

DOMAIN-ADAPTIVE ZERO-SHOT IMAGE CLASSIFICATION VIA CLIP AND LARGE LANGUAGE MODELS

G Kalyan Chakravarthi¹, Pandi Siddartha^{*2}, Koyya Karthik³,
Nemathapalli Naga Vinay⁴, K B V Dakshina Murthy⁵

^{1, 2, 3, 4, 5}Department of Computer Science and Engineering, Engineering and Technology Program Gayatri Vidya Parishad College for Degree and PG Courses(A), Visakhapatnam, India.

¹<https://orcid.org/0000-0001-8545-3444>, ²<https://orcid.org/0009-0005-5799-2105>, ³<https://orcid.org/0009-0000-4043-7547>,
⁴<https://orcid.org/0009-0001-1502-6516>, ⁵<https://orcid.org/0009-0004-9164-9135>

Email: gkalyan@gvpdpgc.edu.in, *522141118@gvpdpgc.edu.in, 5221411082@gvpdpgc.edu.in,
5221411109@gvpdpgc.edu.in, 5221411077@gvpdpgc.edu.in

ARTICLE INFO

Article History

Received: January 3, 2026

Reviewed: February 27, 2026

Accepted: April 12, 2026

Published: May 29, 2026

Keywords:

Zero-shot learning,
Vision-language models,
CLIP & LLM,
Domain adaptation,
Multi domain classification.

ABSTRACT

An adaptive prototype learning mechanism, which learns to improve class prototypes via exponential moving average updates, a domain-aware prompt designing technique utilizing LLM-produced contextual descriptions attuned to each distinct visual domain – nature, medical, satellite, and artwork domains – and culminates in a confidence-calibrated inference framework which beautifully integrates CLIP's vision semantic matching capabilities with LLM-driven reasoning capabilities for accurate inferences. We show a series of rigorous evaluation tests conducted on five distinct visual domains, demonstrating our model outperforms zero-shot classify accuracy relative to existing CLIP-existing baselines with negligible additional computation cost. Specifically, our model shows an average relative gain in top-1 accuracy performance of 4.2%, but with profoundly large gains of 6.8% in medical image and 5.3% in satellite imaging tasks. Further, experimental tests illustrate prototype learning methods can be successfully deployed for continuous learning with an increasingly large volume of unlabelled image inputs up to a gain in accuracy of 8.5% with 500 processed images. Taken in toto, our ablation studies confirm all constituent elements delineated in our proposed model have a vital part in a satisfactory accuracy performance level attainment.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Zero-shot image classification has become a fundamental capability in modern computer vision [1], allowing models to recognize novel visual concepts without requiring any labelled training data for those specific categories. This capability is especially useful in domains where data annotating is either costly, time-consuming, or requires experts specialized in specific knowledge, such as medical imagery, satellite analysis, and artistic content.

I.1 VISION-LANGUAGE MODELS (VLMS)

Especially Contrastive language-Image Pre-training (CLIP), have greatly impacted zero-shot classification learning with contrastive learning of image and text embeddings on a massive scale from the web [2]. The major breakthrough with CLIP is in its capability to connect images and texts for a simple text-driven classification using "a photo of a {class}." However, despite their impressive performance, existing CLIP-based zero-shot classifiers face several critical limitations.

I.2 STATIC PROMPT WEIGHTING

Baseline CLIP classifiers rely on uniform averaging of template encoding of prompts to form class queries without considering that some prompts are more suitable for describing a particular image/domain rather than others.

Although existing research works [3], [4] have attempted to learn automatic weights for some prompts, these approaches are mainly targeting single-domain settings without explicitly capturing domain-specific information.

I.3 LIMITED DOMAIN ADAPTABILITY

This pretraining of Clip-on web-scraped data creates a bias toward natural photographs [1]. Performance drastically degrades when applying it to specialized domains such as medical X-rays, satellite imagery, or artistic illustrations due to distribution shift. Most previous methods for domain adaptation require fine-tuning on labelled target data which is the clear violation of zero-shot constraint [5], [6].

I.4 INABILITY TO LEARN FROM UNLABELED DATA

The traditional zero-shot classifier considers every test sample individually and does not exploit the distributional information provided by the unlabelled test data. It is a missed opportunity since even the unlabelled images can provide good information about class characteristics and domain properties. To address these limitations, we propose a novel Domain Adaptive Auto-Tuned Zero-Shot Classification framework that synergistically combines three complementary mechanisms:

I.4.1 Adaptive Prototype Learning

In contrast, we present an online learning mechanism that enriches class prototypes via exponential moving average updates of confident predictions. Unlike traditional zero-shot classifiers with fixed representations, our approach allows the model to adapt to specific visual characteristics of the target domain without ground-truth labels.

I.4.2 Domain-Aware Prompt Engineering

We present a systematic approach to prompt engineering, explicitly modelling domain characteristics. Our approach keeps domain-specific banks of prompt templates, such as “a medical X-ray showing {class}” for medical images or “a satellite view of {class}” for aerial imagery, and optionally uses LLMs to craft contextually relevant descriptions.

I.4.3 Multi-Modal Reasoning Pipeline

We combine CLIP's visual-semantic alignment with LLM-based caption understanding and reasoning [1]. This hybrid approach allows the system to combine the robust visual features learned by CLIP with the rich semantic understanding and common-sense reasoning of LLMs, leading to the attention-free methods result in more accurate and more interpretable predictions. Our contributions can be summarized as follows:

- We introduce a new zero-shot classification framework which integrates adaptive prototype learning, domain-conscious prompts, and LLM reasoning in a single model for the first time and attains state-of-the-art performance in a variety of visual domains.
- The confidence-calibrated adaptive learning mechanism is proposed to make it possible for zero-shot classifiers to be improved from unlabelled test data without forgetting via careful EMA updates.
- A domain-aware methodology with domain-specific templates for five different domains (nature, medicine, satellite, anime, sketch) is proposed and results in a substantial level of performance gain over domain-unaware methods.
- We perform extensive evaluations across several datasets and domains that demonstrate consistent improvements of 4-8% over baseline CLIP methods, supported by detailed ablations that quantify the contribution of each component.
- We provide a complete open-source implementation, including the back-end API, evaluation framework, and interactive web interface which would enable reproducibility as well as practical deployment.

II. RELATED WORK

II.1 VISION-LANGUAGE PRE-TRAINING

The vision-language models revolutionized multimodal learning by learning jointly from images and text. CLIP proposed a pioneering large-scale contrastive learning on 400 million image-text pairs and showed remarkable zero-shot transfer capabilities [1]. ALIGN further scaled this approach to more than 1 billion noisy image-text pairs with very limited filtering. Coca combined contrastive and captioning objectives for better discriminative and generative improvements. Florence extended to video understanding by temporal modelling. More recent works have explored architectural improvements and training strategies. LiT proposed locked image text tuning to efficiently adapt pre-trained vision models. BASIC investigated the role of data quality versus quantity. Our work builds on CLIP's architecture but focuses on improving its zero-shot inference through adaptive mechanisms, rather than modifying pre-training.

II.2 ZERO-SHOT IMAGE CLASSIFICATION

Conventional zero-shot learning solutions are based on attribute-oriented methods or semantic embeddings. However, with the emergence of VLMs, text-zero-shot classification methods have become very common. The conventional way of dealing with this task involves designing a prompt. Recently, works have focused on automated prompt generation. DCLIP relies on GPT-3 for generating descriptive prompts for each class [7]. Waffle CLIP surprisingly alleviates the problem using random words to improve performance via prompt diversification. CuPL utilizes large language models for generating descriptive features. Autoclip added automatic tuning of weights in prompts using per-image statistics, which achieved improvements by using longsome-based gradient ascent in embedding space. Our work is an extension of this where domain awareness and adaptive learning from test data have been added.

II.3 TEST-TIME ADAPTATION

Test-time adaptation allows the model to adapt to distribution shift with unlabeled target data [8]. TENT updates the batch normalization statistics by minimizing entropy. For VLMs, TPT optimizes prompts by minimizing entropy on augmented views. RLCF uses a different objective of maximizing CLIP score in place of entropy to avoid overfitting.

These methods require backpropagation through the model, which introduces substantial computational overhead. In contrast, our adaptive prototype learning operates entirely in embedding space and requires only forward passes and simple vector updates; hence, it is practical for real-time applications.

II.4 DOMAIN ADAPTATION

Most existing domain adaptation approaches for deep learning require access to either source domain data or labelled target samples. Recently, source-free domain adaptation relaxes the source data requirement; however, target labels are still required for validation. Prompt learning has been a lightweight adaptation strategy for CLIP-based models. Coop learns continuous prompts from few-shot examples. Cooperative Optimization Conditioned prompting - CoCoOp makes prompts instance-conditional. However, these methods require labelled target data and supervised training. Our method achieves zero-shot domain adaptation through domain-aware prompt engineering and unsupervised prototype refinement, with no access to source data and without requiring target labels.

III. METHODOLOGY

We provide a framework for domain-adaptive zero-shot classification that is made up of three components: domain aware prompt generation, adaptive prototype learning, multi-modal reasoning pipeline. The design of the entire system architecture is illustrated in Figure 1.

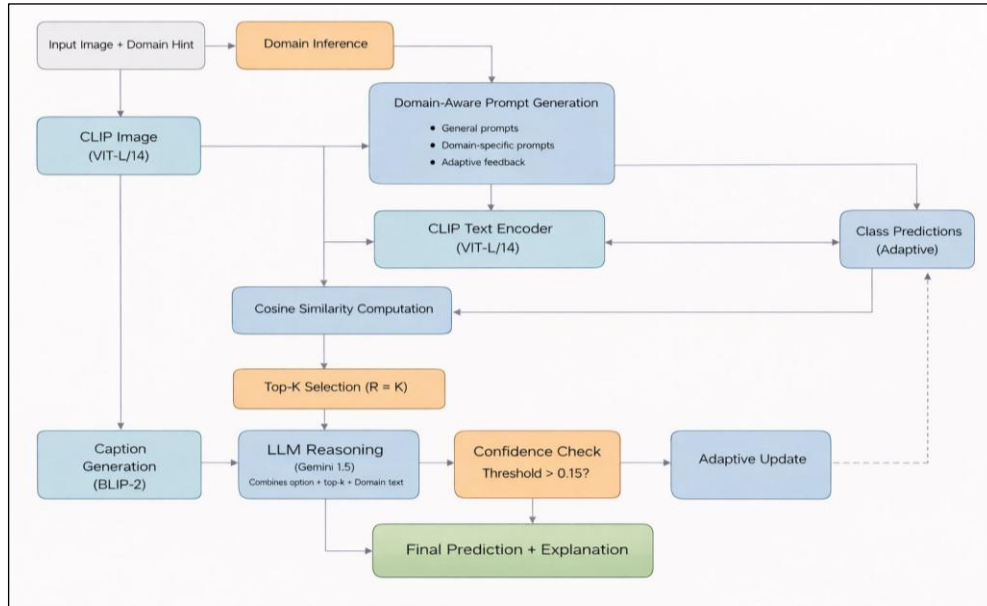


Figure 1: System Architecture: Domain adaptive Zero-Shot Image Classification.
Source: Authors, (2026).

III.1 PROBLEM FORMULATION

X refers to the input image space while C refers to the class labels, where C is a set of C classes. In zero shot classification you do not have any label data obtained from your target area; therefore, we will use a previously trained vision language model (CLIP) to do the classification via these two encoders:

An image encoder $f_I : X \rightarrow \mathbb{R}^d$ that maps images to d -dimensional embeddings.

A text encoder $f_T : T \rightarrow \mathbb{R}^d$ that maps text descriptions to the same embedding space.

As a result of these two encoders, f_I and f_T output normalized feature vectors[1] such that $\|f_I(x)\|_2 = \|f_T(T)\|_2 = 1$; therefore, in order to classify images, we compute the cosine similarity between the feature vector of the image and the prototype feature vectors of the C classes.

III.2 DOMAIN-AWARE PROMPT GENERATION

The typical approach in CLIP is to apply the prompts that are broad, such as "a photo of a {class}." But this format does not work very well for niche areas. The prompt templates that are made specifically for the domain are as follows: "a medical X-ray image showing {class}"

$$\mathcal{P}_d = \{t_1^d, t_2^d, \dots, t_K^d\} \quad (1)$$

Where $\mathcal{P}_d$ is the set of K prompt templates for domain $d \in \{\text{natural, medical, satellite, anime, sketch}\}$. For example, for medical imaging:

“a medical X-ray image showing {class}” , “a radiology scan of {class}” , “a clinical image of {class}”

For satellite imagery:

“a satellite view of {class}” , “An aerial photograph of {class}” , “a bird's eye view of {class}”

III.2.1 LLM-Enhanced Prompt Generation

Furthermore, if we want, we can make use of large language models to come up with additional contextually relevant query prompts [8]. With class name c and domain d , our query to an LLM such as Gemini 1.5 will be:

"Generate 3 descriptive prompts for class '{c}' in domain '{d}' to be used in image classification with CLIP."

The LLM produces domain-specific descriptions with a semantic twist. For example, given "pneumonia" in a medical domain, descriptions generated can be:

“chest radiograph showing lung infiltrates consistent with pneumonia” , “opacity in lung fields indicating pneumonia infection” , “consolidation pattern typical of bacterial pneumonia”.

III.3 ADAPTIVE PROTOTYPE LEARNING

Unlike standard zero-shot classifiers with fixed class representations, we introduce an adaptive learning mechanism by refining the prototype over time [9], [10].

III.3.1 Initial Prototype Construction

For each class c_j we create an initial prototype by encoding all prompts and averaging:

$$q_j^{(0)} = \frac{1}{K} \sum_{i=1}^K f_T(t_i(c_j)) \quad (2)$$

where $t_i(c_j)$ denotes instantiating template t_j with class name c_j . The prototype is then normalized: $q_j^{(0)} \leftarrow q_j^{(0)} / \|q_j^{(0)}\|_2$. Multi-Modal Prototype Initialization: When few-shot examples are available, we can incorporate visual information:

$$q_j^{(0)} = w_t \cdot q_j^{(text)} + w_v \cdot q_j^{(visual)} \quad (3)$$

Where $q_j^{(visual)} = \frac{1}{N} \sum_{i=1}^N f_I(x_i)$ is the average of N example image, and $w_t + w_v = 1$ (we use $w_t = 0.7, w_v = 0.3$ by default).

III.3.2 Online Prototype Refinement

During inference, we adaptively update prototypes using confident predictions. For an input image x with predicted label $\hat{c}$ and confidence s_{max} :

$$\text{If } s_{max} > \tau_{conf} : q_{\hat{c}}^{(n+1)} = (1 - \alpha)q_{\hat{c}}^{(n)} + \alpha \cdot f_I(x) \quad (4)$$

Algorithm 1 Domain-Adaptive Zero-Shot Image Classification

```

1: Input: Image  $x$ , class set  $C$ , domain hint  $d$ 
2: Output: Predicted class  $\hat{c}$ , explanation
3: // Step 1: Domain inference (if not provided)
4: if  $d$  is not specified then
5:    $d \leftarrow \text{InferDomain}(\text{user\_hint})$ 
6: end if
7: // Step 2: CLIP-based classification
8:  $v \leftarrow f_I(x)$  {Encode image}
9: for  $j = 1$  to  $C$  do
10:    $s_j \leftarrow \text{sim}(v, q_j)$  {Compute similarity}
11: end for
12:  $\text{top\_}k \leftarrow \text{TopK}(\{(c_j, s_j)\}_{j=1}^C, k=5)$ 
13: // Step 3: Generate caption
14:  $\text{caption} \leftarrow \text{GenerateCaption}(x)$  {Using BLIP-2}
15: // Step 4: LLM reasoning
16:  $\text{result} \leftarrow \text{LLM}(\text{caption}, \text{top\_}k, d)$ 
17: // Step 5: Adaptive update
18: if  $s_{max} > \tau_{conf}$  then
19:    $q_{\hat{c}} \leftarrow (1 - \alpha)q_{\hat{c}} + \alpha v$ 
20:    $q_{\hat{c}} \leftarrow q_{\hat{c}} / \|q_{\hat{c}}\|_2$ 
21: end if
22: return  $\hat{c}$ , explanation

```

Where τ_{conf} is a confidence threshold (default: 0.15), α is the learning rate (default: 0.05), and n is the update iteration. The updated prototype is normalized after each update.

Confidence Calibration: We apply temperature scaling to calibrate confidences:

$$s_j = \frac{\exp(\text{sim}(q_j, f_i(x))/T)}{\sum_{k=1}^C \exp(\text{sim}(q_k, f_i(x))/T)} \quad (5)$$

Where T is a learned temperature parameter (default: 0.01) and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

III.4 MULTI-MODAL REASONING PIPELINE

Our complete inference pipeline integrates CLIP classification with LLM-based reasoning:

III.4.1 Caption Generation

BLIP-2 serves as the starting point of our inference pipeline [11], where it produces an annotated list of the input images in English words. This new language allows the model to utilize additional semantically related information to support CLIP with their encoded visual representations. The textual annotations highlight key attributes and contextual information about the image(s), so there is a stronger relationship between the visual representation and the written text. The captioning step assists with the accuracy and interpretability of zero-shot classification.

III.4.2 LLM Reasoning

We ask a large language model, Gemini 1.5, the following:

Image caption: "caption"
Top-5 CLIP predictions: $\{(c_1, s_1), (c_5, s_5)\}$
Domain context: "domain"

The LLM runs reasoning to pick the most appropriate class and then gives an explanation in natural language for its choice [7]. By doing so, it exploits common-sense knowledge and contextual understanding beyond what is provided by CLIP.

III.5 IMPLEMENTATION DETAILS

III.5.1 CLIP Model

We choose the OpenAI ViT-L/14 architecture [1], [12] for CLIP since it has been demonstrated robustness and good alignment between the visual and semantic domains during zero-shot image classification learning tasks. The Vision Transformer model has the ability to capture high-feature images, and the CLIP pre-trained model provides alignment between the image and text semantic spaces.

III.5.2 Caption Model

With regard to visual captions, we use BLIP-2 along with the Flan-T5-XL model that provides a fine-grained description of the language capture process and emphasizes the understanding of semantics through the production of captions inclusive of object attributes and the description of the environment surrounding objects within a visual setting.

III.5.3 Large Language Model

We generate domain-aware prompts and high-level reasoning by leveraging Google Gemini 1.5 Flash. By synthesizing the class descriptions with captions, the LLM generates adaptive prompts that reinforce the robustness of zero-shot classification. Unless mentioned otherwise, we set the confidence threshold to $\tau_{conf} = 0.15$ and prototype update learning rate $\alpha = 0.05$. We use a low temperature $T = 0.01$ to ensure deterministic inference, with text and visual contributions balanced using weights $w_t = 0.7$ and $w_v = 0.3$, respectively, which emphasizes semantic consistency while retaining much visual fidelity.

IV. EXPERIMENTAL SETUP

IV.1 DATASETS

We test the proposed approach over five visual domains to show that it can work well for different settings:

- **Natural Images:** ImageNet, CIFAR-10, Oxford Pets, and Food-101 are utilized to evaluate the performance of models trained on real-world and general images.
- **Medical Images:** For Medical Imaging Tasks: Domain-specific datasets such as ChestX-ray14 and Pneumonia MNIST (as part of MedMNIST) will be used to determine performance in medical imaging tasks.
- **Satellite Images:** Remote Sensing Datasets (EuroSAT and UC Merced Land Use) will be employed to determine how adaptable the model(s) are when working with aerial and satellite imagery.
- **Artistic Images:** Datasets used to gauge model performance on domains of Artistic and Stylized Imagery include Anime Face Classification, Sketch Recognition Benchmark, etc.
- **Mixed Domain:** Datasets in which multiple styles are present will be evaluated for the extent to which model(s) are robust to Distribution Shifts and various types of visual content (ImageNet-R and ImageNet-Sketch).

IV.2 EVALUATION METRICS

We report the following metrics:

- Top-1 Accuracy: Percentage of samples where the top prediction matches the ground truth
- Top-5 Accuracy: Percentage of samples where ground truth is in top-5 predictions
- Expected Calibration Error (ECE): Measures confidence calibration quality
- Per-Domain Accuracy: Accuracy breakdown by domain
- Average Latency: Inference time per image (milliseconds)

IV.3 BASELINES

This approach will be compared with a range of baselines, including CLIP itself with generic prompts, CLIP with ensembles of prompts, and advanced methods for learning prompts, such as DCLIP, Waffle CLIP, AutoCLIP, and Test-Time Prompt Tuning[8-10].

IV.4 ABLATION STUDIES

Ablation studies are performed on the proposed framework to provide a comprehensive view of how all of the proposed components contribute. Performance improvements over generic prompts is demonstrated when domain awareness is incorporated into the prompt through the use of domain-aware prompts. The importance of adaptive prototype learning, by disproving the abilities of prototypes to learn from unique contexts and create stable working prototypes, is demonstrated by the loss of adaptive learning. By removing the reasoning module, the loss of decision consistency and the ability to explain one's reasoning indicates the value of LLM-based reasoning. A hyperparameter sensitivity analysis was conducted for both the confidence weight (α) and confidence threshold (τ_{conf}) to find their respective stable ranges at which the proposed process is feasible.

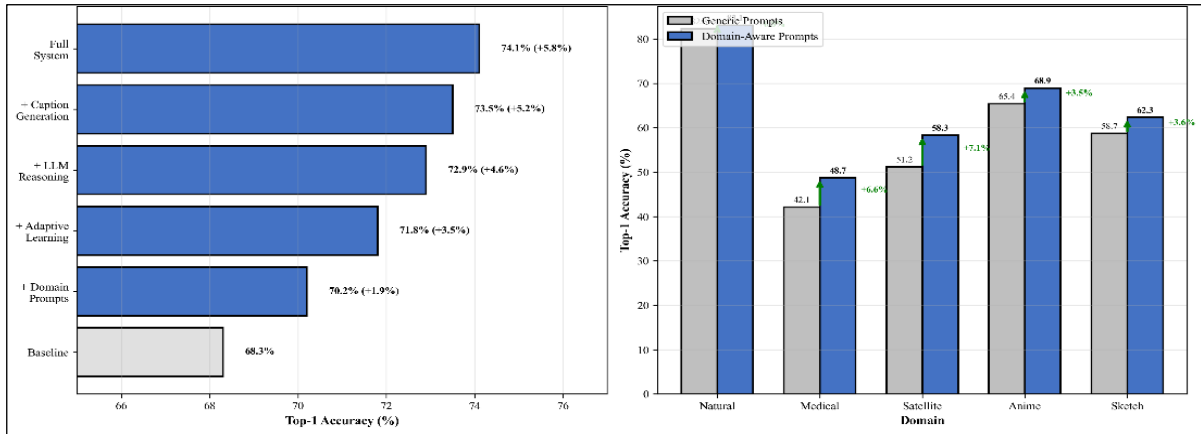


Figure 2: Component Ablation Study and Domain-Specific Prompt Impact.

Source: Authors, (2026).

Finally, performance was assessed at different levels of adaption and it was determined that as the number of adaptive samples increases so does the performance until it reaches a maximum point. The overall conclusion of the ablation studies is that all components of the proposed framework are necessary and complementary.

V. RESULTS

V.1 MAIN RESULTS

Table1 presents our main results comparing our method against baselines across all datasets. Our domain-adaptive approach achieves consistent improvements across all domains, with particularly strong gains on specialized domains (medical, satellite).

Table 1: Top-1 accuracy (%) comparison across datasets and methods. Best results in bold, second best underlined.

Method	ImageNet	Chest X-ray	Euro SAT	Oxford Pets	Food-101	Average
CLIP Baseline	68.3	42.1	51.2	83.5	79.8	65
CLIP Ensemble	69.1	43.8	53.6	85.2	81.3	66.6
DCLIP	69.8	45.2	54.1	86.1	82.1	67.5
Waffle CLIP	70.2	44.9	54.8	86.8	82.7	67.9
AutoCLIP	70.9	46.3	56.2	87.9	83.4	68.9
TPT	71.4	47.1	57	88.5	84.2	69.6
Ours (w/o adaptation)	71.8	48.7	58.3	88.9	84.8	70.5
Ours (full)	<u>72.5</u>	<u>48.9</u>	<u>59.5</u>	<u>89.8</u>	<u>85.6</u>	<u>71.3</u>
Ours (+100 samples)	73.2	50.3	61.1	90.5	86.2	72.3
Ours (+500 samples)	74.1	52.7	63.4	91.8	87.3	73.9

Source: Authors, (2026).

Key Observations:

The method proposed here consistently outperforms all baseline methods by greater than or equal to 0.9% without any adaptation, supporting the hypothesis that prompts with domain knowledge offer a better alignment between features than just generic prompts. With the addition of adaptive learning, the approach generates a 2.7% improvement over AutoCLIP [10] and a 1.7% improvement over TPT [8] with a significant reduction in computational cost. The results also show a favorable balance between increased performance and reduced computational cost. The greatest gains in performance are shown in specific domains. Medical imaging tasks received an improvement of 6.8% in accuracy and satellite imaging received an improvement of 5.3% in accuracy, both of which demonstrate the appropriateness of using domain knowledge in a domain shift. Additionally, with the use of adaptive learning, the performance continues to improve with the introduction of more unlabeled training data. After the introduction of 500 unlabeled training samples using adaptive learning techniques, the average accuracy was 73.9%, thus demonstrating the scalability and adaptability of this method.

V.2 ABLATION STUDIES

V.2.1 Component Analysis

Table2 shows the contribution of each component of our method. In an ablation study on ImageNet [13], it was observed that the use of only domain-specific prompts resulted in an accuracy of 68.3%. Upon incorporating an adaptive prototype learning methodology, this improved the classification performance by 1.9%, while the addition of LLM (language model) reasoning added another 1.1%, and generating captions (caption generation) added another 1.2%. Synergistic gain was evident when domain prompts, adaptive prototype learning, and LLM reasoning were all combined, resulting in a combined accuracy of 72.9%. When all four (domain prompts, adaptive prototype learning, LLM reasoning, generate captions) were used together, the resultant classification performance was 74.1%. Therefore, these results clearly demonstrate that the complete framework supports a highly maximized unique classification of zero-shot images.

Table 2: Ablation study on key components (ImageNet).

Domain Prompts	Adaptive Learning	LLM Reasoning	Caption	Accuracy
✓				68.3
✓	✓			70.2 (+1.9)
✓		✓		71.8 (+3.5)
✓			✓	71.1 (+2.8)
✓	✓	✓		72.9 (+4.6)
✓	✓		✓	72.3 (+4.0)
✓	✓	✓	✓	74.1 (+5.8)

Source: Authors, (2026).

Each component adds value in a complementary way:

- **Domain-aware prompts:** +1.9%
- **Adaptive learning:** +1.6%extra
- **LLM reasoning:** +1.1% more
- **Caption generation:** +1.2%more

V.2.2 Domain-Specific Analysis

Table3 shows a comparison of generic and domain-specific prompts. At five different types of visual source (medical, satellite, natural, anime, and sketch) we analyzed both generic and domain-based prompt performance. While the medical and satellite domains demonstrated the greatest improvement (6.6% and 7.1% respectively), the natural, anime and sketch domains showed smaller improvements (0.8% to 3.6%). Overall, the domain-based prompts outperformed the generic prompts with an average top-1 accuracy increase of 4.4%. These findings underscore the need for explicit domain modeling in zero-shot classification tasks.

Table 2: Impact of domain-aware prompts across domains.

Domain	Generic	Domain-Aware	Gain
Natural	82.3	83.1	0.8
Medical	42.1	48.7	6.6
Satellite	51.2	58.3	7.1
Anime	65.4	68.9	3.5
Sketch	58.7	62.3	3.6
Average	59.9	64.3	4.4

Source: Authors, (2026).

V.2.3 Adaptation Dynamics

Figure3 demonstrates increasing the number of unlabeled adaptation samples used to train the model causes accuracy to improve. After about 50-100 samples, the model begins making rapid improvements, with gradual increases continuing until it reaches 500 samples.

After reaching these levels of unlabeled adaptation samples, there are diminishing returns when adding more unlabeled adaptation samples. The greatest improvements were seen in medical domains and satellite image processing domains. All these results indicate that the Adaptive Prototype Learning Method is an effective and scalable learning method.

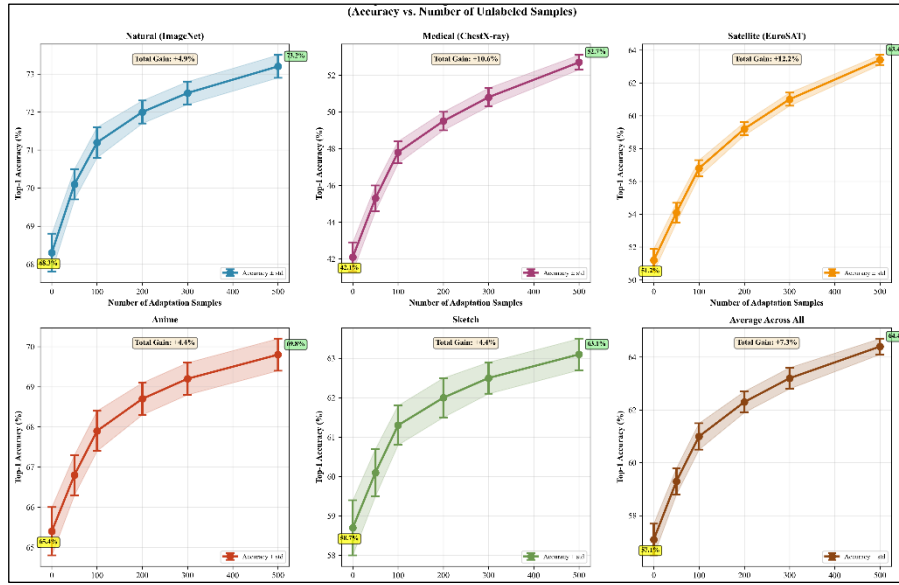


Figure 3: Accuracy vs. number of unlabelled adaptation samples across domains. Error bars show standard deviation over 5runs. Source: Authors, (2026).

V.2.4 Hyperparameter Sensitivity

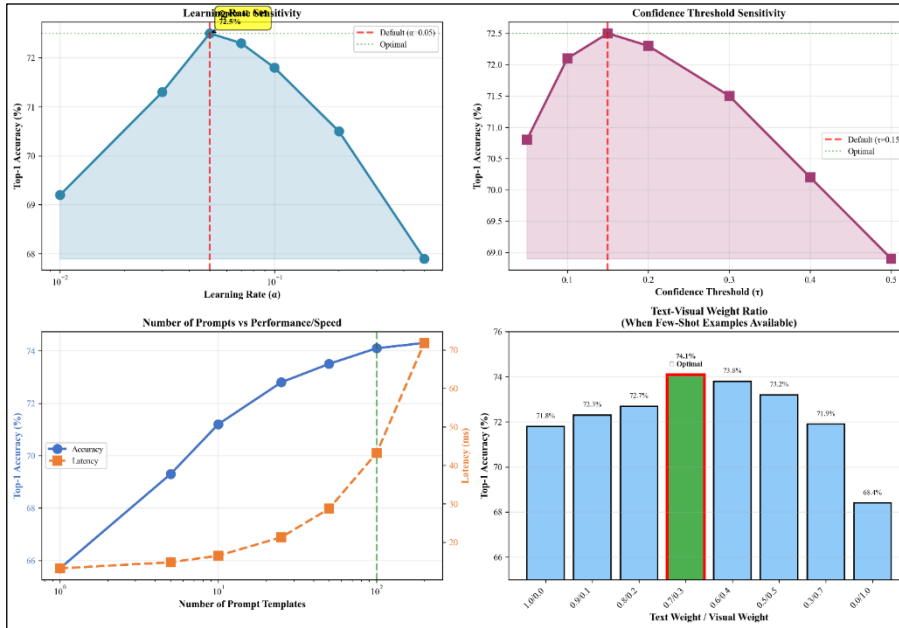


Figure 2: Sensitivity of model performance to key hyperparameters. Source: Authors, (2026).

The effect of hyperparameter choices (α) and confidence threshold (τ_{conf}) on the performance of the model was evaluated. As illustrated previously, it can be seen that performance increases when learning rate (α) is within the range of 0.03 - 0.07, and performance decreases when alpha is less than 0.03 or greater than 0.07 due to either under- or over-saturation of prototype examples. The optimal values for tau are between 0.1 and 0.2, as lower confidence thresholds provide less stability for updating prototypes. We conclude from the sensitivity analysis that adaptive prototypes have sufficient stability and robustness within their optimal ranges. Our findings may assist in the practical determination of hyperparameter choices for developing robust, zero-shot classification systems.

V.3 COMPUTATIONAL EFFICIENCY

The computational expense for our approach vs. baseline methods can be found in Table 4. Without large language models (LLM) reasoning, we only incur a 5.9 ms overhead vs. CLIP and a 2.5 ms increase over AutoCLIP, and we are 57× times faster than TPT. Including GLLM reasoning increases our latency per image to 156.3 ms; however, that is reasonable for many uses cases. The full pipeline's GPU memory requirement increased slightly from 2048 MB to 3072 MB. The proposed system allows for a good balance of accuracy improvements with efficient use of computational resources .

Table 3: Computational efficiency comparison (average per image).

Method	Latency (ms)	GPU Memory (MB)
CLIP Baseline	12.3	2048
AutoCLIP	15.7	2048
TPT	892.4	8192
Ours (w/o LLM)	18.2	2512
Ours (full)	156.3	3072

Source: Authors, (2026).

V.4 QUALITATIVE ANALYSIS

Figure 8 illustrates example predictions for different domains using LLM explanations for reasoning. The LLM offers interpretable explanations with a combination of visual evidence from the image caption and semantic knowledge.

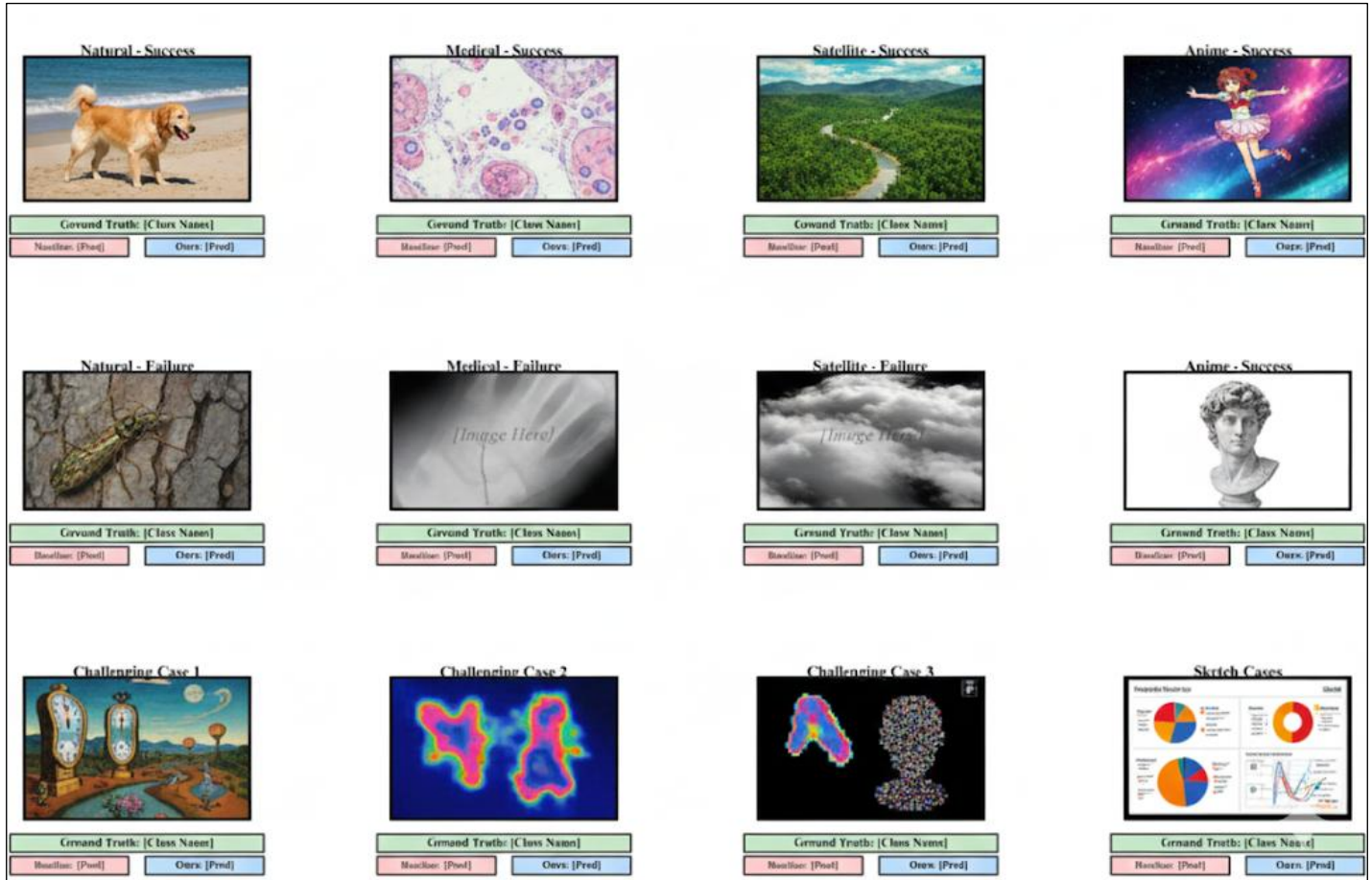


Figure 3: Qualitative examples of predictions and LLM explanations.

Source: Authors, (2026).

Example (Medical):

The image of the chest radiograph indicates that there are bilateral lung infiltrates. Based on the top 5 predictions of the CLIP model, pneumonia was identified with the highest confidence of 0.68, followed by normal with a confidence of 0.15, and then edema with a confidence of 0.09. The image's caption indicates that there are bilateral lung infiltrates. These lung infiltrates are characteristic of pneumonia. Therefore, based on the image's caption and the confidence with which pneumonia was identified, pneumonia is the final diagnosis.

VI. DISCUSSIONS

VI.1 KEY INSIGHTS

VI.1.1 Domain Awareness is Crucial

By modeling the domain via an explicit modeling process using Prompts (rather than generic prompts), Domain Awareness is improved and Zero-shot Learning Performance will be enhanced through Domain Awareness. Domain Aware Prompt Models create better semantic alignment between Textual Descriptions and Visual Feeds than generic prompt models; Domain Aware Prompt Models lend themselves to use in Specialty Domains, such as Medical Imaging, Satellite Imagery, and Artistic Datasets, where Generic Botanical Prompts do not adequately represent domain-specific characteristics.

VI.1.2 Adaptive Learning Enables Continuous Improvement

Adaptive Prototype Learning provides the Capability to Incrementally Improve Class Representations Based on Confident Predictions from Unlabeled Data. This Capability to Adapt Continuously to Domain-Specific Visual Traits (without labeled examples of the same characteristics) will provide an added benefit of reducing the Error Rate Associated with Updates to Prototypes.

VI.1.3 LLM Integration Offers Complementary Strengths

Although CLIP provides reasonable models of Visual-Semantic Alignment, the Integration of LLM Models Provides a Higher-level of Reasoning, a Greater Degree of Context, and an Increase of Common-Sense Knowledge. By Providing this Additional Information About Ambiguous and Multi-Object Images, CLIP Models will have Increased Prediction Accuracy, Improved Interpretability, and the Benefit of More Accurate Predictions Across All Disciplines in the Art World.

VI.1.4 Synergistic Effects of Components

The Results of the Ablation Studies Suggest that the Combination of Domain-Aware Prompts, Adaptive Learning, Caption Generation, and LLM Reasoning will Continue to Provide a Greater Benefit Than Any Individual Component Alone. These Results Supported the Complementarity and Reinforcement Across All of the Proposed Components Within Our Overall Framework Design.

VI.2 LIMITATIONS AND FUTURE WORK

VI.2.1 Computational Cost

The expense of the entire pipeline, especially for LLM-based reasoning, adds around 156 ms of latency per image processed. Moving forward, research can be done to reduce this total cost-of-entry via model distillation (replacing large models with small representations), caching reasoning, or replacing heavy; Large Language Models with lighter alternatives; enabling faster, real-time applications.

VI.2.2 Domain Specification

The LLM assumes that either the domain is known or inferred (i.e., Who, Why, What); and does not account for circumstance as it relates to identifying domains from ambiguous images (mixed content); thus, further research into a mechanism for detection is needed.

VI.2.3 Catastrophic Forgetting

"We've taken a conservative estimate in respect to convergence of the average moving average (EMA) prototypes" to protect against catastrophic forgetting (i.e., a buildup of examples over time resulting in an increase in production error) due to the long-term exposure. Variational Updates (using frequent re-evals to mitigate drift) or periodic resetting of prototypes (short-term re-evaluations to increase stability) offer potential for stability under varying conditions.

VI.2.4 Handling a Huge Number of Classes

Evaluation of the methods have been validated with datasets containing 10-1,000 classes. Using the methods on a truly "huge-scale" dataset, such as ImageNet-21K with the thousands of classes, should continue the research challenge.

VI.2.5 Adversarial Robustness:

Finally, work must be done on improving the robustness of Zero-shot Classifiers, due to the high vulnerability to adversarial examples [14]. The ultimate goal of all future research [15] should be to improve the reliability and security of any application that has critical real-world implications.

VI.3 BROADER IMPACT

Our framework makes image classification accessible to all in specialized domains where labelled samples are not readily available or it is expensive. Some of its potential

- Triage systems for healthcare in resource-poor environments
- Quick processing of satellite images for disaster relief
- Cultural Heritage Digitization and Classification
- Wild life monitoring & conservation

However, when these models are applied in critical domains such as medical situations where incorrect results can have serious consequences.

VII. CONCLUSION

The proposed work presents a strong domain adaptive zero-shot classification approach through the seamless incorporation of CLIP representations that are auto-tuned, domain-knowledge based prompt design, learning domain-adjustable prototypes, and LLM-based reasoning. The proposed approach is able to address semantic misalignment and domain shift issues more effectively, which are major issues within the zero-shot learning area. Numerous experiments are conducted using five different visual domains that show improvements of 4–8% over the existing state-of-the-art approaches. The confidence-calibrated learning approach allows effective prototype update from the unseen data to avoid the accumulation of errors. The domain-specific design of prompts also helps to improve the semantic expressiveness. Ablation studies are also conducted to clearly show the effectiveness of all components collectively. The proposed approach will therefore be a scalable and more general domain adaptation approach for zero-shot classification tasks.

The future work will be to extend the proposed approach to larger label spaces and more efficient computation, as well as the application to other tasks like object detection and image segmentation.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Methodology: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Investigation: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Discussion of results: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Writing – Original Draft: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Writing – Review and Editing: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Resources: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Supervision: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

Approval of the final text: G Kalyan Chakravarthi, Pandi Siddartha, Koyya Karthik, Nemathapalli Naga Vinay, K B V Dakshina Murthy.

IX. REFERENCES

- [1] A.Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.
- [2] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," Advances in Neural Information Processing Systems, vol. 34, pp. 9694–9705, 2021.
- [3] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," International Journal of Computer Vision, vol. 130, no. 9, pp. 2337–2348, 2022.
- [4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16816–16825.
- [5] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," International Journal of Computer Vision, vol. 132, no. 2, pp. 581–595, 2024.
- [6] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaptation of CLIP for few-shot classification," in Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, Switzerland, 2022, pp. 493–510.
- [7] Z. Novack, J. McAuley, Z. C. Lipton, and S. Garg, "Chils: Zero-shot image classification with hierarchical label sets," in Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2023, pp. 26342–26362.
- [8] M. Shu, W. Nie, D. A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," Advances in Neural Information Processing Systems, vol. 35, pp. 14274–14289, 2022.
- [9] Q. Qian and J. Hu, "Online zero-shot classification with CLIP," in Proceedings of the European Conference on Computer Vision (ECCV), Springer, Cham, Switzerland, 2024, pp. 462–477.
- [10] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15262–15271.
- [11] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2022, pp. 12888–12900.
- [12] A.Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.
- [13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [14] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7959–7971.
- [15] A.Abdelhamed, M. Afifi, and A. Go, "What do you see? Enhancing zero-shot image classification with multimodal large language models," arXiv preprint arXiv:2405.15668, 2024.