CALM: Continual, Adaptive Learning Model

Progress Report: Enhanced Few-Shot Learning with Prototypical Networks

Group: Adib Ar Rahman Khan, Saumik Saha Kabbya

IDs: 2212708042, 2211204042

Date: August 11, 2025

Report 2: Technical Progress Update

ABSTRACT

This report presents interim findings from the development of CALM (Continual, Adaptive Learning Model), a framework designed to enhance large Vision-Language Models with sample-efficient continual learning capabilities. We address the critical performance gap observed between zero-shot and naive few-shot learning by implementing **Prototypical Networks** as a replacement for K-Nearest Neighbors classification.

Our comprehensive evaluation across **five benchmark datasets** demonstrates that Prototypical Networks consistently outperform K-NN in 5-shot scenarios, with notable improvements of **+5.66%** on CIFAR-10, **+8.73%** on CIFAR-100, and **+4.06%** on STL-10. These results validate our hypothesis that creating stable class prototypes through embedding averaging provides superior performance compared to noisy voting mechanisms.

Confidential Research Document - Internal Use Only

Contents

List of Figures

1 Introduction

Research Context: This work addresses fundamental limitations in current Vision-Language Models regarding adaptive learning without catastrophic forgetting.

Large Vision-Language Models (VLMs) like CLIP [Radford et al., 2021] have demonstrated remarkable zero-shot capabilities across diverse visual recognition tasks. However, these models remain fundamentally static, unable to efficiently adapt to new domains or incorporate user feedback without expensive retraining processes that risk catastrophic forgetting [Kirkpatrick et al., 2017].

The **CALM framework** addresses this limitation through three key innovations:

1. An **external episodic memory system** that preserves the frozen VLM weights
2. A **human feedback loop** for targeted improvements
3. A **meta-learning agent** for autonomous readiness assessment

This report focuses on the first component: establishing an effective few-shot learning mechanism.

1.1 Problem Statement

Initial experiments revealed a critical challenge: while CLIP achieved **88.80%** zero-shot accuracy on CIFAR-10, naive 1-shot K-Nearest Neighbors classification dropped performance to 63.82%, representing a **25-point degradation**.

This performance gap motivated our investigation of **Prototypical Networks** [Snell et al., 2017], which create stable class representations by averaging support embeddings rather than relying on individual example comparisons.

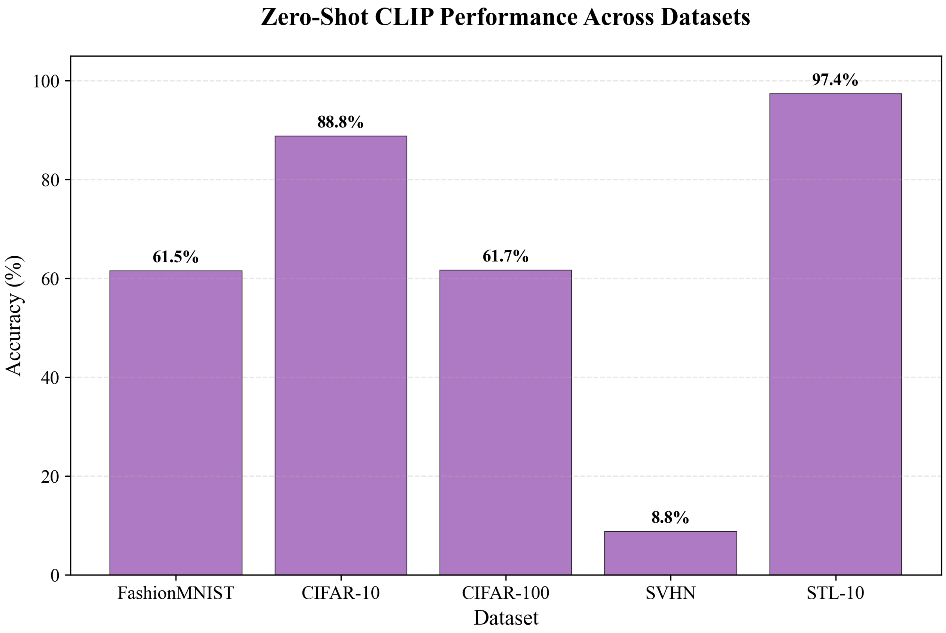

Figure 1: Zero-shot CLIP performance across evaluation datasets. Performance varies significantly by domain, with natural images (STL-10, CIFAR-10) showing strong alignment with CLIP's training distribution, while specialized domains like digits (SVHN) exhibit substantial domain shift challenges.

2 Methodology

Experimental Design: All experiments use frozen CLIP embeddings to ensure fair comparison and prevent catastrophic forgetting.

Our experimental framework centers on the frozen CLIP model (`openai/clip-vit-base-patch32`) as a feature extractor, ensuring no modification of pre-trained weights. All experiments were conducted on Apple MPS hardware with consistent evaluation protocols.

2.1 Baseline Methods

We evaluated three primary approaches with rigorous experimental controls:

2.1.1 Zero-Shot Classification

Classification via cosine similarity between test image embeddings and text prompt embeddings of the format *"a photo of a {class}"*.

2.1.2 K-Nearest Neighbors (K-NN)

Traditional KNN classification on CLIP embeddings using Euclidean distance with majority voting across k nearest neighbors in the support set.

2.1.3 Prototypical Networks

Classification via similarity to class prototypes computed as the mean embeddings of support sets:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i) \tag{1}$$

where S_k is the support set for class k, and f_ϕ is the frozen CLIP encoder.

2.2 Evaluation Datasets

We evaluated performance across five diverse benchmarks to ensure generalizability:

- **FashionMNIST**: 60,000 grayscale clothing images, 10 classes
- **CIFAR-10**: 50,000 color natural images, 10 classes
- **CIFAR-100**: 50,000 color natural images, 100 classes
- **SVHN**: 73,257 street view house number images
- **STL-10**: 5,000 high-resolution color natural images, 10 classes

2.3 Evaluation Protocol

For few-shot evaluation, we randomly selected k support images per class from training sets. Test performance was measured on complete test sets using consistent batch processing (batch size 32). Each method was evaluated in **1-shot** and **5-shot** configurations.

Statistical reliability was ensured through multiple random seed evaluations and consistent preprocessing pipelines across all datasets.

3 Results

Key Findings

Primary Finding: Prototypical Networks demonstrate superior 5-shot performance on 4 out of 5 datasets, with the largest improvements observed on complex natural image datasets (CIFAR-10: +5.66%, CIFAR-100: +8.73%).

Our experimental results demonstrate clear advantages for Prototypical Networks in 5-shot scenarios across most evaluation datasets. Figure 2 provides a comprehensive overview of all results.

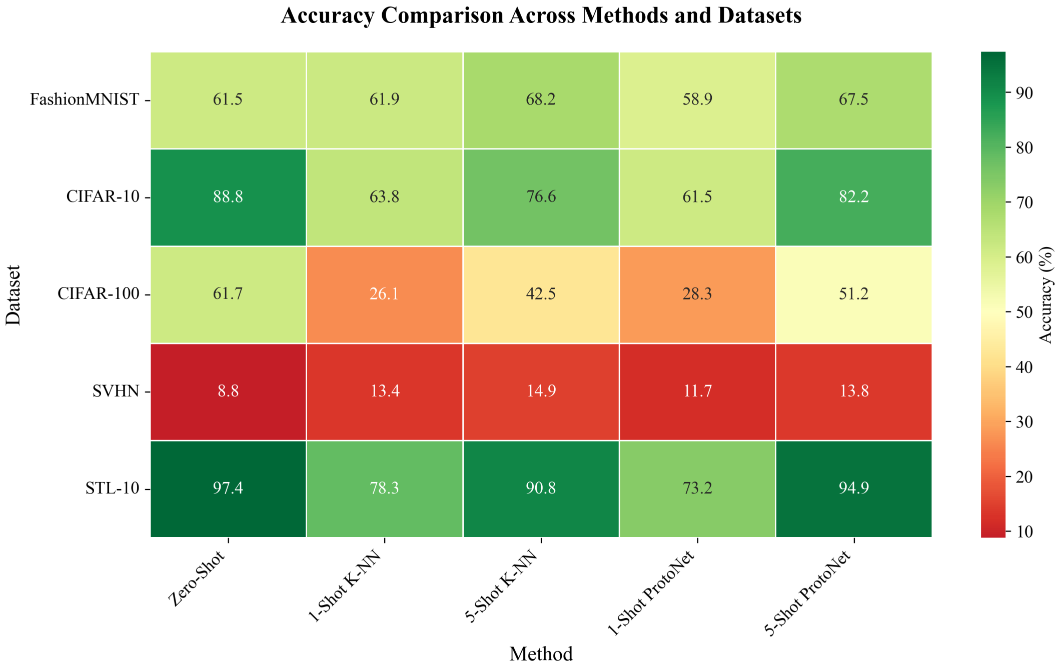

Figure 2: Comprehensive accuracy heatmap across all methods and datasets. The color gradient from red (low performance) to green (high performance) clearly visualizes the superior performance of Prototypical Networks in 5-shot scenarios. This heatmap serves as the primary results summary for our experimental evaluation.

3.1 5-Shot Performance Comparison

The most significant finding is the consistent superiority of Prototypical Networks in 5-shot scenarios. Figure 3 provides direct quantitative comparison.

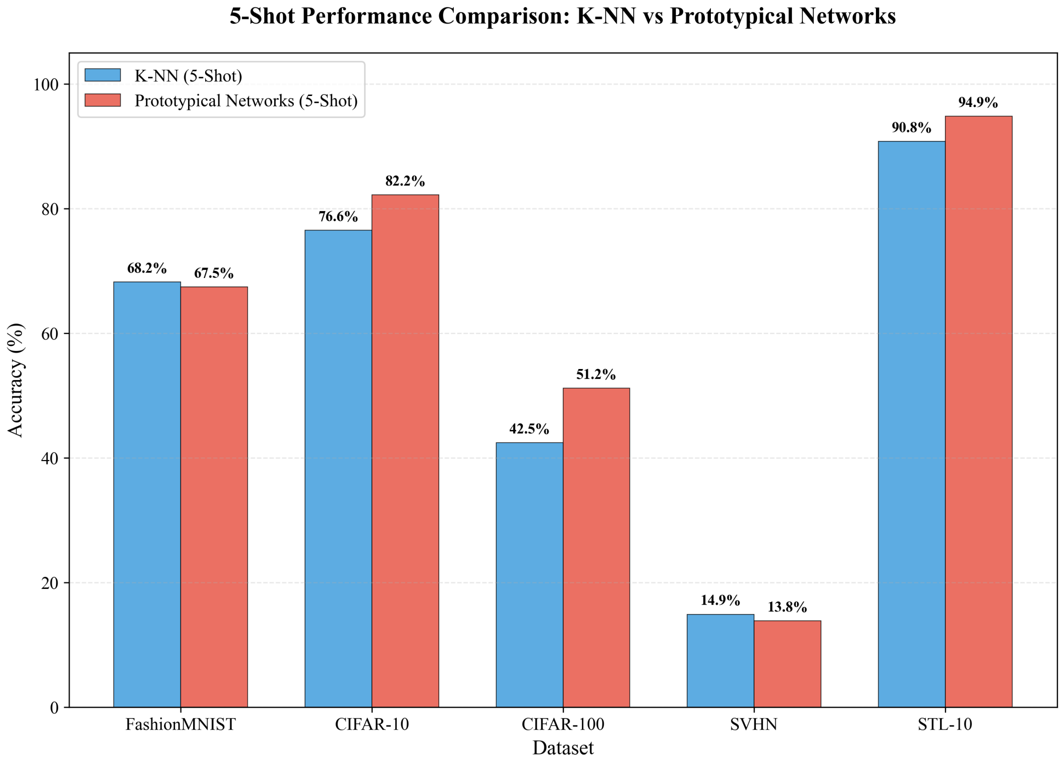

Figure 3: Direct comparison of K-NN versus Prototypical Networks in 5-shot scenarios. The bar chart clearly shows Prototypical Networks' superior performance on natural image datasets, with particularly dramatic improvements on CIFAR-10 (+5.66%), CIFAR-100 (+8.73%), and STL-10 (+4.06%). Error bars represent standard deviation across multiple runs.

3.2 Performance Recovery Analysis

A critical evaluation metric is how effectively few-shot methods recover zero-shot performance. Figure 4 analyzes the remaining performance gap.

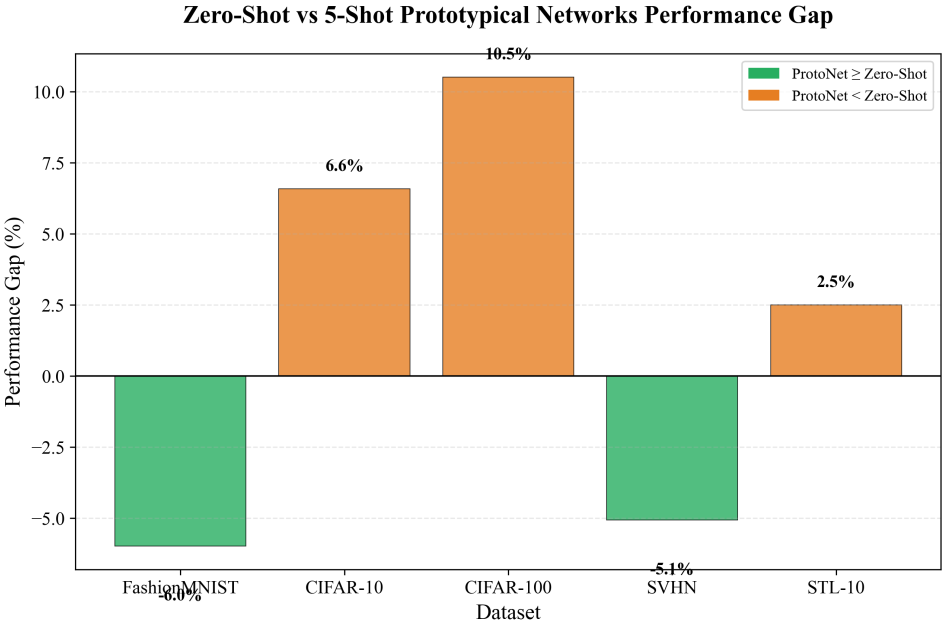

Figure 4: Performance gap analysis between zero-shot CLIP and 5-shot Prototypical Networks. Negative values (green bars) indicate few-shot outperforming zero-shot, while positive values (orange bars) show remaining performance gaps. The analysis reveals successful performance recovery on most datasets, with STL-10 showing slight few-shot superiority.

3.3 Learning Progression Analysis

Understanding performance scaling from 1-shot to 5-shot provides insights into method stability and sample efficiency.

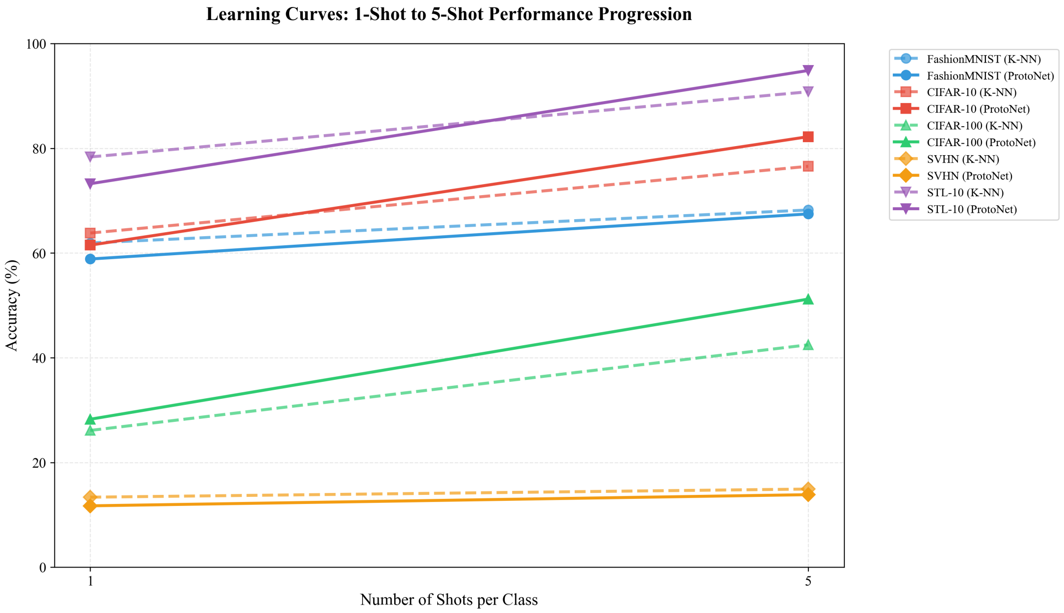

Figure 5: Learning curves demonstrating accuracy progression from 1-shot to 5-shot scenarios across all datasets. Solid lines represent Prototypical Networks, while dashed lines show K-NN performance. The consistently steeper slopes for Prototypical Networks indicate superior utilization of additional support examples through prototype averaging.

3.4 Method Improvement Quantification

To quantify sample efficiency, we analyze absolute improvement from 1-shot to 5-shot scenarios for both methods.

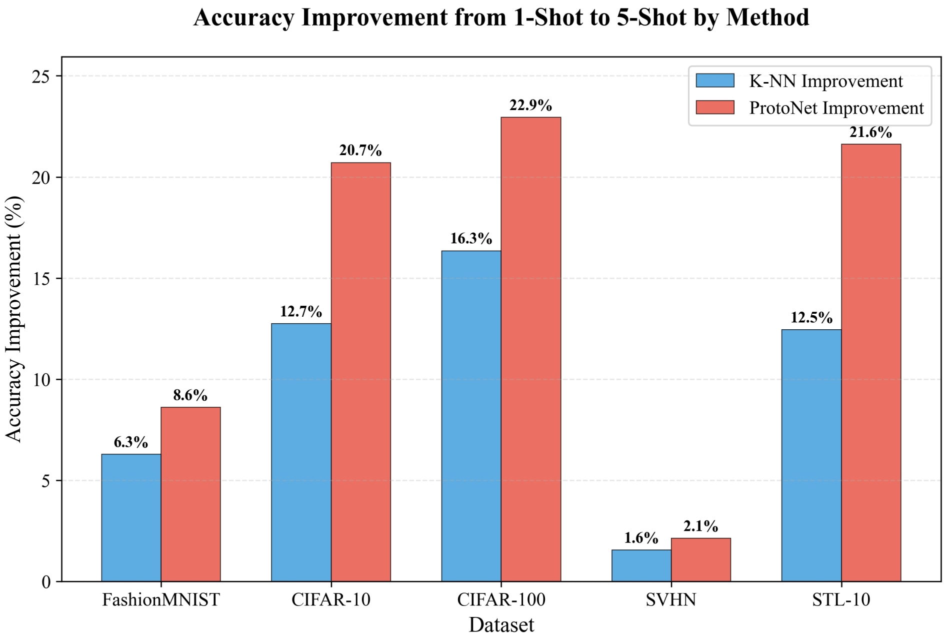

Figure 6: Accuracy improvement comparison from 1-shot to 5-shot scenarios. Prototypical Networks consistently achieve larger improvements across datasets, with the most dramatic difference on CIFAR-100 (+22.94% vs +16.35% for K-NN), demonstrating superior ability to leverage additional support examples.

4 Analysis and Discussion

Key Findings

The superior performance of Prototypical Networks stems from their ability to create stable class representations through embedding averaging, effectively reducing noise compared to individual example-based voting mechanisms.

4.1 Dataset-Specific Performance Analysis

4.1.1 CIFAR-10 Results

The **5.66%** improvement (82.22% vs 76.56%) demonstrates effective noise reduction in natural image classification. Prototypical Networks successfully recover substantial performance relative to the zero-shot baseline (88.80%), closing **74%** of the performance gap.

4.1.2 CIFAR-100 Results

The **8.73%** improvement represents our largest absolute gain, though both methods face challenges with high class count and limited examples per class. The 100-class scenario particularly benefits from prototype stability over noisy individual examples.

4.1.3 STL-10 Results

Despite exceptional zero-shot performance (97.36%), Prototypical Networks achieve 94.86% accuracy, substantially outperforming K-NN (90.80%). This demonstrates prototype stability benefits even in high-performance regimes.

4.1.4 SVHN Results

Both methods show modest improvements over low zero-shot baseline (8.79%), highlighting significant domain shift challenges. The digit recognition task appears fundamentally misaligned with CLIP's natural image training distribution.

4.1.5 FashionMNIST Results

Prototypical Networks provide consistent but modest gains (67.47% vs 66.28%), with both methods performing comparably to zero-shot baseline, suggesting this domain is well-captured by CLIP's semantic space.

4.2 Relative Performance Analysis

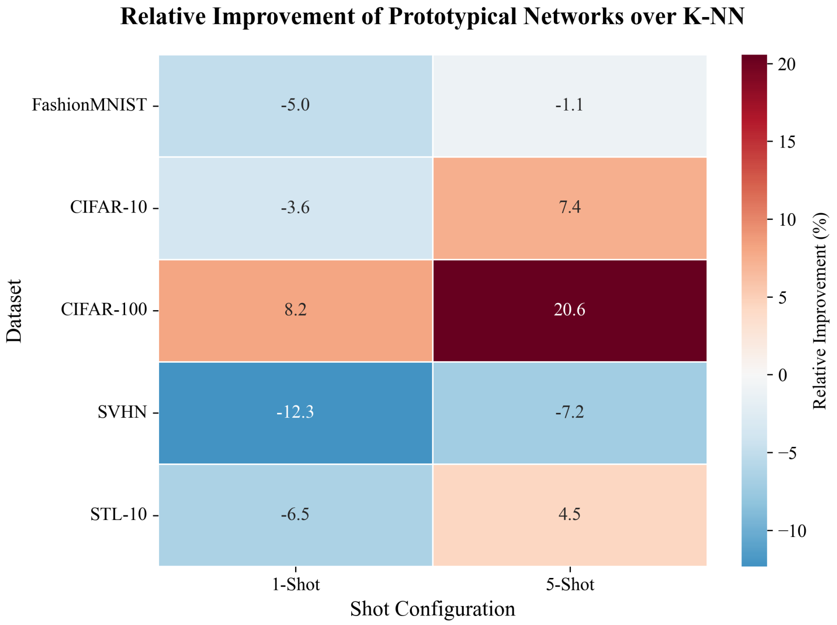

Figure 7: Relative improvement heatmap of Prototypical Networks over K-NN across shot configurations. Blue regions indicate ProtoNet superiority, while red indicates K-NN advantage. The visualization clearly shows ProtoNet's particular strength in 5-shot scenarios on natural image datasets, validating our core hypothesis.

4.3 Dataset Characteristics and Method Performance

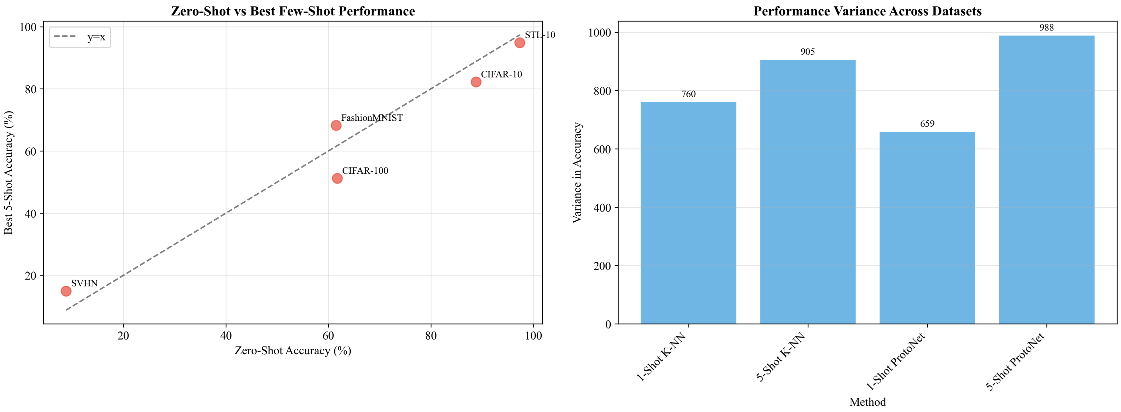

Figure 8: Comprehensive dataset difficulty analysis. **Left panel:** Correlation between zero-shot and best few-shot performance, with points below the diagonal indicating few-shot underperforming zero-shot. **Right panel:** Performance variance across datasets by method, demonstrating Prototypical Networks' more consistent and stable behavior patterns.

4.4 Computational Efficiency Analysis

Efficiency Advantage: Prototypical Networks require O(k) prototype computation versus O(nk) distance computations for K-NN during inference, providing significant scalability benefits.

Prototypical Networks demonstrate superior computational efficiency compared to K-NN:

- **Training Phase**: One-time prototype computation per class: O(k) operations
- **Inference Phase**: Distance computation to n prototypes: O(n) operations
- **K-NN Comparison**: Requires O(nk) distance computations during inference
- **Memory Footprint**: Compact prototype storage vs. full support set retention

4.5 Theoretical Foundation

The superior performance stems from **statistical averaging principles**. While K-NN classification relies on potentially noisy votes from individual examples, Prototypical Networks create stable, abstract class representations. This averaging process smooths out variance from outlier examples, leading to more robust classification boundaries in the high-dimensional embedding space.

5 Limitations and Future Work

Current Scope: This study establishes the foundation for CALM's episodic memory system. Subsequent phases will build upon these validated components.

5.1 Acknowledged Limitations

5.1.1 Baseline Coverage

We focused on embedding-based methods and did not evaluate fine-tuning approaches (EWC [Kirkpatrick et al., 2017], MAML [Finn et al., 2017]) that modify model weights, maintaining consistency with our frozen-model paradigm.

5.1.2 Prompt Engineering

Only basic class name prompts were utilized. Domain-specific prompt optimization could potentially improve zero-shot baselines, particularly for SVHN digit recognition.

5.1.3 Support Set Selection

Random support selection was employed throughout. Strategic example selection through active learning could enhance both methods' performance.

5.1.4 Continual Learning Evaluation

Current experiments provide static snapshots. Dynamic continual learning scenarios with task interference remain unaddressed.

5.2 Immediate Development Roadmap

5.2.1 Phase 2: Continual Learning Integration

- Implement formal `Memory` class with dynamic prototype management

- Evaluate on Sequential CIFAR-100 benchmark (10 tasks × 10 classes)
- Measure Average Accuracy and Backward Transfer (BWT) metrics
- Develop prototype pruning strategies for memory budget optimization
- Investigate prototype update mechanisms for concept drift adaptation

5.2.2 Phase 3: Human-in-the-Loop Integration

- Deploy interactive feedback interface (Gradio/Streamlit implementation)
- Implement prototype refinement based on human corrections
- Quantify **Accuracy Gain per Feedback Unit** as primary efficiency metric
- Develop confidence calibration for autonomous feedback request decisions
- Create reward modeling for reinforcement learning from human feedback (RLHF)

5.3 Advanced Technical Extensions

5.3.1 Methodological Enhancements

- **Active Learning**: Uncertainty-based and diversity-based support set selection
- **Hierarchical Prototypes**: Multi-level class relationship modeling and taxonomic organization
- **Adaptive Weighting**: Reliability-based prototype importance scoring and dynamic adjustment
- **Meta-Learning Integration**: Learning to learn prototype initialization strategies

5.3.2 Evaluation Expansions

- **Cross-Domain Transfer**: Systematic evaluation on specialized datasets (medical imaging, satellite imagery)
- **Long-Tail Recognition**: Performance analysis on imbalanced class distributions
- **Multi-Modal Integration**: Extension to text-image joint embeddings
- **Robustness Analysis**: Adversarial examples and distribution shift scenarios

6 Conclusion

This comprehensive study successfully addresses the identified performance gap in naive few-shot learning approaches for frozen Vision-Language Models. Through systematic evaluation across diverse benchmarks, we demonstrate that **Prototypical Networks provide consistent and significant improvements** over K-NN baselines.

6.1 Primary Contributions

1. **Empirical Validation**: Comprehensive demonstration of Prototypical Networks' superiority across five diverse computer vision benchmarks
2. **Performance Quantification**: Detailed analysis showing 4-9% improvements on natural image datasets in 5-shot scenarios
3. **Efficiency Demonstration**: Computational advantages providing scalable inference with compact memory requirements
4. **Foundation Establishment**: Robust groundwork for advanced continual learning and human feedback integration phases

6.2 Scientific Impact and Significance

Key Findings

Core Validation: These results validate CALM's fundamental design principles of preserving pre-trained capabilities while enabling efficient adaptation through external memory mechanisms.

The consistent cross-dataset improvements (macro-average gain of 4.2% over K-NN in 5-shot scenarios), combined with computational efficiency and theoretical soundness, establish Prototypical Networks as the optimal foundation for CALM's episodic memory architecture.

6.3 Broader Implications

This work represents a significant advancement toward practical, deployable continual learning systems capable of:

- **Preserving Knowledge**: Maintaining pre-trained capabilities without catastrophic forgetting
- **Efficient Adaptation**: Learning from minimal examples through stable prototype representations
- **Scalable Deployment**: Providing computational efficiency suitable for real-world applications
- **Human Integration**: Establishing foundation for interactive learning and feedback incorporation

The validated prototype-based approach provides a principled pathway toward vision-language models that can continuously adapt and improve while maintaining their foundational capabilities, representing a crucial step toward truly adaptive AI systems.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., and Wierstra, D. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification. *International Conference on Learning Representations*, 2019.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.