

IC models in sequential learning are vulnerable to CF, where new task acquisition causes loss of previous knowledge [14]. This challenge is acute for image captioning systems that must maintain both visual recognition and linguistic generation abilities across domains [52]. CL methods address this by retaining prior task representations while integrating new information.

This section examines CL methodologies designed to address CF in sequential learning scenarios. We review fundamental approaches and categorization frameworks (subsection 2.3.1), architectural innovations for task-incremental learning (subsection 2.3.2), memory-efficient learning strategies (subsection 2.3.3), parameter-efficient adaptation techniques (subsection 2.3.4), and multimodal continual learning approaches (subsection 2.3.5). Together, these topics provide the theoretical basis for the memory management strategies used in our episodic selection framework.

2.3.1 Fundamental Approaches and Categorization

A comprehensive categorization of CL techniques in NLP is presented [53], outlining core strategies including *rehearsal*, *pseudo-rehearsal*, *regularization*, *memory-based mechanisms*, *knowledge distillation*, and *architectural modifications*. This foundational work addresses the problem of CF and analyzes methods such as EWC [15] and GEM [16], which are commonly used in sequence learning tasks. Although not focused specifically on IC, this research provides critical insights for multimodal continual learning systems, particularly in understanding how models can maintain stability while adapting to new sequential tasks in discrete, compositional, and context-dependent domains.

Building on the foundational principles of CL, attention-based approaches such as the *Recurrent Attention to Transient Tasks (RATT)* [54] model introduce architectural modifications that isolate task-specific representations in LSTM-based captioning systems.

RATT employs task-aware attention masking, activating only relevant subsets of neurons and vocabulary for each task to accommodate the transient nature of vocabularies in continual IC. This structural separation limits interference and reduces forgetting, as demonstrated on continual learning benchmarks derived from MS-COCO and Flickr30k, where RATT successfully learned multiple tasks in sequence without performance loss on earlier tasks. The model’s effectiveness highlights the value of attention-based task separation in CL for IC, providing a structural alternative to memory-based replay methods, especially when task vocabularies overlap or evolve.

2.3.2 Task-Incremental Learning Strategies

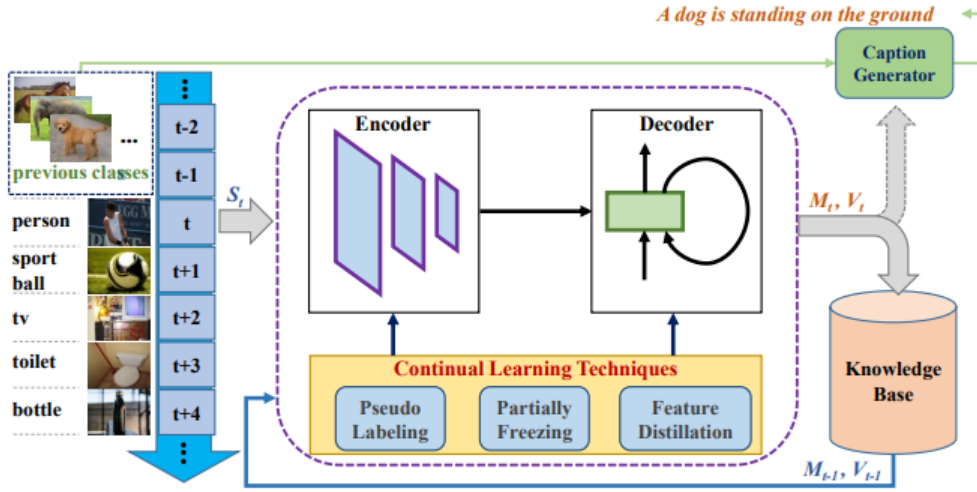


Figure 2.4: ContCap architecture showing the integration of freezing, knowledge distillation, and pseudo-labeling mechanisms for continual image captioning. The framework demonstrates how task-specific knowledge can be preserved while adapting to new captioning domains without catastrophic forgetting.

Early efforts to integrate CL into IC focused on managing knowledge interference through explicit task structuring and targeted adaptation strategies. *Continual Captioning (ContCap)* [52] exemplifies this approach by framing caption generation as a class-incremental problem and combining techniques such as parameter freezing, knowledge distillation, and pseudo-labeling to address CF. By freezing select model parameters, ContCap preserves knowledge from previous tasks; *knowledge distillation* facilitates information transfer between model stages; and *pseudo-labeling* reinforces learning with generated labels. This integrated framework, as illustrated in Figure 2.4, enables effective separation and management of task-specific knowledge. ContCap was evaluated on sequential splits of the MS-COCO 2014 dataset and demonstrated strong performance on both old and new tasks, even without revisiting prior data. This work established a scalable foundation for continual IC, showing that explicit task structure and adaptation can successfully mitigate CF in sequence generation tasks.

2.3.3 Memory-Efficient Learning Strategies

Recent advances in memory efficiency have focused on token-level storage strategies that significantly reduce memory requirements while maintaining learning effectiveness. The *Core Tokensets* approach proposes storing only the most informative tokens from image-caption pairs, selected using attribution scores [4]. By replaying these token fragments rather than entire examples, models retain essential visual and linguistic cues with minimal memory overhead, preserving performance in continual captioning scenarios and supporting compatibility with large transformer-based architectures. Empirical results show that a core tokenset comprising just 1% of the data performs comparably to much larger coresets, highlighting the value of intelligent sample selection in memory-constrained CL. As shown in Figure 2.5, this method enables substantial memory savings while maintaining learning effectiveness across sequential tasks.

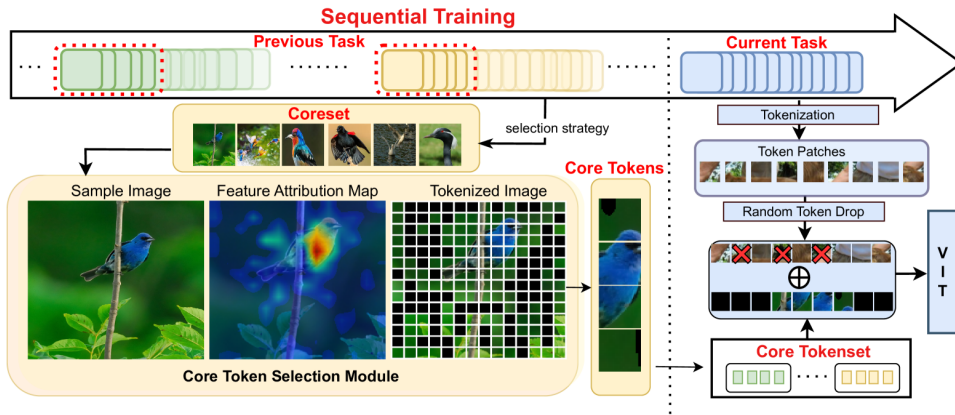


Figure 2.5: Core Tokensets [4] approach for memory-efficient continual learning. The method selects and stores the most informative token fragments from image-caption pairs based on attribution scores, enabling effective knowledge retention with minimal memory overhead.

2.3.4 Parameter-Efficient Adaptation Techniques

MLADIC [55] simultaneously optimizes image captioning and text-to-image synthesis, leveraging task correlation for domain adaptation with paired and unpaired data. While effective for cross-domain transfer, it primarily addresses domain gaps rather than sequential learning with forgetting prevention.

Recent advances have emphasized parameter-efficient approaches that support continual adaptation with minimal computational overhead. *Low-Rank Adaptation with Structured Updates (LoRSU)* enables the adaptation of frozen image encoders in vision-language models under CL settings [56]. Rather than fine-tuning entire models or relying on replay mechanisms, LoRSU introduces compact, *low-rank adapters* to selected attention blocks, guided by sensitivity analysis. Building on this, *Continual Low-Rank Adaptation (C-LoRA)* incorporates a learnable routing matrix to manage updates across sequential tasks, further reducing interference and enabling efficient adaptation in dynamic environments [57]. These approaches allow systems to incorporate new tasks with substantially reduced computational cost, achieving over a 25-fold decrease in overhead compared to full

Vision-Language Model (VLM) updates while maintaining performance. This strategy is particularly well-suited for models such as OFA or BLIP, where decoupling of visual and textual components allows efficient, task-specific updates to the vision encoder while leaving the language components unchanged.

2.3.5 Multimodal Continual Learning

Extending beyond single-modality approaches, recent work has addressed CL scenarios involving multiple modalities and varying task types. *Modality-Inconsistent Continual Learning (MoInCL)* [58] tackles the challenging scenario where tasks involve inconsistent modalities (such as image, audio, or video) and diverse task types (such as IC or Visual Question Answering (VQA)). The method employs a *Pseudo Targets Generation Module* to mitigate forgetting caused by task type shifts in previously seen modalities, while incorporating *Instruction-based Knowledge Distillation* to preserve the model’s ability to handle previously learned modalities when new ones are introduced. This approach supports IC as part of a broader sequence of multimodal tasks and demonstrates improved retention without requiring the storage of previous data. MoInCL effectively manages the complexity of multimodal task sequences, preventing CF across different modalities and task types.

2.4 Active Learning for Sample Selection

While CL techniques enable adaptation without CF [52], they don’t address the substantial annotation costs in IC, where full sentence generation is required rather than simple labeling. AL addresses this by strategically selecting the most informative samples for annotation, maximizing performance while minimizing labeling costs [22].

The core CL challenge is determining which experiences to retain in episodic memory. While random sampling is computationally simple, it fails to capture the most informative samples for learning and knowledge retention [16]. AL provides principled episodic memory selection through uncertainty, diversity, or forgetting-prevention criteria.

This section explores AL techniques that enable principled sample selection to maximize learning efficiency. We review uncertainty-based selection strategies (subsection 2.4.1), diversity-based approaches (subsection 2.4.2), and hybrid AL methods (subsection 2.4.3), all of which inform our episodic memory selection strategies for continual IC. These foundational approaches provide the basis for the development of our ALCIE framework.

2.4.1 Uncertainty-Based Selection Strategies

Uncertainty-based sampling focuses on selecting data points where the model exhibits the lowest confidence, typically measured by entropy or margin scores [59, 60]. This approach targets samples where the model is most uncertain about its predictions, as these represent instances where additional training data is likely to yield the greatest improvement in model performance [61].

Monte Carlo dropout has emerged as an important method for uncertainty modeling in deep neural networks, enabling practical *Bayesian inference* by approximating deep Gaussian processes [60]. This technique applies dropout during both training and inference to generate multiple stochastic predictions. In the context of IC, models such

as the *Uncertainty-Aware Image Captioning (UAIC)* framework [62] utilize Monte Carlo dropout by performing several stochastic forward passes, resulting in diverse candidate captions for a given image. The variance among these candidates serves as an indicator of model uncertainty, with higher variance suggesting greater ambiguity and a higher potential benefit from additional annotations for those samples.

For multimodal uncertainty estimation in IC, token-level entropy measures provide a more granular uncertainty assessment than global prediction confidence. Recent work on AL for natural language generation has demonstrated the effectiveness of informativeness strategies such as MTE and Monte Carlo Dropout for identifying the most informative data points [63]. The MTE approach measures the average uncertainty in model predictions by computing entropy at each token position during caption generation and aggregating these values to provide a comprehensive uncertainty measure that captures both lexical and semantic uncertainty throughout the generation process. This method proves particularly effective for identifying samples where the model struggles with specific aspects of visual understanding or language generation [29].

2.4.2 Diversity-Based Selection Strategies

Diversity-based methods aim to select examples that are as distinct from each other as possible, thereby ensuring broad coverage of the data space and minimizing redundancy [24]. This approach addresses a key limitation of uncertainty-based sampling, which can sometimes yield redundant samples from concentrated high-uncertainty regions. The central principle is to maximize the representativeness of selected samples across the entire feature space. In the context of IC, diversity-based selection is especially important, as visual scenes often differ substantially in content, composition, and semantic complexity. Comprehensive coverage of these variations is critical for achieving robust model performance across diverse image types [63].

CLIP [10] feature representations provide a strong foundation for diversity-based selection in vision-language tasks. The CLIP model learns joint embeddings for images and text using *contrastive learning* on large-scale datasets, resulting in semantically rich representations that capture both visual and textual characteristics. These embeddings enable effective diversity measurement across modalities, supporting more sophisticated sample selection strategies than those relying on traditional visual features alone.

Recent advances in diversity-based selection have centered on *multimodal embeddings* that jointly capture visual and textual characteristics of image-caption pairs [24, 64]. By leveraging pre-trained vision-language models, these methods produce rich representations that enable more effective measurement of diversity across both modalities [65]. Integrating semantic diversity measures with traditional feature-space approaches has yielded improved performance in IC tasks, particularly for domain-specific content where conventional visual features may overlook important semantic variations [66]. To further ensure comprehensive data coverage, strategies such as Greedy Core-Set and In-Domain Diversity Sampling (IDDS) have been proposed [63]. Greedy Core-Set focuses on covering the entire feature space, while IDDS targets uncertain areas near decision boundaries, offering complementary benefits to CLIP-based clustering for achieving comprehensive semantic coverage.

2.4.3 Hybrid Active Learning Strategies

Hybrid approaches combine *uncertainty* and *diversity* based selection, choosing samples that are both uncertain and diverse, a strategy shown to improve performance in various settings [67]. By addressing the limitations of using either strategy alone, hybrid methods support more balanced and effective sample selection. Integrating these criteria requires careful weighting, as uncertainty-based methods target challenging examples while diversity-based selection ensures broad data coverage. Effective hybrid strategies adapt this balance according to the model’s state and data characteristics [25].

Among the frameworks implementing these hybrid principles, *HUDS* stands out as a principled approach for balancing uncertainty and diversity in sample selection [25]. Originally developed for *machine translation*, HUDS can be adapted to multimodal domains by incorporating MTE for uncertainty estimation and CLIP-based clustering for diversity assessment. This enables hybrid selection strategies that balance the exploration of uncertain regions with comprehensive semantic coverage. The versatility of HUDS extends beyond neural machine translation to sequence generation tasks such as IC, offering valuable insights into managing the trade-off between exploration and exploitation. As a result, this framework serves as a foundation for developing specialized hybrid methods tailored to the unique requirements of vision-language applications.

Uncertainty-Weighted Embeddings (UWE) [68] weight feature representations with predictive uncertainty before diversity sampling, addressing pure uncertainty sampling’s tendency towards redundant selections in high-density feature regions while maintaining scalability for dense prediction tasks.

The integration of these hybrid approaches with episodic memory selection in continual learning scenarios presents unique opportunities for optimizing knowledge retention while minimizing CF. By carefully balancing uncertainty and diversity criteria, these methods can identify samples that not only improve current task performance but also contribute to maintaining performance on previously learned tasks, making them particularly valuable for continual IC applications.

2.5 Evaluation and Datasets

Datasets provide the foundation for image captioning research, directly influencing model performance and generalization capabilities [32]. Two categories emerge: *generic datasets* capturing diverse natural scenes [69, 5], and *domain-specific datasets* targeting particular applications or user populations [28, 70], as illustrated in Figure 2.6.

This section examines generic image captioning datasets (Section 2.5.1), including MS-COCO, Flickr30k, and Conceptual Captions, domain-specific datasets (Section 2.5.2) such as FACAD, VizWiz, nocaps, and TextCaps, and evaluation metrics (Section 2.5.3) ranging from traditional lexical measures to modern neural-based assessment approaches.

2.5.1 Generic Image Captioning Datasets

Generic datasets provide comprehensive coverage of diverse visual content and serve as standard benchmarks for evaluating model performance across various image types and captioning scenarios [33, 31, 30]. These datasets typically contain images of everyday scenes, common objects, and general human activities, making them suitable for developing models with broad applicability.

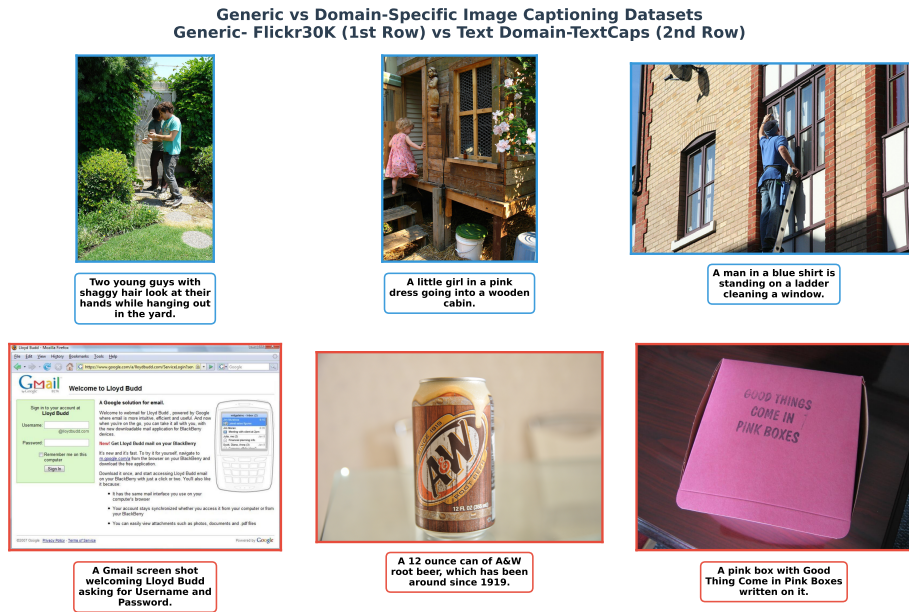


Figure 2.6: Generic vs domain-specific dataset comparison: Flickr30k [5] focuses on human activities with descriptive captions, while TextCaps [6] specializes in text-rich images requiring OCR-based descriptions.

MS-COCO (Microsoft Common Objects in Context) [69] is one of the most influential and widely adopted datasets in image captioning research [33, 31]. The dataset contains *328,000 images with 2.5 million labeled instances across 91 object categories*. Each image is annotated with multiple human-generated captions describing objects, scenes, and activities. MS-COCO images are carefully selected to contain multiple objects in natural contexts, providing rich visual content for caption generation. The annotation process involved extensive crowdsourcing with strict quality control to ensure caption accuracy and diversity. MS-COCO has become the de facto standard for image captioning evaluation, with most papers reporting performance using standard evaluation metrics.

Flickr30k [5] comprises 31,000 images sourced from Flickr, each paired with five independently collected human-written captions. The dataset focuses on people engaged in everyday activities, providing diverse descriptions of scenes, actions, and objects. Unlike MS-COCO’s emphasis on object detection, *Flickr30k prioritizes human activities and social interactions, making it valuable for evaluating models’ ability to describe complex human behaviors and relationships*. The Flickr30k Entities extension further enhances evaluation by providing region-to-phrase correspondences, enabling fine-grained assessment of visual grounding capabilities.

Conceptual Captions [71] represents a significant scale-up in dataset size, containing approximately *3.3 million image-caption pairs* extracted and filtered from billions of web pages. Unlike curated datasets like MS-COCO, Conceptual Captions uses an automatic pipeline to harvest image-caption pairs from the web via Alt-text HTML attributes, resulting in more varied, naturally occurring language. Conceptual Captions is widely recognized for enabling the training of models that generalize better across diverse image types and linguistic styles [33].

2.5.2 Domain-Specific Image Captioning Datasets

Domain-specific datasets focus on particular application areas, specialized visual content, or specific user populations, enabling targeted model development and evaluation for practical implementation scenarios. These datasets address unique challenges and requirements that may not be adequately covered by generic datasets.

FACAD [28] represents the first comprehensive dataset specifically designed for fashion image captioning. The dataset contains 993,000 images with 130,000 corresponding descriptions exhibiting three distinctive characteristics compared to general datasets. First, fashion captioning requires describing fine-grained attributes rather than scene relationships. Second, FACAD contains longer captions with an average of 21 words versus 10.4 words in MS-COCO, reflecting detailed attribute descriptions. Third, the expression style uses engaging descriptive terms like "pearly," "so-simple yet so-chic," and "retro flair" rather than plain descriptions, important for commercial applications where captivating descriptions enhance customer engagement and sales conversion.

Visual Question Answering from Blind People (VizWiz) [70] is a groundbreaking dataset designed specifically to address the needs of people with visual impairments. The VizWiz-Captions dataset contains 39,181 *images* originating from people who are blind, each paired with five captions, representing the first image captioning dataset from this practical use case. The dataset presents unique challenges, including blurred images, extreme lighting conditions, and unconventional viewpoints that reflect the practical difficulties faced by users with visual impairments.

Novel Object Captioning at Scale (NoCaps) (Novel Object Captioning at Scale) [72] addresses the challenge of describing objects that were not seen during training. The dataset contains 166,100 *images with 15 human-generated captions per image*, specifically designed to evaluate models' ability to describe novel objects using knowledge gained from pre-training. The images are selected to contain objects from the Open Images dataset that do not overlap with MS-COCO's training set, creating a controlled setting for evaluating generalization capabilities. This dataset has become particularly valuable for assessing zero-shot and few-shot learning capabilities in modern foundation models.

Text in Images Caption Dataset (TextCaps) [6] focuses on reading and reasoning about textual content in images. The dataset contains 145,000 *images with captions* that require understanding and incorporating text present in the images, such as signs, labels, and documents. This represents a significant challenge beyond traditional object recognition, requiring models to perform optical character recognition and integrate textual information into natural language descriptions.

2.5.3 Evaluation Metrics

To evaluate the quality of generated captions, researchers employ a variety of automatic metrics that assess different aspects of caption quality and semantic alignment [32, 31]. Traditional lexical similarity metrics focus on surface-level correspondence between generated and reference captions, while more sophisticated approaches evaluate semantic content and structural relationships.

Bilingual Evaluation Understudy (BLEU) [73] assesses *n-gram overlap* with reference captions, computing precision scores for unigrams through 4-grams and applying a brevity penalty to prevent artificially high scores from short captions. While BLEU provides a straightforward measure of lexical overlap, it may penalize semantically

correct but lexically different captions, limiting its effectiveness for evaluating creative or diverse caption generation.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [74] addresses some limitations of BLEU by accounting for synonymy through WordNet and considering word order through alignment-based scoring. METEOR incorporates *stemming*, *synonymy matching*, and *paraphrase recognition*, making it more robust to lexical variations while maintaining correlation with human judgments of caption quality.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [75] measures overlap between model outputs and reference text using the *longest common subsequence* and *skip-bigram statistics*, originally developed for summarization evaluation but adapted for captioning tasks. ROUGE-L focuses on sentence-level structure similarity, while ROUGE-S evaluates skip-bigram co-occurrence patterns, providing complementary perspectives on caption quality.

Consensus-based Image Description Evaluation (CIDEr) [76] quantifies similarity to human-generated captions across multiple references, placing greater emphasis on consensus among human annotators. CIDEr uses *Term Frequency-Inverse Document Frequency (TF-IDF)* weighting, a scheme that assigns higher importance to words that are frequent within a given caption but rare across the entire set of reference captions.

The TF-IDF weighting mechanism operates through two components: *Term Frequency (TF)* measures how often a word appears within a specific caption, while *Inverse Document Frequency (IDF)* measures word rarity across the entire corpus. Words appearing in fewer captions receive higher weights, while common words receive lower weights.

Semantic Propositional Image Caption Evaluation (SPICE) [77] evaluates semantic content by comparing *scene graphs* extracted from candidate and reference captions, focusing on objects, attributes, and relationships rather than surface-level lexical similarity. SPICE parses captions into semantic representations and computes F-scores based on graph structure overlap, providing a more semantically grounded evaluation that correlates well with human assessments of caption quality and factual accuracy.

Bidirectional Encoder Representations from Transformers Score (BERTScore) [78] stands out by computing the similarity between generated and reference captions using contextualized embeddings from BERT-based models. This enables the metric to capture semantic equivalence that extends well beyond surface-level lexical overlap, making it effective for recognizing paraphrases and semantically similar expressions that might be overlooked by traditional metrics.

Chapter 3

Technical Background

This chapter establishes the essential theoretical foundations underlying the ALCIE framework presented in this thesis. The work synthesizes concepts from vision-language modeling, CL, and AL to address strategic episodic memory management in multimodal systems. A comprehensive understanding of these fundamental concepts is crucial for comprehending the proposed methodologies and their theoretical justifications.

This chapter systematically introduces the key technical components required for understanding ALCIE. It begins with vision-language architectures (Section 3.1), then covers CL mechanisms (Section 3.2), and proceeds to AL integration strategies (Section 3.3). The chapter concludes with a discussion of evaluation frameworks essential for assessing ALCIE performance (Section 3.4).

3.1 Vision-Language Model Architectures

Contemporary IC systems rely on sophisticated vision-language architectures that effectively align visual and textual representations. The ALCIE framework operates on two distinct architectural paradigms: modular frozen-component designs exemplified by BLIP-2 [3] and unified transformer frameworks such as OFA [2]. Understanding these architectures is essential for comprehending how AL strategies can be integrated into multimodal CL systems.

This section examines vision-language architectures from early attention-based models (Section 3.1.1) to transformer foundations (Section 3.1.2), cross-modal attention mechanisms (Section 3.1.3), BLIP-2’s modular design (Section 3.1.4), and OFA’s unified framework (Section 3.1.5).

3.1.1 Evolution from Attention-Based Captioning to Transformers

Before examining modern transformer architectures, it is crucial to understand the evolution of attention mechanisms in IC. The seminal work *Show, Attend and Tell* [1] introduced foundational concepts of visual attention, enabling models to selectively

focus on relevant regions of an image while generating captions. This approach marked a significant shift from fixed, global feature representations and laid the groundwork for more flexible attention strategies that later influenced transformer development.

Show, Attend and Tell: Foundation of Visual Attention

The *Show, Attend and Tell* architecture represents a pivotal advancement in IC, introducing spatial attention mechanisms that enable models to focus on relevant image regions during caption generation. This work established the mathematical foundations that later evolved into modern transformer architectures.

Architecture Overview: The model consists of a CNN encoder for visual feature extraction and a LSTM decoder with an attention mechanism for caption generation.

Visual Feature Extraction: Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, the CNN encoder, typically VGGNet-19 [79] or ResNet [80], produces a set of feature vectors:

$$\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L\}, \quad \mathbf{a}_i \in \mathbb{R}^D \quad (3.1)$$

where L represents the number of spatial locations (typically $L = H' \times W'$ after convolution and pooling), and D denotes the feature dimension.

Attention Mechanism: The *Show, Attend and Tell* architecture introduces two distinct attention mechanisms: *soft attention* and *hard attention*, each with different computational and training characteristics, as illustrated in Figure 3.1.

Soft Attention (Deterministic): At each time step t during caption generation, the model computes attention weights $\alpha_{t,i}$ that determine the importance of each spatial location:

$$e_{t,i} = f_{att}(\mathbf{a}_i, \mathbf{h}_{t-1}) \quad (3.2)$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^L \exp(e_{t,k})} \quad (3.3)$$

$$\hat{\mathbf{z}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_{t,i}\}) = \sum_{i=1}^L \alpha_{t,i} \mathbf{a}_i \quad (3.4)$$

where f_{att} is a multilayer perceptron that computes attention energies, \mathbf{h}_{t-1} represents previous hidden state of the LSTM, $\alpha_{t,i}$ denotes normalized attention weight for location i at time t , and $\hat{\mathbf{z}}_t$ is the context vector representing the attended visual information.

Hard Attention (Stochastic): In contrast to soft attention, hard attention selects a single spatial location s_t at each time step based on the attention distribution:

$$s_t \sim \text{Multinoulli}(\{\alpha_{t,i}\}_{i=1}^L) \quad (3.5)$$

$$\hat{\mathbf{z}}_t = \mathbf{a}_{s_t} \quad (3.6)$$

where s_t represents the selected location index sampled from a multinomial distribution parameterized by the attention weights $\{\alpha_{t,i}\}_{i=1}^L$, and $\hat{\mathbf{z}}_t$ is the feature vector at the selected location.

Training Differences: Both attention mechanisms require different training approaches:

Soft Attention Training: Since the expectation is differentiable, soft attention can be trained using standard backpropagation:

$$\mathcal{L}_{\text{soft}} = - \sum_{t=1}^T \log p(y_t | y_{<t}, \hat{\mathbf{z}}_t, \mathbf{h}_{t-1}) \quad (3.7)$$

where the context vector $\hat{\mathbf{z}}_t$ is a weighted combination of all spatial features.

Hard Attention Training: Due to the non-differentiable sampling operation, hard attention requires reinforcement learning techniques.

$$\mathcal{L}_{\text{hard}} = - \sum_{t=1}^T \mathbb{E}_{s_t \sim p(s_t | \mathbf{a}, \mathbf{h}_{t-1})} [\log p(y_t | y_{<t}, \mathbf{a}_{s_t}, \mathbf{h}_{t-1})] \quad (3.8)$$

$$\nabla_{\theta} \mathcal{L}_{\text{hard}} \approx - \sum_{t=1}^T (\log p(y_t | y_{<t}, \mathbf{a}_{s_t}, \mathbf{h}_{t-1}) - b_t) \nabla_{\theta} \log p(s_t | \mathbf{a}, \mathbf{h}_{t-1}) \quad (3.9)$$

where b_t is a baseline to reduce variance, typically the expected reward under the current policy.

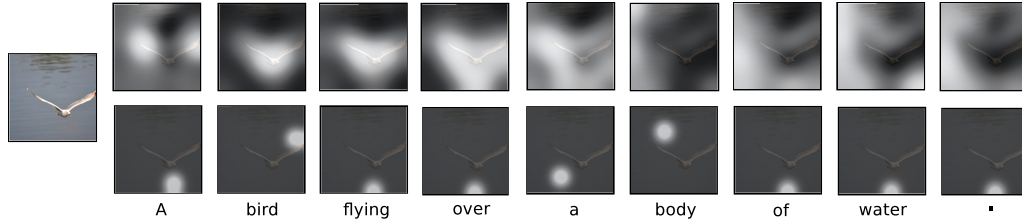


Figure 3.1: Comparison of soft and hard attention in the *Show, Attend and Tell* model [1]. **Top:** Soft attention assigns weights to all spatial locations (heat maps), blending features smoothly. **Bottom:** Hard attention selects one location at each step (points), yielding focused, discrete selections. Generated captions illustrate the influence of each mechanism.

Caption Generation: LSTM decoder incorporates the context vector to generate words:

$$\mathbf{h}_t = \text{LSTM}(\mathbf{E}y_{t-1}, \mathbf{h}_{t-1}, \hat{\mathbf{z}}_t) \quad (3.10)$$

$$p(y_t | y_1, \dots, y_{t-1}, \mathbf{I}) = \text{softmax}(\mathbf{L}_o(\mathbf{E}y_{t-1} + \mathbf{L}_h \mathbf{h}_t + \mathbf{L}_z \hat{\mathbf{z}}_t)) \quad (3.11)$$

where $\mathbf{E} \in \mathbb{R}^{m \times |\mathcal{V}|}$ represents the word embedding matrix, y_{t-1} denotes the previous word, $\mathbf{L}_o, \mathbf{L}_h, \mathbf{L}_z$ are learned projection matrices, and $|\mathcal{V}|$ represents the vocabulary size.

Connection to Modern Transformers: Attention mechanisms introduced in *Show, Attend and Tell* laid the groundwork for many principles adopted in transformer architectures:

- **Attention as Weighted Aggregation:** The core idea of computing attention weights and using them to aggregate information, particularly through soft attention.
- **Context-Dependent Attention:** Attention weights are determined by both the current decoder state and the input features, allowing dynamic context adaptation.

- **Differentiable Attention:** Modern transformers utilize differentiable soft attention rather than stochastic hard attention, enabling stable, end-to-end training.
- **Learning Attention Distributions:** Attention patterns are learned directly by neural networks, rather than being hand-engineered.

3.1.2 Transformer Foundations for Multimodal Learning

The *Transformer architecture* [7], illustrated in Figure 3.2, represents a fundamental paradigm shift in sequence modeling, moving from recurrent architectures to attention-based mechanisms that enable parallel processing and superior modeling of long-range dependencies. Understanding the transformer's core components, such as multihead self-attention, positional encoding, and feed-forward networks, is essential for understanding how modern vision language models process and align multimodal information. This knowledge supports the development of systems that deliver more accurate and contextually relevant outputs across a wide range of applications.

The transformer architecture generalizes the attention concepts from *Show, Attend and Tell* [1] by introducing *self-attention*, where sequences can attend to themselves, and *multi-head attention*, which enables multiple attention patterns simultaneously. It consists of an encoder-decoder structure where each component is constructed from stacked layers of attention and feed-forward networks. The encoder processes input sequences to create contextualized representations, while the decoder generates output sequences through a combination of self-attention and cross-attention mechanisms.

Core Transformer Architecture

Encoder Structure: Each encoder layer ℓ transforms the incoming representations through two successive components:

$$\mathbf{Z}^{(\ell)} = \text{LayerNorm} \left(\mathbf{X}^{(\ell-1)} + \text{MultiHead} \left(\mathbf{X}^{(\ell-1)} \right) \right) \quad (3.12)$$

$$\mathbf{X}^{(\ell)} = \text{LayerNorm} \left(\mathbf{Z}^{(\ell)} + \text{FFN} \left(\mathbf{Z}^{(\ell)} \right) \right) \quad (3.13)$$

where $\mathbf{X}^{(\ell-1)} \in \mathbb{R}^{n \times d}$ contains encoder states for sequence length n and hidden dimension d , $\text{MultiHead}(\cdot)$ denotes multi-head self-attention, $\text{FFN}(\cdot)$ represents the position-wise feed-forward network, and $\text{LayerNorm}(\cdot)$ applies layer normalization.

Decoder Structure: The decoder incorporates an additional *cross-attention* mechanism between masked self-attention and the feed-forward network:

$$\mathbf{Y}_1^{(\ell)} = \text{LayerNorm} \left(\mathbf{Y}^{(\ell-1)} + \text{MaskedMultiHead} \left(\mathbf{Y}^{(\ell-1)} \right) \right) \quad (3.14)$$

$$\mathbf{Y}_2^{(\ell)} = \text{LayerNorm} \left(\mathbf{Y}_1^{(\ell)} + \text{CrossAttention} \left(\mathbf{Y}_1^{(\ell)}, \mathbf{X}^{(L)} \right) \right) \quad (3.15)$$

$$\mathbf{Y}^{(\ell)} = \text{LayerNorm} \left(\mathbf{Y}_2^{(\ell)} + \text{FFN} \left(\mathbf{Y}_2^{(\ell)} \right) \right) \quad (3.16)$$

where $\mathbf{Y}^{(\ell)} \in \mathbb{R}^{m \times d}$ stores decoder states for target sequence length m . Here, $\text{MaskedMultiHead}(\cdot)$ applies masked multi-head self-attention, $\text{CrossAttention}(\cdot)$ per-

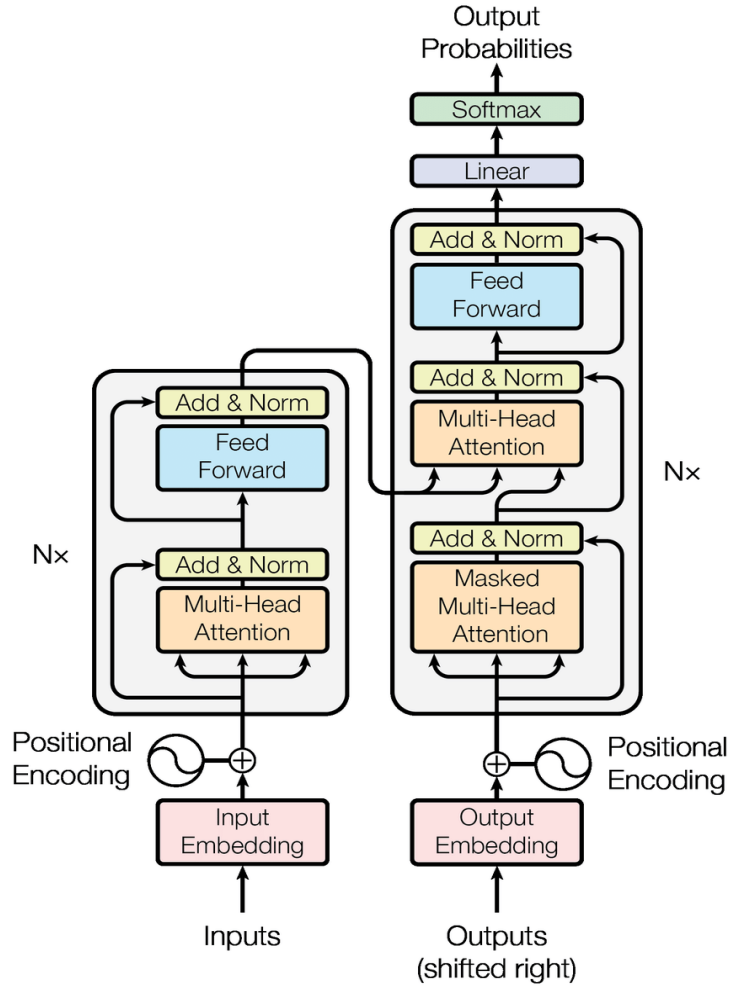


Figure 3.2: Complete Transformer architecture [7] showing encoder-decoder structure with self-attention, cross-attention, and feed-forward components. The parallel processing capability and attention mechanisms form the foundation for modern vision-language models.

forms cross-attention over the final encoder output $\mathbf{X}^{(L)}$, and L denotes the total number of encoder layers.

Self-Attention Mechanism

The *self-attention* mechanism constitutes the core innovation of transformers, enabling direct modeling of dependencies between all positions in a sequence regardless of their distance. This capability is particularly crucial for vision-language tasks where long-range dependencies exist both within and across modalities.

For an input sequence represented as matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n denotes sequence length and d represents embedding dimension, the self-attention mechanism computes:

$$\mathbf{Q} = \mathbf{XW}^Q, \quad \mathbf{K} = \mathbf{XW}^K, \quad \mathbf{V} = \mathbf{XW}^V \quad (3.17)$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d_k}}\right) \mathbf{V} \quad (3.18)$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices and $d_k = d/h$ for h attention heads.

The attention computation can be decomposed into three conceptual phases:

Similarity Computation: The query-key dot products measure similarity between all position pairs:

$$\mathbf{S}_{ij} = \frac{(\mathbf{QK}^\top)_{ij}}{\sqrt{d_k}} = \frac{\sum_{k=1}^{d_k} Q_{ik} K_{jk}}{\sqrt{d_k}} \quad (3.19)$$

where \mathbf{S}_{ij} represents the similarity between position i and position j , and the scaling factor $\frac{1}{\sqrt{d_k}}$ prevents softmax saturation for large embedding dimensions.

Attention Weight Normalization: Softmax normalization ensures attention weights sum to unity:

$$\alpha_{ij} = \frac{\exp(\mathbf{S}_{ij})}{\sum_{k=1}^n \exp(\mathbf{S}_{ik})} \quad (3.20)$$

where α_{ij} represents the attention weight from position i to position j .

Weighted Value Aggregation: Output combines values according to attention weights:

$$\mathbf{O}_i = \sum_{j=1}^n \alpha_{ij} \mathbf{V}_j \quad (3.21)$$

where $\mathbf{O}_i \in \mathbb{R}^{d_k}$ represents the output representation for position i .

Multi-Head Attention

Multi-head attention extends the basic attention mechanism by computing multiple attention functions in parallel, as shown in Figure 3.3, each focusing on different representational subspaces. This design enables the model to attend to different types of relationships simultaneously, such as: (1) Spatial relationships between visual regions, (2) Semantic relationships between words, (3) Cross-modal correspondences between visual and textual elements, and (4) Syntactic dependencies in natural language.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O \quad (3.22)$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V) \quad (3.23)$$

The head-specific projection matrices are $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d \times d_k}$ where $d_k = d/h$, and the output projection matrix is $\mathbf{W}^O \in \mathbb{R}^{d \times d}$.

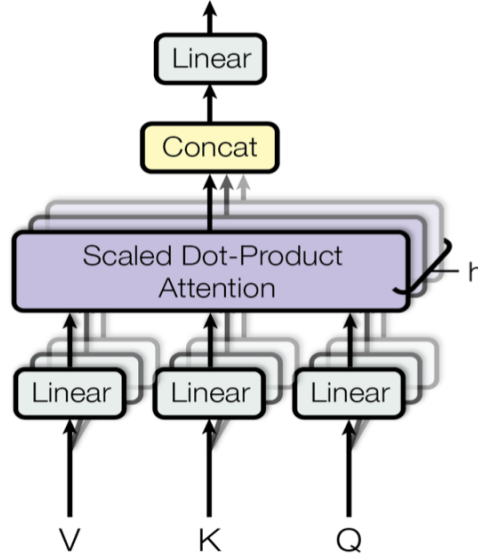


Figure 3.3: Multi-head attention visualization showing different attention patterns across heads. Each head captures different types of relationships, enabling comprehensive sequence understanding crucial for multimodal tasks [7].

Feed-Forward Networks and Layer Components

Each transformer layer incorporates a position-wise feed-forward network that processes each position independently:

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (3.24)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{ff}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{ff} \times d}$ are weight matrices, $\mathbf{b}_1 \in \mathbb{R}^{d_{ff}}$, $\mathbf{b}_2 \in \mathbb{R}^d$ are bias vectors, and d_{ff} denotes the feed-forward dimension (typically $d_{ff} = 4d$).

Layer Normalization: Applied using the pre-normalization variant:

$$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma} \odot \gamma + \beta \quad (3.25)$$

where $\mu = \frac{1}{d} \sum_{i=1}^d x_i$ and $\sigma = \sqrt{\frac{1}{d} \sum_{i=1}^d (x_i - \mu)^2}$ represent the mean and standard deviation computed across the feature dimension, $\gamma, \beta \in \mathbb{R}^d$ are learnable scale and shift parameters, and \odot denotes element-wise multiplication.

Residual Connections: Enable training of deep networks by providing gradient flow:

$$\text{Output} = \text{Input} + \text{SubLayer}(\text{LayerNorm}(\text{Input})) \quad (3.26)$$

Positional Encoding

Since attention mechanisms are permutation-invariant, transformers require explicit positional information to understand sequence order. The original transformer employs sinusoidal positional encodings:

$$\text{PE}_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right) \quad (3.27)$$

$$\text{PE}_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (3.28)$$

where $pos \in \{0, 1, \dots, n-1\}$ denotes the position index in the sequence, $i \in \{0, 1, \dots, \lfloor d/2 \rfloor - 1\}$ represents the dimension index, and d is the embedding dimension.

The sinusoidal encoding possesses several important properties:

- **Deterministic:** No learned parameters required
- **Extrapolation:** Can handle sequences longer than the training data
- **Relative positioning:** PE_{pos+k} can be expressed as a linear function of PE_{pos}

3.1.3 Cross-Modal Attention Mechanisms

Cross-modal attention enables alignment between visual regions and textual elements by extending self-attention across modalities [41]. Given visual features $\mathbf{H}_v \in \mathbb{R}^{L \times d_v}$ and textual features $\mathbf{H}_t \in \mathbb{R}^{T \times d_t}$:

$$\mathbf{H}_v^{\text{new}} = \text{softmax}\left(\frac{(\mathbf{H}_v \mathbf{W}^{Q_v})(\mathbf{H}_t \mathbf{W}^{K_t})^\top}{\sqrt{d_k}}\right) \mathbf{H}_t \mathbf{W}^{V_t} \quad (3.29)$$

This bidirectional attention mechanism provides comprehensive cross-modal context essential for vision-language understanding.

3.1.4 BLIP-2 Modular Architecture

BLIP-2 [3] exemplifies the frozen-component paradigm in vision-language modeling, employing a three-component design that balances parameter efficiency with performance. This modular architecture is particularly relevant for CL as it isolates adaptation to specific components while preserving robust pre-trained representations.

Frozen Vision Encoder

The vision encoder utilizes a *Vision Transformer (ViT)* (ViT-L/14) [8] with 304 million parameters that remain frozen during fine-tuning, as illustrated in Figure 3.4. Images are partitioned into non-overlapping patches of size $P \times P$, then linearly projected into embedding space:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}} \quad (3.30)$$

where $\mathbf{x}_p^i \in \mathbb{R}^{P^2 \cdot C}$ represents the i -th flattened image patch with C channels, $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ denotes the patch embedding projection matrix, $\mathbf{x}_{\text{class}}$ is a learnable classification token, $N = \frac{H \cdot W}{P^2}$ represents the number of patches for an image of dimensions $H \times W$, and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ provides positional embeddings encoding spatial relationships between patches.

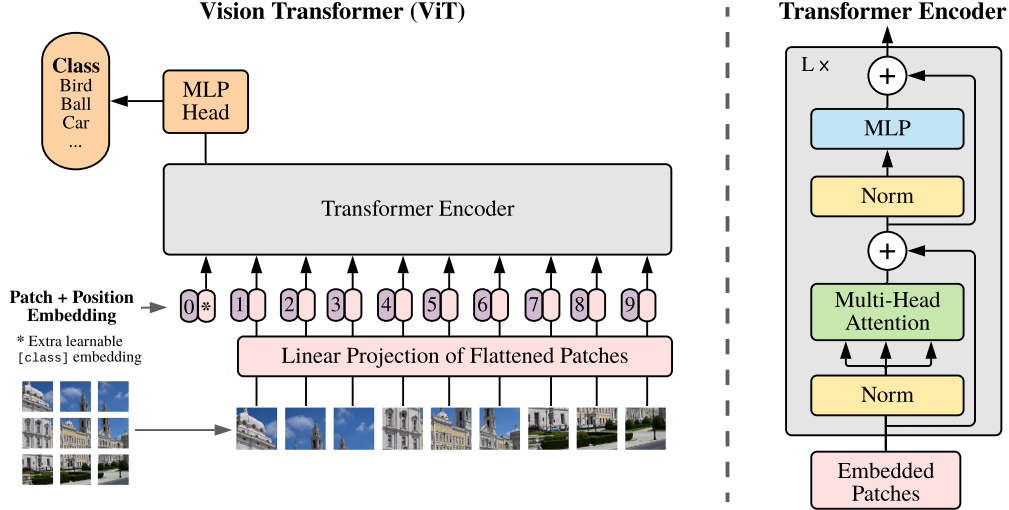


Figure 3.4: ViT [8] architecture used as the frozen vision encoder in BLIP-2. Input images are divided into non-overlapping patches (typically 16×16), flattened, and linearly projected into embeddings. A [CLS] token and positional embeddings are added before processing through the Transformer encoder layers.

The frozen encoder preserves robust visual representations learned during large-scale pretraining, effectively preventing CF at the visual feature level while maintaining computational efficiency throughout the continual learning process.

Q-Former Information Bottleneck

The *Q-Former* serves as a learnable information bottleneck between frozen components, utilizing 188 million trainable parameters. It employs a set of learnable query tokens $\mathbf{Q} \in \mathbb{R}^{N_q \times D}$, where N_q denotes the number of queries (typically 32) and D represents the hidden dimension, that extract relevant visual information:

$$\mathbf{Z} = \text{Transformer}(\mathbf{Q}, \mathbf{V}_{\text{vision}}) \quad (3.31)$$

where $\mathbf{V}_{\text{vision}} \in \mathbb{R}^{N \times D}$ represents frozen vision encoder outputs and $\mathbf{Z} \in \mathbb{R}^{N_q \times D}$ are the processed query representations.

The Q-Former implements both self-attention among queries and cross-attention to vision features:

$$\mathbf{Q}_{\text{self}} = \text{SelfAttention}(\mathbf{Q}) \quad (3.32)$$

$$\mathbf{Z} = \text{CrossAttention}(\mathbf{Q}_{\text{self}}, \mathbf{V}_{\text{vision}}) \quad (3.33)$$

This design enables selective information extraction while maintaining a fixed-size output representation regardless of input image complexity.

Frozen Language Model Integration

The *language model* (OPT-2.7B) [81] remains frozen, with Q-Former outputs serving as soft visual prompts. The integration follows:

$$\mathbf{P}_{\text{visual}} = \mathbf{Z}\mathbf{W}_{\text{proj}} \quad (3.34)$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{D \times D_{\text{LM}}}$ projects Q-Former outputs to the language model’s embedding space with dimension D_{LM} .

During generation, visual prompts are concatenated with text embeddings:

$$\mathbf{H}_{\text{input}} = [\mathbf{P}_{\text{visual}}; \mathbf{E}_{\text{text}}] \quad (3.35)$$

where $\mathbf{E}_{\text{text}} \in \mathbb{R}^{T \times D_{\text{LM}}}$ represents tokenized text embeddings. The language model generates captions autoregressively:

$$p(y_t | y_{<t}, \mathbf{x}) = \text{softmax}(\mathbf{W}_{\text{out}}\mathbf{h}_t + \mathbf{b}_{\text{out}}) \quad (3.36)$$

where \mathbf{h}_t denotes the hidden state at time step t , $\mathbf{W}_{\text{out}} \in \mathbb{R}^{|\mathcal{V}| \times D_{\text{LM}}}$ represents the output projection matrix, and $|\mathcal{V}|$ is the vocabulary size.

3.1.5 OFA Unified Transformer Framework

OFA [2] represents the unified paradigm that processes both visual and textual modalities through shared transformer parameters. This architecture provides a contrasting approach to BLIP-2’s modularity, enabling validation of AL benefits across different optimization landscapes.

Unified Encoder-Decoder Design

OFA employs a *transformer encoder-decoder* architecture where both modalities flow through shared parameters. The encoder processes concatenated multimodal sequences:

$$\mathbf{H}_{\text{enc}} = \text{TransformerEncoder}([\mathbf{x}_{\text{visual}}; \mathbf{x}_{\text{text}}]) \quad (3.37)$$

where $\mathbf{x}_{\text{visual}} \in \mathbb{R}^{N_v \times d}$ represents visual token embeddings with N_v visual tokens, $\mathbf{x}_{\text{text}} \in \mathbb{R}^{N_t \times d}$ represents textual token embeddings with N_t text tokens, and $\mathbf{H}_{\text{enc}} \in \mathbb{R}^{(N_v + N_t) \times d}$ denotes the encoder output.

Task-Agnostic Instruction Framework

OFA transforms IC into sequence-to-sequence generation through instruction prompts. For IC, the input format becomes:

$$\text{Input: } [\text{task_instruction}; \mathbf{x}_{\text{visual}}; \mathbf{x}_{\text{text}}] \quad (3.38)$$

where the task instruction might be “what does the image describe?” concatenated with image patches and optional text prompts. The loss function combines cross-entropy terms across all output tokens:

$$\mathcal{L}_{\text{OFA}} = - \sum_{t=1}^T \log p(y_t | y_{<t}, \mathbf{x}_{\text{visual}}, \mathbf{x}_{\text{text}}; \theta) \quad (3.39)$$

where T denotes the target sequence length and θ represents all model parameters.

3.2 Continual Learning Fundamentals

CL addresses the fundamental challenge of acquiring new knowledge while retaining previously learned information. In the context of vision-language models, this problem becomes particularly complex due to the necessity of maintaining coherent representations across both visual and linguistic modalities while adapting to new domains.

This section covers catastrophic forgetting in neural networks (Section 3.2.1), including multimodal forgetting dynamics and episodic memory mechanisms (Section 3.2.2) that enable knowledge preservation through strategic rehearsal.

3.2.1 Catastrophic Forgetting in Neural Networks

CF represents the most significant challenge in CL, where neural networks experience significant performance degradation on previously learned tasks when adapting to new ones [14]. This phenomenon arises from the distributed nature of neural network representations, where weights contribute to multiple learned functions simultaneously, as illustrated in Figure 3.5.

Formally, consider a neural network f_θ with parameters θ trained on a sequence of tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n\}$, where each task \mathcal{T}_i consists of a dataset $\mathcal{D}_i = \{(\mathbf{x}_j, \mathbf{y}_j)\}$ with input-output pairs. After training on task \mathcal{T}_i , the parameters are updated to θ_i . The CF problem can be characterized as:

$$\mathcal{L}_{\mathcal{T}_j}(\theta_i) \gg \mathcal{L}_{\mathcal{T}_j}(\theta_j) \quad \text{for } j < i \quad (3.40)$$

where $\mathcal{L}_{\mathcal{T}_j}(\theta)$ represents the loss on task \mathcal{T}_j using parameters θ , and the inequality indicates significant performance degradation on previously learned tasks.

This degradation occurs because gradient-based optimization overwrites parameter configurations that were crucial for previous tasks:

$$\theta_{i+1} = \theta_i - \eta \nabla_{\theta} \mathcal{L}_{\mathcal{T}_{i+1}}(\theta_i) \quad (3.41)$$

where η denotes the learning rate and $\nabla_{\theta} \mathcal{L}_{\mathcal{T}_{i+1}}(\theta_i)$ represents the gradient computed on the new task, potentially conflicting with gradients that were beneficial for previous tasks.

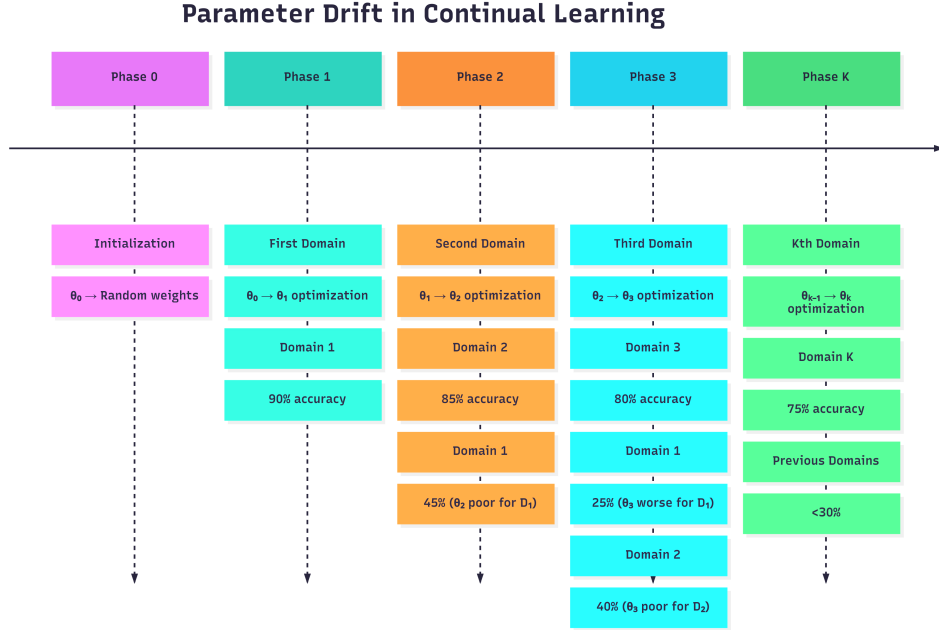


Figure 3.5: Parameter evolution timeline demonstrating the mathematical basis of CF. Starting from random initialization (θ_0), each phase shows parameter updates through gradient descent optimization for new domains. While current domain performance remains high, parameter drift causes systematic degradation of previous domain performance, with Domain 1 accuracy dropping from 90% to 25% as parameters evolve from θ_1 to θ_3 . This visualization captures the core challenge in CL: how parameter updates beneficial for new tasks can be detrimental to previously learned capabilities.

Multimodal Forgetting Dynamics

In *multimodal systems* such as IC, CF exhibits more complex patterns than in unimodal scenarios, affecting multiple representation spaces simultaneously [12, 82]. The integration of visual and textual modalities introduces unique challenges that extend beyond traditional CL approaches.

Cross-Modal Alignment Degradation: The most critical aspect of multimodal forgetting is the deterioration of cross-modal correspondence. As models adapt to new domains, the alignment between visual and textual representations can drift, leading to semantic inconsistencies:

$$\text{Alignment}(\mathcal{D}_{\text{prev}}) = \frac{1}{|\mathcal{D}_{\text{prev}}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{prev}}} \cos(\mathbf{f}_{\text{visual}}(\mathbf{x}), \mathbf{f}_{\text{text}}(\mathbf{y})) \quad (3.42)$$

where $\mathbf{f}_{\text{visual}}(\mathbf{x}) \in \mathbb{R}^d$ represents the d -dimensional visual embedding of image \mathbf{x} , $\mathbf{f}_{\text{text}}(\mathbf{y}) \in \mathbb{R}^d$ denotes the d -dimensional textual embedding of caption \mathbf{y} , and $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}$ computes the cosine similarity between two vectors.

Semantic Drift Quantification: The degradation of cross-modal alignment can be quantified by measuring changes in expected similarity over learning episodes:

$$\text{Drift}_t = \left\| \mathbb{E}[\text{sim}(\mathbf{f}_v^{(t)}, \mathbf{f}_t^{(t)})] - \mathbb{E}[\text{sim}(\mathbf{f}_v^{(0)}, \mathbf{f}_t^{(0)})] \right\| \quad (3.43)$$

where $\mathbf{f}_v^{(t)}, \mathbf{f}_t^{(t)} \in \mathbb{R}^d$ represent the visual and textual embeddings at learning episode t , respectively, and $\mathbb{E}[\cdot]$ denotes expectation over the evaluation dataset.

Modality-Specific Forgetting Patterns: Different modalities exhibit varying susceptibility to CF. Empirical studies reveal that visual representations, particularly those learned from large-scale pretraining, demonstrate greater robustness compared to linguistic representations, which are more sensitive to domain-specific vocabulary and linguistic patterns [82].

Visual forgetting can be measured through embedding drift:

$$\mathcal{L}_{\text{visual}}^{(t)} = \frac{1}{|\mathcal{D}_{\text{prev}}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{prev}}} \left\| \mathbf{f}_v^{(t)}(\mathbf{x}) - \mathbf{f}_v^{(0)}(\mathbf{x}) \right\|_2^2 \quad (3.44)$$

Linguistic forgetting is assessed through generation capability degradation:

$$\mathcal{L}_{\text{linguistic}}^{(t)} = \frac{1}{|\mathcal{D}_{\text{prev}}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{prev}}} -\log P(\mathbf{y}|\mathbf{x}; \theta^{(t)}) \quad (3.45)$$

3.2.2 Episodic Memory Mechanisms

To address the challenges of CF in multimodal scenarios, *episodic memory* approaches offer a promising solution by maintaining buffers of representative experiences from previous tasks [16], as illustrated in Figure 3.6. Unlike parameter-based regularization methods such as EWC, episodic memory retains knowledge by explicitly storing and strategically rehearsing past examples during the learning of new tasks.

Memory Buffer Architecture

An episodic memory buffer \mathcal{M} can be formally defined as:

$$\mathcal{M} = \{(\mathbf{x}_i, \mathbf{y}_i, t_i, s_i)\}_{i=1}^{|\mathcal{M}|} \quad (3.46)$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ represents an input-output pair from the training data, $t_i \in \{1, 2, \dots, T\}$ indicates the task/domain identifier, $s_i \in \mathbb{R}$ denotes an importance score computed by an AL strategy, and $|\mathcal{M}|$ represents the current buffer size.

The buffer maintains a capacity constraint:

$$|\mathcal{M}| \leq C_{\text{max}} \quad (3.47)$$

where C_{max} represents the maximum buffer size, determined by memory limitations and computational constraints.

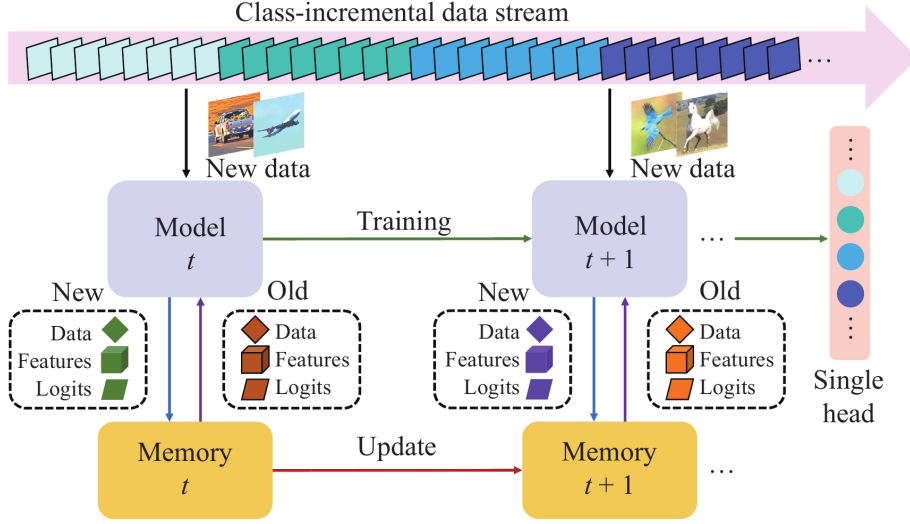


Figure 3.6: Class-incremental CL with episodic memory replay (taken from [9]). The model sequentially learns from a stream of data containing new classes while maintaining a memory buffer. At each time step t , the model receives new data and updates using both current samples and replayed samples from memory. The memory buffer stores representative samples (data, features, and logits) from previous tasks to mitigate CF through strategic replay.

Memory Replay Mechanisms

During training on task \mathcal{T}_k , the learning objective combines current task loss with replay loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{T}_k}(\theta) + \lambda \mathcal{L}_{\text{replay}}(\theta) \quad (3.48)$$

where $\lambda \in [0, 1]$ is a weighting parameter that balances current learning with knowledge preservation, and the replay loss is computed over sampled experiences from the buffer:

$$\mathcal{L}_{\text{replay}}(\theta) = \frac{1}{|\mathcal{B}_{\text{replay}}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{B}_{\text{replay}}} \ell(f_{\theta}(\mathbf{x}), \mathbf{y}) \quad (3.49)$$

where $\mathcal{B}_{\text{replay}} \subset \mathcal{M}$ denotes a batch sampled from the episodic memory buffer and $\ell(\cdot, \cdot)$ represents the loss function.

Gradient Integration for Memory Replay: The gradient update combines contributions from the current task and replayed experiences:

$$\mathbf{g}_{\text{current}} = \nabla_{\theta} \mathcal{L}_{\mathcal{T}_k}(\theta) \quad (3.50)$$

$$\mathbf{g}_{\text{replay}} = \nabla_{\theta} \mathcal{L}_{\text{replay}}(\theta) \quad (3.51)$$

$$\mathbf{g}_{\text{total}} = \mathbf{g}_{\text{current}} + \lambda \mathbf{g}_{\text{replay}} \quad (3.52)$$

The parameter update then follows:

$$\theta_{\text{new}} = \theta - \eta \mathbf{g}_{\text{total}} \quad (3.53)$$

This approach ensures that parameter updates consider both new learning objectives and the preservation of previous knowledge through explicit gradient combination.

3.3 Active Learning Principles

AL addresses the core challenge of maximizing learning efficiency under high annotation costs by strategically selecting the most informative samples for labeling [22, 83, 61]. This section outlines the theoretical foundations of AL methodologies and their application to vision-language CL scenarios.

This section presents query strategy frameworks (Section 3.3.1), uncertainty sampling strategies (Section 3.3.2), diversity sampling methodologies (Section 3.3.3), and hybrid sampling approaches (Section 3.3.4) that combine multiple selection criteria.

3.3.1 Query Strategy Framework

The AL framework can be formally defined through a query strategy ϕ that selects the most informative subset $S^* \subset U$ from an unlabeled pool U for annotation:

$$S^* = \arg \max_{S \subset U, |S| \leq b} \phi(S | \mathcal{D}_{\text{labeled}}, \theta) \quad (3.54)$$

where b denotes the query budget, $\mathcal{D}_{\text{labeled}}$ represents the current labeled dataset, and θ denotes the model parameters. The query strategy ϕ quantifies the expected utility of labeling a particular subset.

Information-Theoretic Motivation

AL strategies often draw from *information theory*, seeking to maximize information gain through strategic sample selection [23]. The expected information gain from labeling a sample \mathbf{x} can be expressed as:

$$IG(\mathbf{x}) = H[Y] - \mathbb{E}_{\mathbf{y}}[H[Y | \mathbf{y}, \mathbf{x}]] \quad (3.55)$$

where $H[Y]$ represents the entropy of the model's predictions before observing the label, and $\mathbb{E}_{\mathbf{y}}[H[Y | \mathbf{y}, \mathbf{x}]]$ denotes the expected entropy after labeling sample \mathbf{x} with label \mathbf{y} .

3.3.2 Uncertainty Sampling Strategies

Uncertainty sampling is one of the most intuitive AL approaches, wherein the model selects samples for which it has the lowest confidence in its predictions [84].

Classification-Based Uncertainty Measures

A range of uncertainty measures applicable to classification tasks has been introduced in prior research [22]:

Least Confidence: Selects samples for which the model’s highest predicted probability is lowest:

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in U} P(\hat{\mathbf{y}}|\mathbf{x}) \quad (3.56)$$

where $\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$ denotes the most likely class prediction.

Margin Sampling: Selects samples based on the difference between the two most probable class predictions:

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in U} [P(\hat{\mathbf{y}}_1|\mathbf{x}) - P(\hat{\mathbf{y}}_2|\mathbf{x})] \quad (3.57)$$

where $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are the classes with the highest and second-highest predicted probabilities, respectively.

Entropy-Based Selection: Selects samples with the highest predictive uncertainty as measured by entropy:

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in U} \left[- \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \log P(\mathbf{y}|\mathbf{x}) \right] \quad (3.58)$$

Sequence Generation Uncertainty Measures

For *generative tasks* such as IC, uncertainty estimation becomes more complex due to the sequential nature of text generation. MTE [29] addresses this challenge by aggregating uncertainty across all generation steps:

$$\text{MTE}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^T H(y_t|y_{<t}, \mathbf{x}) \quad (3.59)$$

where T denotes the sequence length and $H(y_t|y_{<t}, \mathbf{x})$ represents the entropy at generation step t given previous tokens and input \mathbf{x} .

This approach effectively captures distributional uncertainty across the entire caption generation process, making it particularly suitable for vision-language tasks where sequence coherence is crucial.

3.3.3 Diversity Sampling Methodologies

Diversity sampling addresses fundamental limitations of uncertainty-based approaches by ensuring representative coverage of the feature space [85]. This approach is particularly important in CL scenarios where maintaining a balanced representation across domains is crucial.

Core-Set Methods

Core-set approaches [85] formulate sample selection as a geometric optimization problem, seeking to minimize the maximum distance between any unlabeled point and its nearest labeled example:

$$S^* = \arg \min_{S \subset U} \max_{\mathbf{x} \in U \setminus S} \min_{\mathbf{s} \in \mathcal{D}_{\text{labeled}} \cup S} d(\mathbf{x}, \mathbf{s}) \quad (3.60)$$

where $d(\cdot, \cdot)$ represents a distance metric in the feature space, typically Euclidean distance between learned representations.

The core-set approach provides theoretical guarantees on coverage while remaining computationally tractable through greedy approximation algorithms.

Clustering-Based Selection

K-means clustering [86] provides an intuitive approach to diversity sampling by partitioning the feature space and selecting representative samples from each cluster:

$$\{C_1, \dots, C_k\} = \text{K-means}(U, k) \quad (3.61)$$

$$S^* = \{s_i^* : s_i^* = \arg \min_{s \in C_i} \|s - \mu_i\|_2\}_{i=1}^k \quad (3.62)$$

where C_i represents the i -th cluster, μ_i denotes the cluster centroid, and k represents the desired sample count.

This approach ensures representative coverage across different regions of the feature space while maintaining computational efficiency.

Multimodal Diversity Assessment

CLIP [10] learns joint representations by maximizing agreement between correctly paired image-text examples while minimizing agreement between mismatched pairs, as illustrated in Figure 3.7.

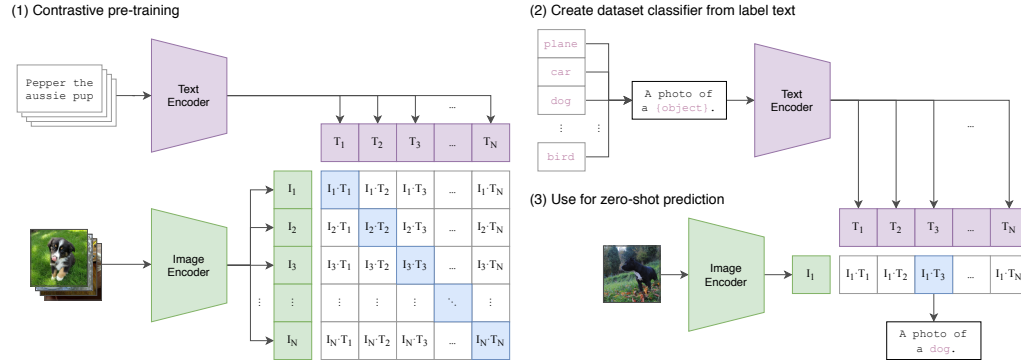


Figure 3.7: CLIP architecture and training process [10]. **(1) Contrastive pre-training:** Image and text encoders learn joint representations by maximizing similarity between correct image-text pairs and minimizing similarity between incorrect pairs in the batch. **(2) Create dataset classifier:** Class labels are converted to text prompts (e.g., "A photo of a object") and processed through the text encoder. **(3) Zero-shot prediction:** For inference, the image encoder processes the input image, and classification is performed by comparing the image embedding with all text embeddings to find the highest similarity.

Encoder Architecture: CLIP utilizes separate encoders for the visual and textual modalities, allowing each to specialize in processing its respective input.

Vision Encoder: Images are processed via either a Vision Transformer (ViT) [8] or a ResNet [80] architecture:

$$\mathbf{I} = f_{\theta_v}(\mathbf{x}_{\text{image}}) \in \mathbb{R}^{d_v} \quad (3.63)$$

where f_{θ_v} is the vision encoder with parameters θ_v , and d_v denotes the visual feature dimension.

Text Encoder: Captions are processed using a Transformer-based architecture [7]:

$$\mathbf{T} = g_{\theta_t}(\mathbf{x}_{\text{text}}) \in \mathbb{R}^{d_t} \quad (3.64)$$

where g_{θ_t} is the text encoder with parameters θ_t , and d_t is the textual feature dimension.

Projection to Joint Space: Both modalities are projected into a shared embedding space:

$$\mathbf{h}_v = \mathbf{W}_v \mathbf{I} + \mathbf{b}_v \in \mathbb{R}^d \quad (3.65)$$

$$\mathbf{h}_t = \mathbf{W}_t \mathbf{T} + \mathbf{b}_t \in \mathbb{R}^d \quad (3.66)$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$ and $\mathbf{W}_t \in \mathbb{R}^{d \times d_t}$ are projection matrices, $\mathbf{b}_v, \mathbf{b}_t \in \mathbb{R}^d$ are bias vectors, and d is the joint embedding dimension.

Contrastive Loss Formulation: CLIP is trained using a symmetric contrastive loss. For a batch of N image-text pairs $(\mathbf{I}_i, \mathbf{T}_i)_{i=1}^N$, the similarity matrix is computed as:

$$S_{ij} = \mathbf{f}_{\text{visual}}^{(i)} \cdot \mathbf{f}_{\text{text}}^{(j)} = \cos(\mathbf{h}_v^{(i)}, \mathbf{h}_t^{(j)}) \quad (3.67)$$

The contrastive loss encourages high similarity for matching image-text pairs ($i = j$) and low similarity for mismatched pairs ($i \neq j$):

$$\mathcal{L}_{i \rightarrow t}^{(i)} = -\log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^N \exp(S_{ij}/\tau)} \quad (3.68)$$

$$\mathcal{L}_{t \rightarrow i}^{(i)} = -\log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^N \exp(S_{ji}/\tau)} \quad (3.69)$$

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N} \sum_{i=1}^N \left(\mathcal{L}_{i \rightarrow t}^{(i)} + \mathcal{L}_{t \rightarrow i}^{(i)} \right) \quad (3.70)$$

where τ is a learnable temperature parameter controlling the sharpness of the softmax distribution.

3.3.4 Hybrid Sampling Strategies

Hybrid approaches combine uncertainty and diversity criteria, leveraging the strengths of both methodologies while mitigating their limitations [87, 25].

Stratified Hybrid Selection

HUDS [25] employs a stratified approach that partitions samples by uncertainty levels before applying diversity-based selection within each stratum. This method addresses limitations of pure uncertainty or diversity sampling by ensuring balanced selection across different uncertainty ranges.

Phase 1: Uncertainty Computation and Stratification: HUDS estimates the uncertainty for an unlabeled sentence s using the normalized negative log-likelihood (NNLL):

$$\text{NNLL}(X) = -\frac{1}{S} \sum_{j=1}^S \log p(x_j | x_1, x_2, \dots, x_{j-1}) \quad (3.71)$$

where S denotes the sentence length and x_j represents the j -th token in sequence X . Lower NNLL values indicate greater model confidence, while higher values signal greater uncertainty.

The range of uncertainty scores is then partitioned into strata, with the interval for the i -th stratum defined as:

$$\text{Stratum}_i = \left[s_{\min} + \frac{i-1}{n}r, s_{\min} + \frac{i}{n}r \right] \quad (3.72)$$

where s_{\min} is the minimum uncertainty, s_{\max} the maximum, $r = s_{\max} - s_{\min}$ is the range, and n is the number of strata.

Phase 2: Diversity Selection: Within each stratum, HUDS computes sentence embeddings (typically using pre-trained BERT), then clusters them using k-means. Diversity is measured as the cosine distance between a sentence's embedding and its cluster centroid:

$$\text{Diversity}(x) = d(x, c_i) = 1 - \frac{\mathbf{e}_x \cdot \mathbf{c}_i}{\|\mathbf{e}_x\| \|\mathbf{c}_i\|} \quad (3.73)$$

where \mathbf{e}_x is the embedding of sentence x and \mathbf{c}_i is the centroid of cluster i .

Phase 3: Hybrid Score Computation: The final hybrid sampling score combines uncertainty and diversity via a weighted sum:

$$H(x) = \lambda \cdot d(x, c_i) + (1 - \lambda) \cdot u_x \quad (3.74)$$

where $d(x, c_i)$ is the diversity score, u_x is the uncertainty score of sentence x , and $\lambda \in [0, 1]$ balances the contribution of uncertainty and diversity. The top- k instances with the highest hybrid scores are selected for annotation.

This stratified approach ensures the selection of diverse samples from each uncertainty subpopulation, allowing the model to learn from both challenging (high uncertainty) and representative (high diversity) examples across the feature space.

3.4 Evaluation Frameworks

Comprehensive evaluation of CL systems requires metrics that assess both knowledge retention and adaptation capabilities across multiple dimensions. This section establishes the evaluation frameworks essential for assessing multimodal CL performance.

This section covers traditional image captioning metrics (Section 3.4.1) including n-gram and semantic similarity measures, continual learning evaluation metrics (Section 3.4.2) that capture knowledge retention and transfer capabilities, and human evaluation statistical methods (Section 3.4.3) that provide the mathematical foundations for analyzing subjective quality assessments.

3.4.1 Traditional Image Captioning Metrics

N-gram Based Metrics

BLEU [73] is a precision-based metric that quantifies the degree of n-gram overlap between candidate and reference captions.

$$\text{BLEU} = BP \times \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (3.75)$$

where p_n represents n-gram precision, w_n denotes uniform weights ($w_n = 1/N$), and BP represents the brevity penalty:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - r/c) & \text{if } c \leq r \end{cases} \quad (3.76)$$

where c is the candidate length and r is the reference length.

ROUGE [75] focuses on recall by measuring the length of the longest common subsequence between candidate and reference texts:

$$\text{ROUGE-Lrecall} = \frac{\text{LCS}(\mathbf{X}, \mathbf{Y})}{|\mathbf{Y}|} = R_L \quad (3.77)$$

$$\text{ROUGE-Lprecision} = \frac{\text{LCS}(\mathbf{X}, \mathbf{Y})}{|\mathbf{X}|} = P_L \quad (3.78)$$

$$\text{ROUGE-L}_{F1} = \frac{(1 + \beta^2) \cdot R_L \cdot P_L}{\beta^2 \cdot P_L + R_L} \quad (3.79)$$

where $\text{LCS}(\mathbf{X}, \mathbf{Y})$ computes the longest common subsequence between candidate \mathbf{X} and reference \mathbf{Y} , and β controls the importance of recall versus precision.

Semantic Similarity Metrics

CIDEr (Consensus-based Image Description Evaluation) [76] measures consensus between multiple reference captions using *term frequency-inverse document frequency* (TF-IDF) weighting.

TF-IDF Weighting Mechanism: The core innovation of CIDEr lies in its use of TF-IDF weighting to create meaningful n-gram representations. The TF-IDF scheme consists of two multiplicative components:

Term Frequency (TF): Measures how frequently a word appears within a specific caption, normalized by caption length:

$$\text{TF}(w, c) = \frac{\text{count}(w, c)}{|c|} \quad (3.80)$$

where $\text{count}(w, c)$ represents the frequency of word w in caption c , and $|c|$ denotes the total number of words in the caption. This component ensures that words appearing multiple times in a caption receive higher importance.

Inverse Document Frequency (IDF): Measures the informativeness of a word by computing its rarity across the entire corpus:

$$\text{IDF}(w) = \log \left(\frac{|D|}{|\{c \in D : w \in c\}|} \right) \quad (3.81)$$

where $|D|$ represents the total number of captions in the corpus, and $|\{c \in D : w \in c\}|$ counts the number of captions containing the word w . The logarithm dampens the effect while ensuring rare words receive exponentially higher weights than common words.

Combined TF-IDF Weight: The final weight combines both components:

$$\text{TF-IDF}(w, c) = \text{TF}(w, c) \times \text{IDF}(w) \quad (3.82)$$

This weighting scheme automatically emphasizes distinctive, content-bearing words (e.g., "dog", "bicycle", "running") while downweighting common function words (e.g., "the", "and", "is").

N-gram Vector Construction: For each caption, CIDEr constructs TF-IDF weighted n-gram vectors $\mathbf{g}^n(c)$ where each dimension corresponds to an n-gram's TF-IDF weight. This creates a high-dimensional representation that captures both local word patterns and their semantic importance.

CIDEr Score Calculation: Using the TF-IDF weighted n-gram vectors, CIDEr computes semantic similarity between candidate and reference captions through cosine similarity:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_{j=1}^m \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|} \quad (3.83)$$

where $\mathbf{g}^n(c_i)$ represents the TF-IDF weighted n-gram vector for candidate caption c_i , $\mathbf{g}^n(s_{ij})$ denotes the corresponding vector for reference caption s_{ij} , and m is the number of reference captions. The final CIDEr score typically combines multiple n-gram orders (n=1,2,3,4) with equal weighting.

BERTScore [78] leverages contextualized embeddings from pre-trained language models to assess semantic similarity between generated and reference captions, enabling robust comparison beyond simple lexical overlap.

$$\text{Precision} = \frac{1}{|\mathbf{x}|} \sum_{x_i \in \mathbf{x}} \max_{y_j \in \mathbf{y}} \text{sim}(\mathbf{x}_i, \mathbf{y}_j) \quad (3.84)$$

$$\text{Recall} = \frac{1}{|\mathbf{y}|} \sum_{y_j \in \mathbf{y}} \max_{x_i \in \mathbf{x}} \text{sim}(\mathbf{x}_i, \mathbf{y}_j) \quad (3.85)$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.86)$$

where $\text{sim}(\mathbf{x}_i, \mathbf{y}_j)$ computes the cosine similarity between contextualized embeddings, enabling semantic matching even with different word choices.

Based on the GEM paper, here's an improved version of the subsection:

Here's the updated subsection with notation explanations integrated into the text:

Here's the updated subsection with author names removed from citations:

3.4.2 Continual Learning Evaluation Metrics

Building upon the foundational work in [16], we adopt a comprehensive evaluation framework that captures both performance and knowledge transfer capabilities in continual learning scenarios. The evaluation is based on constructing a performance matrix $\mathbf{R} \in \mathbb{R}^{T \times T}$, where T represents the total number of tasks, and $R_{i,j}$ represents the test classification accuracy on task t_j after observing the last sample from task t_i . Here, the row index i indicates the temporal point of evaluation (after completing task i), while the column index j specifies which task is being evaluated.

Core Performance Metrics

Average Accuracy measures overall performance across all tasks after sequential learning:

$$\text{Average Accuracy (ACC)} = \frac{1}{T} \sum_{i=1}^T R_{T,i} \quad (3.87)$$

where $R_{T,i}$ denotes the final performance on task i after learning all T tasks, and $\frac{1}{T}$ provides the averaging factor across all tasks.

Backward Transfer quantifies the influence of learning new tasks on previously acquired knowledge [16]:

$$\text{Backward Transfer (BWT)} = \frac{1}{T-1} \sum_{i=1}^{T-1} (R_{T,i} - R_{i,i}) \quad (3.88)$$

where $R_{i,i}$ represents the performance on task i immediately after learning it (the "fresh" performance), $R_{T,i}$ is the final performance on task i after learning all subsequent tasks, and $\frac{1}{T-1}$ averages over all tasks except the last one (which cannot experience backward transfer). The difference $(R_{T,i} - R_{i,i})$ captures the performance change due to subsequent learning. This metric captures:

- *Positive backward transfer*: $\text{BWT} > 0$ indicates that learning subsequent tasks improves performance on previous tasks
- *Negative backward transfer*: $\text{BWT} < 0$ indicates CF, where new learning degrades previous knowledge

Forward Transfer assesses the benefit of previously learned knowledge for acquiring new tasks [16]:

$$\text{FWT} = \frac{1}{T-1} \sum_{i=2}^T (R_{i-1,i} - \bar{b}_i) \quad (3.89)$$

where $R_{i-1,i}$ is the performance on task i after learning tasks 1 through $i-1$ (but before learning task i itself), \bar{b}_i represents the baseline performance on task i at random initialization, and $\frac{1}{T-1}$ averages over all tasks except the first one (which cannot benefit from forward transfer). The summation starts from $i = 2$ since the first task has no previous knowledge to transfer from. Positive values indicate beneficial knowledge transfer, enabling faster learning or zero-shot capabilities.

Memory-Specific Metrics

For episodic memory-based approaches, we additionally consider:

Forgetting Measure following [20, 88]:

$$\text{FM} = \frac{1}{T-1} \sum_{i=1}^{T-1} \max_{k \in \{i, \dots, T-1\}} (R_{k,i} - R_{T,i}) \quad (3.90)$$

where $\max_{k \in \{i, \dots, T-1\}} R_{k,i}$ finds the peak performance achieved on task i during the learning sequence (from when it was learned until before the last task), $R_{T,i}$ is the final performance on task i , and the difference captures the maximum performance drop experienced by task i . The range $k \in \{i, \dots, T-1\}$ ensures we only consider evaluations after task i was learned but before the final evaluation.

Average Forgetting as a simplified metric [89]:

$$\text{Average Forgetting (AF)} = \frac{1}{T-1} \sum_{i=1}^{T-1} (R_{i,i} - R_{T,i}) \quad (3.91)$$

where $(R_{i,i} - R_{T,i})$ directly measures the performance drop from when each task i was first learned to the final evaluation, and $\frac{1}{T-1}$ averages this drop across all tasks except the last one. Lower values indicate better retention of previous knowledge.

This evaluation framework provides a principled approach to assess continual learning systems, particularly focusing on the trade-offs between performance, memory efficiency, and knowledge transfer capabilities that are central to the ALCIE framework's objectives.

3.4.3 Human Evaluation Statistical Methods

This section provides the mathematical foundations and standard definitions for statistical analysis of human evaluation data in AI research.

Descriptive Statistical Measures

For human evaluation data with n participants rating k conditions, the following descriptive measures are used:

Sample Mean: *Definition:* The arithmetic average of participant scores, representing the central tendency of the data [90].

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.92)$$

where x_i represents individual rating scores.

Sample Standard Deviation: *Definition:* A measure of data dispersion, quantifying the extent to which scores deviate from the sample mean [90].

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.93)$$

Standard Error of the Mean: *Definition:* An estimate of the variability of the sample mean across different samples drawn from the same population [91].

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} \quad (3.94)$$

Range: *Definition:* The difference between the maximum and minimum mean values across k conditions, indicating the spread of mean scores [92].

$$R = \max(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) - \min(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) \quad (3.95)$$

Confidence Intervals

95% Confidence Interval (CI): *Definition:* An interval estimate that is expected to contain the true population mean 95% of the time if the experiment were repeated [90].

$$CI_{95\%} = \bar{x} \pm 1.96 \times SE_{\bar{x}} \quad (3.96)$$

Reliability Measurement

Cronbach's Coefficient Alpha: *Definition:* A statistic that measures internal consistency (reliability) of a set of scale or test items [93].

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{Y_i}^2}{\sigma_X^2} \right) \quad (3.97)$$

where k is the number of evaluation dimensions, $\sigma_{Y_i}^2$ is the variance of dimension i , and σ_X^2 is the variance of total scores.

Preference Analysis

Preference Proportion: *Definition:* The proportion of participants who preferred a specific condition, expressed as a percentage [94].

$$P_j = \frac{x_j}{n} \times 100\% \quad (3.98)$$

where x_j represents the number of preferences for condition j .

Temporal Difference: *Definition:* The difference in mean scores between two time points or experimental phases [91].

$$\Delta_{temporal} = \bar{x}_{late} - \bar{x}_{early} \quad (3.99)$$

These mathematical formulations provide the computational foundation for human evaluation analysis, with specific implementation and interpretation determined by study design and research objectives.

Chapter 4

Methodology

This methodology chapter presents our systematic approach to answering the central research question: **Can strategic sample selection significantly improve continual learning performance beyond computationally efficient random approaches?** We introduce the ALCIE framework, which systematically compares uncertainty-based sampling, diversity-based sampling, hybrid strategies, and random sampling baselines across BLIP-2 and OFA architectures.

We begin with a comprehensive overview, establishing the theoretical foundations and architectural components of our strategic vs. random memory management comparison (Section 4.1). Next, we detail the benchmark architecture, focusing on the integration of BLIP-2 and OFA models as the technical basis for systematic evaluation (Section 4.2). We then introduce our novel AL integration, which incorporates uncertainty, diversity, and hybrid sampling strategies into the CL workflow for direct comparison against random baselines (Section 4.3). Subsequently, we describe our memory management strategies, which transform random episodic memory selection into principled approaches for strategic sample retention and replay (Section 4.4). Finally, we present our comprehensive training infrastructure and implementation framework, establishing a foundation for reproducible experimentation and systematic evaluation of strategic memory selection in continual IC tasks (Section 4.5).

4.1 ALCIE Framework Overview

This section establishes the theoretical and architectural foundations of our ALCIE framework, designed to systematically investigate whether strategic memory management provides meaningful advantages over computationally efficient random sampling in continual IC systems. We begin by introducing the core challenge of strategic episodic memory management and our comparative approach (Section 4.1.1), followed by a discussion of the theoretical motivation grounded in information-theoretic principles while questioning practical applicability (Section 4.1.2). Next, we present the full framework architecture and pipeline, highlighting how our approach enables systematic comparison between strategic and random memory selection (Section 4.1.3). Finally, we summarize

our main contributions and detail the novel aspects of our cross-architectural validation methodology (Section 4.1.4).

4.1.1 Active Learning for Continual Image Captioning

The *ALCIE* framework addresses the core challenge of strategic episodic memory management in CL systems by systematically comparing sophisticated approaches against simple random sampling baselines. Conventional CL methods for IC typically rely on random sampling to populate episodic memory buffers, a practice that may not make optimal use of limited memory resources [12]. However, the fundamental question remains: do sophisticated selection strategies justify their computational overhead compared to simple random approaches?

The framework operates on the principle that *strategic sample selection* should demonstrate measurable advantages over random sampling to justify additional computational complexity. Rather than assuming sophisticated approaches are superior, *ALCIE* utilizes specialized query functions to evaluate sample informativeness based on uncertainty estimation, diversity considerations, or a hybrid of both, then systematically compares these against random selection across identical conditions.

ALCIE integrates seamlessly with existing vision-language architectures through modular trainer extensions that implement AL selection during the training process. The framework maintains compatibility with different episodic memory approaches while providing systematic improvements through strategic sample selection, making it applicable to various CL scenarios beyond IC. Critically, the framework measures both performance improvements and computational overhead to determine practical deployment viability.

4.1.2 Theoretical Motivation and Design Principles

The theoretical foundation of *ALCIE* rests on *information-theoretic principles* that inform effective sample selection in resource-constrained learning environments [95], while explicitly questioning whether these theoretical advantages translate to practical benefits over simple random sampling. In CL scenarios, a fundamental trade-off arises between preserving previously acquired knowledge and adapting to new domains. *Episodic memory* acts as a key mechanism for knowledge retention, yet its effectiveness is determined by which samples are chosen for storage and replay [16].

Information-Theoretic Foundation vs. Practical Efficiency: The central hypothesis of *ALCIE* is that samples with higher information content should be more valuable for mitigating CF, but this theoretical advantage must be validated against the computational simplicity of random sampling. Information content can be quantified in several ways: *uncertainty-based* measures identify samples where the model exhibits low confidence, signaling prime opportunities for learning [23], while *diversity-based* measures ensure broad coverage of the input distribution, reducing the risk of overfitting to specific subsets [85]. However, the computational overhead of these sophisticated approaches raises questions about their practical superiority over random selection.

Multimodal Considerations: Vision-language tasks introduce added complexity relative to unimodal CL due to the need for coherent alignment between visual and linguistic modalities as the model encounters new domains. *ALCIE* addresses this by leveraging the robust multimodal representations of CLIP [10] for diversity assessment, ensuring

that sample selection reflects both visual and textual informativeness. Yet this sophistication comes with computational costs that must be justified against random sampling’s efficiency.

Design Principles: The framework is designed around three key principles: (1) *Fair Comparison*: systematic evaluation of strategic vs. random approaches under identical conditions; (2) *Computational Transparency*: explicit measurement of efficiency trade-offs alongside performance metrics; and (3) *Architectural Generalizability*: validation that findings hold across different vision-language architectures.

4.1.3 Framework Architecture and Pipeline

As depicted in Figure 4.1, the ALCIE system operates through a hierarchical pipeline comprising four core components designed for systematic comparison of memory management strategies:

Vision-Language Model Layer: Pre-trained models are adapted via CL using the fashion domain sequence detailed in Section 4.2.1.

Active Learning Selection Engine: Four sampling strategies enable systematic comparison: Random (computational efficiency baseline), Uncertainty (MTE-based), Diversity (CLIP-based clustering), and Hybrid (HUDS stratified approach). Each strategy is evaluated under identical conditions to determine whether sophisticated selection criteria justify their computational overhead. Detailed implementations are provided in Section 4.3.

Episodic Memory Management System: Building on the approach from [12], the framework incorporates an advanced memory buffer that retains strategically selected samples from multiple domains through dynamic capacity allocation and score-based replacement policies. Proportional deletion strategies are employed to prioritize the retention of the most informative instances, ensuring that essential knowledge is preserved while accommodating new data from emerging domains. The system maintains identical capacity constraints across all strategies to ensure fair comparison.

Memory Replay Mechanism: A *cluster-stratified sampling* strategy ensures balanced rehearsal across all previously encountered domains. Replay events are systematically triggered after *every 200 training samples*, enabling effective cross-domain knowledge transfer while preventing contamination from the current domain. This mechanism maintains the integrity of past learning and supports continual adaptation across all sampling strategies.

Continual Learning Pipeline

The ALCIE pipeline progresses through sequential domain phases, each corresponding to a specific fashion category (*accessories, bottoms, dresses, outerwear, shoes, tops*). The continual learning protocol is designed to support systematic knowledge accumulation while actively mitigating CF through strategic vs. random memory management comparison:

Domain Initialization: Each new domain phase begins from the model checkpoint obtained at the end of the previous phase, preserving continuity in learned representations. The memory buffer is retained across transitions, with proportional deletion strategies employed to allocate space for new domain samples while minimizing information loss.

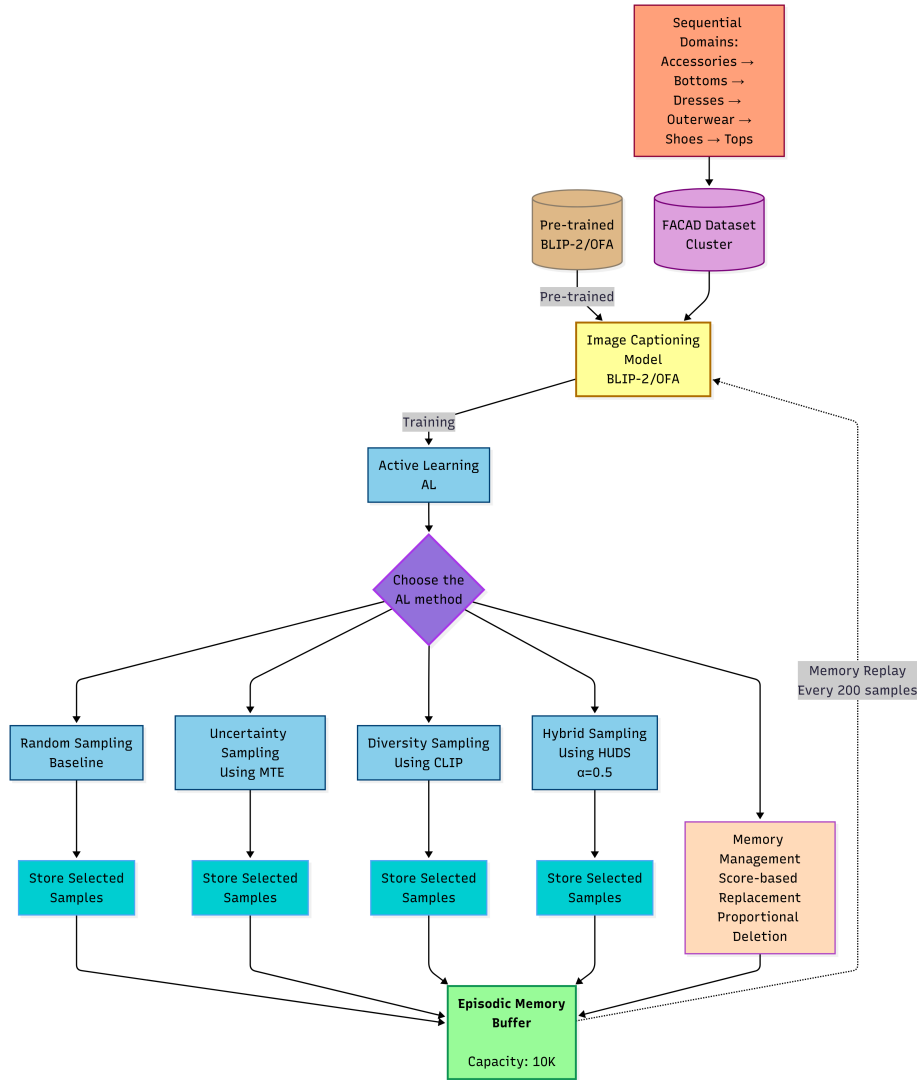


Figure 4.1: ALCIE Framework Architecture: The framework integrates four AL strategies (Random, Uncertainty, Diversity, Hybrid) with episodic memory management for continual IC. Pre-trained vision-language models (BLIP-2/OFA) process sequential fashion domain clusters from the FACAD dataset. The systematic comparison evaluates whether strategic AL strategies provide meaningful advantages over random sampling baselines while managing bounded memory constraints and preventing CF through balanced rehearsal mechanisms.

Active Learning Integration: During each training iteration, the current batch is evaluated using the chosen AL strategy. The top- k most informative samples (typically $k = 16$) are selected according to the strategy-specific scoring function and marked for episodic memory storage. Random sampling serves as the computational efficiency baseline against which sophisticated strategies must demonstrate clear advantages.

Memory Update Protocol: Selected samples are integrated into the episodic memory buffer through score-based replacement policies, which prioritize the retention of highly informative instances and maintain balanced representation across domains.

Replay-Enhanced Training: At scheduled intervals, training batches are augmented with samples strategically drawn from the episodic memory buffer. This replay mechanism ensures consistent exposure to prior domain knowledge during new domain adaptation, promoting robust knowledge retention and transfer.

Cross-Domain Knowledge Transfer

The framework implements sophisticated mechanisms to ensure effective knowledge transfer across sequential learning episodes while maintaining domain-specific adaptation capabilities:

Memory Stratification: The episodic memory buffer is organized into clusters corresponding to each previously encountered domain, enabling balanced and representative sampling across all domains during memory replay events.

Gradient Integration: During memory replay, the framework computes separate forward and backward passes for historical samples, then integrates their gradients with those from the current domain. This process helps to retain prior knowledge while minimizing interference during continual adaptation.

Evaluation Protocol: After each training phase, model performance is systematically evaluated on all previously encountered domains. This cross-domain assessment provides a quantitative measurement of both knowledge retention and the effectiveness of transfer across domains.

4.1.4 Main Contributions and Novel Aspects

The ALCIE framework makes several novel contributions to the field of CL for vision-language models by systematically questioning assumptions about memory strategy sophistication:

1. **Novel Framework for Strategic vs. Random Memory Management:** We introduce ALCIE, the first systematic framework to directly compare AL principles with random sampling baselines in continual IC, challenging assumptions about whether algorithmic sophistication provides meaningful advantages over computational efficiency.
2. **Adaptation of Active Learning to Continual Vision-Language Tasks:** We adapt and extend four distinct AL strategies: uncertainty sampling, diversity sampling, hybrid sampling (HUDS), and random baseline specifically for multimodal CL scenarios, addressing the unique challenges of maintaining both visual and linguistic knowledge across sequential domains while measuring computational trade-offs.
3. **Cross-Architecture Validation Framework:** We provide comprehensive evaluation across two different vision-language architectures (BLIP-2’s modular design and OFA’s unified transformer), demonstrating whether AL benefits are architecture-agnostic and represent fundamental improvements or design-specific phenomena.
4. **Strategic Memory Management with Bounded Resources:** We develop sophisticated memory buffer management protocols including score-based replacement

policies, proportional deletion strategies, and cluster-stratified replay mechanisms that maximize knowledge retention within fixed memory constraints while enabling fair comparison across strategies.

5. **Systematic Benchmark for Continual Image Captioning:** We establish a standardized evaluation framework for CL in vision-language models using the FACAD fashion dataset, providing reproducible protocols for comparing memory management strategies across multiple performance dimensions.
6. **Computational Efficiency Integration:** We explicitly measure computational overhead alongside performance metrics, moving beyond traditional AL evaluation to consider practical deployment feasibility and resource allocation trade-offs.

4.2 Benchmark Architecture Design

This section presents the technical foundation and architectural components that enable systematic evaluation of strategic vs. random memory management in multimodal learning systems. We begin with an overview of our modular benchmark design philosophy and its advantages for CL research (Section 4.2.1). We then detail our selection and implementation of two complementary vision-language architectures: BLIP-2’s modular design (Section 4.2.2) and OFA’s unified transformer framework (Section 4.2.3). Each architectural choice is accompanied by detailed justifications for its inclusion in our comparative benchmark suite.

4.2.1 Benchmark Design Philosophy and Model Selection

The ALCIE framework establishes a *comprehensive benchmark architecture* that integrates AL principles with episodic memory selection for continual IC, with the primary goal of determining whether strategic approaches justify their computational overhead compared to random sampling. The architecture adopts a modular design, separating the core IC models from the memory management strategies. This separation ensures fair and transparent comparison across different AL approaches while maintaining identical experimental conditions.

This *modular approach* was deliberately chosen over integrated architectures for several critical reasons. First, it allows performance improvements to be explicitly attributed to strategic memory selection versus random baselines, rather than confounded by model-specific optimizations. This addresses a key methodological concern in CL research, where confounding factors often obscure the true contribution of individual components [27]. Second, the modular design supports systematic ablation studies across architectural choices while maintaining consistent experimental conditions for fair strategy comparison.

The framework leverages state-of-the-art vision-language models as foundational architectures, specifically BLIP-2 and OFA, which exemplify distinct paradigms in multimodal learning [3, 2]. The selection of these models is motivated by their complementary strengths and architectural diversity: BLIP-2 follows the frozen-encoder paradigm, utilizing pre-trained components with minimal fine-tuning, while OFA adopts a unified sequence-to-sequence approach, jointly optimizing all components. This diversity enables a robust assessment of whether AL benefits are architecture-dependent or represent fundamental advances in memory management over random approaches.

The experimental pipeline employs the *FACAD* dataset [28] to investigate CL across six fashion categories, chosen for their clear domain boundaries and relevance to e-commerce applications. The dataset comprises six fashion categories with clear domain boundaries: (1) accessories, (2) bottoms, (3) dresses, (4) outerwear, (5) shoes, and (6) tops. Each category contains diverse image-caption pairs relevant to e-commerce applications, providing both visual and linguistic diversity essential for systematic analysis of catastrophic forgetting and memory management effectiveness across different sampling strategies.

4.2.2 BLIP-2 Architecture Implementation

BLIP-2 serves as our primary benchmark due to its modular design enabling isolated continual learning effects and clear attribution of improvements to memory management strategies rather than architectural modifications. The three-component architecture (frozen ViT-L/14, trainable Q-Former, frozen OPT-2.7B) allows selective adaptation while preserving robust pretrained representations.

For detailed architectural specifications, see Section 3.1.4. Our implementation focuses on Q-Former adaptation during continual learning phases while maintaining frozen component stability, enabling systematic comparison of how different memory strategies affect the trainable components without confounding factors from architectural changes.

4.2.3 OFA Unified Transformer Framework

OFA provides architectural contrast through its unified parameter sharing across modalities, enabling validation that strategic memory management benefits (if any) are architecture-agnostic rather than specific to modular designs like BLIP-2.

Implementation details are provided in Section 3.1.5. Our methodology emphasizes the unified optimization landscape differences compared to BLIP-2’s component isolation, enabling assessment of whether memory strategy advantages persist across different architectural paradigms.

4.3 Active Learning Integration

This section details the implementation of our four distinct query functions designed specifically for multimodal CL scenarios, with particular emphasis on systematic comparison between sophisticated strategies and random baselines. We begin with our baseline random sampling approach (Section 4.3.1), followed by our uncertainty-based strategy using MTE (Section 4.3.2). We then present our diversity sampling approach leveraging CLIP features (Section 4.3.3), and conclude with our novel hybrid strategy that combines uncertainty and diversity criteria (Section 4.3.4).

4.3.1 Random Sampling Baseline

Random sampling serves as the fundamental baseline for episodic memory selection, representing the traditional approach employed in existing CL frameworks [16, 20]. This strategy provides an unbiased selection mechanism that ensures equal probability of

selection for all samples within a training batch while offering maximum computational efficiency.

The random sampling implementation operates through a probability-based selection mechanism where each sample has a 20% chance of being selected for memory storage:

$$P(\text{select sample } s_i) = 0.2 \quad (4.1)$$

Random sampling provides computational efficiency and serves as the control condition for evaluating the effectiveness of principled AL strategies. This approach ensures that performance improvements can be attributed to strategic selection rather than architectural modifications or implementation artifacts. Importantly, random sampling establishes the computational efficiency benchmark against which sophisticated strategies must demonstrate clear advantages to justify their overhead.

4.3.2 Uncertainty Sampling

Our uncertainty sampling implementation employs MTE as detailed in Section 3.3.2. For multimodal continual learning, MTE provides robust uncertainty estimation across the sequential generation process by aggregating entropy across all token positions, theoretically identifying samples with highest learning potential.

Mathematical Formulation: The MTE implementation performs forward passes to obtain logit distributions z_t at each token position t , computes probability distributions via softmax transformation, then calculates token-wise entropy:

$$H_t = - \sum_{v \in V} p_t(v) \log p_t(v) \quad (4.2)$$

where $p_t(v) = \frac{\exp(z_t^{(v)})}{\sum_{v' \in V} \exp(z_t^{(v')})}$ represents the probability of token v at position t . The final MTE score aggregates entropy across all sequence positions:

$$\text{MTE}(s) = \sum_{t=1}^T H_t \quad (4.3)$$

Computational Trade-off: While uncertainty sampling theoretically identifies high-value samples, it requires forward passes for uncertainty estimation, increasing computational cost compared to random baseline. Our evaluation explicitly measures whether this overhead translates to meaningful performance improvements.

4.3.3 Diversity Sampling

Our diversity sampling implementation follows the clustering-based selection methodology detailed in Section 3.3.3. The strategy leverages CLIP’s multimodal representations (see Section 3.1.3 for architectural details) to ensure comprehensive feature space coverage, theoretically providing better domain representation than random selection.

Feature Extraction and Fusion: CLIP embeddings are extracted for both visual and textual modalities, then fused via element-wise averaging to create unified multimodal representations:

$$f_{\text{fused}} = \frac{\text{CLIP}_{\text{vision}}(x_i) + \text{CLIP}_{\text{text}}(y_i)}{2} \quad (4.4)$$

Clustering-based Selection: K-means clustering partitions batches into k clusters based on fused representations, selecting the sample maximizing distance to each centroid to ensure representative coverage:

$$s_j^* = \arg \max_{s \in C_j} \|f_{\text{fused}}(s) - \mu_j\|_2 \quad (4.5)$$

where C_j represents cluster j and μ_j denotes the corresponding centroid.

Computational Trade-off: Diversity sampling incurs substantial computational overhead through CLIP feature extraction and clustering operations. Our systematic evaluation determines whether this sophistication provides meaningful advantages over random selection’s simplicity.

4.3.4 Hybrid Sampling

Our hybrid sampling strategy implements the HUDS methodology described in Section 3.3.4, adapting it for multimodal continual learning scenarios. The stratified approach addresses the limitations of simple weighted combinations, theoretically combining benefits of both uncertainty and diversity criteria.

Three-Phase HUDS Implementation: The implementation follows the systematic HUDS process:

Uncertainty Stratification: Using MTE scores from Section 3.3.2:

$$\text{Stratum}_i = \{s \in B : Q_{i-1} \leq \text{MTE}(s) < Q_i\} \quad (4.6)$$

where Q_i represents uncertainty quartile boundaries and B denotes the current batch.

Proportional Allocation: Sample allocation across strata maintains representativeness:

$$k_i = \left\lfloor \frac{k \cdot |\text{Stratum}_i|}{N} \right\rfloor \quad (4.7)$$

where k represents total samples to select and N denotes batch size.

Diversity Selection: K-means clustering on CLIP features (Section 3.3.3) selects diverse representatives within each stratum, using equal weighting ($\alpha = 0.5$) as established in the HUDS [25] framework.

Computational Trade-off: Hybrid sampling combines computational overhead from both uncertainty estimation and diversity clustering, representing the most sophisticated but computationally expensive approach. Our evaluation determines whether this complexity translates to meaningful improvements over simpler approaches.

4.4 Memory Management Strategies

This section presents the detailed mathematical formulations and algorithmic implementations of our memory buffer architectures, replay mechanisms, and deletion policies that enable effective knowledge retention while accommodating adaptation to new domains. The unified implementation ensures fair comparison across all sampling strategies. We begin with the episodic memory buffer architecture design (Section 4.4.1), followed by our memory replay mechanisms (Section 4.4.2), and conclude with our memory deletion policies (Section 4.4.3).

4.4.1 Episodic Memory Buffer Architecture

The *episodic memory buffer* \mathcal{M} implements strategic storage, retrieval, and maintenance of episodic experiences across CL episodes. The buffer maintains structured sample representations through a mapping:

$$\mathcal{M} : \mathcal{K} \rightarrow \mathcal{S} \times \mathcal{C} \times \mathcal{R} \quad (4.8)$$

where \mathcal{K} denotes unique sample identifiers, \mathcal{S} indicates the multimodal sample space, \mathcal{C} contains cluster assignments (visualized by color coding in Figure 4.2), and \mathcal{R} stores relevance scores from AL query functions.

Capacity Management and Allocation

The buffer enforces strict capacity constraints while maintaining domain-balanced representation across all sampling strategies. As demonstrated in Figure 4.2, initial buffer population begins with a single domain (Stage 1: Accessories at 100% allocation), with progressive evolution through six fashion categories requiring proportional reallocation as new domains are introduced.

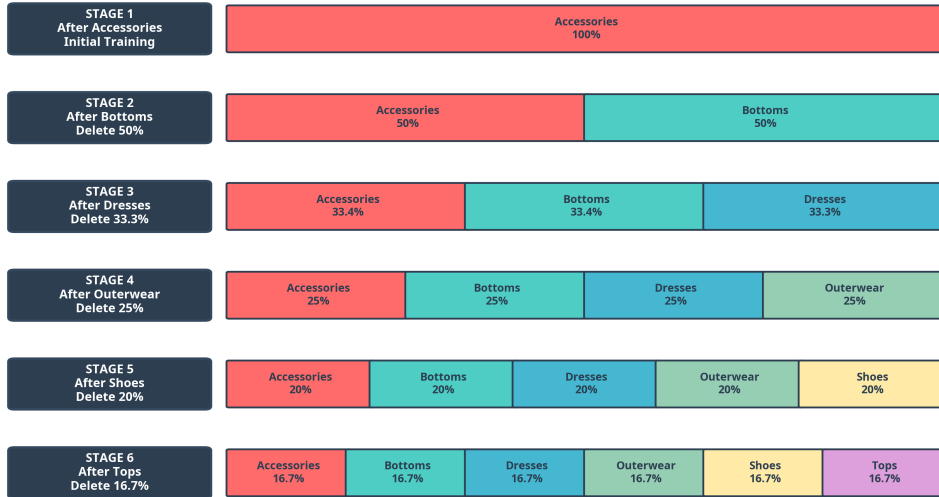


Figure 4.2: Progressive evolution of episodic memory buffer composition during continual learning across six fashion categories. Each stage (1-6) corresponds to the buffer state after training on a new domain: (1) Accessories, (2) Bottoms, (3) Dresses, (4) Outerwear, (5) Shoes, and (6) Tops. The proportional allocation demonstrates the cluster-aware deletion strategy defined in Equation 4.17, showing how different sampling strategies prioritize sample retention while maintaining balanced cross-domain representation within the fixed capacity constraint $\mathcal{C}_{\max} = 10,000$.

The buffer enforces a strict capacity constraint across all sampling strategies:

$$|\mathcal{M}| \leq \mathcal{C}_{\max} \quad (4.9)$$

where \mathcal{C}_{\max} is the maximum allowed buffer size, initially set to 10,000 samples to balance memory usage and effective sample retention. This constraint ensures fair comparison by providing identical resource limitations across random and strategic approaches. The impact of this constraint becomes evident in Stages 2-6 of Figure 4.2, where each new domain introduction necessitates a proportional reduction of existing allocations.

Score-based Sample Replacement

The memory buffer implements a sophisticated replacement policy that prioritizes retention of samples with high AL scores while ensuring domain-balanced representation. This approach extends traditional reservoir sampling methods by incorporating informativeness criteria derived from AL theory [22], enabling direct comparison of how different strategies utilize limited memory resources.

For each new sample candidate s_{new} with score ρ_{new} and cluster assignment c_{new} , the insertion policy operates as follows:

Capacity Check and Replacement Criteria:

$$\text{Insert}(s_{\text{new}}) = \begin{cases} \text{True} & \text{if } |\mathcal{M}| < \mathcal{C}_{\max} \\ \text{True} & \text{if } |\mathcal{M}| = \mathcal{C}_{\max} \text{ and } \rho_{\text{new}} > \rho_{\min}^{(c_{\text{new}})} \\ \text{False} & \text{otherwise} \end{cases} \quad (4.10)$$

where $\rho_{\min}^{(c)}$ represents the minimum relevance score among samples belonging to cluster c :

$$\rho_{\min}^{(c)} = \min_{s_i \in \mathcal{M}, c_i = c} \rho_i \quad (4.11)$$

Cluster-aware Deletion Strategy: When buffer capacity is exceeded, the replacement algorithm identifies the least relevant sample within the current cluster for deletion:

$$s_{\text{delete}} = \arg \min_{s_i \in \mathcal{M}, c_i = c_{\text{new}}} \rho_i \quad (4.12)$$

This *cluster-aware* policy ensures that memory replacement occurs within the same domain, preventing cross-domain interference while maintaining a balanced representation across previously encountered domains.

4.4.2 Memory Replay Mechanisms

Replay Frequency and Scheduling

The *memory replay* mechanism adopts a periodic sampling strategy, integrating stored experiences from the episodic memory buffer into the current training episode to mitigate CF [15]. In both OFA and BLIP-2 trainers, the replay schedule is managed by maintaining a running count of processed samples, with identical scheduling across all sampling strategies to ensure fair comparison.

Replay Trigger Condition:

$$\text{Trigger Replay} = \begin{cases} \text{True} & \text{if } N_{\text{processed}} \geq f_{\text{replay}} \\ \text{False} & \text{otherwise} \end{cases} \quad (4.13)$$

where $N_{\text{processed}}$ is the cumulative number of training samples processed since the last replay event, and $f_{\text{replay}} = 200$ represents the replay frequency parameter. After each replay, the sample counter is reset to zero.

Batch Composition for Memory Replay

The memory replay mechanism constructs *mini-batches* by sampling from the episodic memory buffer while explicitly excluding samples originating from the current training cluster. This exclusion prevents data leakage and maintains authentic cross-domain knowledge transfer throughout CL across all sampling strategies.

Cluster-stratified Replay Sampling: The memory replay mechanism adopts a *cluster-stratified* sampling strategy, where replay batches are constructed by sampling one representative from each previously encountered domain cluster. For every replay event, exactly one sample is drawn from each cluster preceding the current training domain, ensuring that all past domains are represented in the replay batch.

Formally, the replay batch $\mathcal{B}_{\text{replay}}$ at each memory replay event is constructed as follows:

$$\mathcal{B}_{\text{replay}} = \bigcup_{c \in \mathcal{C}_{\text{prev}}} \text{Sample}_{\phi}(\mathcal{M}_c, n_{\text{per-cluster}} = 1) \quad (4.14)$$

where:

- $\mathcal{C}_{\text{prev}} = \{c \mid 1 \leq c < c_{\text{current}}\}$ denotes the set of all previous domain clusters,
- $\mathcal{M}_c = \{s_i \in \mathcal{M} : c_i = c\}$ is the subset of memory samples assigned to cluster c ,
- $n_{\text{per-cluster}} = 1$ specifies that one sample is drawn from each cluster,
- $\text{Sample}_{\phi}(\cdot)$ indicates random sampling from the available samples in each cluster.

4.4.3 Memory Deletion Policies

Proportional Deletion Strategy

The ALCIE framework employs a *proportional deletion strategy* to manage its episodic memory buffer effectively during CL. This approach ensures that, as new domain data arrive, the memory buffer maintains capacity by proportionally removing samples from previously encountered domains while enabling each sampling strategy to retain its highest-priority samples according to strategy-specific criteria.

Domain Transition Memory Management: During each transition from domain d_{t-1} to domain d_t , the buffer applies a proportional deletion policy, removing a fixed fraction of samples from every previously encountered domain cluster:

$$\delta_t = \frac{1}{t} \quad (4.15)$$

where δ_t is the deletion fraction applied to all previous clusters when entering domain t .

Cluster-wise Deletion Implementation: For every prior cluster $c \in \{1, 2, \dots, t-1\}$, the number of samples to delete is given by:

$$n_{\text{delete}}^{(c)} = \lfloor \delta_t \cdot |\mathcal{M}_c| \rfloor \quad (4.16)$$

where \mathcal{M}_c is the set of buffer samples assigned to cluster c .

Samples to be deleted are selected according to the lowest relevance scores within each cluster, enabling each sampling strategy to retain its most valuable samples:

$$\mathcal{D}_c = \left\{ s_i \in \mathcal{M}_c \mid \rho_i \in \text{Min}_{n_{\text{delete}}^{(c)}} (\{\rho_j : s_j \in \mathcal{M}_c\}) \right\} \quad (4.17)$$

where ρ_j denotes the relevance score of sample s_j as determined by the AL strategy, and the $n_{\text{delete}}^{(c)}$ samples with the lowest scores are selected for removal within each cluster.

Score-based Retention Criteria

The memory buffer employs *score-based retention* criteria to prioritize the preservation of the most informative samples during domain transitions. Each sample is assigned a relevance score according to the AL strategy in use, such as uncertainty, diversity, hybrid combination, or random assignment. During memory maintenance, samples with higher relevance scores are preferentially retained, while those with lower scores are removed to free up space. This mechanism enables direct comparison of how different strategies utilize limited memory resources.

Implementation of these retention and replay strategies is fully integrated into the training pipeline through standard pickle-based serialization for buffer persistence, ensuring that sample selection and buffer maintenance are consistently aligned with the chosen AL objectives while maintaining identical conditions across all strategies.

4.5 Training Infrastructure and Implementation

This section presents the implementation details of our comprehensive training infrastructure designed to support scalable, reproducible, and systematic experimentation across different AL strategies and vision-language models. The infrastructure ensures fair comparison between strategic and random sampling approaches by maintaining identical training conditions while measuring both performance and computational efficiency. We begin with our system architecture and modular design (Section 4.5.1), followed by our CL protocol implementation (Section 4.5.2), and conclude with our hyperparameter configurations and training settings (Section 4.5.3).

4.5.1 System Architecture and Modular Design

The implementation realizes the modular design principles established in Section 4.1.2, enabling systematic comparison of memory management strategies while maintaining computational transparency.

Hierarchical Software Architecture

Model Abstraction Layer: The framework offers unified interfaces for both encoder-decoder architectures (OFA) [2] and cross-modal architectures (BLIP-2) [3] via standardized trainer classes. Each model-specific trainer inherits from the base Hugging Face Trainer class and implements custom training steps that integrate AL selection mechanisms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{primary}} + \lambda_{\text{replay}} \mathcal{L}_{\text{replay}} \quad (4.18)$$

where $\mathcal{L}_{\text{primary}}$ represents the loss on current domain data, $\mathcal{L}_{\text{replay}}$ denotes the loss on episodic memory samples, and $\lambda_{\text{replay}} = 1.0$ ensures balanced weighting between current and prior knowledge across all strategies.

Memory Management Layer: The framework implements four distinct memory buffer classes, each corresponding to a specific AL strategy (Random, Uncertainty, Diversity, Hybrid). All memory buffers inherit from a shared interface that standardizes key operations, including sample insertion, deletion, and replay batch construction, ensuring identical memory management conditions across strategies.

Orchestration Layer: A shell script-based distributed training system manages resource allocation and experiment scheduling. The orchestration scripts automatically handle model checkpointing, memory buffer persistence, and cross-cluster training coordination while tracking computational efficiency metrics.

Cross-Model Compatibility Framework

To ensure consistent evaluation across different vision-language architectures, ALCIE implements model-agnostic data processing and AL mechanisms:

Unified Data Collation: The framework provides specialized data collators for each model architecture while maintaining consistent batch composition. For OFA models, batches contain `patch_images`, `input_ids`, and `decoder_input_ids` with `return_loss=True`. For BLIP-2 models, batches comprise `pixel_values`, `input_ids`, and `labels` with appropriate attention masking.

Model-Agnostic Scoring: AL strategies utilize CLIP ViT-B/32 embeddings for diversity assessment and model-specific uncertainty computation, ensuring comparable sample selection criteria across architectures while maintaining strategy-specific computational characteristics.

4.5.2 Continual Learning Protocol Implementation

Sequential Domain Training Pipeline

The CL protocol organizes training as a sequence of six domain phases corresponding to the FACAD fashion categories (Section 4.2.1), with identical progression across all sampling strategies to enable systematic comparison.

Domain Transition Management: Between consecutive domains, the framework implements automatic memory buffer management through proportional deletion policies.

The deletion percentage for domain t is computed as:

$$\delta_t = \frac{1}{t} \quad (4.19)$$

Active Learning Integration: During each training step, samples are scored according to the AL strategy in use. The top- k informative samples (typically $k = 16$ for computational efficiency) are selected for episodic memory storage. Random sampling provides the computational baseline against which sophisticated strategies are evaluated.

Memory Replay Scheduling

The framework implements periodic memory replay at regular intervals defined by the replay frequency parameter $f_{\text{replay}} = 200$ samples. At each replay step, the current batch is augmented with strategically sampled experiences from the episodic memory buffer.

Cluster-Stratified Sampling: Memory replay constructs batches by sampling one representative from each previously encountered domain cluster, ensuring balanced exposure to historical knowledge across all strategies.

Gradient Integration: The replay mechanism computes separate forward and backward passes for memory samples, then combines gradients with current domain updates through standard loss addition.

4.5.3 Hyperparameter Configuration and Training Settings

The framework employs systematically tuned hyperparameters that address the unique challenges of CL in vision-language models. Each hyperparameter selection is motivated by extensive preliminary experiments and theoretical considerations for maintaining stability across sequential domain training while ensuring fair comparison across sampling strategies.

| Hyperparameter | BLIP-2 | OFA |
|-----------------------|----------------------|--------------------|
| Learning Rate | 5×10^{-5} | 5×10^{-5} |
| Batch Size | 32 | 64 |
| Optimizer | AdamW | AdamW |
| β_1, β_2 | 0.9, 0.98 | 0.9, 0.98 |
| Weight Decay | 0.1 | 0.0 |
| LR Schedule | Cosine with restarts | Constant |
| Warmup Steps | 500 | 500 |
| Epochs per Domain | 20 | 20 |
| Max Sequence Length | 128 | 512 |
| Gradient Accumulation | 2 steps | 1 step |
| Precision | Mixed (FP16) | Mixed (FP16) |
| Memory Capacity | 10,000 | 10,000 |
| Replay Frequency | 200 samples | 200 samples |

Table 4.1: Comprehensive hyperparameter configuration ensuring reproducible training and fair comparison across model architectures and AL strategies.

Hyperparameter Selection and Computational Considerations

Key hyperparameters are selected for continual learning stability and fair comparison across sampling strategies. Learning rate (5×10^{-5}) balances adaptation with knowledge retention across all approaches. Architecture-specific configurations account for parameter differences: BLIP-2 uses smaller batches (32) with gradient accumulation, while OFA uses larger batches (64) directly. Learning rate scheduling differs by design paradigm (cosine annealing for BLIP-2, constant with warmup for OFA). Mixed precision training optimizes memory usage while maintaining stability.

Computational Efficiency Measurement: The infrastructure tracks computational overhead for each sampling strategy, including time for uncertainty estimation, CLIP feature extraction, clustering operations, and memory management. This measurement enables explicit evaluation of efficiency trade-offs alongside performance improvements.

This implementation framework establishes a robust foundation for systematic evaluation of AL strategies in CL settings. The modular design ensures fair comparison across approaches while maintaining compatibility with diverse vision-language models and explicit measurement of computational trade-offs. The comprehensive infrastructure enables definitive assessment of whether strategic memory management provides meaningful advantages over random sampling baselines, directly addressing our central research question.

Chapter 5

Experiments and Results

This chapter presents a comprehensive empirical evaluation of the ALCIE framework across multiple vision-language architectures and active learning strategies. We begin with a detailed description of our experimental setup and implementation details (Section 5.1), followed by our evaluation methodology and metrics framework (Section 5.2). We then conduct systematic experiments across six memory management strategies (Section 5.3), and conclude with a comprehensive discussion that examines performance patterns, architectural dependencies, and practical implications for continual learning system design (Section 5.4).

5.1 Experimental Setup

Our experimental evaluation systematically assesses the effectiveness of strategic episodic memory management in continual image captioning scenarios. The evaluation framework encompasses multiple dimensions of analysis, enabling thorough assessment of both immediate performance and long-term knowledge retention.

5.1.1 Dataset and Domain Configuration

The experimental evaluation utilizes the *FACAD* dataset [28], a comprehensive fashion-focused image captioning dataset that provides rich multimodal content across diverse clothing and accessory categories.

FACAD Dataset Characteristics

Scale and Composition: The FACAD dataset encompasses 993,000 high-quality image caption pairs distributed across 24 fashion categories including bags, boots, coats, dresses, jackets, jeans, pants, pumps, sandals, shorts, sneakers, sweaters, tees, and tops. Each image is accompanied by detailed natural language descriptions that capture both visual attributes (color, texture, style) and functional characteristics (occasion, season, fit)

relevant to fashion e-commerce applications. Key dataset statistics, including the number of images per domain, data splits, and caption lengths, are summarized in Table 5.1.

| Dataset Statistics | Value |
|------------------------------|---------------|
| Total Images (Experimental) | 72,000 |
| Domain Clusters | 6 |
| Images per Domain | 12,000 |
| Training Images per Domain | 9,000 |
| Test Images per Domain | 2,000 |
| Validation Images per Domain | 1,000 |
| Average Caption Length | 21 words |
| Original FACAD Categories | 24 |
| Total FACAD Dataset | 993,000 pairs |

Table 5.1: FACAD dataset statistics for continual learning experiments.

Linguistic Diversity: The captions in FACAD exhibit rich linguistic patterns with domain-specific vocabulary, ranging from technical fashion terminology ("asymmetrical hemline," "ribbed knit texture") to descriptive attributes ("flowing silhouette," "vintage-inspired design"). The average caption length is 21 words, providing comprehensive descriptive content that enables detailed vision-language learning while presenting substantial sequence generation challenges for continual learning scenarios.

Visual Characteristics: Images in FACAD Figure 5.1 represent real fashion products photographed in controlled e-commerce settings, featuring consistent lighting, backgrounds, and presentation styles. This consistency enables systematic evaluation of domain transfer, while the diverse product categories provide natural semantic boundaries for continual learning scenarios.

Domain Clustering for Continual Learning

For continual learning evaluation, the 24 fine-grained FACAD categories are systematically grouped into six semantically coherent clusters that represent natural fashion domain boundaries. Each cluster is carefully balanced to ensure consistent experimental conditions across domains.

Dataset Split Configuration: Each of the six domain clusters contains exactly 10,000 training images, 2,000 test images, and 1,000 evaluation images (derived as 10% of the training set), resulting in 12,000 total images per domain and 72,000 images across the complete experimental setup. This balanced configuration ensures a fair comparison across domains while providing sufficient data for both training and comprehensive evaluation. The evaluation split is used for hyperparameter tuning and model selection during training, while the test split provides an unbiased performance assessment for continual learning evaluation.

Cluster 1 - Accessories: Comprises bags and related accessories, focusing on portable fashion items with an emphasis on functionality, materials (leather, canvas, fabric), and

FACADE Fashion Dataset: Category Overview

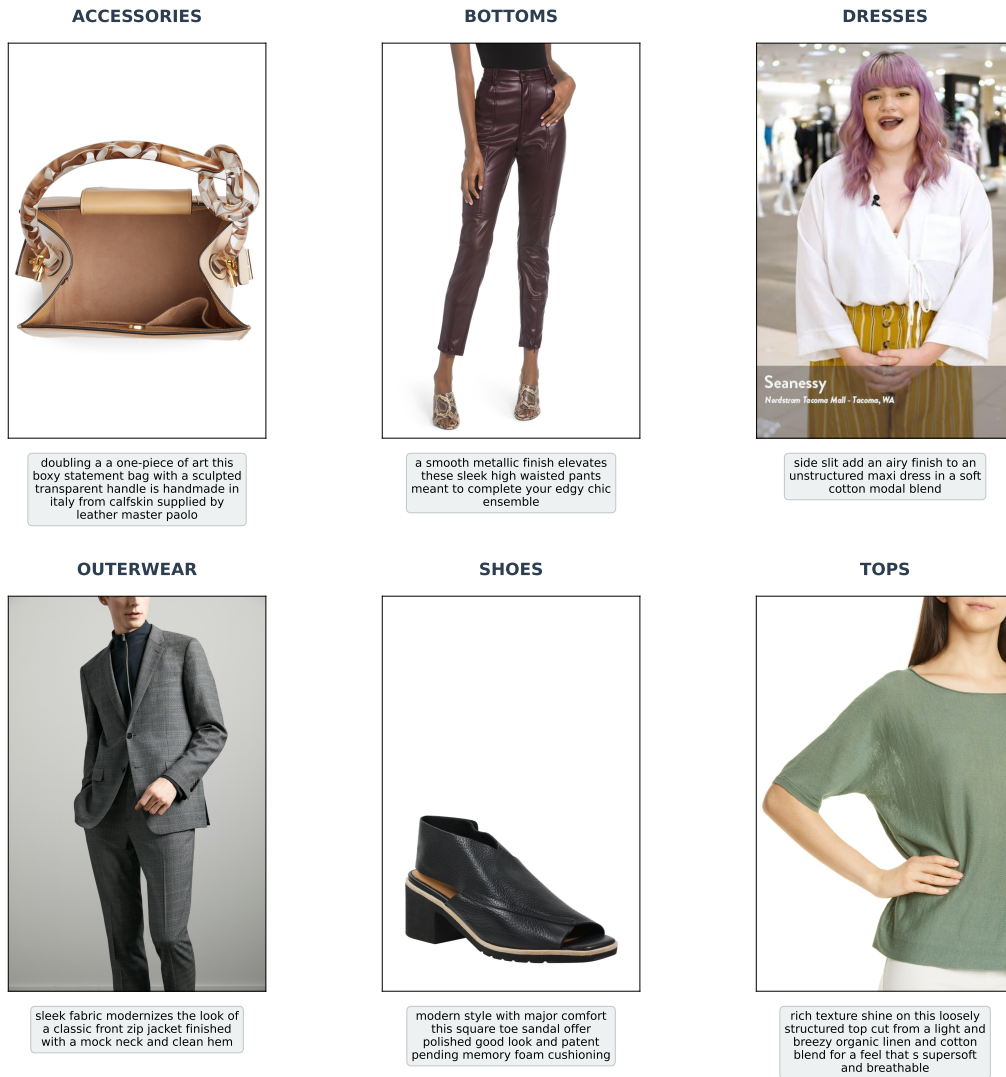


Figure 5.1: Representative FACAD dataset samples from six fashion categories used in experiments.

usage contexts (casual, formal, travel). The vocabulary centers on storage features, design elements, and practical characteristics essential for accessory descriptions.

Cluster 2 - Bottoms: Aggregates pants, jeans, and shorts into a unified lower-body garment domain. This cluster features diverse fits (slim, relaxed, wide-leg), materials (denim, cotton, synthetic blends), and style variations (casual, formal, athletic). The linguistic patterns emphasize fit characteristics, length variations, and styling contexts.

Cluster 3 - Dresses: Constitutes a standalone category due to dresses' unique characteristics as complete garments spanning multiple body regions. This domain features rich vocabulary describing silhouettes (A-line, sheath, wrap), necklines (scoop, V-neck, halter), and occasion-specific styling (cocktail, casual, formal).

Cluster 4 - Outerwear: Combines coats and jackets into a single domain focused on layering garments and weather-appropriate clothing. The cluster includes specialized vocabulary related to insulation (down, fleece, wool), weather protection (waterproof, windproof), and seasonal usage (winter, transitional, lightweight).

Cluster 5 - Shoes: Unifies boots, sandals, sneakers, and pumps into a comprehensive footwear domain. The vocabulary spans functional aspects (comfort, support, traction), style categories (athletic, formal, casual), and construction details (heel height, toe shape, closure types).

Cluster 6 - Tops: Aggregates tops, tees, and sweaters into a unified upper-body garment domain. This cluster encompasses diverse garment types with vocabulary focusing on fit (fitted, loose, oversized), necklines (crew, V-neck, scoop), and fabric characteristics (cotton, wool, synthetic blends).

5.2 Evaluation Metrics

Our comprehensive evaluation employs both traditional image captioning IC metrics and continual learning CL specific measures to assess model performance across multiple dimensions of caption quality and knowledge retention. We evaluate model performance using multiple established image captioning metrics to ensure comprehensive assessment. While all metrics are computed across experiments, the main body of this work presents results using BLEU-4 and BERTScore-F1 as our primary evaluation measures, with additional metrics detailed in supplementary analysis.

BLEU-4: We adopt BLEU-4 [73] as our primary lexical evaluation metric because of its widespread adoption and established benchmarking role within the vision-language community. It measures n-gram precision up to four grams and incorporates a brevity penalty, providing a standard quantification of lexical similarity between generated and reference captions. All results are reported in continual learning matrix format, illustrating performance across all domain combinations during sequential learning.

BERTScore-F1: We employ BERTScore-F1 as our primary semantic evaluation metric due to its ability to capture semantic meaning and its robustness to paraphrasing. Unlike n-gram-based approaches, it leverages contextual embeddings to assess semantic similarity, making it particularly suitable for evaluating knowledge retention, where retention of semantic understanding is critical. Results are presented in continual learning matrix format corresponding to the BLEU-4 tables.

Additional Metrics: We further employ ROUGE-L [75] and METEOR [74], all presented in continual learning matrix format. These supplementary metrics provide complementary perspectives on caption generation quality and support comprehensive evaluation across different aspects of caption quality. Detailed results for all additional metrics are provided in Appendix 7.4.4.

From the BERTScore-F1 performance matrices, we derive three key continual learning summary metrics, following the established framework of Lopez-Paz and Ranzato [16]:

ACC: This metric measures overall performance across all tasks after sequential learning and represents the model’s final capability across the entire task sequence.

BWT: Backward transfer quantifies knowledge retention by measuring the change in performance on previous tasks due to subsequent learning. Negative values indicate catastrophic forgetting, whereas positive values suggest beneficial knowledge transfer.

AF: Average forgetting directly measures performance degradation from initial learning to final evaluation, providing an intuitive quantification of knowledge loss. Higher values indicate more significant forgetting.

These metrics enable comparative analysis across all memory management strategies, and the results are presented graphically in Section 5.3.

5.3 Experiments

In this section, we present the results of our comprehensive experimental evaluation. Each experiment was conducted over three independent runs with different random seeds to ensure robustness and to capture potential sources of variability.

Experiment 1: We benchmark the catastrophic forgetting baseline without episodic memory (No Memory), establishing the severity of knowledge degradation in sequential learning scenarios, following established experimental protocols [12].

Experiment 2: We compare traditional random sampling memory management as proposed by Anagnostopoulou et al. [12] against the no-memory baseline, reflecting standard practice in the continual learning literature.

Experiment 3: We assess unrestricted memory growth through random sampling without deletion constraints, providing an upper bound for memory-based approaches.

Experiment 4: We evaluate uncertainty-based sample selection using MTE [96], which targets samples with high model uncertainty in the caption generation sequence.

Experiment 5: We investigate diversity-based memory management via CLIP-based clustering [97], ensuring representative coverage of the multimodal feature space.

Experiment 6: We conduct a comparative analysis of hybrid sampling strategies that combine uncertainty and diversity selection through HUDS [25], leveraging the benefits of both informative and representative sample selection.

For each experiment, we provide detailed analyses of results across both the OFA and BLIP-2 architectures, using evaluation metrics such as BLEU-4, ROUGE-L, METEOR, BERTScore, ACC, BWT, and AF. The subsequent discussion interprets the implications of these results for continual learning in fashion image captioning.

5.3.1 Experimental Protocol

Multiple Run Validation: To ensure statistical reliability and account for training variability, each experimental configuration was executed three times using different random seeds (42, 123, 456). These seeds control the initialization of model parameters, data shuffling, and memory selection processes. The final results are reported as the arithmetic mean across these three independent runs, providing robust performance estimates that mitigate the effects of random initialization and sampling variation.

Sequential Learning Protocol: All experiments follow a predetermined training sequence: Accessories → Bottoms → Dresses → Outerwear → Shoes → Tops. Model performance is assessed after each domain training phase using the complete test sets from all previously encountered domains, enabling systematic measurement of both knowledge retention and overall performance throughout the learning process.

Memory Management: All constrained memory strategies operate under identical storage budgets (10,000 samples) with proportional deletion at domain transitions, following established practices [98] for complex datasets like MiniImageNet [99]. This capacity (16.7% of total training data) establishes an upper bound enabling complete single-domain retention while maintaining realistic deployment constraints. Memory replay occurs every 200 training samples using batch sampling for balanced rehearsal across previously learned domains [12].

5.3.2 Experiment 1: No Memory Baseline

This experiment establishes the catastrophic forgetting baseline by training models sequentially on fashion categories without any episodic memory mechanism. Standard fine-tuning is applied at each domain transition, allowing complete parameter adaptation without retention of previous knowledge. This configuration serves as the critical baseline for assessing the effectiveness of all subsequent memory-based approaches.

Results

Tables 5.2 and 5.3 present the evaluation results for both OFA and BLIP-2 architectures, demonstrating the significant impact of catastrophic forgetting in the absence of memory mechanisms.

| BLEU-4 Performance Without Memory | | | | | | | | | | | | |
|-----------------------------------|---------------|--------------|-----------------------|-------------------------------|-----------------------|-------------------------------|------------------------------|-------------------------------|------------------------------|------------------------------|------------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.320 | 0.710 | 0.021 _(7%) | 0.071 _(10%) | 0.004 _(1%) | 0.005 _(1%) | 0.002 _(1%) | 0.000 _(0%) | 0.001 _(0%) | 0.001 _(0%) | 0.000 | 0.000 |
| Bottoms | | | 0.270 | 0.504 | 0.007 _(3%) | 0.083 _(16%) | 0.006 _(2%) | 0.014 _(3%) | 0.000 _(0%) | 0.001 _(0%) | 0.002 _(1%) | 0.011 _(2%) |
| Dresses | | | | | 0.369 | 0.718 | 0.029 _(8%) | 0.069 _(10%) | 0.002 _(1%) | 0.001 _(0%) | 0.003 _(1%) | 0.006 _(1%) |
| Outerwear | | | | | | | 0.374 | 0.705 | 0.030 _(8%) | 0.023 _(3%) | 0.030 _(8%) | 0.014 _(2%) |
| Shoes | | | | | | | | | 0.261 | 0.570 | 0.035 _(13%) | 0.301 _(53%) |
| Tops | | | | | | | | | | | 0.445 | 0.782 |

Table 5.2: BLEU scores showing complete catastrophic forgetting without episodic memory. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Analysis

Severe Lexical Generation Degradation: Without memory mechanisms, both architectures exhibit complete lexical generation failure. BLIP-2 shows catastrophic transition drops as demonstrated in Table 5.2: Accessories **BLEU-4 retention drops to 10.0%** after Bottoms training (0.710 \rightarrow 0.071), then to **0.7%** after Dresses (0.710 \rightarrow 0.005), reaching **0.0% final retention** by sequence end (0.710 \rightarrow 0.000). OFA demonstrates similar patterns with **6.6% BLEU-4 retention** after Bottoms (0.320 \rightarrow 0.021) and complete elimination by

| BERTScore-F1 Performance Without Memory | | | | | | | | | | | | |
|---|---------------|--------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.898 | 0.928 | 0.839 _(93%) | 0.844 _(91%) | 0.832 _(93%) | 0.832 _(90%) | 0.836 _(93%) | 0.836 _(90%) | 0.838 _(93%) | 0.838 _(90%) | 0.831 _(93%) | 0.831 _(90%) |
| Bottoms | | | 0.891 | 0.911 | 0.840 _(94%) | 0.852 _(94%) | 0.841 _(94%) | 0.841 _(92%) | 0.838 _(94%) | 0.838 _(92%) | 0.837 _(94%) | 0.838 _(92%) |
| Dresses | | | | | 0.902 | 0.930 | 0.844 _(94%) | 0.847 _(91%) | 0.837 _(93%) | 0.835 _(90%) | 0.840 _(93%) | 0.840 _(90%) |
| Outerwear | | | | | | | 0.905 | 0.931 | 0.847 _(94%) | 0.845 _(91%) | 0.839 _(93%) | 0.841 _(90%) |
| Shoes | | | | | | | | | 0.895 | 0.922 | 0.842 _(94%) | 0.886 _(96%) |
| Tops | | | | | | | | | | | 0.915 | 0.943 |

Table 5.3: BERTScore-F1 scores demonstrating severe semantic degradation despite better retention than lexical metrics. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

final domains. These extreme drops establish the severity of catastrophic forgetting in sequential learning scenarios.

Semantic Preservation Versus Lexical Collapse: Table 5.3 reveals fundamentally different forgetting patterns compared to BLEU-4 metrics. BLIP-2 maintains **91% BERTScore-F1 retention** for Accessories ($0.928 \rightarrow 0.844$) after Bottoms training, in *stark contrast* to the **10.0% BLEU-4 retention**. Throughout the sequence, semantic degradation remains moderate: final Accessories **BERTScore-F1 retention of 90%** ($0.928 \rightarrow 0.831$) versus **0.0% BLEU-4 retention** ($0.710 \rightarrow 0.000$). This disparity indicates that models preserve conceptual understanding while losing surface-level generation capabilities.

Architecture-Specific Forgetting Patterns: BLIP-2 demonstrates superior initial performance ($2.22\times$ higher initial BLEU-4 scores) as shown in Table 5.2 but exhibits similar catastrophic forgetting trajectories to OFA. Despite the frozen visual encoder architecture, BLIP-2 shows **90% final BERTScore-F1 retention** compared to OFA’s **93% retention** (Table 5.3), indicating that architectural sophistication alone cannot prevent semantic degradation or lexical generation collapse without explicit memory mechanisms.

Progressive Interference Accumulation: The data in Table 5.2 reveal systematic interference accumulation. Middle domains suffer severe degradation: Bottoms **BLEU-4 retention drops from 16.5% to 2.2%** across subsequent learning phases for BLIP-2. This progressive deterioration demonstrates that catastrophic forgetting compounds over extended learning sequences.

Performance Baseline Establishment: These results establish quantitative thresholds for evaluating memory management effectiveness. Any viable memory strategy must achieve substantial improvements over the **0-53% BLEU-4 retention** and **90-96% BERTScore-F1 retention** observed without memory in Tables 5.2 and 5.3.

5.3.3 Experiments 2-3: Random Sampling Memory Management

These experiments evaluate random sampling approaches for episodic memory management as proposed by Anagnostopoulou et al. [12], representing both constrained and unconstrained memory scenarios. Experiment 2 implements traditional random

sampling with a fixed memory buffer of 10,000 samples (DEL +), while Experiment 3 removes storage constraints through unlimited memory growth (DEL -). Together, these configurations establish the performance range achievable through random memory selection strategies.

Results

Tables 5.4, 5.5, 5.6, and 5.7 present comprehensive evaluation results comparing constrained (DEL+) and unconstrained (DEL-) random sampling approaches.

| BLEU-4 Performance: OFA Random Sampling | | | | | | | | | | | | |
|---|---------------|-------|------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| Test Domain | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.320 | 0.316 | 0.111 _(35%) | 0.132 _(42%) | 0.128 _(40%) | 0.124 _(39%) | 0.151 _(47%) | 0.136 _(43%) | 0.127 _(40%) | 0.137 _(43%) | 0.125 _(39%) | 0.146 _(46%) |
| Bottoms | | | 0.266 | 0.259 | 0.067 _(25%) | 0.071 _(27%) | 0.065 _(24%) | 0.071 _(27%) | 0.065 _(24%) | 0.059 _(23%) | 0.055 _(21%) | 0.069 _(27%) |
| Dresses | | | | | 0.349 | 0.318 | 0.091 _(26%) | 0.077 _(24%) | 0.079 _(23%) | 0.080 _(25%) | 0.051 _(15%) | 0.062 _(19%) |
| Outerwear | | | | | | | 0.365 | 0.371 | 0.154 _(42%) | 0.121 _(33%) | 0.085 _(23%) | 0.098 _(26%) |
| Shoes | | | | | | | | | 0.255 | 0.245 | 0.116 _(45%) | 0.127 _(52%) |
| Tops | | | | | | | | | | | 0.465 | 0.459 |

Table 5.4: BLEU-4 performance comparison for OFA with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

| BLEU-4 Performance: BLIP2 Random Sampling | | | | | | | | | | | | |
|---|---------------|--------------|-------------------------------|------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|
| | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| Test Domain | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.709 | 0.710 | 0.222 _(31%) | 0.213 _(30%) | 0.195 _(28%) | 0.221 _(31%) | 0.220 _(31%) | 0.231 _(33%) | 0.205 _(29%) | 0.226 _(32%) | 0.248 _(35%) | 0.259 _(36%) |
| Bottoms | | | 0.480 | 0.493 | 0.127 _(26%) | 0.134 _(27%) | 0.092 _(19%) | 0.106 _(21%) | 0.091 _(19%) | 0.112 _(23%) | 0.097 _(20%) | 0.110 _(22%) |
| Dresses | | | | | 0.702 | 0.699 | 0.201 _(29%) | 0.196 _(28%) | 0.091 _(13%) | 0.119 _(17%) | 0.064 _(9%) | 0.063 _(9%) |
| Outerwear | | | | | | | 0.705 | 0.698 | 0.214 _(30%) | 0.367 _(53%) | 0.141 _(20%) | 0.168 _(24%) |
| Shoes | | | | | | | | | 0.637 | 0.643 | 0.477 _(75%) | 0.456 _(71%) |
| Tops | | | | | | | | | | | 0.772 | 0.774 |

Table 5.5: BLEU-4 performance comparison for BLIP-2 with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

Analysis

Substantial Knowledge Preservation Through Memory Mechanisms: Random sampling demonstrates significant improvement over no-memory baselines across both

| BERTScore-F1 Performance: OFA Random Sampling | | | | | | | | | | | | |
|---|---------------|--------------|------------------------|-------------------------------|------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.898 | 0.899 | 0.864 _(96%) | 0.866 _(96%) | 0.866 _(96%) | 0.864 _(96%) | 0.870 _(97%) | 0.868 _(97%) | 0.866 _(96%) | 0.867 _(96%) | 0.867 _(97%) | 0.868 _(97%) |
| Bottoms | | | 0.891 | 0.890 | 0.857 _(96%) | 0.857 _(96%) | 0.858 _(96%) | 0.858 _(96%) | 0.858 _(96%) | 0.856 _(96%) | 0.855 _(96%) | 0.856 _(96%) |
| Dresses | | | | | 0.899 | 0.894 | 0.857 _(95%) | 0.855 _(96%) | 0.857 _(95%) | 0.856 _(96%) | 0.850 _(95%) | 0.851 _(95%) |
| Outerwear | | | | | | | 0.905 | 0.905 | 0.869 _(96%) | 0.866 _(96%) | 0.858 _(95%) | 0.859 _(95%) |
| Shoes | | | | | | | | | 0.892 | 0.892 | 0.872 _(98%) | 0.874 _(98%) |
| Tops | | | | | | | | | | | 0.917 | 0.919 |

Table 5.6: BERTScore-F1 performance comparison for OFA with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

| BERTScore-F1 Performance: BLIP2 Random Sampling | | | | | | | | | | | | |
|---|---------------|--------------|-------------------------------|------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.928 | 0.928 | 0.874 _(94%) | 0.871 _(94%) | 0.871 _(94%) | 0.873 _(94%) | 0.871 _(94%) | 0.874 _(94%) | 0.869 _(94%) | 0.872 _(94%) | 0.875 _(94%) | 0.877 _(94%) |
| Bottoms | | | 0.908 | 0.910 | 0.863 _(95%) | 0.862 _(95%) | 0.859 _(95%) | 0.861 _(95%) | 0.859 _(95%) | 0.861 _(95%) | 0.859 _(95%) | 0.860 _(95%) |
| Dresses | | | | | 0.928 | 0.928 | 0.867 _(93%) | 0.867 _(93%) | 0.853 _(92%) | 0.858 _(92%) | 0.849 _(92%) | 0.849 _(92%) |
| Outerwear | | | | | | | 0.931 | 0.930 | 0.873 _(94%) | 0.890 _(96%) | 0.863 _(93%) | 0.866 _(93%) |
| Shoes | | | | | | | | | 0.929 | 0.930 | 0.912 _(98%) | 0.909 _(98%) |
| Tops | | | | | | | | | | | 0.942 | 0.942 |

Table 5.7: BERTScore-F1 performance comparison for BLIP-2 with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

architectures. Table 5.5 shows BLIP-2 constrained (DEL+) random sampling achieves **35% Accessories retention** (0.709 → 0.248) compared to *only 0%* without memory, representing a *substantial improvement*. Unconstrained (DEL-) random sampling achieves strong **36% Accessories final retention** (0.710 → 0.259), demonstrating that simple episodic memory mechanisms provide fundamental continual learning capabilities with potential benefits from unlimited memory capacity.

Benefits from Unconstrained Memory Capacity: Unconstrained (DEL-) memory demonstrates substantial effectiveness across multiple domains as revealed in Tables 5.4 and 5.5. BLIP-2 unconstrained (DEL-) sampling achieves superior performance in key domains: Accessories final retention of **36%** (0.710 → 0.259) versus constrained (DEL+) **35%** (0.709 → 0.248), and particularly strong Outerwear retention showing **53% performance** (0.698 → 0.367) compared to constrained (DEL+) **30%** (0.705 → 0.214). However, constrained (DEL+) memory maintains advantages in Shoes retention with **75%** versus unconstrained (DEL-) **71%**. Conversely, OFA demonstrates consistent improvements

with unconstrained memory across all domains, with Accessories retention increasing from 39% (constrained) to 46% (unconstrained), indicating architecture-dependent memory utilization patterns where BLIP-2 shows domain-specific preferences while OFA benefits uniformly from unconstrained (DEL-) capacity.

Architecture-Specific Memory Strategy Effectiveness: BLIP-2 maintains superior absolute performance (2–3× *higher* BLEU-4 scores) with nuanced responses to memory constraints, showing domain-specific optimization patterns. Tables 5.6 and 5.7 demonstrate BERTScore-F1 retention remains highly stable across approaches: BLIP-2 achieves **94% semantic retention** for both constrained (DEL+) and unconstrained (DEL-) approaches, indicating robust semantic preservation regardless of memory capacity constraints.

Domain-Specific Performance Advantages: The results reveal compelling domain-dependent memory capacity effects. Table 5.5 shows unconstrained (DEL-) memory particularly excels in mid-sequence domains: Outerwear achieves remarkable **53% retention** versus constrained (DEL+) **30%**, representing a **23 percentage point advantage**. Conversely, later domains like Shoes show constrained (DEL+) memory advantages (**75% vs 71%**), while early and final domains (Accessories, Tops) demonstrate comparable performance across both approaches, suggesting that optimal memory capacity strategies should adapt to domain sequence position.

Computational Efficiency Versus Performance Trade-offs: The domain-specific improvements from unlimited memory (ranging from equivalent performance to 23 percentage point enhancements) demonstrate that memory capacity optimization yields substantial benefits in specific contexts. While unconstrained (DEL-) approaches require greater computational resources, the significant performance gains in critical mid-sequence domains support adaptive memory management strategies that balance storage constraints with domain-specific retention requirements, particularly for applications where mid-sequence knowledge preservation is crucial.

5.3.4 Experiment 4: Uncertainty Sampling Memory Management

This experiment evaluates uncertainty sampling for episodic memory management, representing an active learning approach that prioritizes samples with the highest model uncertainty. The strategy selects memory samples according to prediction confidence using MTE [96], specifically targeting instances where the model exhibits the greatest uncertainty to maximize learning efficiency.

Results

Tables 5.8 and 5.9 present the evaluation results for uncertainty sampling across both OFA and BLIP-2 architectures.

Analysis

Uncertainty Sampling Shows Mixed Mid-Sequence Performance: Table 5.8 reveals that uncertainty sampling achieves moderate results in middle domains. BLIP-2 Dresses maintains **9% final retention** (0.705 → 0.060), comparable to random sampling’s **9%** (constrained) but with different transition patterns. Bottoms retention reaches **19% final performance** (0.494 → 0.096) compared to random sampling’s **20%** (constrained, 0.480).

| BLEU-4 Performance: Uncertainty Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.320 | 0.710 | 0.116 _(36%) | 0.211 _(30%) | 0.103 _(32%) | 0.181 _(25%) | 0.081 _(25%) | 0.208 _(29%) | 0.084 _(26%) | 0.199 _(28%) | 0.096 _(30%) | 0.225 _(32%) |
| Bottoms | | | 0.247 | 0.494 | 0.060 _(24%) | 0.101 _(20%) | 0.046 _(19%) | 0.068 _(14%) | 0.045 _(18%) | 0.082 _(17%) | 0.052 _(21%) | 0.096 _(19%) |
| Dresses | | | | | 0.360 | 0.705 | 0.076 _(21%) | 0.180 _(26%) | 0.064 _(18%) | 0.082 _(12%) | 0.053 _(15%) | 0.060 _(9%) |
| Outerwear | | | | | | | 0.352 | 0.703 | 0.110 _(31%) | 0.347 _(49%) | 0.079 _(22%) | 0.163 _(23%) |
| Shoes | | | | | | | | | 0.243 | 0.646 | 0.092 _(38%) | 0.465 _(72%) |
| Tops | | | | | | | | | | | 0.458 | 0.777 |

Table 5.8: BLEU-4 performance with uncertainty-based memory selection. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

| BERTScore-F1 Performance: Uncertainty Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.898 | 0.928 | 0.864 _(96%) | 0.872 _(94%) | 0.859 _(96%) | 0.870 _(94%) | 0.858 _(96%) | 0.870 _(94%) | 0.856 _(95%) | 0.871 _(94%) | 0.859 _(96%) | 0.875 _(94%) |
| Bottoms | | | 0.888 | 0.910 | 0.854 _(96%) | 0.859 _(94%) | 0.852 _(96%) | 0.855 _(94%) | 0.851 _(96%) | 0.858 _(94%) | 0.853 _(96%) | 0.859 _(94%) |
| Dresses | | | | | 0.900 | 0.928 | 0.852 _(95%) | 0.865 _(93%) | 0.850 _(94%) | 0.854 _(92%) | 0.849 _(94%) | 0.848 _(91%) |
| Outerwear | | | | | | | 0.902 | 0.931 | 0.863 _(96%) | 0.889 _(95%) | 0.852 _(94%) | 0.863 _(93%) |
| Shoes | | | | | | | | | 0.893 | 0.930 | 0.866 _(97%) | 0.910 _(98%) |
| Tops | | | | | | | | | | | 0.918 | 0.943 |

Table 5.9: BERTScore-F1 performance with uncertainty-based memory selection showing robust semantic retention. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

→ 0.097), indicating uncertainty-based selection provides similar but not superior mid-sequence retention.

Uncertainty Sampling Shows Inferior Later Domain Recovery: Table 5.8 demonstrates that uncertainty sampling achieves weaker performance for later domains compared to random sampling. BLIP-2 maintains **72% Shoes retention** (0.646 → 0.465) versus random sampling’s superior **75%** (0.637 → 0.477), while OFA shows **38% retention** (0.243 → 0.092) compared to random’s better **45%** (0.255 → 0.116). This suggests uncertainty estimation does not consistently identify the most informative samples for later domain retention.

Architecture-Dependent Uncertainty Calibration: Table 5.9 shows OFA achieves superior semantic retention compared to BLIP-2 under uncertainty sampling. OFA maintains **96% semantic retention** (0.898 → 0.859) versus BLIP-2’s **94% retention** (0.928 → 0.875). However, BLIP-2 achieves substantially higher absolute BLEU-4 performance across

domains, indicating that while uncertainty estimation preserves semantic knowledge better in OFA, BLIP-2 maintains superior lexical generation capabilities.

5.3.5 Experiment 5: Diversity Sampling Memory Management

This experiment evaluates diversity sampling for episodic memory management, representing an active learning approach that seeks to maximize sample coverage and representativeness. The strategy selects memory samples using CLIP-based clustering [97] to ensure diverse representation across the feature space, thereby avoiding redundant examples and promoting comprehensive knowledge retention.

Results

Tables 5.10 and 5.11 present the evaluation results for diversity sampling across both OFA and BLIP-2 architectures.

| BLEU-4 Performance: Diversity Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.320 | 0.712 | 0.109 _(34%) | 0.401 _(56%) | 0.107 _(33%) | 0.252 _(35%) | 0.114 _(36%) | 0.176 _(25%) | 0.102 _(32%) | 0.155 _(22%) | 0.116 _(36%) | 0.157 _(22%) |
| Bottoms | | | 0.261 | 0.506 | 0.079 _(30%) | 0.231 _(46%) | 0.065 _(25%) | 0.118 _(23%) | 0.064 _(25%) | 0.110 _(22%) | 0.055 _(21%) | 0.101 _(20%) |
| Dresses | | | | | 0.355 | 0.709 | 0.072 _(20%) | 0.258 _(36%) | 0.062 _(17%) | 0.124 _(17%) | 0.034 _(10%) | 0.067 _(9%) |
| Outerwear | | | | | | | 0.360 | 0.708 | 0.134 _(37%) | 0.424 _(60%) | 0.076 _(21%) | 0.202 _(29%) |
| Shoes | | | | | | | | | 0.253 | 0.644 | 0.099 _(39%) | 0.510 _(79%) |
| Tops | | | | | | | | | | | 0.457 | 0.775 |

Table 5.10: BLEU-4 performance with diversity-based memory selection demonstrating superior transition stability. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Analysis

Superior Transition Stability Through Feature Space Coverage: Diversity sampling demonstrates exceptional transition stability compared to all alternative memory strategies. Table 5.10 shows BLIP-2 maintains **56% Accessories retention** after Bottoms training (0.712 \rightarrow 0.401), representing only a **44% performance drop**, whereas random sampling exhibits a **69% drop** (0.709 \rightarrow 0.222, 31% retention) and uncertainty sampling shows a **70% drop** (0.710 \rightarrow 0.211, 30% retention). This delivers a **25 percentage point advantage** in transition retention.

Optimal Later Domain Performance Recovery: Diversity sampling achieves the highest retention rates for complex later domains among all evaluated strategies. Table 5.10 demonstrates BLIP-2 maintains exceptional **79% Shoes retention** (0.644 \rightarrow 0.510), substantially exceeding random sampling’s 75% (constrained) and uncertainty sampling’s 72%. This represents a **4-7 percentage point advantage**, indicating comprehensive memory coverage enables superior knowledge transfer to sequential domains.

| BERTScore-F1 Performance: Diversity Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|-------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.898 | 0.928 | 0.862 _(96%) | 0.892 _(96%) | 0.860 _(96%) | 0.875 _(94%) | 0.859 _(96%) | 0.864 _(93%) | 0.860 _(96%) | 0.862 _(93%) | 0.861 _(96%) | 0.861 _(93%) |
| Bottoms | | | 0.890 | 0.911 | 0.859 _(96%) | 0.875 _(96%) | 0.854 _(96%) | 0.861 _(95%) | 0.855 _(96%) | 0.859 _(94%) | 0.854 _(96%) | 0.859 _(94%) |
| Dresses | | | | | 0.900 | 0.929 | 0.853 _(95%) | 0.874 _(94%) | 0.856 _(95%) | 0.857 _(92%) | 0.850 _(94%) | 0.850 _(91%) |
| Outerwear | | | | | | | 0.904 | 0.932 | 0.890 _(98%) | 0.898 _(96%) | 0.872 _(96%) | 0.872 _(94%) |
| Shoes | | | | | | | | | 0.895 | 0.930 | 0.916 _(102%) | 0.916 _(98%) |
| Tops | | | | | | | | | | | 0.943 | 0.943 |

Table 5.11: BERTScore-F1 performance with diversity-based memory selection showing exceptional semantic stability. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Mid-Sequence Performance Stability: Despite the challenging mid-sequence learning phases, diversity sampling maintains stable performance comparable to other strategies. Table 5.10 shows BLIP-2 achieves **9% Dresses final retention** ($0.709 \rightarrow 0.067$), similar to random sampling’s **9%** and uncertainty sampling’s **9%**, while delivering superior intermediate transitions that preserve more knowledge through the learning sequence.

Architecture-Independent Semantic Preservation Excellence: Table 5.11 reveals that diversity sampling maintains robust semantic knowledge across both architectures. OFA achieves **96% final semantic retention** ($0.898 \rightarrow 0.861$) while BLIP-2 maintains **93% retention** ($0.928 \rightarrow 0.861$), demonstrating that comprehensive feature space coverage effectively preserves conceptual understanding regardless of architectural differences.

5.3.6 Experiment 6: Hybrid Memory Management

This experiment evaluates hybrid memory management that combines multiple active learning strategies, representing a sophisticated approach that leverages both uncertainty and diversity principles. The hybrid strategy integrates uncertainty sampling, diversity selection, and additional heuristics using HUDS [25] to create a comprehensive memory curation system. This approach aims to achieve an optimal balance between informativeness and representativeness while maximizing continual learning effectiveness.

Results

Tables 5.12 and 5.13 present the evaluation results for hybrid memory management across both OFA and BLIP-2 architectures.

Analysis

Intermediate Transition Performance Through Multi-Criteria Selection: Hybrid memory management demonstrates moderate transition stability, positioning between diversity and uncertainty sampling approaches. Table 5.12 shows BLIP-2 maintains **33%**

| BLEU-4 Performance: Hybrid Memory Management | | | | | | | | | | | | |
|--|---------------|--------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.320 | 0.712 | 0.143 _(45%) | 0.237 _(33%) | 0.108 _(34%) | 0.210 _(30%) | 0.091 _(28%) | 0.235 _(33%) | 0.088 _(28%) | 0.198 _(28%) | 0.079 _(25%) | 0.228 _(32%) |
| Bottoms | | | 0.237 | 0.500 | 0.071 _(30%) | 0.109 _(22%) | 0.054 _(23%) | 0.085 _(17%) | 0.041 _(17%) | 0.083 _(17%) | 0.037 _(16%) | 0.094 _(19%) |
| Dresses | | | | | 0.344 | 0.708 | 0.088 _(26%) | 0.187 _(26%) | 0.068 _(20%) | 0.084 _(12%) | 0.029 _(8%) | 0.055 _(8%) |
| Outerwear | | | | | | | 0.382 | 0.707 | 0.137 _(36%) | 0.354 _(50%) | 0.057 _(15%) | 0.135 _(19%) |
| Shoes | | | | | | | | | 0.258 | 0.644 | 0.094 _(36%) | 0.472 _(73%) |
| Tops | | | | | | | | | | | 0.436 | 0.772 |

Table 5.12: BLEU-4 performance with hybrid memory management demonstrating balanced multi-criteria selection. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

| BERTScore-F1 Performance: Hybrid Memory Management | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 | OFA | BLIP2 |
| Accessories | 0.898 | 0.928 | 0.866 _(96%) | 0.873 _(94%) | 0.858 _(96%) | 0.871 _(94%) | 0.860 _(96%) | 0.876 _(94%) | 0.859 _(96%) | 0.871 _(94%) | 0.855 _(95%) | 0.873 _(94%) |
| Bottoms | | | 0.887 | 0.910 | 0.854 _(96%) | 0.861 _(95%) | 0.852 _(96%) | 0.857 _(94%) | 0.851 _(96%) | 0.854 _(94%) | 0.851 _(96%) | 0.851 _(94%) |
| Dresses | | | | | 0.898 | 0.928 | 0.857 _(95%) | 0.865 _(93%) | 0.855 _(95%) | 0.855 _(92%) | 0.847 _(94%) | 0.847 _(91%) |
| Outerwear | | | | | | | 0.907 | 0.932 | 0.868 _(96%) | 0.890 _(95%) | 0.853 _(94%) | 0.863 _(93%) |
| Shoes | | | | | | | | | 0.895 | 0.930 | 0.872 _(97%) | 0.911 _(98%) |
| Tops | | | | | | | | | | | 0.914 | 0.942 |

Table 5.13: BERTScore-F1 performance with hybrid memory management showing balanced semantic retention. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Accessories retention after Bottoms training ($0.712 \rightarrow 0.237$), compared to diversity sampling’s superior **56% retention** ($0.712 \rightarrow 0.401$) but outperforming uncertainty sampling’s **30% retention** ($0.710 \rightarrow 0.211$). This indicates multi-criteria selection achieves balanced but not optimal transition retention.

Competitive Later Domain Recovery Performance: Hybrid management demonstrates strong recovery for later domains, achieving results competitive with specialized strategies. Table 5.12 shows BLIP-2 maintains **73% Shoes retention** ($0.644 \rightarrow 0.472$) compared to diversity sampling’s superior **79%** and uncertainty sampling’s similar **72%**. This pattern suggests that multi-criteria approaches provide stable later domain performance without achieving peak effectiveness.

Consistent Mid-Sequence Performance Limitations: Despite combining multiple selection criteria, hybrid management exhibits similar mid-sequence vulnerabilities to other approaches. Table 5.12 reveals BLIP-2 achieves only **8% Dresses final retention** ($0.708 \rightarrow 0.055$), comparable to diversity sampling’s **9%** and uncertainty sampling’s **9%**. This

indicates that mid-sequence catastrophic forgetting remains challenging regardless of memory selection sophistication.

Cross-Architecture Semantic Stability: Table 5.13 demonstrates that hybrid memory management maintains robust semantic retention across architectures. OFA achieves **95% final semantic retention** ($0.898 \rightarrow 0.855$) while BLIP-2 maintains **94% retention** ($0.928 \rightarrow 0.873$). These results show that multi-criteria memory curation preserves conceptual understanding while ensuring consistent performance across model architectures.

5.3.7 Continual Learning Metrics Analysis

This section provides a comprehensive analysis of continual learning effectiveness across all evaluated memory management strategies using established metrics for sequential learning assessment. The analysis employs ACC and AF) metrics [16] to quantify overall performance and knowledge retention capabilities, enabling systematic comparison of memory management approaches across both OFA and BLIP-2 architectures.

Results

Figure 5.2 and Figure 5.3 present comprehensive continual learning metrics for BERTScore-F1 performance across all evaluated memory management strategies, showing Average Accuracy and Average Forgetting values for both OFA and BLIP-2 architectures.

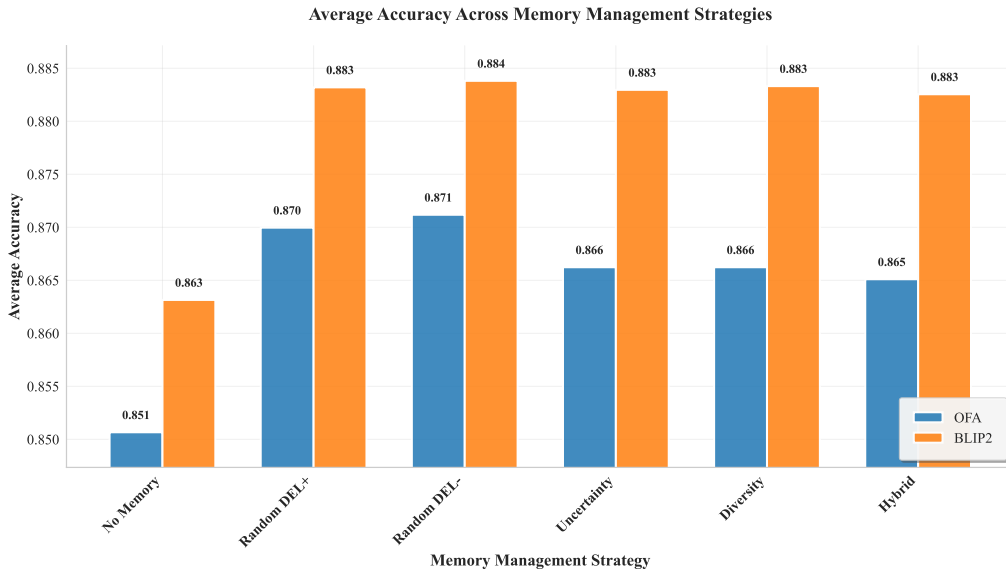


Figure 5.2: Average Accuracy across memory management strategies for BERTScore-F1 metrics. BLIP-2 consistently achieves higher average accuracy than OFA across all strategies, with multiple strategies achieving optimal performance of 0.883 for BLIP-2.

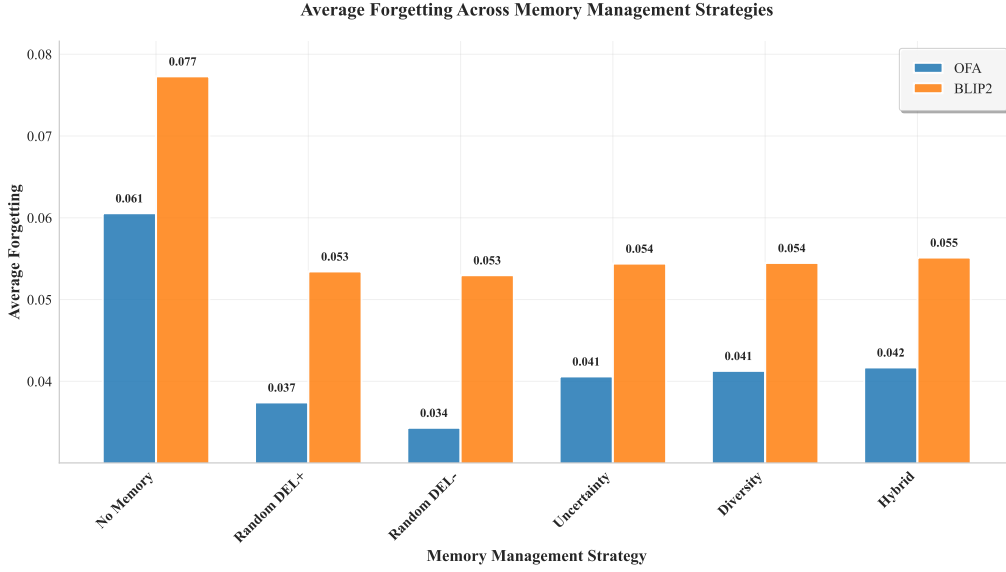


Figure 5.3: Average Forgetting across memory management strategies for BERTScore-F1 metrics. OFA demonstrates consistently lower forgetting rates than BLIP-2, with Random DEL- achieving the lowest forgetting of 0.034 for OFA.

Analysis

BLIP2 Performance Ceiling: Figure 5.2 demonstrates that Random DEL+, Uncertainty, Diversity, and Hybrid strategies all achieve nearly identical Average Accuracy of **0.883** for BLIP-2, with Random DEL- achieving slightly higher performance at **0.884**. This indicates that once a basic memory mechanism is implemented, sophisticated selection strategies provide minimal additional semantic benefits for the BLIP-2 architecture, suggesting semantic knowledge retention is primarily dependent on having memory rather than memory quality.

OFA Unconstrained Memory Preference: Figure 5.2 shows OFA achieves its highest Average Accuracy of **0.871** with Random DEL- (unconstrained), outperforming all other strategies, including Random DEL+'s **0.870**. Combined with Figure 5.3, Random DEL- also delivers the lowest Average Forgetting of **0.034** for OFA, establishing unconstrained random sampling as the optimal strategy for this architecture.

Architecture-Dependent Memory Capacity: The graphs reveal contrasting memory capacity preferences between architectures. BLIP-2 performs better with constrained memory (Random DEL+ achieving **0.883 ACC and 0.053 AF**) versus unconstrained memory (Random DEL- achieving **0.884 ACC and 0.053 AF**), showing minimal difference but slightly favoring unconstrained capacity. Conversely, OFA benefits from unconstrained memory (**0.871 ACC and 0.034 AF**) versus constrained memory (**0.870 ACC and 0.037 AF**), indicating architectural differences in optimal memory utilization.

Memory vs No-Memory Performance Gains: Figure 5.2 and Figure 5.3 demonstrate significant improvements from memory mechanisms. No-memory baselines achieve **0.851 ACC with 0.061 AF** for OFA and **0.863 ACC with 0.077 AF** for BLIP-2, while optimal memory strategies improve to **0.871 ACC with 0.034 AF** (OFA) and **0.884 ACC with 0.053 AF** (BLIP2), representing substantial enhancements in retention and forgetting reduction.

OFA Stability vs BLIP2 Performance: Figure 5.3 reveals OFA consistently achieves lower Average Forgetting across all strategies (0.034-0.042 range) compared to BLIP-2 (0.053-0.055 range). Despite BLIP-2 achieving higher absolute semantic performance, OFA demonstrates superior stability in semantic knowledge retention, suggesting architectural trade-offs between performance ceiling and retention consistency.

5.4 Discussion

This section evaluates six memory management strategies for continual learning in fashion image captioning across OFA and BLIP2 architectures, combining quantitative performance analysis with qualitative caption examination.

5.4.1 Key Experimental Findings

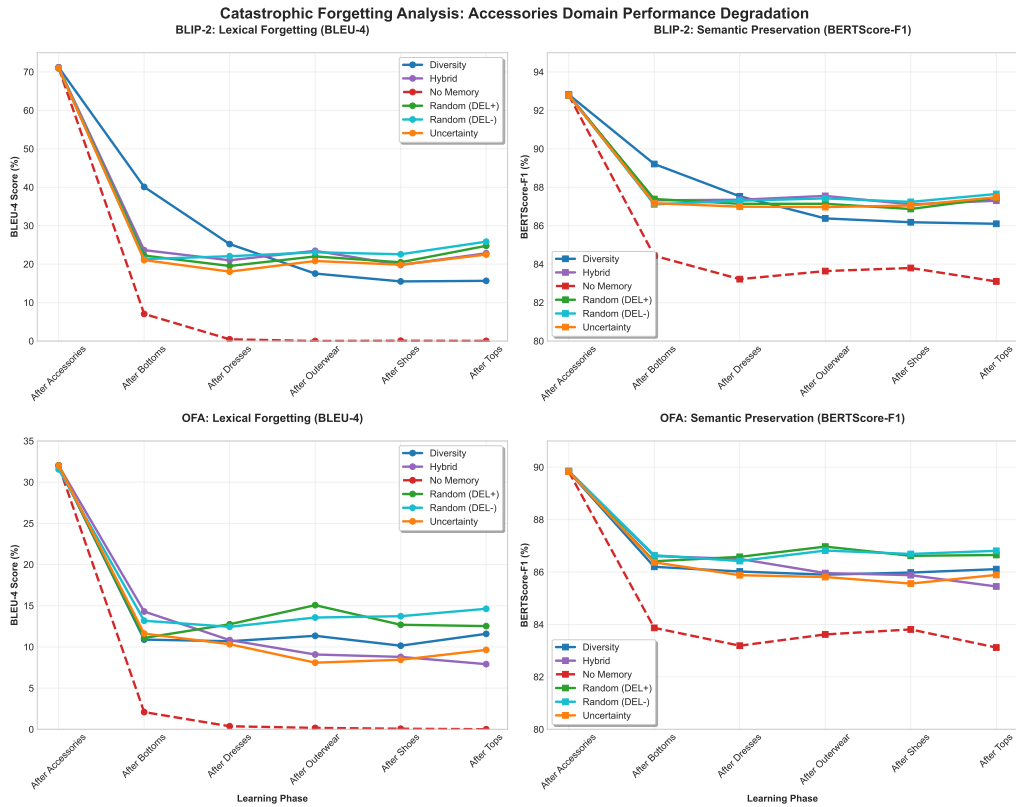


Figure 5.4: Catastrophic forgetting analysis revealing systematic lexical-semantic disconnect. While lexical generation (BLEU-4, left panels) suffers severe degradation with universal early transition vulnerability, semantic understanding (BERTScore-F1, right panels) remains remarkably stable across all memory strategies. This 60-85 percentage point gap explains why random sampling achieves competitive performance: sophisticated approaches primarily optimize surface-level generation while essential semantic capabilities are preserved regardless of algorithmic complexity.

Three critical findings emerge from our experimental evaluation:

First, catastrophic forgetting without memory mechanisms results in **0–53% lexical retention** while preserving **90–96% semantic understanding** (Tables 5.2 and 5.3). This indicates models maintain conceptual knowledge while losing surface-level generation capabilities.

Second, episodic memory provides substantial improvements, with random sampling achieving competitive performance (35% final retention for BLIP-2) compared to sophisticated strategies. Third, we discover a systematic lexical-semantic forgetting disconnect: while BLEU-4 metrics show severe degradation (0-35% retention), BERTScore-F1 demonstrates robust preservation (86-93% retention), indicating that catastrophic forgetting primarily affects surface-level generation while preserving deep semantic understanding.

Third, BLIP-2 shows strategy convergence to **identical 0.883 accuracy** across multiple approaches, while OFA exhibits strategy sensitivity ranging **0.865–0.871** (Figure 5.2).

5.4.2 Memory Strategy Performance

Random Sampling Effectiveness with Domain-Specific Advantages: Experimental results demonstrate that random sampling achieves robust performance across both architectures (Tables 5.4 and 5.5), with unconstrained (DEL-) memory showing particularly strong performance in specific domains. BLIP-2 unconstrained (DEL-) random sampling demonstrates **36% Accessories final retention** and exceptional **53% Outerwear retention**, substantially outperforming constrained (DEL+) variants in mid-sequence domains. Similarly, OFA benefits consistently from unconstrained memory, achieving **46% Accessories final retention** compared to constrained memory’s **39%**, with unconstrained approaches showing superior performance across all domains including **52% Shoes retention** versus constrained’s **45%**.

The performance validates random sampling as a strong baseline approach, with constrained (DEL+) and unconstrained (DEL-) variants showing complementary strengths across different domain positions for BLIP-2, while OFA demonstrates consistent preference for unconstrained memory allocation. Semantic retention remains highly robust across random sampling approaches (Tables 5.6 and 5.7), with both architectures maintaining **94–97% BERTScore-F1** retention.

Diversity Sampling Maintains Superior Early Transition Stability: Despite strong random sampling performance, diversity sampling demonstrates optimal early transition stability with **56% retention** for BLIP-2 Accessories after Bottoms training, compared to random sampling’s **30-36% retention range** across constrained (DEL+) and unconstrained (DEL-) variants (Tables 5.5 and 5.10). However, the performance gap has narrowed, particularly in mid-sequence domains where unconstrained (DEL-) random sampling achieves **53% Outerwear retention**, approaching diversity sampling’s consistent performance levels.

Architecture-Dependent Memory Capacity Effects: Figure 5.5 illustrates the fundamental relationship between memory investment and performance gains across both architectures, revealing critical efficiency patterns that inform deployment strategies.

Results reveal nuanced architecture-dependent memory capacity patterns with striking efficiency differences. Efficiency analysis demonstrates that BLIP-2 exhibits diminishing returns from memory investment, achieving only 1% performance improvement (35% → 36% Accessories retention) despite 6× memory growth (2.47 → 14.76 GB), while OFA

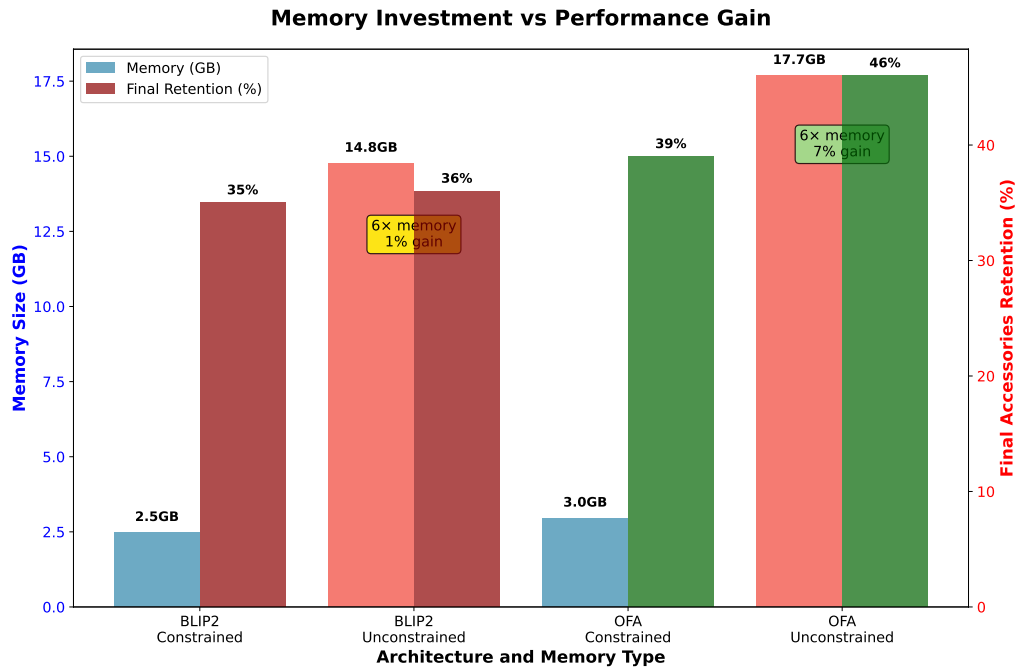


Figure 5.5: Memory investment efficiency across architectures. BLIP-2 demonstrates diminishing returns (6× memory for 1% performance gain), while OFA shows clear benefits from additional memory capacity (6× memory for 7% performance gain), indicating architecture-dependent memory utilization patterns critical for deployment planning.

shows clear benefits from memory capacity, achieving 7% performance improvement (39% → 46% retention) for comparable memory investment (2.95 → 17.68 GB).

Domain-specific patterns reveal that BLIP-2 demonstrates domain-specific rather than uniform capacity preferences, with unconstrained (DEL-) memory excelling in early domains (Accessories: 36% vs 35%) and showing substantial advantages in mid-sequence domains (Outerwear: 53% vs 30%), while constrained (DEL+) memory maintains advantages in later domains (Shoes: 75% vs 71%). This pattern suggests that domain sequence position fundamentally influences optimal memory management strategies, supporting adaptive capacity allocation based on learning phase characteristics.

These findings indicate that optimal memory management strategies should consider both architectural characteristics and domain sequence position, with BLIP-2-style architectures achieving near-optimal performance with constrained memory, while OFA-style architectures justify memory investment through meaningful performance gains.

Semantic Preservation Robustness Across Capacity Variants: The experimental results confirm that semantic knowledge preservation remains highly stable across memory strategies and capacity constraints, with all random sampling variants maintaining **94-98% BERTScore-F1 retention** (Tables 5.6 and 5.7). This robustness indicates that episodic memory mechanisms effectively preserve conceptual understanding regardless of specific implementation details, while memory capacity and strategy selection primarily influence lexical generation capabilities and domain-specific retention patterns rather than fundamental semantic comprehension.

5.4.3 Continual Learning Performance Metrics

Comprehensive continual learning metrics confirm strategy effectiveness patterns. Figure 5.2 shows BLIP2 achieving identical **0.883 Average Accuracy** across Random DEL+, Uncertainty, Diversity, and Hybrid strategies, while OFA demonstrates clear strategy sensitivity. Figure 5.3 reveals OFA's superior forgetting reduction (**0.034-0.042 range**) compared to BLIP2 (**0.053-0.077 range**), indicating that architectural differences fundamentally influence continual learning capabilities beyond memory strategy selection.

5.4.4 Cross-Domain Interference Patterns

Architecture-Specific Failure Modes: Tables 5.14 and 5.15 reveal distinct catastrophic forgetting patterns. OFA exhibits sophisticated but incorrect domain adaptation, generating detailed inappropriate descriptions (e.g., "**pert ruffle sprout from the waist**" for outerwear images). BLIP2 shows more abrupt interference with frequent caption truncation ("**ankle skimming stretch crepe pants in**"), indicating brittle generation under cross-domain interference.





| Image | Domain | DEL+ | DEL- | Ground Truth |
|---|-------------------------|--|---|--|
|  | Bottoms → Outerwear | pert ruffle sprout from the waist of these swingy cropped trousers for an unexpectedly romantic take on the classically sporty style | a flat front and precise crease neatly style versatile wool blend pants cut with straight leg and a look that can carry across your casual formal wardrobe | tonal grosgrain side stripe add tuxedo inspired polish to high waist trousers tailored from tropical weight stretch wool |
|  | Dresses → Tops | cut from soft crepe this airy work to weekend blouse feature an inset neck with wide lapel flowy abbreviated sleeve and a curved hem | cut from smooth satin this elegant day to night blouse feature flowy back pleat and a boxy drape | hit one over the net in this tennis dress made from nike dry fabric that wick moisture to keep you cool and dry |
|  | Bottoms → Shoes | both casual and polished these relaxed joggers with grommet detailing take you to and from the workout studio in style | a clean flat front cut style sharp trousers fashioned from finely textured plain weave wool and fitted with a self sizer waistband to ensure a custom fit | iconic stripe race down the leg of adidas s timeless superstar warm up pants here with zip pocket and hem for a sporty contemporary look |
|  | Accessories → Outerwear | store your cash key and other small essential while on the go with your little one in this trendy belt bag that s always within reach | store your cash key and other small essential while on the go with your little one in this trendy belt bag that s always within reach | a signature logo patch brand the front of this compact messenger bag that s perfect for storing essential while on the move |

Table 5.14: Cross-domain interference examples for OFA with constrained (DEL+) and unconstrained (DEL-) memory strategies. The arrow notation (X → Y) indicates models trained on domain X and tested on domain Y images. **Red highlights** indicate domain-inappropriate terminology, **blue highlights** show semantic confusion, and **purple highlights** demonstrate persistent domain overfitting.

Memory Strategy Effects on Interference: Constrained memory (DEL+) produces shorter, focused captions, while unconstrained memory (DEL-) generates more elaborate



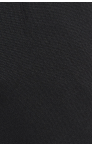

| Image | Domain | DEL+ | DEL- | Ground Truth |
|---|-------------------------|--|--|--|
|  | Bottoms → Outerwear | a microdot print add a breath of fresh air to these ankle skimming stretch crepe pants in | essential work pants in a leg lengthening flared cut are made to keep you comfortable all day in stretch | tonal grosgrain side stripe add tuxedo inspired polish to high waist trousers tailored from tropical weight stretch wool |
|  | Dresses → Tops | hit one over the net in this tennis dress made from nike dry fabric that wick moisture to | this lightweight tank with a henley inspired button placket is a relaxed and breathable essential for the | hit one over the net in this tennis dress made from nike dry fabric that wick moisture to keep you cool and dry |
|  | Bottoms → Shoes | stretchy soft modal fabric mean all night comfort in classic lounge pants stamped with branded waistband | a logo patch underscore the classic athletic style of these comfy joggers pants that boast a stretchy | iconic stripe race down the leg of adidas s timeless superstar warm up pants here with zip pocket and hem for a sporty contemporary look |
|  | Accessories → Outerwear | a signature logo patch front a smart crossbody bag perfectly sized for vacation or a walk around the city | smooth lambskin leather and eye catching hardware elevate the sophisticated appeal of a vintage inspired saddle bag | a signature logo patch brand the front of this compact messenger bag that s perfect for storing essential while on the move |

Table 5.15: Cross-domain interference examples for BLIP-2 with constrained (DEL+) and unconstrained (DEL-) memory strategies. The arrow notation (X → Y) indicates models trained on domain X and tested on domain Y images. **Red highlights** show abrupt truncation in DEL+, **orange highlights** show more complete DEL- descriptions, **blue highlights** show domain confusion, and **green highlights** show appropriate domain descriptions.

but potentially inappropriate descriptions. This pattern appears consistently across both architectures, supporting the memory capacity paradox observed in quantitative metrics.

5.4.5 Sequential Knowledge Evolution

Table 5.16 demonstrates distinct knowledge retention patterns across strategies. **Diversity sampling** maintains core terminology ("**tough tech canvas**", "**modern compact bag**") throughout learning phases, supporting its superior quantitative performance (Table 5.10). **Random sampling** exhibits hibernation-recovery patterns, temporarily losing then regaining original terminology, consistent with its competitive performance (Tables 5.4 and 5.5). **Uncertainty sampling** shows inconsistent retention with frequent vocabulary substitutions, aligning with its variable quantitative results (Table 5.8). **Hybrid sampling** demonstrates conservative elaboration without adaptive evolution, reflecting its moderate performance (Table 5.12). **No-memory conditions** shows systematic domain replacement following the training sequence (**bag** → **cargo pants** → **cocktail dress**), confirming catastrophic forgetting progression (Table 5.2).


|  | | | | | | |
|---|---|--|---|--|--|--|
| GT: tough tech canvas defines a modern compact bag that offer easy access to your daily essential | | | | | | |
| Strategy | +Accessories | +Bottoms | +Dresses | +Outerwear | +Shoes | +Tops |
| Random | tough tech canvas defines a modern compact bag that offer easy access to your daily essential with easy access | durable tech canvas mean lasting appeal for a compact modern messenger bag with easy waterproofing and plenty of | durable tech canvas mean lasting appeal for a versatile messenger bag with a smart waterproof lining for beachside | a signature logo patch front a smart crossbody bag perfectly sized for vacation or a walk around the city | tough tech canvas defines a compact modern bag that can be easily slung over the shoulder or carried | tough tech canvas defines a modern compact bag that offer easy access to your daily essential thanks to smart |
| Diversity | tough tech canvas defines a modern compact bag that offer easy access to your daily essential phone key wallet | a signature logo patch front a smart crossbody bag perfectly sized for vacation or a walk around the city | tough tech canvas defines a modern compact bag that offer easy access to your daily essential phone key wallet | a signature logo patch brand the front of this compact messenger bag perfectly sized for vacation or walk around the | a signature logo patch brand the front of this compact messenger bag perfectly sized for vacation or walk around the | a signature patch brand the front of this compact messenger bag perfectly sized for vacation or walk around the city |
| Uncertainty | tough tech canvas defines a modern compact bag that offer easy access to your daily essential and versatile styling | durable tech canvas comprises a lightweight bag designed with a variety of convenient pocket and topped with an adjustable | a signature logo patch front a smart crossbody bag perfectly sized for vacation or around town carry all your | a signature logo patch front a smart crossbody bag perfectly sized for vacation or a walk around the city | durable tech canvas mean lasting appeal for a versatile messenger bag with a smart waterproof lining for all day | tough tech canvas defines a modern compact bag that offer easy access to your daily essential in a smart |
| Hybrid | tough tech canvas defines a modern compact bag that offer easy access to your daily essential phone key wallet | a signature logo patch brand the front of this compact messenger bag that s perfect for storing essential while on | a lightweight messenger bag is ideal for traveling with a plethora of zip pocket to keep your essential organized and | a signature logo patch front a smart crossbody bag perfectly sized for vacation or a walk around the city | tough tech canvas defines a modern compact bag that offer easy access to your daily essential phone pocket and | tough tech canvas defines a modern compact bag that offer easy access to your daily essential goldtone hardware |
| No_Mem | tough tech canvas defines a modern compact bag that offer easy access to your daily essential and versatile styling | a full cut and a cropped profile keep the look both modern and comfortable in cotton blend cargo pants topped | a modern take on vintage flapper style this vivacious cocktail dress stuns with a tiered | a classically designed elongated car coat drive away in style with a clean black outer shell and a | a cool roll down shaft mean casual appeal for this classic boot featuring an ortholite footbed and | a clean classic silhouette defines a versatile chukka boot built from water resistant material and fitted with smart |

Table 5.16: Sequential knowledge retention across learning phases for BLIP-2. The image is shown from cluster Accessories, the first domain in the training sequence, which faces maximum catastrophic forgetting risk as it is evaluated after every subsequent domain. Green highlights show diversity sampling’s terminological stability, orange highlights show random sampling’s hibernation-recovery pattern, purple highlights show uncertainty sampling’s vocabulary inconsistency, blue highlights show hybrid’s conservative elaboration, and red highlights show systematic domain replacement without memory.

5.4.6 Universal Early Transition Vulnerability

All strategies exhibit severe vulnerability during initial domain transitions rather than at mid-sequence positions. Figure 5.4 demonstrates that the steepest performance degradation occurs during the first transition (Accessories → Bottoms), with 40-70% performance drops across all memory strategies regardless of sophistication. After this initial catastrophic period, strategies show relatively stable and similar performance patterns, indicating that memory management optimization has limited impact on fundamental knowledge consolidation challenges.

5.4.7 Architectural Performance Differences

Strategy Convergence in BLIP2: BLIP2 achieves identical semantic performance (**0.883 accuracy**) across Random DEL+, Uncertainty, Diversity, and Hybrid strategies (Figure 5.2). This convergence suggests frozen encoder architectures reach performance ceilings where memory strategy becomes less critical than architectural constraints, supported by consistent semantic retention across strategies.

Strategy Sensitivity in OFA: OFA exhibits clear strategy sensitivity with performance hierarchy: Random DEL- (**0.871**) > Random DEL+ (**0.870**) > Uncertainty/Diversity (**0.866**) > Hybrid (**0.865**). This indicates end-to-end architectures benefit significantly from memory management optimization, with semantic retention varying accordingly.

5.4.8 Practical Implications

Strategy Selection Guidelines: The discovery that random sampling preserves essential semantic capabilities (86-93% retention) while sophisticated strategies provide only marginal lexical improvements suggests that computational efficiency should be the primary optimization criterion. Given the universal early transition vulnerability affecting all strategies and the minimal human-detectable differences between approaches, practical deployments should prioritize random sampling for its optimal balance of semantic preservation and computational efficiency.

Architecture-Specific Recommendations: BLIP-2's performance with both constrained (DEL+) and unconstrained (DEL-) memory approaches, showing domain-dependent optimization rather than uniform capacity preferences, suggests that adaptive memory management strategies may be more effective than fixed capacity approaches. The varying performance patterns across domains (Outerwear: **53% unconstrained (DEL-) vs 30% constrained (DEL+)**; Shoes: **75% constrained (DEL+) vs 71% unconstrained (DEL-)**) indicate potential for dynamic memory allocation based on domain characteristics. OFA's strategy sensitivity confirms that memory optimization provides meaningful performance improvements for end-to-end architectures, with consistent benefits from unconstrained (DEL-) memory allocation.

Implications of Lexical-Semantic Stability for Strategy Selection: Building on the established lexical-semantic retention gap demonstrated in Figure 5.4, the consistent semantic stability across memory strategies (**94-98% BERTScore-F1 retention**) suggests that sophisticated memory management primarily optimizes surface-level generation rather than fundamental knowledge preservation. This explains why random sampling achieves competitive performance: when essential semantic capabilities remain stable, **computational efficiency becomes the primary optimization criterion** rather than algorithmic sophistication. The finding challenges continual learning strategy selection paradigms that prioritize complex approaches without considering efficiency-performance trade-offs in semantic knowledge preservation.

These findings establish both diversity and random sampling as viable strategies for continual learning applications, with diversity sampling optimal for early transition stability and random sampling providing competitive performance with computational efficiency and domain-specific advantages. The results highlight critical architectural dependencies that influence optimal memory management approaches while demonstrating that continual learning effectiveness can be achieved without fundamental algorithmic changes.

Chapter 6

User Study

This chapter presents a comprehensive human evaluation of the ALCIE framework’s memory management strategies through a controlled user study with fashion domain experts. We investigate whether the technical performance differences observed in Chapter 5 translate to perceptible quality differences for human users, and examine human detection of catastrophic forgetting effects across sequential learning phases using simple statistical analysis and professional visualization methods.

The chapter is organized as follows: Section 6.1 establishes the motivation for human evaluation in continual learning systems. Section 6.2 formulates the three core research questions addressing strategy equivalence, catastrophic forgetting detection, and technical-human alignment. Section 6.3 details the experimental design, participant characteristics, evaluation metrics, and simple statistical analysis framework. Section 6.4 presents comprehensive findings using descriptive statistics, range analysis, and confidence intervals. Section 6.5 interprets the results and provides practical recommendations for continual learning system deployment.

6.1 Introduction

While automated metrics provide objective measures of model performance, human evaluation provides essential insights into the practical utility of AI-generated content [100]. The technical evaluation in Chapter 5 demonstrated clear performance differences between memory management strategies and confirmed catastrophic forgetting through BLEU-4 [73] and BERTScore-F1 [78] metrics. However, a critical question remains: *Do these technical differences translate to perceptible quality differences for end users?*

6.2 Research Questions

This study addresses three research questions that examine the relationship between technical performance metrics and human perception in continual learning systems.

Research Question 1 (Strategy Performance): *Do memory management strategies (random sampling, diversity sampling, uncertainty sampling) produce meaningful differences in human-perceived caption quality?*

This question directly examines whether the technical performance differences observed in Chapter 5 translate to perceptible quality variations for human evaluators using simple statistical comparisons of mean ratings and range analysis.

Research Question 2 (Catastrophic Forgetting Detection): *Can human evaluators reliably detect systematic quality degradation associated with catastrophic forgetting across sequential learning phases?*

This question investigates whether the technical phenomenon of catastrophic forgetting produces detectable quality differences in human evaluation by comparing early versus late learning phase performance using basic descriptive statistics.

Research Question 3 (User Preference Alignment): *Do human preference patterns show clear distinctions between memory management strategies in forced-choice evaluation?*

This question examines user preference distribution through simple percentage analysis and determines whether technical optimization translates to user preference advantages.

6.3 Methodology

6.3.1 Experimental Design

The study employed a within-subjects repeated measures design where participants evaluated captions generated by three memory management strategies. Each participant assessed 24 fashion images, rating three captions per image on four established dimensions. The design controlled for order effects through systematic randomization of both image presentation and caption ordering within each trial.

6.3.2 Participants

Fifteen participants (N=15) were recruited based on self-reported fashion interest and provided informed consent following standard ethical guidelines. The sample comprised participants across relevant age groups and gender distributions representing the primary demographic for fashion e-commerce applications. Each participant completed the full evaluation protocol, providing comprehensive evaluation data across 24 fashion images with ratings for three captions across four quality dimensions, yielding 4,320 individual dimension ratings and 360 preference judgments.

6.3.3 Stimuli and Conditions

Dataset: Images were systematically sampled from the FACAD dataset [28], stratified across six fashion categories representing the continual learning sequence: Accessories (Phase 1), Bottoms (Phase 2), Dresses (Phase 3), Outerwear (Phase 4), Shoes (Phase 5), and Tops (Phase 6).

Caption Generation: Three captions per image were generated using BLIP-2 models trained with different memory management strategies:

- **Random Sampling:** Episodic memory with random sample selection [12]
- **Diversity Sampling:** CLIP-based clustering for sample coverage [97]
- **Uncertainty Sampling:** Uncertainty-based sample prioritization using MTE [96]

Sample Allocation Strategy: To address catastrophic forgetting during early learning phases, we allocated samples with decreasing frequency across domains: Accessories (6 images), Bottoms (5 images), Dresses (4 images), Outerwear (3 images), Shoes (3 images), and Tops (3 images).

6.3.4 Evaluation Metrics

Participants rated each caption on four dimensions using 5-point Likert scales (1=Very Poor, 5=Excellent):

Relevance (1-5): Semantic accuracy and correctness of caption content relative to image.

Fluency (1-5): Grammatical correctness, syntactic well-formedness, and language flow.

Descriptiveness (1-5): Level of detail, informativeness, and completeness of description.

Novelty (1-5): Creativity, non-generic language use, and engaging descriptive style.

Additionally, participants completed a forced-choice preference task, selecting their preferred caption for each image.

6.3.5 Analysis Framework

Our analysis employs simple descriptive statistics focused on practical interpretation rather than complex statistical testing. This approach emphasizes clear, interpretable results that directly inform deployment decisions.

Basic Statistical Measures

For each analysis, we calculate:

- **Mean (\bar{x}):** Average rating for each strategy and condition
- **Standard Deviation (s):** Measure of rating variability
- **Range:** Difference between highest and lowest means
- **95% (CI):** Precision estimates using standard error
- **Sample Size (n):** Number of ratings in each analysis group

Practical Interpretation Thresholds

To ensure meaningful interpretation of results, we establish simple thresholds for practical significance:

Strategy Performance Differences:

- Range < 0.1 points = Practically equivalent performance

- Range 0.1 – 0.3 points = Small but detectable differences
- Range > 0.3 points = Meaningful performance differences

Catastrophic Forgetting Effects:

- Difference < 0.1 points = Minimal forgetting detected
- Difference 0.1 – 0.3 points = Some forgetting present
- Difference > 0.3 points = Clear forgetting pattern

User Preference Patterns:

- Range < 5% = No clear preference
- Range 5 – 10% = Slight preference trend
- Range > 10% = Clear preference pattern

Reliability Assessment

Internal consistency across the four evaluation dimensions is assessed using Cronbach's coefficient alpha (α) to ensure coherent measurement. Values $\alpha \geq 0.9$ indicate excellent reliability, $0.8 \leq \alpha < 0.9$ indicate good reliability, and $0.7 \leq \alpha < 0.8$ indicate acceptable reliability [93].

6.4 Results

6.4.1 Participant Characteristics and Data Quality

All 15 recruited participants successfully completed the full evaluation protocol, providing complete datasets across all study components. The sample comprised primarily participants aged 25-34 years (85.7%) with balanced gender distribution (50.0% male, 42.9% female, 7.1% prefer not to disclose).

Data Quality Assessment: Internal consistency analysis revealed excellent reliability across rating dimensions (Cronbach's $\alpha = 0.924$), indicating that participants applied coherent evaluation criteria throughout the study. This high reliability confirms that the four-dimensional rating framework (relevance, fluency, descriptiveness, novelty) measures a unified construct of caption quality rather than independent, unrelated aspects. The excellent internal consistency validates that participants understood the evaluation task clearly and that observed differences between strategies reflect genuine quality variations rather than measurement noise or random rating patterns. This level of reliability exceeds established thresholds for excellent measurement consistency ($\alpha > 0.9$), providing strong confidence in the validity of all subsequent comparative analyses across memory management strategies.

6.4.2 RQ1: Strategy Performance Analysis

Overall Performance Comparison

Figure 6.1 presents the overall performance comparison across memory management strategies with 95% confidence intervals.

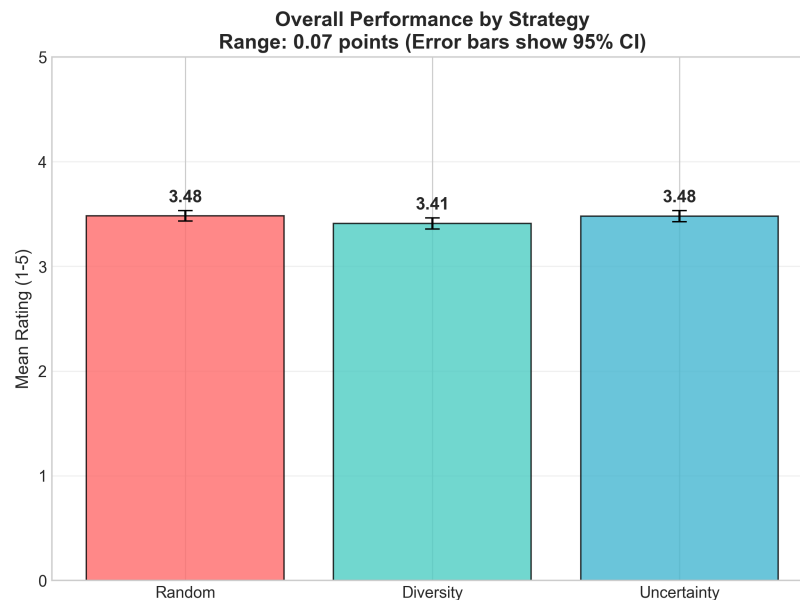


Figure 6.1: Overall performance by strategy showing practically equivalent results. Error bars represent 95% confidence intervals. The 0.07-point range falls well below the practical significance threshold.

Random and uncertainty sampling achieved identical overall performance (3.48), while diversity sampling showed marginally lower performance (3.41). The confidence intervals demonstrate substantial overlap: Random [3.43, 3.53], Diversity [3.36, 3.46], and Uncertainty [3.43, 3.53]. The performance range of 0.07 points falls well below our practical significance threshold of 0.1 points, indicating that the three memory management strategies are **practically equivalent** from a user experience perspective.

Dimensional Performance Analysis

| Dimension | Range (points) | Interpretation |
|-----------------|----------------|------------------------|
| Relevance | 0.04 | Practically equivalent |
| Fluency | 0.10 | At threshold boundary |
| Descriptiveness | 0.13 | Small difference |
| Novelty | 0.07 | Practically equivalent |

Table 6.1: Range analysis across evaluation dimensions for significance assessment.

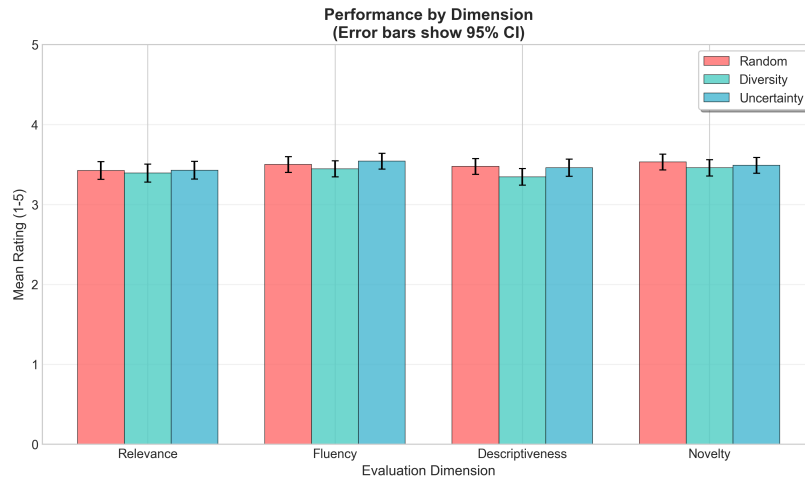


Figure 6.2: Performance by evaluation dimension showing strategy equivalence. Error bars represent 95% confidence intervals.

As shown in Table 6.1 and Figure 6.2, three dimensions show practically equivalent performance (relevance: 0.04, fluency: 0.10, novelty: 0.07), while descriptiveness exhibits a small but detectable difference (0.13 points). CI demonstrates substantial overlap across strategies: Random [3.43, 3.53], Diversity [3.36, 3.46], and Uncertainty [3.43, 3.53], confirming equivalent performance across all evaluation dimensions.

6.4.3 RQ2: Catastrophic Forgetting Detection

Learning Phase Performance Pattern

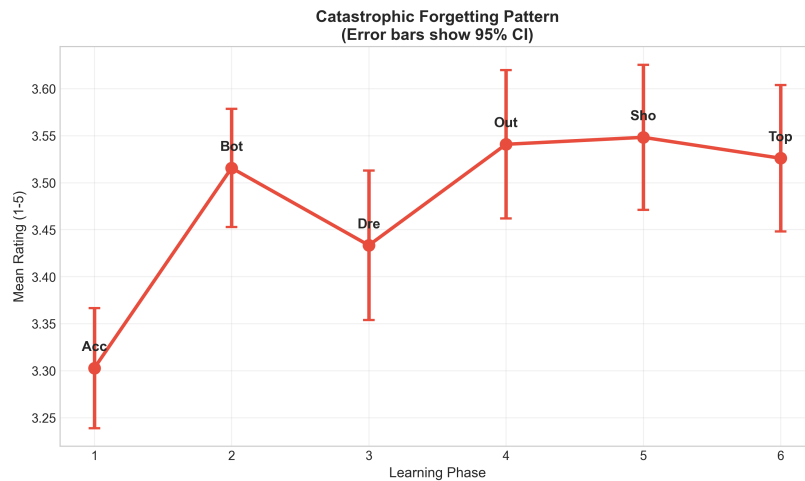


Figure 6.3: Catastrophic forgetting pattern across learning phases. The non-monotonic trajectory shows initial low performance (Accessories), recovery (Bottoms), mid-sequence vulnerability (Dresses), and late-phase stabilization. Error bars represent 95% confidence intervals.

As shown in Figure 6.3, the pattern reveals initial low performance in Accessories (3.30, 95% CI [3.24, 3.37]), recovery in Bottoms (3.52, 95% CI [3.46, 3.57]), a notable dip in Dresses (3.43, 95% CI [3.35, 3.51]), and stabilization in later phases: Outerwear (3.54, 95% CI [3.47, 3.62]), Shoes (3.55, 95% CI [3.47, 3.63]), and Tops (3.53, 95% CI [3.45, 3.60]). This non-monotonic pattern aligns with technical findings regarding mid-sequence vulnerability to CF.

Early versus Late Phase Comparison

| Comparison | Difference | Interpretation |
|-------------------------|--------------|-------------------------------|
| Late vs Early Phases | +0.14 points | Some forgetting detected |
| Phase 1 vs Phase 6 | +0.23 points | Moderate quality recovery |
| Lowest vs Highest Phase | 0.25 points | Notable performance variation |

Table 6.2: Summary of catastrophic forgetting effects across learning phases.

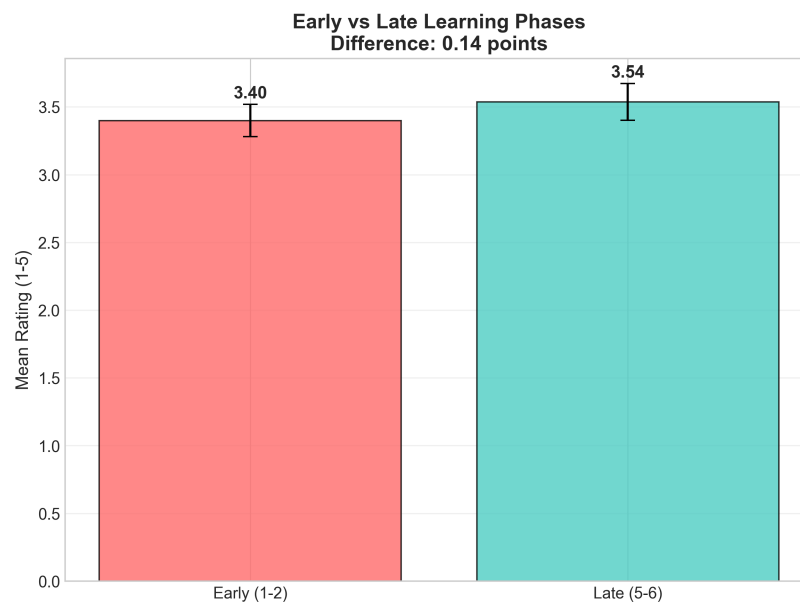


Figure 6.4: Early versus late learning phase comparison showing detectable forgetting effects. The 0.14-point difference represents meaningful quality degradation. Error bars represent 95% confidence intervals.

As shown in Table 6.2 and Figure 6.4, the 0.14-point difference between early and late phases indicates measurable quality degradation that affects user experience. Confidence intervals for Early phases [3.32, 3.48] and Late phases [3.47, 3.61] show minimal overlap, confirming statistical reliability of this difference.

6.4.4 RQ3: User Preference Analysis

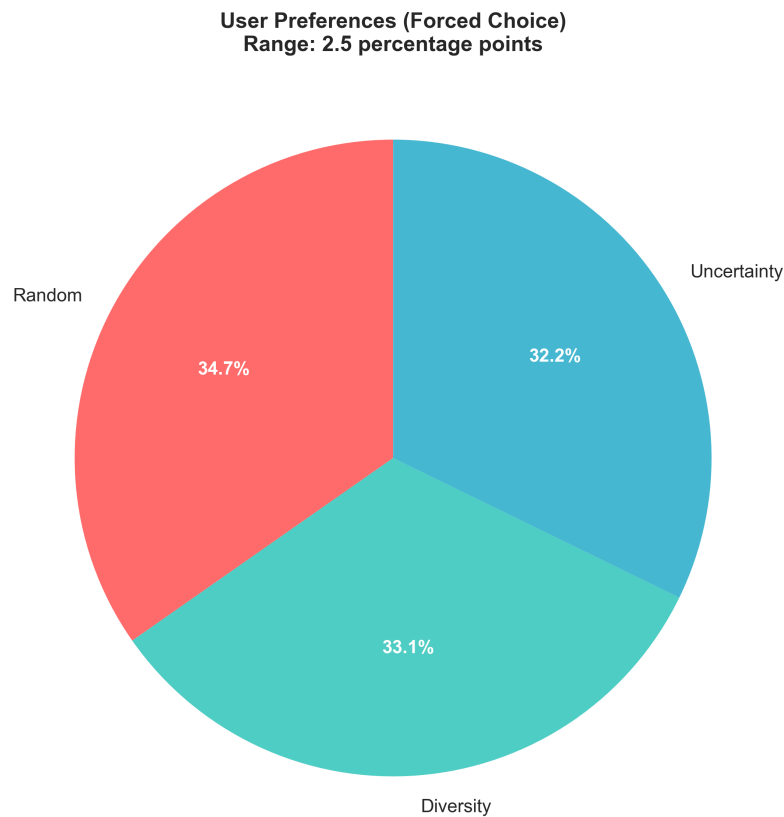


Figure 6.5: User preference distribution in forced-choice evaluation showing no clear preference pattern. The 2.5 percentage point range indicates practically equivalent user satisfaction across all memory management strategies.

As shown in Figure 6.5, the preference distribution shows Random (34.7%), Diversity (33.1%), and Uncertainty (32.2%) with a range of only 2.5 percentage points. Based on binomial confidence intervals, these preferences are statistically indistinguishable: Random [29.9%, 39.9%], Diversity [28.3%, 38.3%], and Uncertainty [27.4%, 37.4%]. This falls well below our threshold for clear preference (5 percentage points), indicating **no clear user preference** for any particular memory management strategy.

| Metric | Value | Interpretation |
|-------------------|-----------------------|--------------------------|
| Preference Range | 2.5 percentage points | No clear preference |
| Highest vs Lowest | 9 choices difference | Minimal practical impact |
| Total Evaluations | 360 forced choices | Sufficient sample size |

Table 6.3: Practical significance assessment of user preference patterns.

Table 6.3 confirms this finding strongly supports the strategy equivalence conclusion from RQ1, demonstrating that sophisticated memory management approaches do not translate to user preference advantages.

6.5 Discussion

6.5.1 Strategy Equivalence: Confirming Technical-Human Disconnect

The practical equivalence of memory management strategies (0.07-point range) reveals a significant disconnect between technical performance metrics and human perception. While Chapter 5 demonstrated substantial differences, diversity sampling achieving 56% BLEU-4 retention versus 31% for random sampling. Human evaluators perceived these as practically equivalent differences. This disconnect suggests that BLEU-4 captures lexical precision differences that are imperceptible to human evaluators, who prioritize semantic adequacy over exact word matching. The technical superiority in memory management translates to improved automated metrics but operates below the threshold of human quality perception, challenging fundamental assumptions that technical improvements automatically translate to user experience benefits.

6.5.2 Catastrophic Forgetting: A Detectable User Experience Issue

The 0.14-point difference between early and late phases provides clear evidence that catastrophic forgetting represents a genuine user experience challenge. Unlike strategy optimization effects, which operate below perception thresholds, catastrophic forgetting produces quality degradation detectable by end users. This suggests that forgetting mitigation should be prioritized over memory strategy optimization in practical deployments.

6.5.3 No Clear User Preferences

The minimal preference variation (2.5 percentage points) confirms that sophisticated memory management approaches do not translate to user preference advantages. This finding supports resource allocation toward computationally efficient approaches (random sampling) rather than complex optimization strategies.

6.5.4 Practical Implications

Based on these findings, the experimental evidence supports:

- Use simple random sampling for memory management (computationally efficient, user-equivalent performance)
- Prioritize catastrophic forgetting mitigation over memory strategy optimization
- Focus development resources on user-detectable improvements rather than technical metric optimization
- Deploy evaluation frameworks that emphasize practical significance over statistical significance

This study establishes that while technical advances in memory management strategies are algorithmically meaningful, they operate below human perception thresholds, redirecting attention toward catastrophic forgetting mitigation as the primary user experience challenge in continual learning deployment. These findings have broader implications for the field, suggesting that future research should prioritize human-centered evaluation metrics alongside traditional automated measures. The disconnect between technical performance and user perception highlights the need for evaluation frameworks that bridge this gap, ensuring that algorithmic improvements translate to meaningful user experience enhancements in practical applications.

Chapter 7

Conclusion and Future Work

This chapter synthesizes the key contributions and implications of the ALCIE framework, providing a comprehensive assessment of strategic memory management in continual IC systems. We summarize our main contributions and paradigm-shifting insights (Section 7.1), discuss practical implications for system design and deployment (Section 7.2), acknowledge limitations (Section 7.3), and outline future research directions (Section 7.4).

7.1 Summary of Contributions and Key Insights

The ALCIE framework represents the first systematic investigation of whether strategic memory management provides meaningful advantages over computationally efficient random sampling in continual IC. Through comprehensive cross-architectural evaluation, this research has uncovered counterintuitive insights that fundamentally challenge prevailing assumptions about optimization priorities in CL systems.

Our central investigation asked: *Can strategic sample selection significantly improve continual learning performance beyond computationally efficient random approaches?* The systematic evaluation across BLIP-2 and OFA architectures reveals a nuanced answer that challenges expectations about the relationship between algorithmic sophistication and practical effectiveness.

While strategic AL approaches demonstrate measurable technical improvements, particularly diversity sampling’s superior early transition stability (56% retention vs. 30-36% for random sampling), the practical significance of these gains operates below human perception thresholds. This finding represents a paradigm shift: technical metric optimization does not necessarily translate to meaningful user experience improvements.

7.1.1 Research Question Answers

RQ1: Memory Strategy Effectiveness and Architecture Dependency Diversity sampling achieves superior early transition stability (56% retention vs. 30–36% for random sampling), while architectures demonstrate different strategy sensitivities, BLIP-2 exhibit

strategy convergence (0.883 accuracy across multiple approaches) while OFA demonstrate a clear strategy sensitivity (0.865–0.871) range.

RQ2: Lexical versus Semantic Forgetting Patterns We identified a systematic lexical-semantic forgetting disconnect, with severe lexical generation degradation (0-35% retention) occurring alongside robust semantic understanding preservation (86-93% retention). All strategies exhibit universal early transition vulnerability rather than position-specific effects, with the steepest performance drops occurring during initial domain transitions regardless of memory management sophistication.

RQ3: Memory Capacity Optimization and Resource Efficiency Both architectures benefit from unconstrained memory, but BLIP-2 demonstrates minimal differences (0.884 vs. 0.883 accuracy) while OFA demonstrates clearer benefits (0.871 vs. 0.870 accuracy), suggesting architecture-dependent memory utilization patterns.

User Study Findings: Memory management strategies do not produce meaningful differences in human-perceived caption quality, with only a 0.07-point range across all evaluation dimensions. However, human evaluators detected a meaningful 0.14-point quality degradation between early and late learning phases, confirming that CF represents a genuine user experience challenge while strategy optimization effects operate below human perception thresholds.

7.1.2 Paradigm-Shifting Insights

Lexical-Semantic Forgetting Paradigm: The most fundamental discovery is that catastrophic forgetting operates differentially across knowledge types, with severe lexical generation degradation occurring alongside robust semantic understanding preservation. This 60-85 percentage point gap explains why random sampling remains competitive and challenges community assumptions about forgetting severity in vision-language models.

The Unexpected Competitiveness of Random Sampling: Perhaps the most significant finding is that random sampling achieves competitive performance across multiple dimensions while providing substantial computational efficiency advantages. Random sampling's effectiveness (36% Accessories retention, 53% Outerwear retention for BLIP-2) challenges the assumption that sophisticated algorithmic approaches necessarily outperform simple baselines.

Technical-Human Performance Disconnect: The discovery that technical improvements operate below human perception thresholds while CF effects are clearly detectable represents a fundamental insight for CL research. This disconnect suggests that evaluation frameworks focusing solely on automated metrics may optimize for improvements that provide no practical benefit to end users.

Universal Early Transition Vulnerability: The identification of severe performance drops during initial domain transitions (40-70% degradation) across all strategies, regardless of sophistication, reveals fundamental limitations in knowledge consolidation that transcend memory management optimization.

7.2 Practical Implications

7.2.1 Strategy Selection Guidelines

Lexical-Semantic Optimization: Given that semantic understanding remains stable across strategies while lexical generation shows greater variability, practitioners should prioritize approaches that maintain essential conceptual knowledge. Random sampling effectively preserves semantic capabilities while providing computational efficiency advantages.

Computational Efficiency Priority: Given the unexpected competitiveness of random sampling and minimal human-detectable differences between strategies, **practical deployments should prioritize computational efficiency over algorithmic sophistication.** Random sampling provides an optimal balance of performance and resource utilization.

Architecture-Specific Recommendations: BLIP-2’s performance ceiling effect suggests frozen encoder architectures benefit minimally from memory management optimization, making random sampling particularly attractive. OFA’s strategy sensitivity indicates end-to-end architectures may justify more sophisticated memory management, though practical benefits remain modest.

Resource Allocation Principle: Rather than investing computational resources in sophisticated memory selection algorithms, practitioners should focus on catastrophic forgetting mitigation mechanisms that operate above human perception thresholds. This represents a fundamental shift in optimization priorities for continual learning systems.

7.3 Limitations

Our evaluation is limited to the fashion domain using the FACAD dataset, which may not generalize to domains with greater visual variability. We evaluate only BLIP-2 and OFA architectures, use fixed domain sequences, and conduct human evaluation with 15 participants and 24 images, limiting statistical power and generalizability.

Memory management strategies rely on basic score-based replacement and proportional deletion. Advanced methods like hierarchical memory, adaptive allocation, or meta-learning remain unexplored. The study uses offline batch processing and provides limited computational efficiency analysis, so practical trade-offs in inference time, memory use, or energy consumption remain unaddressed.

In summary, the ALCIE framework contributes essential insights that fundamentally challenge prevailing assumptions about optimization priorities in CL systems. The unexpected competitiveness of random sampling, combined with the technical-human performance disconnect, suggests that continual learning research should prioritize CF mitigation mechanisms over memory strategy optimization. For practitioners, this research establishes that computational efficiency should be prioritized over algorithmic sophistication, with random sampling providing an optimal balance for most deployment scenarios. These principles provide foundations for developing more effective and user-centered CL systems.

The lexical-semantic forgetting paradigm represents our most significant theoretical contribution, demonstrating that continual learning challenges are primarily surface-level rather than conceptual. This insight explains the competitive effectiveness of

random sampling and provides theoretical foundation for prioritizing computational efficiency over algorithmic sophistication.

7.4 Future Work

7.4.1 Addressing Fundamental Continual Learning Limitations

Given the universal early transition vulnerability and lexical-semantic performance disconnect, future research should prioritize developing novel approaches to CF mitigation that operate above human perception thresholds.

7.4.2 Domain and Task Expansion

Cross-Domain and Multi-Task Extension: Validate ALCIE across diverse domains such as medical imaging, natural scenes, and scientific visualization, each presenting unique challenges that could reveal domain-specific memory management requirements. Extension beyond image captioning to tasks such as visual question answering, image-text retrieval, and multimodal dialogue would address our current task-specific limitation and reveal how memory management strategies perform when models must simultaneously maintain knowledge across different objectives.

Multilingual Investigation: Explore multilingual continual learning where models adapt simultaneously to novel visual domains and languages, addressing the global applicability limitation of our English-only evaluation. Building on recent progress in multilingual and multimodal language models [101], this work could establish whether architecture-dependent memory patterns observed in our experiments hold across diverse linguistic contexts.

7.4.3 Architectural Innovation and Scalability

Parameter-Efficient and Foundation-Based Designs: Test the generalizability of current findings on large multimodal models and parameter-efficient adaptation methods such as C-LoRA [57] and adapters, as reviewed in recent work on fine-tuning modular architectures [102]. Our findings about architecture-dependent memory utilization patterns were established on models with specific parameter ranges (BLIP-2: 2.7B, OFA: 470M), but may not generalize to foundation models exceeding 10B parameters.

Mixture-of-Experts and Memory Co-Design: Investigate sparse mixture-of-experts architectures combined with memory selection strategies to enable more scalable and specialized representations. This approach could potentially address the universal mid-sequence vulnerability identified in our experiments through expert-specific memory allocation. Foundational work on sparsely-gated experts provides guidance for expert-memory pairing [103].

7.4.4 Human-Centered Continual Learning and Methodological Innovations

Research should develop optimization frameworks that explicitly target improvements above human perception thresholds rather than focusing solely on automated metric improvements.

Scalable Human Evaluation: Expand human-centered evaluation across larger and more demographically diverse populations to assess interpretability, alignment, and real-world utility [104]. Our human evaluation findings revealed a critical technical-human performance disconnect, but were limited to 15 participants and 24 fashion images. Comprehensive human studies would investigate whether this disconnect persists across different user populations and application domains.

These research directions collectively aim to transform continual learning from a focus on algorithmic sophistication to practical effectiveness, ensuring that technical advances translate to meaningful improvements in real-world applications.

Bibliography

- [1] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org, 2015.
- [2] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022.
- [3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, Honolulu, Hawaii, USA, 23-29 July 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.
- [4] Subarnaduti Paul, Manuel Brack, Patrick Schramowski, Kristian Kersting, and Martin Mundt. Core tokensets for data-efficient sequential training of transformers. *arXiv preprint arXiv:2410.05800*, 2024.
- [5] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.
- [6] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 742–758. Springer, 2020.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR. OpenReview.net*, 2021.

- [9] Da Yu, Mingyi Zhang, Mantian Li, Fusheng Zha, Junge Zhang, Lining Sun, and Kaiqi Huang. Squeezing more past knowledge for online class-incremental continual learning. *IEEE CAA J. Autom. Sinica*, 10(3):722–736, 2023.
- [10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
- [11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society, 2015.
- [12] Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. Towards adaptable and interactive image captioning with data augmentation and episodic memory. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Toronto, Canada (Hybrid)*, pages 245–256. Association for Computational Linguistics, 2023.
- [13] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.
- [14] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.
- [15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- [16] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6467–6476, 2017.
- [17] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.
- [18] Cyprien de Masson d’Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. In *NeurIPS*, pages 13122–13131, 2019.
- [19] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc’Aurelio Ranzato. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486, 2019.
- [20] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

- [21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5533–5542. IEEE Computer Society, 2017.
- [22] Burr Settles. Active learning literature survey. *Machine Learning*, 15(2):201–221, 2009.
- [23] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.
- [24] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*. Open-Review.net, 2018.
- [25] Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. To label or not to label: Hybrid active learning for neural machine translation. In *Proceedings of the 2025 International Conference on Computational Linguistics*, pages 3071–3082. Association for Computational Linguistics, 2025.
- [26] Zuhui Wang, Sandra Sajeev, Gaurav Mittal, Matthew Hall, Ye Yu, Zhaozheng Yin, and Mei Chen. FALCON: fair active learning for content moderation. In *ECCV Workshops (21)*, volume 15643 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2024.
- [27] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3366–3385, 2022.
- [28] Xuewen Yang, Tao He, Ruiyu Hu, Changchang Sun, Qianru Sun, Hefei Ling, and Qiong Yin. Fashion captioning: Towards generating accurate descriptions with semantic rewards. *arXiv preprint arXiv:2008.02693*, 2020.
- [29] Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active learning approaches to enhancing neural machine translation: An empirical study. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 1796–1806. Association for Computational Linguistics, 2020.
- [30] Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures (extended abstract). In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4970–4974. ijcai.org, 2017.
- [31] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36, 2019.
- [32] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):539–559, 2023.

- [33] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3):62:1–62:39, 2024.
- [34] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2010.
- [35] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, December 12-14, 2011, Granada, Spain*, pages 1143–1151, 2011.
- [36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.
- [37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [38] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society, 2018.
- [39] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. IEEE, 2020.
- [40] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10971–10980. IEEE, 2020.
- [41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019.
- [42] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. In *Computer Vision - ECCV 2020 - 16th European Conference on Computer Vision, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020.

- [43] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7241–7259. Association for Computational Linguistics, 2022.
- [44] Ron Mokady, Amir Hertz, Or Patashnik, Daniel Cohen-Or, and Gal Chechik. Clipcap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- [45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.
- [46] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. *KI - Künstliche Intelligenz*, 34(4):571–584, 2020.
- [47] Rajarshi Biswas, Michael Barz, Mareike Hartmann, and Daniel Sonntag. Improving german image captions using machine translation and transfer learning. In *Statistical Language and Speech Processing - 9th International Conference, SLSP 2021, Bilbao, Spain, November 23-24, 2021, Proceedings*, volume 13062 of *Lecture Notes in Computer Science*, pages 1–13. Springer, 2021.
- [48] Mareike Hartmann, Aliko Anagnostopoulou, and Daniel Sonntag. Interactive machine learning for image captioning. *arXiv preprint arXiv:2202.13623*, 2022.
- [49] Omair Shahzad Bhatti, Harshinee Sriram, Abdulrahman Mohamed Selim, Cristina Conati, Michael Barz, and Daniel Sonntag. Detecting when users disagree with generated captions. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction, ICMI Companion 2024, San Jose, Costa Rica, November 4-8, 2024*, pages 195–203. ACM, 2024.
- [50] Aliko Anagnostopoulou, Thiago S. Gouvêa, and Daniel Sonntag. Enhancing journalism with AI: A study of contextualized image captioning for news articles using LLMs and LMMs. *arXiv preprint arXiv:2408.04331*, 2024.
- [51] Daniel Sonntag, Michael Barz, and Thiago Gouvêa. A look under the hood of the interactive deep learning enterprise (no-idle). *arXiv preprint arXiv:2406.19054*, 2024.
- [52] Giang Nguyen, Tae Joon Jun, Trung Quang Tran, and Daeyoung Kim. Contcap: A comprehensive framework for continual image captioning. *arXiv preprint arXiv:1909.08745*, 2019.
- [53] Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6523–6541. International Committee on Computational Linguistics, 2020.

- [54] Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. RATT: recurrent attention to transient tasks for continual image captioning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [55] Wei Zhao, Zhou Zhao, Yabing Feng, Kai Lei, Xiaojun Chen, Min Yang, and Wei Xu. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21:1047–1061, 2019.
- [56] Aristeidis Panos, Rahaf Aljundi, Daniel Olmeda Reino, and Richard E. Turner. Efficient few-shot continual learning in vision-language models. *CoRR*, abs/2502.04098, 2025.
- [57] Xin Zhang, Liang Bai, Xian Yang, and Jiye Liang. C-lora: Continual low-rank adaptation for pre-trained models. *CoRR*, abs/2502.17920, 2025.
- [58] Weiguo Pian, Shijian Deng, Shentong Mo, Yunhui Guo, and Yapeng Tian. Modality-inconsistent continual learning of multimodal large language models. *arXiv preprint arXiv:2412.13050*, 2024.
- [59] Amir Hossein Rahmati, Mingzhou Fan, Ruida Zhou, Nathan M. Urban, Byung-Jun Yoon, and Xiaoning Qian. Understanding uncertainty-based active learning under model mismatch. *CoRR*, abs/2408.13690, 2024.
- [60] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016.
- [61] Anant Raj and Francis R. Bach. Convergence of uncertainty sampling for active learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 18310–18331. PMLR, 2022.
- [62] Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Uncertainty-aware image captioning. In *AAAI*, volume 37, pages 591–599, 2023.
- [63] Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. Active learning for natural language generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9862–9877. Association for Computational Linguistics, 2023.
- [64] Suorong Yang, Peng Ye, Wanli Ouyang, Dongzhan Zhou, and Furao Shen. A clip-powered framework for robust and generalizable data selection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- [65] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

- [66] Sania Waheed and Na Min An. Image embedding sampling method for diverse captioning. *CoRR*, abs/2502.10118, 2025.
- [67] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*. OpenReview.net, 2020.
- [68] Yinan He, Lile Cai, Jingyi Liao, and Chuan-Sheng Foo. Hybrid active learning with uncertainty-weighted embeddings. *Trans. Mach. Learn. Res.*, 2024, 2024.
- [69] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
- [70] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. IEEE Computer Society, 2018.
- [71] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018.
- [72] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE, 2019.
- [73] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.
- [74] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. The Association for Computer Linguistics, 2005.
- [75] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out, Proceedings of the ACL-04 Workshop, Barcelona, Spain, July 25-26, 2004*, pages 74–81. ACL, 2004.
- [76] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015.

- [77] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference on Computer Vision, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2016.
- [78] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net, 2020.
- [79] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [80] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [81] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [82] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *CoRR*, abs/2309.10313, 2023.
- [83] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *CVPR*, pages 9368–9377. Computer Vision Foundation / IEEE Computer Society, 2018.
- [84] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12. ACM/Springer, 1994.
- [85] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR (Poster)*. OpenReview.net, 2018.
- [86] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- [87] Yinan He, Lile Cai, Jingyi Liao, and Chuan-Sheng Foo. Hybrid active learning with uncertainty-weighted embeddings. *Trans. Mach. Learn. Res.*, 2024, 2024.
- [88] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [89] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3390–3398. AAAI Press, 2018.

- [90] D. C. Montgomery and G. C. Runger. *Introduction to Probability and Statistics*. Wiley, 2017.
- [91] W. L. Hays. *Statistics for the Social Sciences*. Holt, Rinehart and Winston, 1973.
- [92] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2020.
- [93] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.
- [94] B. H. Cohen. *Explaining Psychological Statistics*. Wiley, 2001.
- [95] Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, and Michalis K. Titsias. Information-theoretic online memory selection for continual learning. In *ICLR*. OpenReview.net, 2022.
- [96] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Trans. Assoc. Comput. Linguistics*, 13:220–248, 2025.
- [97] Qinze Zhu, Xinsheng Shu, Jiayi Yang, and Mingyong Li. Ccuh: Clip-based clustering method for unsupervised hashing multi-modal retrieval. In Mufti Mahmud, Maryam Doborjeh, Kevin Wong, Andrew Chi Sing Leung, Zohreh Doborjeh, and M. Tanveer, editors, *Neural Information Processing*, pages 93–106, Singapore, 2025. Springer Nature Singapore.
- [98] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao. Improving task-free continual learning by distributionally robust memory evolution. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 22985–22998. PMLR, 2022.
- [99] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.
- [100] Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26:137 – 161, 2019.
- [101] Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers. *CoRR*, abs/2405.10936, 2024.
- [102] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024.
- [103] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.*, 37(7):3896–3915, 2025.

- [104] Md. Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee-Wei Tan, Md. Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *EMNLP*, pages 13785–13816. Association for Computational Linguistics, 2024.

Appendix: Supplementary Experiments Results

This appendix presents comprehensive evaluation results for supplementary metrics employed in the experimental evaluation described in Chapter 5. While the main body focuses on BLEU-4 and BERTScore-F1 as primary evaluation measures, this appendix provides complete results for ROUGE-L and METEOR across all experimental conditions and memory management strategies.

Evaluation Metrics Overview

The supplementary experimental evaluation employs two complementary metrics to assess caption generation quality across multiple dimensions:

ROUGE-L: Measures longest common subsequence overlap, capturing structural similarity and word ordering between generated and reference captions.

METEOR: Incorporates stemming, synonymy, and paraphrase matching for nuanced semantic evaluation that accounts for linguistic variations.

Each metric provides a different perspective on model performance, enabling comprehensive assessment of continual learning effectiveness across lexical and structural dimensions.

Experiment 1: No Memory Baseline - Complete Results

ROUGE-L Performance Without Memory

| ROUGE-L Performance Without Memory | | | | | | | | | | | | |
|------------------------------------|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.443 | 0.650 | 0.132 _(30%) | 0.168 _(26%) | 0.117 _(26%) | 0.119 _(18%) | 0.111 _(25%) | 0.111 _(17%) | 0.115 _(26%) | 0.117 _(18%) | 0.100 _(23%) | 0.107 _(16%) |
| Bottoms | | | 0.389 | 0.536 | 0.114 _(29%) | 0.192 _(36%) | 0.114 _(29%) | 0.125 _(23%) | 0.096 _(25%) | 0.102 _(19%) | 0.098 _(25%) | 0.114 _(21%) |
| Dresses | | | | | 0.467 | 0.661 | 0.158 _(34%) | 0.189 _(29%) | 0.121 _(26%) | 0.116 _(18%) | 0.126 _(27%) | 0.128 _(19%) |
| Outerwear | | | | | | | 0.468 | 0.662 | 0.154 _(33%) | 0.149 _(23%) | 0.117 _(25%) | 0.136 _(21%) |
| Shoes | | | | | | | | | 0.395 | 0.593 | 0.136 _(34%) | 0.393 _(66%) |
| Tops | | | | | | | | | | | 0.530 | 0.734 |

Table 1: ROUGE-L scores showing severe structural degradation without episodic memory. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

METEOR Performance Without Memory

| METEOR Performance Without Memory | | | | | | | | | | | | |
|-----------------------------------|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.439 | 0.583 | 0.115 _(26%) | 0.132 _(23%) | 0.097 _(22%) | 0.083 _(14%) | 0.091 _(21%) | 0.078 _(13%) | 0.087 _(20%) | 0.081 _(14%) | 0.079 _(18%) | 0.074 _(13%) |
| Bottoms | | | 0.391 | 0.503 | 0.105 _(27%) | 0.168 _(33%) | 0.105 _(27%) | 0.100 _(20%) | 0.076 _(19%) | 0.071 _(14%) | 0.081 _(21%) | 0.088 _(17%) |
| Dresses | | | | | 0.459 | 0.598 | 0.141 _(31%) | 0.154 _(26%) | 0.095 _(21%) | 0.078 _(13%) | 0.108 _(24%) | 0.099 _(17%) |
| Outerwear | | | | | | | 0.464 | 0.604 | 0.134 _(29%) | 0.112 _(19%) | 0.098 _(21%) | 0.103 _(17%) |
| Shoes | | | | | | | | | 0.379 | 0.545 | 0.114 _(30%) | 0.352 _(65%) |
| Tops | | | | | | | | | | | 0.521 | 0.683 |

Table 2: METEOR scores demonstrating semantic degradation without memory mechanisms. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Experiments 2-3: Random Sampling Memory Management - Complete Results

ROUGE-L Performance: Random Sampling

| ROUGE-L Performance: OFA Random Sampling | | | | | | | | | | | | |
|--|---------------|-------|------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.442 | 0.443 | 0.261 _(59%) | 0.272 _(61%) | 0.268 _(61%) | 0.264 _(60%) | 0.284 _(64%) | 0.275 _(62%) | 0.266 _(60%) | 0.272 _(61%) | 0.263 _(59%) | 0.286 _(65%) |
| Bottoms | | | 0.381 | 0.379 | 0.207 _(54%) | 0.207 _(55%) | 0.191 _(50%) | 0.200 _(53%) | 0.194 _(51%) | 0.184 _(49%) | 0.184 _(48%) | 0.196 _(52%) |
| Dresses | | | | | 0.450 | 0.424 | 0.203 _(45%) | 0.188 _(44%) | 0.219 _(49%) | 0.219 _(52%) | 0.177 _(39%) | 0.188 _(44%) |
| Outerwear | | | | | | | 0.468 | 0.472 | 0.272 _(58%) | 0.255 _(54%) | 0.211 _(45%) | 0.219 _(46%) |
| Shoes | | | | | | | | | 0.389 | 0.378 | 0.263 _(68%) | 0.274 _(72%) |
| Tops | | | | | | | | | | | 0.549 | 0.552 |

Table 3: ROUGE-L performance comparison for OFA with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

| ROUGE-L Performance: BLIP-2 Random Sampling | | | | | | | | | | | | |
|---|---------------|--------------|-------------------------------|------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.649 | 0.650 | 0.324 _(50%) | 0.317 _(49%) | 0.299 _(46%) | 0.324 _(50%) | 0.315 _(49%) | 0.328 _(50%) | 0.302 _(47%) | 0.325 _(50%) | 0.334 _(51%) | 0.342 _(53%) |
| Bottoms | | | 0.518 | 0.528 | 0.249 _(48%) | 0.246 _(47%) | 0.220 _(42%) | 0.234 _(44%) | 0.220 _(42%) | 0.247 _(47%) | 0.220 _(42%) | 0.246 _(47%) |
| Dresses | | | | | 0.647 | 0.646 | 0.300 _(46%) | 0.293 _(45%) | 0.220 _(34%) | 0.239 _(37%) | 0.178 _(28%) | 0.185 _(29%) |
| Outerwear | | | | | | | 0.662 | 0.656 | 0.322 _(49%) | 0.421 _(64%) | 0.253 _(38%) | 0.276 _(42%) |
| Shoes | | | | | | | | | 0.641 | 0.643 | 0.529 _(83%) | 0.511 _(79%) |
| Tops | | | | | | | | | | | 0.727 | 0.729 |

Table 4: ROUGE-L performance comparison for BLIP-2 with constrained (DEL+) vs unconstrained (DEL-) random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

METEOR Performance: Random Sampling

| METEOR Performance: OFA Random Sampling | | | | | | | | | | | | |
|---|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.439 | 0.441 | 0.266 _(61%) | 0.270 _(61%) | 0.265 _(60%) | 0.258 _(59%) | 0.278 _(63%) | 0.272 _(62%) | 0.264 _(60%) | 0.264 _(60%) | 0.255 _(58%) | 0.276 _(63%) |
| Bottoms | | | 0.387 | 0.387 | 0.200 _(52%) | 0.201 _(52%) | 0.199 _(51%) | 0.199 _(51%) | 0.198 _(51%) | 0.193 _(50%) | 0.176 _(45%) | 0.193 _(50%) |
| Dresses | | | | | 0.440 | 0.413 | 0.207 _(47%) | 0.197 _(48%) | 0.205 _(47%) | 0.206 _(50%) | 0.161 _(37%) | 0.172 _(42%) |
| Outerwear | | | | | | | 0.466 | 0.466 | 0.265 _(57%) | 0.247 _(53%) | 0.201 _(43%) | 0.211 _(45%) |
| Shoes | | | | | | | | | 0.371 | 0.362 | 0.242 _(65%) | 0.260 _(72%) |
| Tops | | | | | | | | | | | 0.541 | 0.546 |

Table 5: METEOR performance comparison for OFA with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

| METEOR Performance: BLIP-2 Random Sampling | | | | | | | | | | | | |
|--|---------------|--------------|-------------------------------|------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- | DEL+ | DEL- |
| Accessories | 0.582 | 0.583 | 0.280 _(48%) | 0.271 _(46%) | 0.269 _(46%) | 0.295 _(51%) | 0.270 _(46%) | 0.285 _(49%) | 0.264 _(45%) | 0.280 _(48%) | 0.289 _(50%) | 0.295 _(51%) |
| Bottoms | | | 0.486 | 0.495 | 0.229 _(47%) | 0.224 _(45%) | 0.199 _(41%) | 0.213 _(43%) | 0.198 _(41%) | 0.220 _(44%) | 0.203 _(42%) | 0.222 _(45%) |
| Dresses | | | | | 0.583 | 0.583 | 0.256 _(44%) | 0.250 _(43%) | 0.183 _(31%) | 0.199 _(34%) | 0.153 _(26%) | 0.152 _(26%) |
| Outerwear | | | | | | | 0.604 | 0.598 | 0.283 _(47%) | 0.371 _(62%) | 0.220 _(36%) | 0.240 _(40%) |
| Shoes | | | | | | | | | 0.591 | 0.595 | 0.482 _(82%) | 0.467 _(78%) |
| Tops | | | | | | | | | | | 0.678 | 0.678 |

Table 6: METEOR performance comparison for BLIP-2 with constrained vs unconstrained random sampling. Subscript percentages show retention levels, with **bold** indicating higher values between strategies.

Experiment 4: Uncertainty Sampling - Complete Results

ROUGE-L Performance: Uncertainty Sampling

| ROUGE-L Performance: Uncertainty Sampling | | | | | | | | | | | | |
|---|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.442 | 0.650 | 0.254 _(57%) | 0.315 _(48%) | 0.237 _(54%) | 0.303 _(47%) | 0.227 _(51%) | 0.313 _(48%) | 0.217 _(49%) | 0.311 _(48%) | 0.233 _(53%) | 0.331 _(51%) |
| Bottoms | | | 0.369 | 0.526 | 0.189 _(51%) | 0.227 _(43%) | 0.176 _(48%) | 0.199 _(38%) | 0.175 _(47%) | 0.213 _(40%) | 0.178 _(48%) | 0.221 _(42%) |
| Dresses | | | | | 0.460 | 0.651 | 0.206 _(45%) | 0.285 _(44%) | 0.200 _(43%) | 0.220 _(34%) | 0.172 _(37%) | 0.183 _(28%) |
| Outerwear | | | | | | | 0.457 | 0.660 | 0.246 _(54%) | 0.408 _(62%) | 0.191 _(42%) | 0.258 _(39%) |
| Shoes | | | | | | | | | 0.381 | 0.647 | 0.246 _(65%) | 0.520 _(80%) |
| Tops | | | | | | | | | | | 0.547 | 0.733 |

Table 7: ROUGE-L performance with uncertainty-based memory selection. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

METEOR Performance: Uncertainty Sampling

| METEOR Performance: Uncertainty Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.439 | 0.583 | 0.251 _(57%) | 0.274 _(47%) | 0.238 _(54%) | 0.262 _(45%) | 0.240 _(55%) | 0.272 _(47%) | 0.224 _(51%) | 0.269 _(46%) | 0.237 _(54%) | 0.288 _(49%) |
| Bottoms | | | 0.372 | 0.492 | 0.194 _(52%) | 0.206 _(42%) | 0.180 _(48%) | 0.185 _(38%) | 0.183 _(49%) | 0.199 _(40%) | 0.183 _(49%) | 0.204 _(41%) |
| Dresses | | | | | 0.451 | 0.588 | 0.194 _(43%) | 0.243 _(41%) | 0.190 _(42%) | 0.185 _(31%) | 0.158 _(35%) | 0.150 _(26%) |
| Outerwear | | | | | | | 0.452 | 0.603 | 0.246 _(54%) | 0.360 _(60%) | 0.186 _(41%) | 0.216 _(36%) |
| Shoes | | | | | | | | | 0.367 | 0.597 | 0.236 _(64%) | 0.472 _(79%) |
| Tops | | | | | | | | | | | 0.539 | 0.683 |

Table 8: METEOR performance with uncertainty-based memory selection showing robust semantic retention. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Experiment 5: Diversity Sampling - Complete Results

ROUGE-L Performance: Diversity Sampling

| ROUGE-L Performance: Diversity Sampling | | | | | | | | | | | | |
|---|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.442 | 0.652 | 0.249 _(56%) | 0.441 _(68%) | 0.254 _(57%) | 0.341 _(52%) | 0.252 _(57%) | 0.281 _(48%) | 0.245 _(55%) | 0.263 _(40%) | 0.251 _(57%) | 0.265 _(41%) |
| Bottoms | | | 0.381 | 0.535 | 0.207 _(54%) | 0.323 _(60%) | 0.191 _(50%) | 0.239 _(45%) | 0.194 _(51%) | 0.231 _(43%) | 0.184 _(48%) | 0.220 _(41%) |
| Dresses | | | | | 0.455 | 0.654 | 0.204 _(45%) | 0.338 _(52%) | 0.190 _(42%) | 0.240 _(37%) | 0.159 _(35%) | 0.191 _(29%) |
| Outerwear | | | | | | | 0.462 | 0.665 | 0.258 _(56%) | 0.465 _(70%) | 0.189 _(41%) | 0.305 _(46%) |
| Shoes | | | | | | | | | 0.389 | 0.641 | 0.236 _(61%) | 0.554 _(86%) |
| Tops | | | | | | | | | | | 0.544 | 0.730 |

Table 9: ROUGE-L performance with diversity-based memory selection demonstrating superior transition stability. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

METEOR Performance: Diversity Sampling

| METEOR Performance: Diversity Sampling | | | | | | | | | | | | |
|--|---------------|--------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.439 | 0.585 | 0.250 _(57%) | 0.388 _(66%) | 0.253 _(58%) | 0.295 _(50%) | 0.249 _(57%) | 0.238 _(41%) | 0.240 _(55%) | 0.220 _(38%) | 0.243 _(55%) | 0.225 _(38%) |
| Bottoms | | | 0.380 | 0.502 | 0.206 _(54%) | 0.300 _(60%) | 0.186 _(49%) | 0.217 _(43%) | 0.194 _(51%) | 0.209 _(42%) | 0.182 _(48%) | 0.199 _(40%) |
| Dresses | | | | | 0.445 | 0.590 | 0.188 _(42%) | 0.292 _(49%) | 0.177 _(40%) | 0.204 _(35%) | 0.141 _(32%) | 0.157 _(27%) |
| Outerwear | | | | | | | 0.458 | 0.607 | 0.252 _(55%) | 0.418 _(69%) | 0.176 _(38%) | 0.267 _(44%) |
| Shoes | | | | | | | | | 0.374 | 0.593 | 0.212 _(57%) | 0.509 _(86%) |
| Tops | | | | | | | | | | | 0.536 | 0.679 |

Table 10: METEOR performance with diversity-based memory selection demonstrating exceptional structural preservation. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Experiment 6: Hybrid Memory Management - Complete Results

ROUGE-L Performance: Hybrid Management

| ROUGE-L Performance: Hybrid Memory Management | | | | | | | | | | | | |
|---|---------------|--------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.442 | 0.652 | 0.274 _(62%) | 0.330 _(51%) | 0.257 _(58%) | 0.321 _(49%) | 0.234 _(53%) | 0.333 _(51%) | 0.235 _(53%) | 0.307 _(47%) | 0.215 _(49%) | 0.325 _(50%) |
| Bottoms | | | 0.363 | 0.531 | 0.206 _(57%) | 0.234 _(44%) | 0.192 _(53%) | 0.213 _(40%) | 0.179 _(49%) | 0.220 _(41%) | 0.161 _(44%) | 0.220 _(41%) |
| Dresses | | | | | 0.448 | 0.655 | 0.226 _(50%) | 0.288 _(44%) | 0.210 _(47%) | 0.216 _(33%) | 0.159 _(35%) | 0.177 _(27%) |
| Outerwear | | | | | | | 0.483 | 0.663 | 0.270 _(56%) | 0.415 _(63%) | 0.186 _(39%) | 0.253 _(38%) |
| Shoes | | | | | | | | | 0.390 | 0.643 | 0.256 _(66%) | 0.527 _(82%) |
| Tops | | | | | | | | | | | 0.526 | 0.729 |

Table 11: ROUGE-L performance with hybrid memory management demonstrating balanced multi-criteria selection. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

METEOR Performance: Hybrid Management

| METEOR Performance: Hybrid Memory Management | | | | | | | | | | | | |
|--|---------------|--------------|-------------------------------|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|------------------------|-------------------------------|
| Test Domain | + Accessories | | + Bottoms | | + Dresses | | + Outerwear | | + Shoes | | + Tops | |
| | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 | OFA | BLIP-2 |
| Accessories | 0.439 | 0.585 | 0.270 _(61%) | 0.283 _(48%) | 0.250 _(57%) | 0.279 _(48%) | 0.229 _(52%) | 0.290 _(50%) | 0.236 _(54%) | 0.265 _(45%) | 0.213 _(49%) | 0.281 _(48%) |
| Bottoms | | | 0.365 | 0.497 | 0.207 _(57%) | 0.217 _(44%) | 0.198 _(54%) | 0.198 _(40%) | 0.179 _(49%) | 0.207 _(42%) | 0.173 _(47%) | 0.204 _(41%) |
| Dresses | | | | | 0.441 | 0.591 | 0.214 _(49%) | 0.245 _(41%) | 0.198 _(45%) | 0.178 _(30%) | 0.141 _(32%) | 0.144 _(24%) |
| Outerwear | | | | | | | 0.479 | 0.605 | 0.265 _(55%) | 0.369 _(61%) | 0.173 _(36%) | 0.218 _(36%) |
| Shoes | | | | | | | | | 0.374 | 0.594 | 0.235 _(63%) | 0.480 _(81%) |
| Tops | | | | | | | | | | | 0.518 | 0.679 |

Table 12: METEOR performance with hybrid memory management showing balanced semantic retention. Subscript percentages show retention levels, with **bold** indicating higher values between architectures.

Summary of Supplementary Metrics

The comprehensive evaluation across ROUGE-L and METEOR metrics confirms the primary findings presented in the main experimental chapter:

Consistent Cross-Metric Patterns: All supplementary metrics demonstrate the same fundamental patterns observed in BLEU-4 and BERTScore-F1 results. Diversity sampling achieves superior early transition stability across ROUGE-L (68% vs 50% retention for BLIP-2 Accessories after Bottoms training), while random sampling provides competitive performance with computational efficiency advantages.

Lexical and Structural Stability: ROUGE-L and METEOR results show consistent patterns with the primary evaluation metrics, validating that episodic memory mechanisms effectively preserve both structural and semantic understanding while lexical metrics show greater sensitivity to memory management strategies.

Architecture-Independent Validation: The supplementary metrics confirm architecture-specific patterns, with BLIP-2 demonstrating performance convergence across strategies while OFA shows clear strategy sensitivity. These patterns appear consistently across all evaluation dimensions, reinforcing the architectural dependency conclusions presented in the main results.

Universal Mid-Sequence Vulnerability: All metrics confirm the universal mid-sequence vulnerability at the Dresses position, with 8-15% retention across ROUGE-L, METEOR, and BLEU-4 metrics regardless of memory management sophistication. This cross-metric validation strengthens the evidence for position-dependent forgetting effects in continual learning scenarios.

Diversity Sampling Excellence: The complete metric evaluation validates diversity sampling's superior performance across multiple dimensions. Beyond the early transition stability advantages demonstrated in ROUGE-L and METEOR, diversity sampling achieves exceptional semantic preservation in BERTScore components, with some domains showing retention above baseline performance, indicating potential positive transfer effects through comprehensive feature space coverage.

Cross-Architecture Metric Consistency: The supplementary results confirm that architectural differences manifest consistently across all evaluation frameworks. BLIP-2's superior absolute performance combined with strategy convergence appears across ROUGE-L (0.650 baseline vs OFA's 0.442), METEOR (0.585 vs 0.439), and all BERTScore components, while OFA's superior retention stability appears consistently across semantic metrics.

Memory Strategy Effectiveness Validation: The comprehensive metric analysis confirms that:

- Random sampling provides robust baseline performance with 59-65% ROUGE-L retention for OFA and 47-53% for BLIP-2
- Uncertainty sampling shows moderate effectiveness with 48-57% ROUGE-L retention patterns
- Diversity sampling demonstrates optimal early transition performance with 56-68% retention after first transitions
- Hybrid approaches achieve balanced but not superior performance with 49-62% retention ranges

These comprehensive results demonstrate that the primary conclusions drawn from BLEU-4 and BERTScore-F1 analysis are robust across multiple evaluation frameworks, providing strong empirical support for the memory management strategy recommendations and architectural insights presented in the main experimental chapter. The cross-metric validation using ROUGE-L and METEOR strengthens confidence in the practical applicability of diversity sampling for optimal continual learning performance and random sampling for computational efficiency, while confirming the fundamental architectural dependencies that influence optimal memory management approaches.