FineCAP: Fine-grained Cyclic Augmentation Prompting with Zoom-in Enhancement for Few-shot Learning

Zhongjiang He[1,2], Hongbo Sun[2], Hao Sun[2], Han Fang[2], An Zhao[2*],
Ye Yuan[2], Kongming Liang[1,3], Zhanyu Ma[1,3*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd, Beijing, China
[3]Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance

***Abstract*—Vision-Language models (VLMs) have garnered significant attention in Few-Shot Learning (FSL) tasks for powerful generic representation and generalization abilities. Existing FSL methods typically add either learnable prompts or adapters on VLMs, which generally face two key limitations, i.e., an inadequate mechanism for encoding fine-grained features of objects, and misalignment between textual class information and visual objects. Motivated by these observations, this paper proposes a Fine-grained Cyclic Augmentation Prompting (FineCAP) method of Vision-Language Models for few-shot learning, with two core designs: the Fine-grained Visual Enhancement Mechanism (FVEM) and the Dynamic Textual Enhancement Mechanism (DTEM). Specifically, (1) The FVEM module first leverages the hybrid attention to guide the model to focus on and zoom in on visual objects, extracting discriminative features by visual prompting. (2) The DTEM module proposes a learnable textual prompting method to optimize textual class features with objects' attribute information by cross-modal mapping of fine-grained object features. (3) These two modules fully exploit object information through a cycling augmentation manner, achieving adaptive fine-grained alignment. Extensive experiments on 11 widely-used benchmarks demonstrate that FineCAP achieves new state-of-the-art in few-shot learning.**

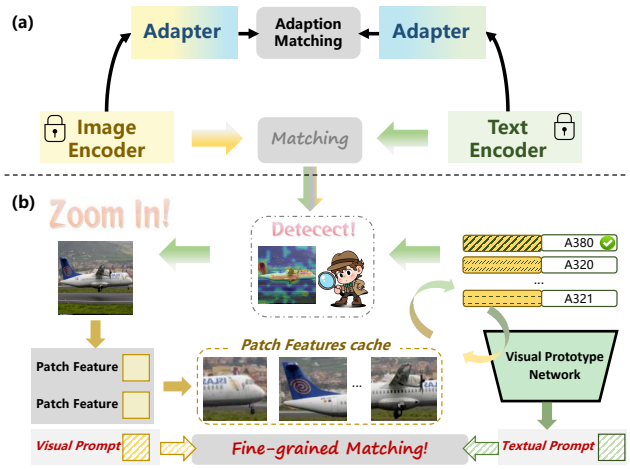

Fig. 1. Illustration of our proposed FineCAP. (a) Existing methods typically propose the adapter to align image and text. (b) In contrast, our FineCAP focuses on and zooms in on visual objects with text-guided cross-modal attention and self-attention from the image itself. Adaptive fine-grained matching and recognition are achieved through a cycling augmentation manner.

I. Introduction

Few-shot learning (FSL) aims to enable models to recognize various categories using only a small number of labeled training samples. Due to the scarcity of training data, traditional image recognition models such as ResNet [1] and ViT [2] exhibit limited performance in the FSL scenario. By contrast, Vision-Language Models (VLMs) have demonstrated remarkable generalization capabilities in visual understanding and have attracted growing interest in FSL. As a representative example, CLIP (Contrastive Language-Image Pretraining) [3] employs visual and textual encoders to align images and text in a shared embedding space, which is pre-trained by large-scale contrastive learning. Inspired by the powerful alignment ability, CLIP-based few-shot learning methods have gradually emerged and attracted significant attention.

Existing CLIP-based FSL methods primarily introduce prompt learning or trainable adapters to adapt CLIP to learn more robust classifiers. For instance, CoOp [4] introduced learnable vectors into the text prompt to generate adaptive classification weights. CoCoOp [5] proposed image-conditioned prompts to improve the generalization ability, which dynamically depend on the input image. Tip-Adapter [6] proposed to construct caches of support set images and use adapters for aligning. MMA [7] proposed the adapters to aggregate features from different branches into a shared space for better matching. Besides, some methods follow the paradigm of traditional image recognition by learning more compact classification boundaries, specifically, by minimizing intra-class variance and maximizing inter-class separation. For example, LDC [8] identifies and removes inter-class confusion to achieve clearer

*Corresponding authors: Zhanyu Ma (mazhanyu@bupt.edu.cn) and An Zhao (zhaoa@chinatelecom.cn). This work was supported in part by the National Natural Science Foundation of China (Grant 62225601, U23B2052), in part by the Beijing Natural Science Foundation Project No. L242025, and in part by the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

category distinctions. However, the above methods overlook the intrinsic correlation of the text and visual modalities during the matching process, failing to identify which image regions are most informative for representing visual objects. The usefulness of the the visual object attributes on enhancing textual representations is also neglected.

To address the above limitations, we propose a **Fine-grained Cyclic Augmentation Prompting (FineCAP)** of Vision-Language Models for few-shot learning, which aims to locate the main object that reflects cross-modal semantic alignment, serving as fine-grained visual explanations to enhance both image and text features. FineCAP is composed of the Fine-grained Visual Enhancement Mechanism (FVEM) and the Dynamic Textual Enhancement Mechanism (DTEM). Concretely, (1) in FVEM, by leveraging self-attention in the Vision Transformer (ViT) of CLIP and text-guided cross-attention, i.e., the hybrid attention,the model can identify the crucial visual object in the image, whose tokens receive the highest attention weights reflecting both salient visual regions and patches semantically aligned with the text. The visual object is thus focused on and zoomed in for highlighting distinctive visual clues and extracting discriminative visual features by visual prompting. (2) In DTEM, a learnable textual prompt is built based on the objects' attributes information, which is contained in the visual object patch features. Specifically, the patch features from the zoom-in object image are reserved as caches, which involves plenty of attribute information, such as texture and color. A cross-modal network is then constructed to map the attribute information into the textual space and combined with a learnable textual token, enhancing the textual class features. These two modules fully utilize the object information in multimodal data by a cycling augmentation manner for fine-grained alignment, significantly improving image recognition performance.

The main contributions of this work are as follows:

- We propose Fine-grained Cyclic Augmentation Prompting for few-shot learning, focusing on salient objects for fine-grained visual and textual feature alignment.
- In FineCAP, the Fine-grained Visual Enhancement Mechanism (FVEM) generates saliency maps via hybrid attention to amplify the main object for object-focused visual features. The Dynamic Textual Enhancement Mechanism (DTEM) uses learnable textual prompting to optimize textual class features with object attributes.
- Extensive experiments on 11 benchmarks show our FineCAP's consistent improvements over existing FSL methods, demonstrating superior performance.

The code of FineCAP is available at https://github.com/telecomhzj/FineCAP.

II. Related Work

Vision-Language Models (VLMs). Vision-Language Models (VLMs) [3], [9] are distinguished for their powerful generalization capabilities. Architecturally, they primarily follow two paradigms. Fusion encoders [10] process image and text jointly within one network, enabling deep interaction but requiring paired inputs during inference, which limits efficiency. Conversely, dual-encoder models like CLIP [3] and ALIGN [11] use separate encoders to align images and text via contrastive learning in a shared embedding space. This decoupling paradigm enables high efficiency and strong zero-shot transfer ability, benefiting from large-scale pretraining on web data. Subsequent models have further scaled this method into various downstream tasks. However, applying VLMs to downstream tasks, such as few-shot learning, often reveals a performance gap due to a distribution mismatch between pretraining data and the target data. To bridge this gap with limited training data, the prompt learning and adapter methods have emerged as key adaptation strategies.

A. Few Shot Learning with VLMs

Inspired by natural language processing, prompt learning adapts VLMs by modifying inputs rather than model weights. CoOp [4] pioneered the replacement of hand-crafted text prompts with learnable context vectors, boosting few-shot recognition performance. Moreover, Tip-Adapter [6] proposed a non-parametric, cache-based method using stored support set features and labels to provide auxiliary predictions, which are then fused with CLIP's output for final classification. Recent advances in VLM adaptation for FSL include attention mechanisms, visual-guided prompting [12], strategic parameter adaptation 2SFS [13], and logit deconfusion LDC [8]. However, a key limitation remains: these methods often fail to utilize the correlation and establish effective synergy between visual and textual modalities. In contrast, our FineCAP attempts to identify and amplify the image regions that represent visual objects for discriminative visual features, while enhancing textual class features with the object attributes in a cycling augmentation manner. Thus, fine-grained vision-language alignment can be achieved for few-shot learning.

III. Method

As illustrated in Figure 2, we employ the well-known two-stream encoder CLIP as the backbone. Our method incorporates a collaborative matching framework composed of a global branch and a object-aware branch. In the global branch, the image encoder of CLIP transforms the input image I into an image feature $f^G(I)$. To adapt VL model for few-shot image classification, we construct textual prompts using class names, such as "a photo of a $class_i$", and use the text encoder to derive the text features $f^t(TP_i)$. Let k denote the number of classes, where $i \in \{1, 2, \ldots, k\}$. The prediction probability for the input image based on the global branch is formulated as:

$$p(y=i|I)_G = \frac{\exp(\text{sim}(f^t(TP_i), f^G(I)))}{\sum_{j=1}^{k} \exp(\text{sim}(f^t(TP_j), f^G(I)))}, \quad (1)$$

where y is the prediction label and sim represents the cosine similarity. To further enhance the model's visual object perception capabilities, we elaborate on the **Fine-grained Visual Enhancement Mechanism**. Additionally, the **Dynamic Textual Enhancement Mechanism** is utilized to incorporate object attributes into the textual side, enhancing textual features.

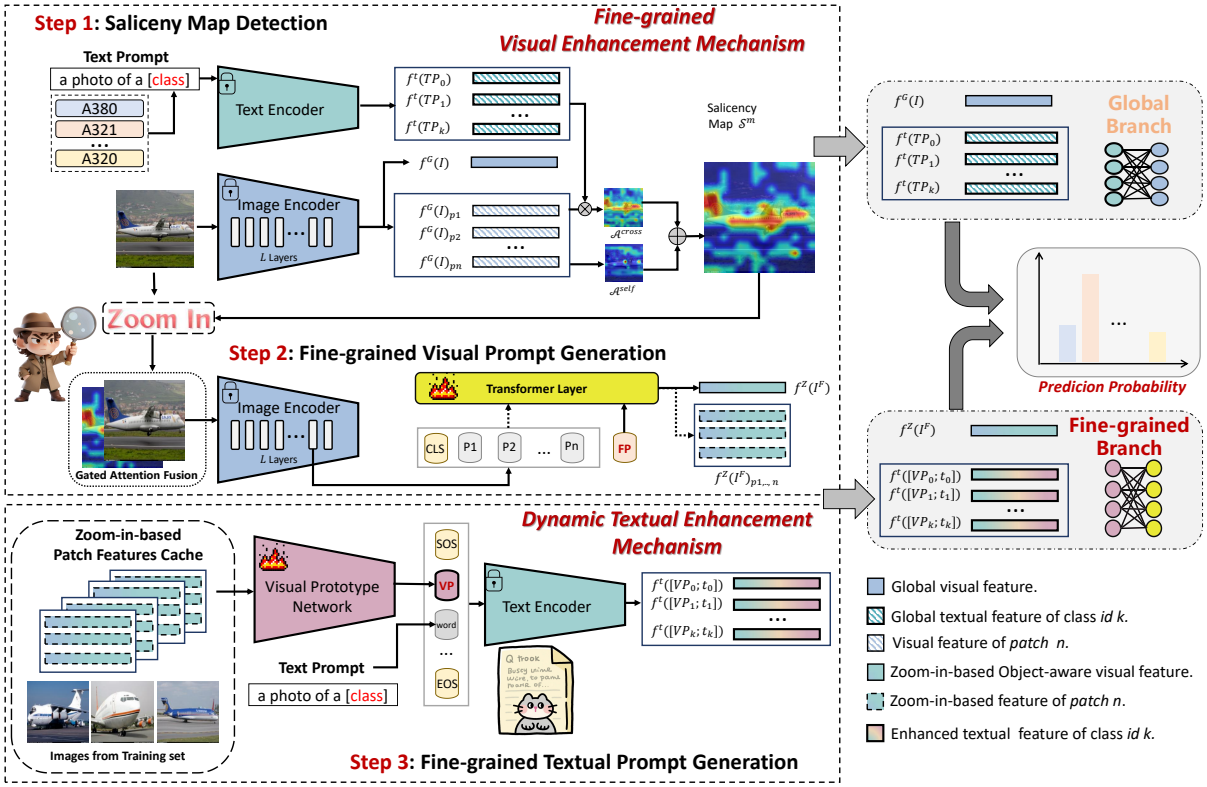

Fig. 2. The framework of our FineCAP model. In the Fine-grained Visual Enhancement Mechanism (FVEM) module, the crucial object is focused on and zoomed in to extract discriminative object-aware visual features. In the Dynamic Textual Enhancement Mechanism (DTEM), a learnable textual prompt is built based on the attribute information of objects to optimize textual class features, which is contained in the object patch features. Fine-grained alignment is thus achieved for classification.

A. Fine-grained Visual Enhancement

The vision encoder of CLIP primarily captures global semantics, often overlooking fine-grained visual cues critical for image recognition. To address this limitation, we propose a Fine-grained Visual Enhancement mechanism to enhance the representation of the target object. This enables the extraction of discriminative visual features, thereby improving recognition performance. As shown in step 1 of Figure 2, we adopt the Vision Transformer as the image encoder. The input image is divided into $n = hw/p_s^2$ non-overlapping patches, where h and w denote the height and width of the image, and p_s is the patch size. Each patch is linearly projected into a d-dimensional token. A learnable [CLS] token, denoted z_{cls}, is prepended to the sequence of patch tokens, resulting in a feature matrix $Z \in \mathbb{R}^{(n+1)\times d} = [z_{\text{cls}}, z_1^p, z_2^p, \ldots, z_n^p]$. Subsequently, multiple multi-head self-attention blocks in the transformer layers are applied. The output of the l-th MSA block is computed as:

$$\mathcal{A}^l = \frac{1}{H}\sum_{i=1}^{H} \text{softmax}\left(Q_i^l {K_i^l}^\top / \sqrt{d/H}\right), \tag{2}$$

where Q_i^l and K_i^l are the query and key matrices derived from the token embeddings Z^l at layer l, H is the number of attention heads. The dimension of attention map $\mathcal{A}^l$ is $(n+1)\times(n+1)$. The average attention weight is as:

$$\mathcal{A}^{act} = \frac{1}{L}\sum_{l=1}^{L} \mathcal{A}^l, \tag{3}$$

where $\mathcal{A}^{act}$ denotes the self-attention correlations among image patches. The attention weights between the class token and each image patch token reflect the importance of the corresponding patch. Therefore, we extract the first row of the self-attention matrix (excluding the class token) to represent the spatial attention distribution: $\mathcal{A}^{\text{self}} = \{\mathcal{A}^{\text{act}}_{\text{cls},1}, \mathcal{A}^{\text{act}}_{\text{cls},2}, \ldots, \mathcal{A}^{\text{act}}_{\text{cls},n}\}$, where $\mathcal{A}^{\text{self}}$ represents self-attention weights. Additionally, we compute the cosine similarity between the text feature from the text template "object" and the patch features from the final layer to obtain cross-attention weights. The saliency map $\mathcal{S}^m$, i.e., the hybrid attention is computed by fusing the self-attention and cross-attention weights, highlighting target object patches while suppressing background regions:

$$\mathcal{S}^m = \frac{1}{2}\left(\mathcal{A}^{\text{self}} + \mathcal{A}^{\text{cross}}\right). \tag{4}$$

To obtain the object-focused representation, we employ a grid sampling that amplifies salient regions while preserving

contextual information, guided by the previously computed saliency map. Specifically, for an input image $I(x, y)$, object amplification involves constructing a coordinate mapping that scales up high-saliency areas and suppresses low-saliency ones through two functions, $h(x, y)$ and $g(x, y)$. The resulting object-focused image is then obtained as:

$$I^F(x, y) = I(h(x, y), g(x, y)).$$

The functions $h(x, y)$ and $g(x, y)$ are designed to remap pixel coordinates proportionally according to the normalized weights of the saliency map $\mathcal{S}^m(u, v)$. An approximate solution requires $h(x, y)$ and $g(x, y)$ to satisfy as:

$$\int_0^{h(x,y)} \int_0^{g(x,y)} \mathcal{S}^m(u, v)\, du\, dv = xy. \tag{5}$$

Following [14], the solution is given by:

$$h(x, y) = \frac{\sum_{u,v} \mathcal{S}^m(u, v)\, k((u, v), (x, y))\, u}{\sum_{u,v} \mathcal{S}^m(u, v)\, k((u, v), (x, y))}, \tag{6}$$

$$g(x, y) = \frac{\sum_{u,v} \mathcal{S}^m(u, v)\, k((u, v), (x, y))\, v}{\sum_{u,v} \mathcal{S}^m(u, v)\, k((u, v), (x, y))}, \tag{7}$$

where $k(\cdot, \cdot)$ denotes a Gaussian distance kernel. This kernel acts as a regularizer to prevent degenerate cases, such as uniform pixel values across the image. In this manner, the salient object is magnified according to the saliency map.

The image I^F and hybrid attention $\mathcal{S}^m$ are fused like Alpha-CLIP [15], which is then projected into patch tokens using the same operations, with an additional learnable fine-grained visual prompt z_{FP} appended to the end of the token sequence, resulting in $Z^F \in \mathbb{R}^{(n+2)\times d} = [z^F_{\text{cls}}, z^F_1, z^F_2, \ldots, z^F_n, z^F_{\text{FG}}]$. We then pass this sequence through $L - 1$ frozen Transformer layers, followed by a newly added trainable Transformer layer to model the token representations. The output of z_{FP}, denoted as $f^Z(I^F)$, serves as the Zoom-in-based Object-aware visual feature that fully capture discriminative visual clues.

B. Dynamic Textual Enhancement

In few-shot image classification, the semantic discriminability between category labels is often limited due to sparse and highly similar textual descriptions. For instance, labels such as "A380" and "A320" refer to different aircraft but exhibit minimal semantic variation, making it difficult for text encoders to generate sufficiently discriminative classifier space. As a result, relying solely on textual information may lead to ambiguous and inadequate semantic representations. In contrast, the object attribute information from the training set provides critical cues that help ground textual descriptions, thereby enhancing model interpretability.

To this end, we propose incorporating zoom-in-based feature of patches into text prompts to enrich the fine-grained representation of textual features. Specifically, in Step 2, the zoomed-in image features, particularly the patch-level features $f^Z(I^F)_{p1,\ldots,n}$, have already been extracted and capture object-focused visual content. We aggregate the patch features from training images belonging to the same category by averaging, yielding the class-level patch feature $f^Z_{\text{avg}}(I^F)_{p1,\ldots,n}$. These features are then projected through a learnable prototype network, consisting of two convolutional layers followed by a linear layer, to obtain the object attribute prototype VP_k, where k denotes the class index. The visual attribute prototype VP_k is prepended to the text prompt token sequence to form the input to the text encoder:

$$f^t(VP_k; t_k) = f^t([t_{\text{SOS}}, VP_k, t_1, \ldots, t_m, t_{\text{EOS}}]), \tag{8}$$

where t_{SOS} and t_{EOS} are special tokens indicating the start and end of the sequence. Following CLIP, the output token embedding of t_{EOS} is then used as the prototype-augmented textual prompt $f^t(VP_k; t_k)$, serving as enhanced textual features for fine-grained matching. Therefore, the prediction probability of the object-aware branch is as:

$$p(y = i|I)_z = \frac{\exp(\text{sim}(f^t(VP_i; t_i), f^Z(I^F)))}{\sum_{j=1}^{k} \exp(\text{sim}(f^t(VP_j; t_j), f^Z(I^F)))}, \tag{9}$$

In our method, we aggregate the prediction probability of two branches as the final prediction probability as:

$$p(y = i|I) = p(y = i|I)_g + \alpha \times p(y = i|I)_f, \tag{10}$$

IV. Experiments

A. Experimental Setting

Dataset Setup and Evaluation Metric. To evaluate our FineCAP method, we conduct experiments on 11 widely used few-shot image recognition benchmarks, following the State-Of-The-Art (SOTA) FSL methods Tip-Adapter [6] and 2SFS [13]. The datasets include ImageNet [18], Caltech101 [19], OxfordPets [20], StanfordCars [21], Flowers102 [22], Food101 [23], FGVCAircraft [24], SUN397 [25], DTD [26], EuroSAT [27], and UCF101 [28], covering diverse tasks. For fair comparison with existing methods, all models, except zero-shot CLIP are trained using 1, 2, 4, 8, and 16 images per class as the training set, corresponding to the 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot settings. We adopt top-1 classification accuracy as the evaluation metric to evaluate the performance of FineCAP and all compared methods.

Implementation Details. We adopt CLIP as the backbone, using ViT-B/16 as the image encoder and a Transformer network as the text encoder. The pre-trained CLIP weights are loaded and kept frozen during training. We follow the same data pre-processing method as Tip-Adapter, including image resizing and random horizontal flipping. The handcrafted prompt ensemble is used for ImageNet, while the single handcrafted prompt is adopted for the other 10 datasets. The hyperparameter α, which controls the fine-grained branch weight, is set based on the results on the validation set. During training, the batch size is set to 256, and the number of epochs is set to 75 for ImageNet, 400 for EuroSAT, and 200 for all other datasets. We use AdamW as optimizer, with a learning rate of 2e-3 for the added network and 1e-3 for the others. All experiments are conducted on one NVIDIA A800 GPU.

TABLE I

COMPARISON WITH SOTA FSL METHODS ON THE IMAGENET UNDER 1, 2, 4, 8, AND 16-SHOT SETTINGS. OUR FINECAP CONSISTENTLY OUTPERFORMS ALL COMPARED METHODS ACROSS ALL SETTINGS. BOLD AND UNDERLINED VALUES DENOTE THE BEST AND SECOND-BEST ACCURACY.

Methods	**Venue**	**Backbone**	**K-Shot Accuracy** (%)				
			1	2	4	8	16
Zero-shot CLIP [3]	ICML'21	ResNet 50			60.3		
Zero-shot CLIP [3]	ICML'21	ViT-B/16			68.7		
CoOp [4]	IJCV'22	ResNet 50	57.2	57.8	60.0	61.6	63.0
CLIP-Adapter [16]	IJCV'24	ResNet 50	61.2	61.5	61.8	62.7	63.6
LDC [8]	CVPR'25	ResNet 50	60.5	61.3	62.5	64.4	66.6
CoOp [4]	IJCV'22	ViT-B/16	66.3	67.1	68.7	70.6	71.9
Tip-Adapter [6]	ECCV'22	ViT-B/16	67.6	68.6	69.8	71.4	73.3
PromptSRC [17]	ICCV'23	ViT-B/16	68.1	69.8	71.1	72.3	73.2
LDC [8]	CVPR'25	ViT-B/16	69.5	69.9	71.0	<u>72.5</u>	<u>73.9</u>
2SFS [13]	CVPR'25	ViT-B/16	68.8	69.8	<u>71.2</u>	72.2	73.6
FineCAP (Ours)	This Paper	**ViT-B/16**	**70.3**	**71.2**	**72.5**	**73.8**	**75.1**

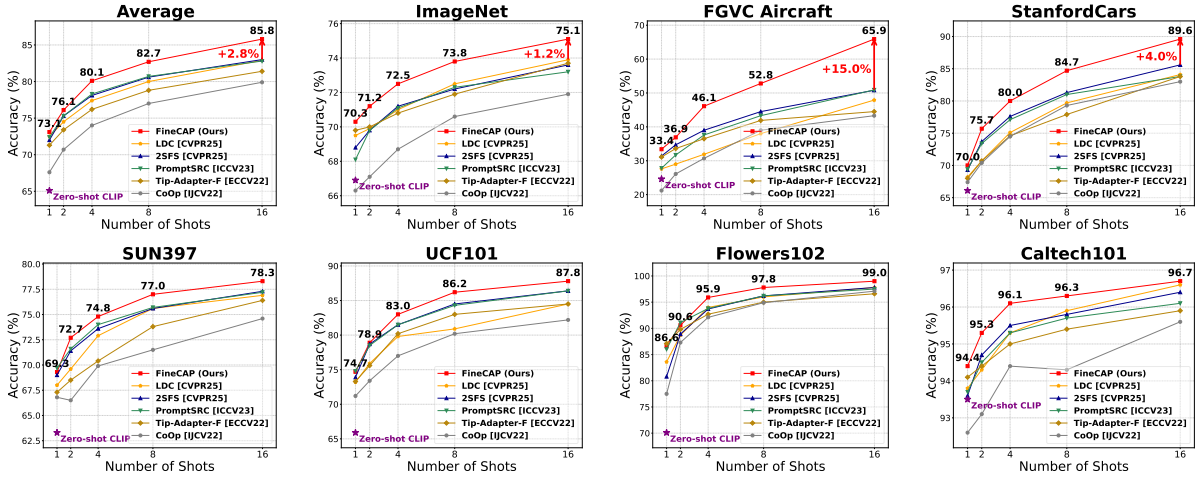

Fig. 3. Comparisons with SOTA FSL methods on average performance across 11 datasets and seven representative datasets, under 1, 2, 4, 8, and 16-shot settings. Our FineCAP model achieves significant performance improvements across all settings, especially on the fine-grained dataset FGVC, where it surpasses the suboptimal method by 15% in the 16-shot setting.

TABLE II

COMPARISON OF THE AVERAGE RESULTS OF DIFFERENT FSL METHODS OVER 11 DATASETS USING THE VIT-B/16 BACKBONE.

Methods (ViT-B/16)	**K-Shot Accuracy** (%)				
	1	2	4	8	16
Zero-shot CLIP (ICML'21)			65.1		
CoOp (IJCV'22)	67.6	70.7	74.0	77.0	79.9
Tip-Adapter-F (ECCV'22)	71.3	73.4	76.2	78.8	81.4
PromptSRC (ICCV'23)	72.4	75.3	<u>78.3</u>	<u>80.7</u>	82.8
LP++ (CVPR'24)	-	-	75.6	78.2	80.5
MMA (CVPR'24)	-	-	72.5	77.4	81.0
LDC (CVPR'25)	<u>72.5</u>	74.5	77.4	80.0	82.9
2SFS (CVPR'25)	72.0	<u>75.3</u>	78.1	80.6	<u>83.0</u>
FineCAP (Ours)	**73.1**	**76.1**	**80.1**	**82.7**	**85.8**

B. Comparisons with State-of-The-Art Models

We conduct extensive experiments comparing our FineCAP with SOTA FSL methods across 11 benchmark datasets under 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot settings. The results are reported in Table I, Table II, and Figure 3. As shown in Table I, on the challenging large-scale ImageNet dataset comprising 50,000 test images from 1000 classes, our FineCAP outperforms all compared methods across all few-shot settings. This performance gain demonstrates FineCAP's ability to capture object-centric visual content and jointly refine visual and textual features.

Table II presents a comparison of image recognition performance across all 11 datasets. FineCAP achieves the highest accuracy in every few-shot setting, outperforming the optimal compared method by an average of 2.8% in the 16-shot setting. As shown in Figure 3, our method consistently delivers superior results across all datasets and shot levels. It significantly outperforms other FSL methods, especially on fine-grained

TABLE III
ABLATION STUDIES ON DIFFERENT PROPOSED COMPONENTS. **VEM** INDICATES FVEM WITHOUT ZOOM-IN OPERATION.

Methods (ViT-B/16)	K-Shot Accuracy (%)				
	1	2	4	8	16
Zero-shot CLIP (ICML'21)	24.6				
+VEM	30.5	35.1	44.2	50.4	62.7
+FVEM	31.6	36.3	45.1	52.3	64.2
+DTEM	32.1	36.5	45.5	52.5	64.4
+FVEM+DTEM (FineCAP)	**33.4**	**36.9**	**46.1**	**52.8**	**65.9**

benchmarks such as FGVC Aircraft and StanfordCars. This highlights its enhanced ability to capture subtle visual object details and learn more robust classification boundaries through visual and textual prompts guidance.

C. Ablation Experiments

We investigate the effects of different components and report results on FGVC in Table III. **VEM** denotes the performance gain achieved by introducing a learnable visual prompt without the zoom-in operation. Notably, by leveraging saliency map detection and enhancing the content of the main object, the performance of **FVEM** is significantly improved. As observed, the dynamic text enhancement mechanism **DTEM** achieves greater improvement than **FVEM**. This is because image features already possess strong discriminative capabilities. In contrast, textual prompts often rely on minor token-level differences that provide limited performance gains. By introducing object attributes into the textual prompt, we obtain more robust classification boundaries. Furthermore, when both mechanisms are utilized, the object-focused visual content provides stronger visual grounding for the textual prompt, while the enriched textual prompt in turn guides the refinement of visual features, forming a cyclic optimization process.

D. Experiments on Different Vision Backbones

As shown in Table IV, our proposed FineCAP achieves the best performance on both vision backbones compared to other SOTA FSL methods. By leveraging a stronger vision backbone, FineCAP demonstrates improved recognition accuracy, indicating its extensibility across different architectures. With a stronger backbone, FineCAP can capture more precise object distributions, enabling enhanced object-focused representation. This, in turn, enriches the textual features with discriminative object attributes information.

E. Parameter Experiments

We conduct parameter experiments on α in Eq. 10 using the FGVC dataset under the 16-shot setting to explore the importance of fine-grained branch. The results are summarized in Figure 4. Performance steadily improves as α increases, reaching a maximum of 65.9% at $\alpha = 1$. For $\alpha > 1$, performance declines, as the pre-trained knowledge from global branch is progressively diminished. Thus, we set $\alpha = 1$ to achieve an optimal balance between the two branches.

TABLE IV
IMAGE CLASSIFICATION ACCURACY (%) COMPARISONS WITH STATE-OF-THE-ART METHODS ON FGVC [24] USING DIFFERENT VISION BACKBONES IN THE 16-SHOT SETTING.

Methods	ViT-B/16	ViT-L/14
Zero-shot CLIP (ICML'21)	24.7	32.6
CoOp (IJCV'22)	43.3	55.2
Tip-Adapter (ECCV'22)	44.5	55.8
KgCoOp (CVPR'23)	36.5	47.5
2SFS (CVPR'25)	50.8	64.1
FineCAP (Ours)	**65.9**	**70.1**

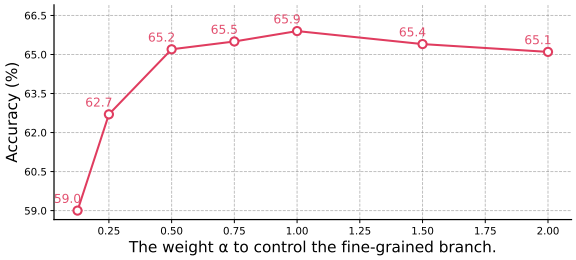

Fig. 4. Hyperparameter sweep on α conducted on FGVC.

F. Qualitative Results

We present a visualization of the zoom-in operation in Figure 5. The self-attention and cross-attention maps, denoted as $\mathcal{A}^{self}$ and $\mathcal{A}^{cross}$, are extracted from the global branch. As observed, $\mathcal{A}^{cross}$ localizes the main object while suppressing irrelevant background regions, whereas $\mathcal{A}^{self}$ attends to fine-grained visual details within the object. By combining these two attention maps to get the hybrid attention, a saliency map is generated, which highlights the principal region of the visual object. The salient regions are then amplified to produce an object-focused image, which serves as a visual prior to guide the image encoder in learning more discriminative features for fine-grained recognition.

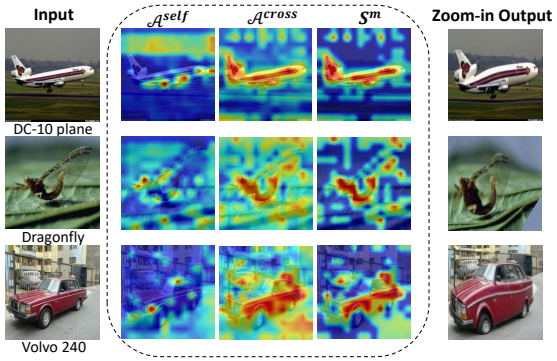

Fig. 5. Visualizations of saliency map detection and zoom-in operations.

V. CONCLUSION

This paper proposes FineCAP, a cyclic augmentation prompting framework for fine-grained CLIP-based few-shot

learning. We identify two key issues: weak fine-grained visual encoding and vision-text misalignment. To address these issues, FineCAP introduces a Fine-grained Visual Enhancement Mechanism to focus on objects and a Dynamic Textual Enhancement Mechanism to enhance textual features with object attributes. These components work synergistically for superior few-shot classification performance.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[4] 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.

[5] 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*, 2022, pp. 16 816–16 825.

[6] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.

[7] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie, "Mma: Multi-modal adapter for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 826–23 837.

[8] S. Li, F. Liu, Z. Hao, X. Wang, L. Li, X. Liu, P. Chen, and W. Ma, "Logits deconfusion with clip for few-shot learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 411–25 421.

[9] B. Wu, W. Shi, J. Wang, and M. Ye, "Synthetic data is an elegant gift for continual vision-language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2813–2823.

[10] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[11] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[12] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," *arXiv preprint arXiv:2110.05208*, 2021.

[13] M. Farina, M. Mancini, G. Iacca, and E. Ricci, "Rethinking few-shot adaptation of vision-language models in two stages," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 989–29 998.

[14] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: a saliency-based sampling layer for neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 51–66.

[15] Z. Sun, Y. Fang, T. Wu, P. Zhang, Y. Zang, S. Kong, Y. Xiong, D. Lin, and J. Wang, "Alpha-clip: A clip model focusing on wherever you want," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 13 019–13 029.

[16] 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.

[17] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15 190–15 200.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[21] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[22] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.

[23] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European conference on computer vision*. Springer, 2014, pp. 446–461.

[24] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[26] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[27] P. Helber, B. Bischke, A. Dengel, and D. 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[28] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.