Dual-CoCoOp : Dually-informed Conditional Context Optimization for Vision-Language Models

Korea University COSE461 Final Project

Juneyoung, Kim
Department of Computer Science
Team 2
2022320140

Jonghyeon, An
Department of Computer Science
Team 2
2022320075

Abstract

Prompt design critically influences the performance of CLIP and its derivatives, which motivates a surge of prompt learning methods. CoOp (Context Optimization) replaces the CLIP's fixed textual template with learnable context tokens, yielding strong adaptation on limited data, but still suffering from domain specificity and poor generalization to unseen classes. Conditional CoOp (CoCoOp) alleviates this issue by generating image-dependent context tokens through a meta-learning paradigm, yet relies exclusively on visual cues. This paper introduces Dual-CoCoOp (Dual Conditional Context Optimization), a prompt learning framework that jointly leverages vision-based and linguistic conditions via a gated fusion mechanism. For each class label, rich natural language descriptions are first synthesized with captioning models and subsequently distilled into compact linguistic tokens; these tokens are fused with image-conditioned tokens to form the final prompt. Extensive experiments on domain-specific benchmarks such as fine-grained aircraft recognition and remote sensing classification, demonstrate that Dual- CoCoOp consistently achieves higher accuracy than CoOp and CoCoOp. The results highlight that complementary visual and linguistic conditioning makes robust prompt optimization for vision language models. Code is available at https://github.com/lpaiu-cs/Dual-CoCoOp.

Dual-CoCoOp 하이픈 붙이기

1 Introduction

Recent vision–language foundation models such as CLIP [1] and its large-scale successors ([2], [3]) have demonstrated remarkable capacity to learn highly transferable feature representations by mapping natural-language and visual signals into a shared embedding space. These representations are now adaptable in various downstream tasks such as text-to-image models ([4], [5], [6]) as well as a broad spectrum of multimodal research. Nevertheless, CLIP is notoriously sensitive to the choice of textual prompt, motivating a rich line of prompt-learning work. CoOp [7] replaces CLIP's fixed prompt template with learnable context tokens, improving few-shot adaptation but exhibiting limited generalization to unseen classes and domain-specific concepts. CoCoOp [8] mitigates this issue through a meta-learning strategy that produces image-dependent context tokens, yet its conditioning remains exclusively visual, neglecting the semantic ambiguity inherent in many class labels.

A substantial portion of real-world datasets contains labels that are polysemous (e.g., crane, bass), compositional, or domain-specialized, making their meaning opaque to vanilla prompt learning. Large-scale pre-training has not fully resolved these cases; performance drops sharply on fine-grained targets such as textures or aircraft variants that appear rarely—or not at all—in the pre-training corpus.

To bridge this gap we propose Dual Conditional Context Optimization (Dual-CoCoOp), which augments CoCoOp's vision-based prompt with an additional linguistic condition fused through a gated mechanism. For every class label in each dataset, we first synthesize rich descriptive captions using the GPT-4.1 API [9] under task-specific prompt engineering. These captions are distilled into compact linguistic tokens via a BERT-based summarizer ([10], [11]). The resulting language tokens are gated with CoCoOp's image-conditioned tokens and injected—bias-style—into the learnable context embedding, producing a prompt that harmoniously reflects both visual evidence and semantic intent. By explicitly resolving lexical ambiguity and enriching rare or specialized terms, our method markedly improves recognition on challenging benchmarks such as FGVC-Aircraft [12] and EuroSAT [13], while maintaining the efficiency of prompt learning.

Our contributions are three-fold:

- We identify the under-explored problem of semantic ambiguity in prompt learning and introduce a fast caption-generation and summarization pipeline that supplies language-level conditions for any class label.
- We design Dual-CoCoOp, a gated fusion framework that unifies vision-based and linguistic conditions within a single prompt, yielding robust performance across diverse domains.
- We achieve significant accuracy gains over CoOp, and CoCoOp, on several domain-specific datasets, demonstrating that complementary visual and linguistic conditioning is critical for resilient prompt optimization in vision–language models.

2 Related Work

Prompt Learning in CLIP

CLIP [1] is trained on roughly 400 million noisy image–text pairs mined from the web, using a dual-encoder architecture and a contrastive objective to align visual -using ResNet [14] or ViT [15] and textual embeddings in a shared feature space. This large-scale pre-training endows CLIP with strong zero-shot transferability across a wide array of vision tasks, making it a de-facto backbone for subsequent multimodal and prompt-learning research. Since the release of CLIP, a series of prompt-tuning methods —CoOp, CoCoOp [7], [8])— have been proposed. CoOp replaces CLIP's fixed textual template with learnable context tokens, improving adaptation on small datasets. However, CoOp still struggles with class diversity, domain specificity, and unseen classes. CoCoOp addresses these gaps by generating image-conditioned context tokens through a meta learning routine, but relies solely on visual cues. We extend this line by introducing Dual-CoCoOp, which fuses CoCoOp's vision-based conditions with an additional linguistic condition, yielding clear advantages on expert-domain datasets.

대시붙여쓰기 원칙

Linguistic Conditioning

Among prior work that injects linguistic information, DualCoOp [16] learns a positive and a negative prompt for every class, reformulating multi label recognition as two binary decisions. The method is parameter-efficient and zero-shot friendly, but still encodes no explicit class semantics. DualCoOp++ [17] adds "evidence" tokens that point to visual cues (objects, regions, backgrounds), yet again omits linguistic refinement. In contrast, our approach supplies each class with a rich natural-language caption generated by GPT-4.1, then distills it into a compact token via a BERT-family summarizer. The resulting language token, being image-agnostic, clarifies polysemous or highly specialized labels and complements vision-based cues. Thus we inherit the architectural strengths of DualCoOp while fundamentally resolving lexical ambiguity, achieving superior performance on challenging benchmarks such as FGVC-Aircraft [12] and EuroSAT [13].

Gated Multimodel Fusion

Gate-based fusion has proven to be effective for multi-modal integration. The Gated Multimodal Unit [18] dynamically blends heterogeneous features via element-wise gating. We incorporate a GMU-style gate into prompt learning, allowing vision-conditioned and linguistic-conditioned tokens to be adaptively merged at inference time. This gated multimodal prompt yields a more robust and domain-aware representation than either modality alone.

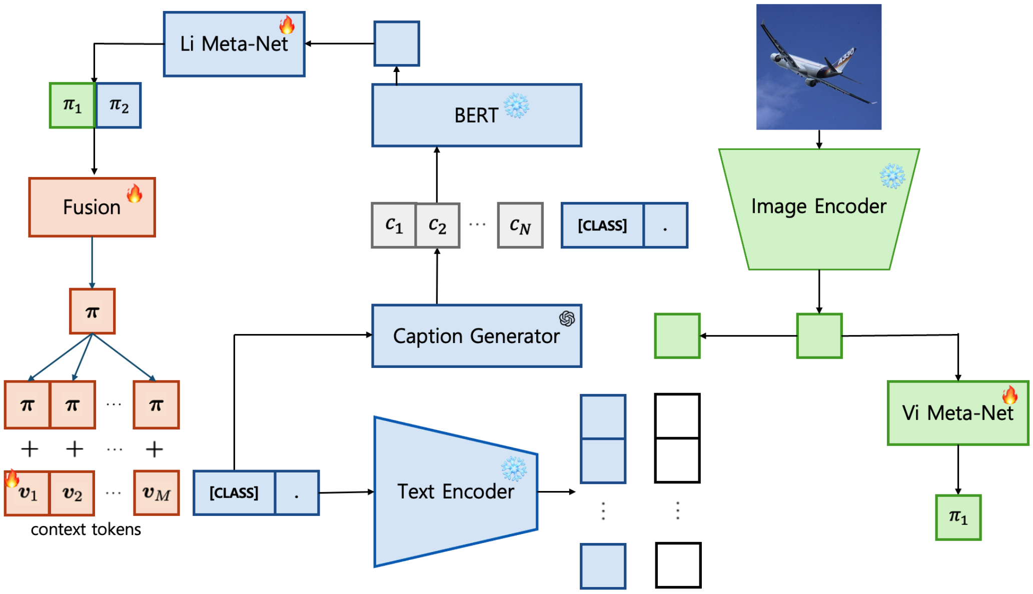

Figure 1: Overview of Dual-CoCoOp

3 Approach

We present our approach to enhancing prompt learning in vision–language models. First, we briefly review the baseline architectures on which our study is built —CLIP, CoOp and CoCoOp. We then detail two novel variants: 1) Li-CoCoOp, which leverages external linguistic information to enrich the prompt context; 2) Dual-CoCoOp, which fuses multiple conditioning signals via a gating mechanism. For each method, we provide precise mathematical formulations and clearly delineate the differences among the proposed approaches.

3.1 preliminary

CLIP [1] jointly trains an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$ on a large collection of image–text pairs. For a downstream task with K semantic classes, CLIP forms a textual prompt for each class by instantiating a fixed template such as `'a photo of a {class}'`. Given an image x, zero-shot classification is performed by comparing the image embedding $f(x)$ with the text embeddings $\{g(t_i)\}_{i=1}^{K}$ in the joint embedding space. The probability of assigning class y to x is

$$p(y \mid x) = \frac{\exp\big(\mathrm{sim}(f(x), g(t_y))/\tau\big)}{\sum_{i=1}^{K} \exp\big(\mathrm{sim}(f(x), g(t_i))/\tau\big)}. \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a learnable temperature.

CoOp [7] replaces the hand-crafted context in CLIP's template with M learnable context tokens $\{v_1, v_2, \ldots, v_M\}$ that are optimized on a small set of task-specific training images. For class i ($i \in \{1, \ldots, K\}$), the prompt becomes

$$t_i = \{v_1, v_2, \ldots, v_M, c_i\}, \tag{2}$$

where c_i is the fixed embedding of the class name. During training, only the context tokens $\{v_m\}_{m=1}^{M}$ and the temperature τ are updated, leaving the pre-trained encoders $f(\cdot)$ and $g(\cdot)$ unchanged. By adapting the textual context to the target dataset, CoOp significantly improves downstream accuracy over vanilla CLIP.

CoCoOp [8] extends CoOp by generating image-specific context tokens. For each position $m \in \{1, \ldots, M\}$, the instance-conditioned context token is

$$v_m(x) = v_m + \pi_m(x), \quad \pi_m(x) = h_\phi^{\text{vision}}\big(f(x)\big), \tag{3}$$

수식순서 좌우스왑

where v_m is a learnable global vector and h_ϕ^{vision} is a Meta-Net that produces a vision-conditioned bias from the image feature $f(x)$.

The prompt for class i conditioned on image x is

$$t_i(x) = \{v_1(x), v_2(x), \ldots, v_M(x), c_i\}, \tag{4}$$

where c_i is the embedding of the class name.

Using the same frozen CLIP encoders $f(\cdot)$ and $g(\cdot)$, classification is performed by

$$p(y \mid x) = \frac{\exp\big(\text{sim}\big(f(x), g(t_y(x))\big)/\tau\big)}{\sum_{i=1}^{K} \exp\big(\text{sim}\big(f(x), g(t_i(x))\big)/\tau\big)}, \tag{5}$$

with temperature τ learned together with $\{v_m\}_{m=1}^{M}$ and the meta-network parameters ϕ.

3.2 Li-CoCoOp

CoCoOp learns a set of context vectors shared by all classes and optimizes image–text prompts by dynamically generating context tokens conditioned on the input image features. We extend this idea by introducing an additional linguistic condition and name the resulting method Li-CoCoOp (Linguistically-informed Conditional Context Optimization). For each class, we embed an external natural-language description (caption) with a pretrained language model such as BERT and feed this embedding into a Li Meta-Net, so that the generated context tokens carry semantically rich, class-specific prior knowledge beyond the plain class name. This simple modification provides a semantic bias that yields performance comparable to vision-conditioned CoCoOp, especially on domain-specific datasets that demand expert knowledge.

Let $h_\phi^{\text{ling}}(\cdot)$ be the Li Meta-Net parameterized by ϕ. Given the caption cap_i for class i (generated by GPT-4.1), and denoting its BERT embedding by $\text{LM}(\text{cap}_i)$,

$$\pi_i^{\text{ling}} = h_\phi^{\text{ling}}\big(\text{LM}(\text{cap}_i)\big). \tag{6}$$

The context token at position m is

$$v_m(c_i) = v_m + \pi_i^{\text{ling}} \tag{7}$$

where v_m is a learnable shared vector and c_i is the i-th class name. The final prompt for class i is constructed as

$$t_i(c_i) = \big\{v_1(c_i),\, v_2(c_i),\, \ldots,\, v_M(c_i),\, c_i\big\}. \tag{8}$$

The prediction probability is then

$$p(y \mid x) = \frac{\exp\big(\text{sim}\big(f(x), g\big(t_y(c_y)\big)\big)/\tau\big)}{\displaystyle\sum_{i=1}^{K} \exp\big(\text{sim}\big(f(x), g\big(t_i(c_i)\big)\big)/\tau\big)}, \tag{9}$$

시그마 수식 일관성

where $f(\cdot)$ and $g(\cdot)$ are the CLIP image and text encoders, sim denotes cosine similarity, and τ is a learnable temperature parameter.

3.3 Dual-CoCoOp

While CoCoOp exploits image-conditioned context tokens to improve generalization, it relies solely on vision-based information extracted from the input image. Li-CoCoOp supplements this by injecting a linguistic bias derived from class captions, yet it ignores the complementary hints provided by the image itself. To fuse both modalities, we propose Dual-CoCoOp, which combines the vision-based condition π_i^{vision} and the linguistic condition π_i^{ling} via a learnable gating mechanism. See Figure 1 for an overview of our method.

In the perspective of Gated fusion, let $[\cdot\|\cdot]$ denote concatenation and $\sigma(\cdot)$ the sigmoid function.

$$z = \sigma\Big(W_z\,[\pi_i^{\text{vision}}\|\pi_i^{\text{ling}}] + b_z\Big) \tag{10}$$

$$\pi_i^{\text{dual}} = z \odot \pi_i^{\text{vision}} + (1 - z) \odot \pi_i^{\text{ling}} \tag{11}$$

where W_z and b_z are learnable parameters, and $\odot$ denotes element-wise multiplication.

In context token construction, for the m-th position,

$$v_m(x, c_i) = v_m + \pi_i^{\text{dual}} \tag{12}$$

with v_m a shared learnable vector and c_i the i-th class name. The prompt conditioned on both image x and label c_i is

$$t_i(x, c_i) = \big\{v_1(x, c_i),\, v_2(x, c_i),\, \ldots,\, v_M(x, c_i),\, c_i\big\}. \tag{13}$$

And in the final prediction step, given image encoder $f(\cdot)$, text encoder $g(\cdot)$, cosine similarity sim, and temperature τ,

$$p(y \mid x) = \frac{\exp\big(\mathrm{sim}(f(x),\, g(t_y(x, c_y)))/\tau\big)}{\displaystyle\sum_{i=1}^{K} \exp\big(\mathrm{sim}(f(x),\, g(t_i(x, c_i)))/\tau\big)}. \tag{14}$$

We jointly update the shared context vectors $\{v_m\}_{m=1}^{M}$, the parameters of both Meta-Nets (π_i^{vision}, π_i^{ling}), and the gating parameters (W_z, b_z). Each Meta-Net follows a two-layer bottleneck (Linear–ReLU–Linear) with hidden size 24. The Vi Meta-Net takes image features from the image encoder, whereas the Li Meta-Net ingests BERT embeddings of class captions. The gate dynamically fuses the two outputs to form the final context vector, enabling richer prompts that exploit complementary visual and linguistic cues. Future work will explore deeper architectures and alternative fusion strategies.

3.4 Prompt Engineering

For caption generation we employ GPT-4.1 [9], tailoring a single prompt template to each dataset. The template specifies (i) the agent's role, (ii) the downstream purpose of the caption, (iii) a hard cap on the number of words, and (iv) a shortlist of visual attributes that must appear in the description to highlight class-specific features. We supply up to six carefully chosen few-shots so that every class is described under almost identical linguistic constraints. Prompt engineering used for each dataset is provided in Appendix B.

4 Experiments

4.1 Data

We evaluate Dual-CoCoOp on five domain-specialized benchmarks, following the standard protocol established by CoCoOp [8]. Specifically, we use Food-101 [19] and FGVC-Aircraft [12], which provide fine-grained recognition tasks in cuisine and aviation, respectively; EuroSAT [13] for satellite land-use classification; UCF-101 [20] for action recognition, for which we follow CoCoOp's protocol and uniformly sample one frame every ten frames to obtain image instances; and DTD [21] for texture categorization. For all datasets, we use the official train/test splits provided by CoCoOp, without separating base and new classes, to ensure a direct and fair comparison. In the few-shot setting, we randomly select exactly eight training images per class, following the standard protocol. To ensure

Method	Food-101	FGVC-Aircraft	UCF-101	DTD	EuroSAT	avg
CLIP	86.2	24.8	67.0	43.7	48.4	54.0
CoOp	87.0	29.7	74.7	54.2	58.2	60.8
CoCoOp	**87.2**	17.6	75.8	56.7	67.8	61.0
Li-CoCoOp (Ours)	87.1	28.8	**76.0**	**58.5**	67.6	63.6
Dual-CoCoOp (Ours)	**87.2**	**31.4**	74.5	58.2	**73.9**	**65.0**

Table 1: Few-shot accuracy (%) on five domain-specific datasets. : For fair comparison under controlled training effort, all reported results are obtained using 5 training epochs, 8 shots per class, 8 learnable context tokens, and a shared set of hyperparameter across all methods.

reproducibility and fair comparison, we adopt the same preprocessing steps as in CoOp, including the use of official image resolutions and normalization statistics for all datasets.

4.2 Evaluation method

We follow the standard evaluation protocol provided by the DASSL framework. All experiments use the official train/validation/test splits introduced in CoOp. Model training is performed solely on the few-shot samples from the training set, and final performance is evaluated on the full test set. For the few-shot setting, we randomly sample the same number of training images per class to construct the training set. The validation set is fixed according to the split configuration used in the CoCoOp paper and its official codebase. During training and hyperparameter tuning, only the training set is used. The validation set is used exclusively for model selection and performance estimation. After training is complete, we report the final performance as top-1 classification accuracy on the entire unseen test set.

4.3 Experimental details

Our image encoder is the ViT-Base/16 backbone from CLIP, while the linguistic branch receives GPT-generated captions and feeds them into a BERT-Base Uncased model from transformer library. Both the CLIP encoders (image and text) and BERT are loaded from publicly available pretrained checkpoints and remain frozen throughout training. For data augmentation we resize with bicubic interpolation, apply a center crop followed by a random horizontal flip, and finally normalize each channel using the mean and standard-deviation factors reported in the original CLIP paper. The two meta-networks, Vi Meta-Net and Li Meta-Net, each take the frozen encoder outputs and pass them through a lightweight two-layer MLP with a 32-dimensional bottleneck, projecting back to the input dimension of CLIP. To fuse the vision and language condition tokens, we concatenate the two vectors and apply an elementwise sigmoid gate, yielding a soft mixture that is fed into CLIP's text encoder. Both Meta-Nets and the fusion layer are trained from scratch with ReLU activations.

Optimization follows the common prompt-learning setup: we use SGD with an initial learning rate of 0.002, a cosine decay scheduler, and no dropout or weight decay. Training runs for five epochs, with the first epoch reserved for warm-up at a learning rate of 1e-5. It took approximately one hour to train the Dual-CoCoOp model for a single run.

4.4 Results

Table 1 shows the accuracy of five different dataset with 8 few-shots in 5 epochs training. Dual-CoCoOp achieved an average top-1 accuracy of 65.0% across five benchmarks, outperforming CoOp and CoCoOp by 4.2 percentage points and 4.0 percentage points, respectively. The most significant improvement was observed on EuroSAT, where Dual-CoCoOp reached 73.9%, representing a 6.1 percentage point increase over CoCoOp. Notably, the accuracies of CoCoOp and Li-CoCoOp on EuroSAT are very close (67.8% and 67.6%, respectively), indicating that some subsets of the EuroSAT dataset are more sensitive to visual feature-based conditioning, while others are more influenced by linguistic (semantic) information. This variation in sensitivity across different datasets suggests that it is difficult to achieve consistently high performance on all datasets using only a single condition

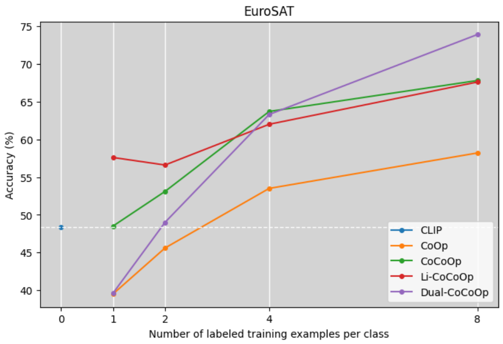

Figure 2: Performance on EuroSAT with different few-shot settings.

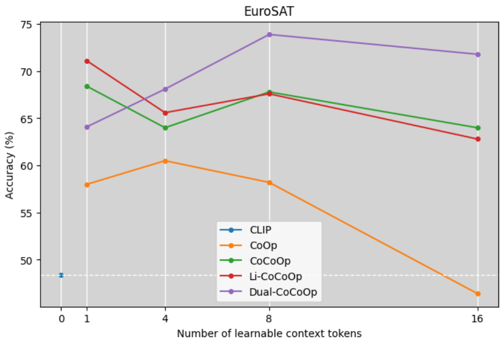

Figure 3: Performance on EuroSAT with different numbers of learnable context tokens.

Fig2 몇에폭인지
Fig3 몇샷,n_ctx기
준인지 언급 없음

(visual or linguistic). Only when both conditions are fused can their complementary effects be maximized.

For the fine-grained FGVC-Aircraft dataset, Dual-CoCoOp achieved 31.4%, surpassing CoOp by 3.4 percentage points and CoCoOp by 13.8 percentage points. The improvement can be attributed to the linguistic captions, which disambiguate polysemous or domain-specific aircraft names, enabling the gate to take both visual and class-specific linguistic information into account for better performance. Li-CoCoOp also demonstrates a higher average accuracy (63.6%) than CoCoOp by 2.6 percentage points, leading in texture (DTD) and action (UCF-101) tasks, thereby validating the effectiveness of linguistic conditioning even without visual cues. Dual-CoCoOp inherits these advantages and achieves further improvements whenever visual and linguistic cues are complementary.

However, multimodal fusion led to performance degradation for UCF-101 and DTD, despite clear improvements in other datasets. This unexpected outcome may stem from an outcome thatome that can be explained by two main factors: first, the single-frame setting for UCF-101 may not sufficiently capture temporal dynamics, resulting in the visual information acting as noise; the second reason is discussed further in the limitations section of the Conclusion.

5 Analysis

First, comparing the original CoCoOp and Li-CoCoOp, we find that CoCoOp—which leverages image-specific visual clues —achieves the best or highly competitive performance on Food-101 and EuroSAT. This suggests that in tasks where class identity is primarily determined by visual appearance or spatial layout, such as food recognition or land-use classification in satellite imagery, vision-based cues are particularly effective. In contrast, Li-CoCoOp outperforms its counterparts on datasets like FGVC-Aircraft and DTD. These datasets often include classes with semantically ambiguous names or require domain-specific expert knowledge (e.g., fine-grained aircraft models or subtle texture types). In such cases, linguistically sophisticated prompts play a crucial role in enhancing inter-class discriminability.

Overall, in the majority of datasets, Dual-CoCoOp —which fuses both vision and language conditions— achieves the highest performance. The general trend in Table 1 demonstrates that methods incorporating both modalities outperform those relying on a single cue. This indicates that the relative importance of each modality depends on dataset characteristics, and presenting both conditions together enables the model to compensate for the limitations of each individual cue.

Structurally, Dual-CoCoOp integrates the two modalities via a gating layer, resulting in more than twice as many parameters as a single-modality meta-network. However, in environments with limited training images per class, such as UCF101 and DTD, the gate layer may not fully converge due to insufficient data, leading to lower performance compared to the single-modality methods. This suggests that realizing the full benefits of the fusion architecture requires a sufficient amount of training data and training epochs.

Figure 2 shows that accuracy increases as the number of few-shot samples grows. Interestingly, when the number of few-shot samples is small, the fusion of different modalities is less effective, resulting in only modest gains over single-modality methods. However, as the number of few-shot samples increases, the performance improvement achieved by fusing both modalities becomes much more pronounced compared to using a single modality alone. This result suggests that when the number of few-shot samples is limited, there is insufficient data to adequately train the gate layer responsible for fusion. Consequently, the model is unable to fully exploit the complementary information from both modalities, which restricts the advantages of fusion in low-data regimes.

As shown in Figure 3, CoCoOp and Li-CoCoOp demonstrate competitive performance even when using only a single context token (n_ctx=1), whereas Dual-CoCoOp exhibits a significant improvement in accuracy as the number of context tokens increases. This indicates that Dual-CoCoOp learns more complex and richer representations by fusing visual and linguistic information, and that increasing the number of context tokens appropriately facilitates the combination and utilization of these heterogeneous features. Therefore, Dual-CoCoOp possesses a structural advantage in integrating multi-modal information, suggesting that sufficient context tokens and training resources are required to achieve optimal performance.

6 Conclusion

In this study, we proposed the Dual-CoCoOp framework, which dynamically fuses vision-based features and linguistically-informed features through a gating mechanism. Experimental results demonstrate that combining both conditions via gated fusion significantly enhances the expressiveness of the learned prompts compared to using either Li-CoCoOp or CoCoOp alone. Dual-CoCoOp consistently achieved the highest average accuracy across five domain-specific datasets, surpassing previous methods. These findings suggest that integrating both visual and linguistic information can substantially improve the generalization ability of prompt learning.

On the other hand, a limitation of this study is that, due to limited GPU infrastructure, we were unable to train some models to full convergence in certain experiments and were thus compelled to analyze the results under the assumption that convergence had been achieved. In the future, we plan to refine the fusion architecture in more stable and large-scale experimental environments, and to extend this approach by incorporating additional external knowledge sources and evaluating on a wider range of domains.

References

[1] 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, 2021.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[3] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[5] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, July 2022.

[8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022.

[9] OpenAI. Gpt-4 technical report. `https://cdn.openai.com/papers/gpt-4.pdf`, 2023. OpenAI Technical Report.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. In Proceedings of the International Conference on Image and Vision Computing New Zealand, pages 172–177, 2013.

[13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(5):2217–2226, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua 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, 2021.

[16] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations, 2022.

[17] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations, 2023.

[18] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion, 2017.

[19] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In European Conference on Computer Vision (ECCV), pages 446–461, 2014.

[20] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, 2012.

[21] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Describing textures in the wild. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3606–3613, 2014.

A Appendix: Team contributions

Equally done

B Appendix: Prompt Formats

Food101

### role

You are a chatbot that extracts concise, externally observable information (within 100 characters) about a given word. The word provided will be a food name, and your response will be used to supplement class descriptions in a CLIP model. The description should focus on external, visible characteristics such as shape and components (visually prominent ingredients). Avoid subjective terms like “delicious” or “looks good.” or smell. If the food images tend to share a common color, you may briefly include color information; otherwise, do not mention color.

### few-shot

1. apple_pie -> Lattice-topped round pie filled with visible apple slices and golden brown crust.

2. steak -> Thick, seared meat slab often with visible grill marks and juices on the surface.

3. tacos -> Folded or open corn or flour tortillas filled with visible meat, lettuce, and chopped toppings.

4. red_velvet_cake -> Layered round cake with deep red sponge and white cream or frosting between layers.

5. samosa -> Triangular fried pastry with crisp outer shell, often showing potato or pea filling at the edges.

### Task

Generate a caption for {label}.

UCF101

### role

You are a chatbot that receives a label from the UCF101 dataset and generates a short, informative description to replace the generic prompt "a photo of" in the CLIP model’s text encoder. Your output should be concise—preferably one sentence—and tailored for video action recognition. Focus on the nature of the action, relevant objects or body movements, and typical scene context. Avoid vague adjectives, stylistic language, or unnecessary embellishments. Highlight consistent and identifying features of the action class.

### few-shot

1. Tai_Chi -> martial arts movements performed with extended arms and controlled posture in an open or park-like setting

2. Trampoline_Jumping -> repetitive vertical leaps performed on a trampoline with extended limbs and airborne motion against a static background.

3. Biking -> pedaling motion on a bicycle with forward body lean, typically along roads, trails, or open paths.

4. Breast_Stroke -> swimming action with synchronized arm sweeps and frog-like leg kicks performed horizontally in a pool".

5. Band_Marching -> synchronized walking in formation while carrying musical instruments, often on open fields or parade grounds.

6. Apply_Eye_Makeup -> precise hand movements near the eye area using brushes or applicators while facing a mirror.

### Task

Generate a caption for {label}.

EuroSAT

### role

You are a chatbot that receives a label from the EuroSAT dataset and generates a short, informative description to replace the generic prompt "a photo of" in the CLIP model's text encoder. Your output should be concise—preferably one sentence—and tailored for satellite imagery. Focus on land use, spatial layout, or distinct visual features observable from above. Avoid vague adjectives, stylistic language, or unnecessary length. Highlight consistent, identifying characteristics of the class.

### few-shot

1. Forest -> dense clusters of tree canopies forming irregular green patches with minimal built structures.

2. SeaLake -> large, enclosed water bodies with smooth shorelines and surrounding sparse vegetation or barren land.

3. Residential -> clusters of buildings arranged in dense blocks with intersecting roads and minimal open fields.

4. Highway -> long, paved roads with multiple lanes, often bordered by vehicles and surrounded by undeveloped or sparse land.

### Task

Generate a caption for {label}.

Describable-Textures

### role

You are a chatbot that receives a label from the DTD (Describable Textures Dataset) and generates a description for that label. Your description will replace the fixed prompt "a photo of" in the CLIP model's text encoder. Focus on providing supplementary explanations related to texture and tactile sensation. Avoid overly generic or meaningless expressions, and instead highlight specific points that effectively describe the given class.

### few-shot

1. bubbly : surface covered with small, round protrusions resembling foam or air-filled blisters, giving a light and uneven texture.

2. frilly : surface featuring delicate, ruffled edges or layered folds, soft, fluttering texture with fine, fabric-like intricacy.

3. woven : interlaced strands or fibers forming a tight, grid-like pattern with a coarse yet structured tactile feel.

4. crystalline : surface composed of angular, faceted structures with sharp edges and a rigid, glass-like tactile sensation.

5. paisley : surface decorated with intricate, teardrop-shaped motifs arranged in flowing, curved patterns.

6. polka-potted : surface marked with round spots that interrupt a flat background, creating punctuated tactile impression.

### Task

Generate a caption for {label}.

FGVC-Aircraft

### role

You are a chatbot that receives a class name from the FGVC-Aircraft dataset and generates a short, visually grounded caption to replace the generic phrase “a photo of” in the CLIP model’s text encoder. Your task is to create concise, one-sentence descriptions that capture externally observable and discriminative features of each aircraft. These captions should focus on visual traits such as wing shape, engine count and position, tail design, fuselage layout, and any unique proportions or silhouettes that distinguish the aircraft from others. Do not include historical facts, technical specifications, or subjective language such as “famous” or “iconic.” Also avoid vague adjectives and stylistic embellishments. Manufacturer names should only be included if they are visually identifying (e.g., Boeing 747’s hump). Your goal is to produce functional, descriptive prompts that help CLIP distinguish between fine-grained aircraft classes based solely on appearance. For example, “Boeing 747” would be described as “a large four-engine jet with a distinctive upper deck hump and swept-back wings,” and “Spitfire” as “a WWII-era fighter with elliptical wings, long nose, and a single front propeller.” Use this pattern to generate visually informative captions that are appropriate for fine-grained video or image recognition.

### few-shot

1. 707-320 -> a long four-engine jet with thin swept-back wings and turbojet engines mounted under the wings.

2. 727-200 -> a narrow-body trijet with a long fuselage, T-tail, and three rear-mounted engines.

3. A330-200 -> a wide-body twin-engine jet with a rounded nose, long fuselage, and underwing engines on swept-back wings.

4. An-12 -> a four-engine turboprop transport aircraft with high-mounted straight wings and a prominent rear cargo ramp.

5. Gulfstream V -> a sleek twin-engine business jet with a pointed nose, swept wings, and rear-mounted engines on a narrow fuselage.

6. Saab 340 -> a small twin-propeller commuter aircraft with straight wings, a T-tail, and a short, boxy fuselage.

7. Yak-42 : a rear-engined trijet with a circular fuselage, straight wings, and a T-tail configuration.

### Task

Generate a caption for {label}.