
VT-FSL: Bridging Vision and Text with LLMs for Few-Shot Learning

Wenhao Li^{1,2}, Qiangchang Wang^{1*}, Xianjing Meng³, Zhibin Wu¹, Yilong Yin^{1*}

¹School of Software, Shandong University ²Shenzhen Loop Area Institute

³School of Computing and Artificial Intelligence, Shandong University of Finance and Economics

{wenhao.li, zhibinwu}@mail.sdu.edu.cn,

rongmengyuan@gmail.com, {qiangchang.wang, ylyin}@sdu.edu.cn

Abstract

Few-shot learning (FSL) aims to recognize novel concepts from only a few labeled support samples. Recent studies enhance support features by incorporating additional semantic information (e.g., class descriptions) or designing complex semantic fusion modules. However, these methods still suffer from hallucinating semantics that contradict the visual evidence due to the lack of grounding in actual instances, resulting in noisy guidance and costly corrections. To address these issues, we propose a novel framework, bridging Vision and Text with LLMs for Few-Shot Learning (VT-FSL), which constructs precise cross-modal prompts conditioned on Large Language Models (LLMs) and support images, seamlessly integrating them through a geometry-aware alignment mechanism. It mainly consists of Cross-modal Iterative Prompting (CIP) and Cross-modal Geometric Alignment (CGA). Specifically, the CIP conditions an LLM on both class names and support images to generate precise class descriptions iteratively in a single structured reasoning pass. These descriptions not only enrich the semantic understanding of novel classes but also enable the zero-shot synthesis of semantically consistent images. The descriptions and synthetic images act respectively as complementary textual and visual prompts, providing high-level class semantics and low-level intra-class diversity to compensate for limited support data. Furthermore, the CGA jointly aligns the fused textual, support, and synthetic visual representations by minimizing the kernelized volume of the 3-dimensional parallelotope they span. It captures global and nonlinear relationships among all representations, enabling structured and consistent multimodal integration. The proposed VT-FSL method establishes new state-of-the-art performance across ten diverse benchmarks, including standard, cross-domain, and fine-grained few-shot learning scenarios. Code is available at <https://github.com/peacelwh/VT-FSL>.

1 Introduction

Deep learning has achieved remarkable success in computer vision [1–3] and natural language processing [4] thanks to the availability of large-scale annotated datasets. However, collecting such data is often expensive or infeasible in many real-world scenarios [5–10]. Few-shot learning (FSL) [11] aims to address this challenge by enabling models to generalize from only a few labeled samples.

In FSL, the support set provides N novel classes, each with K labeled samples. The model learned from the support set is required to classify test samples from the query set. Among various approaches, metric-based methods [11–22] are widely adopted due to their superior scalability and performance.

*Corresponding authors.

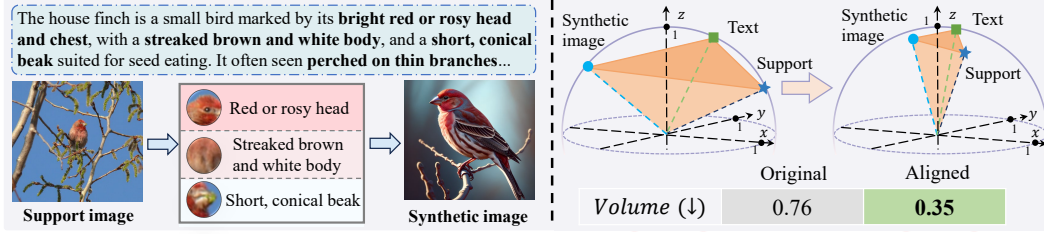


Figure 1: Illustration of the VT-FSL intuition. Left: The generated text and synthetic images provide high-level class semantics and low-level sample diversity. Right: By minimizing the volume of the 3-dimensional parallelotope spanned by all embeddings, they lie closer, indicating better alignment.

These methods embed both support and query samples into a shared feature space and classify each query by finding the nearest support sample. A common strategy is to construct class prototypes by averaging support features within each class to represent its semantic center [11]. However, the limited number of labeled examples makes it challenging to learn discriminative prototype representations, resulting in a semantic deviation from the true class center.

To address this issue, an alternative is to integrate additional semantic information from textual modality. Following this perspective, several studies [23–28] reveal that integrating semantics from class names can improve prototype representation. However, class names offer minimal contextual information. Several recent [29–31] works attempt to leverage Large Language Models (LLMs) [4] or external knowledge bases [32] to generate richer descriptions or attribute-based semantics to replace names. Despite their success, they rely solely on class names, neglecting valuable visual patterns of the support images. As a result, the generated text may lead to semantic hallucinations, i.e., misalignment with the actual corresponding object, which requires costly manual or algorithmic corrections. Moreover, naive input prompting hinders LLMs from fully utilizing the reasoning and generation capabilities, limiting the quality of semantics in few-shot scenarios.

In this paper, we propose a novel framework, bridging Vision and Text with LLMs for Few-Shot Learning (VT-FSL). First, a Cross-modal Iterative Prompting (CIP) module is introduced to condition LLMs [33, 34] on both class names and support images, obtaining precise and visually grounded descriptions. This is accomplished by a single structured inference pass, comprising four distinct stages for iteratively optimizing text quality: strategy, perception, refinement, and conclusion. Next, to enrich intra-class sample diversity, a text-to-image model [35, 36] is utilized to generate synthetic images with semantic consistency based on these descriptions in a zero-shot manner. As shown in Fig. 1 (left), the resulting text and images serve as complementary textual and visual prompts to compensate for limited support data. To fully utilize these cross-modal prompts, we incorporate textual embeddings from CLIP [37] into support features along spatial and channel dimensions via a lightweight two-layer perceptron. Furthermore, we propose a novel Cross-modal Geometric Alignment (CGA) that leverages the volume of a 3-dimensional parallelotope spanned by the fused support, textual, and synthetic visual embeddings for consistent alignment, as shown in Fig. 1 (right). CGA establishes a contrastive learning objective to enhance global and nonlinear relationships among all features by minimizing their volume in a kernelized parallelotope embedding space. Finally, VT-FSL achieves comprehensive cross-modal integration, enabling the extraction of generalized class prototypes enriched with discriminative information.

Overall, our contributions are summarized as follows:

- A VT-FSL framework is proposed to construct complementary cross-modal prompts with large language models, seamlessly integrating them through a geometry-aware alignment.
- A CIP module is proposed to generate precise descriptions conditioned on both class names and support images, driving zero-shot synthesis of semantically consistent examples.
- A CGA module is proposed to achieve comprehensive alignment across all representations and capture nonlinear semantic relations by kernelized volume-based contrastive learning.
- The proposed method achieves state-of-the-art performance on ten standard, cross-domain, and fine-grained FSL benchmarks, significantly improving accuracy by 4.2% on average.

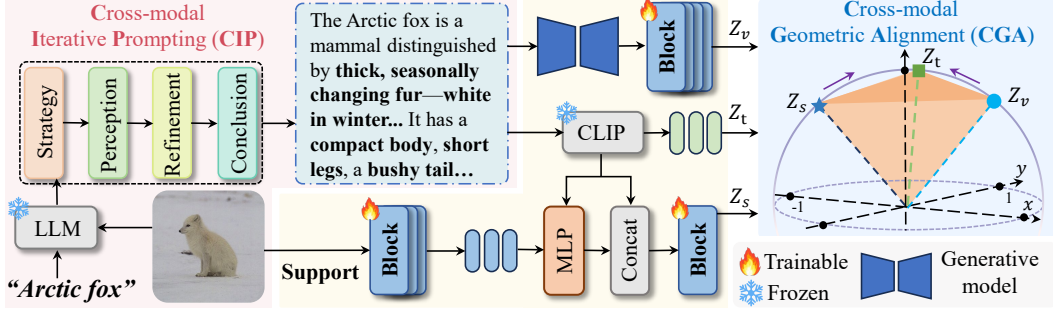


Figure 2: Overview of the proposed VT-FSL framework. First, given both class names and support images, CIP guides an LLM to generate precise descriptions via four structured stages. Synthetic images with semantic consistency are then generated based on these descriptions to expand the limited data. They are extracted to obtain features Z_v by a feature extractor consisting of multiple Transformer blocks with shared weights. Next, the textual features Z_t encoded by CLIP are injected into the support features z_s via a two-layer MLP, enhancing the support embeddings Z_s . Finally, Z_s , Z_t , and Z_v are jointly aligned through CGA, enabling global and nonlinear cross-modal interactions.

2 Related Work

Few-shot Learning. Existing few-shot learning (FSL) methods can be broadly categorized into two types. Optimization-based methods [38–41] adapt models to novel tasks with a few optimization steps, but may lead to meta-overfitting [42–44] due to the scarcity of task-specific supervision. In contrast, metric-based methods [11–22] learn a generalized embedding space where inter-class distances are maximized and intra-class distances are minimized. Several works also incorporate self-supervised learning to refine feature representations [20, 45–48]. To compensate for rare representative features within limited support data, many methods [23–28, 49] incorporate additional semantic information. For example, AM3 [25] fuses semantic features from class names with visual prototypes through an adaptive fusion mechanism. CaFo [28] generates synthetic images based on class names to expand the few-shot data. SIFT [49] generates high-quality features via a semantic transformation process. Several recent methods [29–31] explore the use of Large Language Models (LLMs) to replace class names with richer textual information. Conditioned on class names, ECER [30] extracts attribute-level textual information, and SemFew [31] constructs coherent descriptions to enhance prototype learning. In contrast, our approach jointly leverages both class names and support images to generate visually grounded textual descriptions via structured LLM reasoning and further produces semantically consistent synthetic images, forming complementary cross-modal prompts. Moreover, we fully integrate these prompts with support features in a geometry-aware alignment manner, capturing global and nonlinear cross-modal relationships.

Contrastive Learning. Most existing methods [50–55] adopt CLIP-based pairwise contrastive learning [37]. However, aligning each representation only to a single anchor neglects the interactions among the remaining points, making it difficult to capture global structural relationships and limiting the full integration of cross-modal semantics. The Gram matrix, which characterizes the mutual geometry among sets of vectors, has shown promise in theoretical analyses of deep learning networks [56] and various downstream tasks [57, 58]. To the best of our knowledge, we are the first to consider volume-based contrastive learning for global consistent alignment and further capture nonlinear relationships within a kernelized parallelotope embedding space in few-shot learning.

3 Methodology

We begin by introducing the preliminaries of FSL, and then present two components of our method: (1) explaining how textual and visual prompts are generated through CIP, and (2) how all embeddings are simultaneously aligned via CGA. Fig 2 shows the overview of the proposed VT-FSL.

3.1 Preliminaries

Few-Shot Learning (FSL) focuses on generalizing the knowledge learned from training set C_{train} to test set C_{test} , where both sets are disjoint ($C_{\text{train}} \cap C_{\text{test}} = \emptyset$). FSL is typically formalized as an

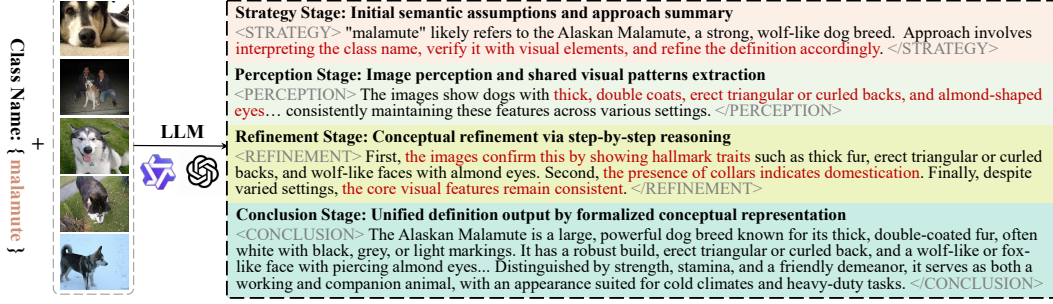


Figure 3: Illustration of Cross-modal Iterative Prompting (CIP). Given the class name and 5-shot support samples, CIP exploits the LLM through four structured reasoning stages: Strategy, Perception, Refinement, and Conclusion, to generate class-specific, precise class descriptions.

N -way K -shot classification task using an episodic strategy [12]. Each task consists of a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N \times K}$ and a query set $\mathcal{Q} = \{(x_i, y_i)\}_{i=1}^{N \times M}$ for performance evaluation.

In the FSL approach, prototype-based inference is commonly utilized, computing a prototype embedding c_i for each class i by averaging support features from the feature extractor f_ϕ as follows:

$$c_i = \frac{1}{K} \sum_{k=1}^K f_\phi(x_k), \quad x_k \in S_i. \quad (1)$$

For a query sample $\mathbf{x}^q$, the similarity score between $\mathbf{x}^q$ and all support classes is determined by calculating the distance function between $\mathbf{x}^q$ and N class prototypes as follows:

$$p(y^q = i | \mathbf{x}^q) = \frac{\exp(\cos(f_\phi(\mathbf{x}^q), c_i)/\tau)}{\sum_{j=1}^N \exp(\cos(f_\phi(\mathbf{x}^q), c_j)/\tau)}, \quad (2)$$

where τ is a temperature parameter. $\cos(\cdot, \cdot)$ denotes the cosine distance function. The support class with the maximum score is regarded as the classification result.

3.2 Cross-modal Iterative Prompting

To address the ambiguity in class names and capture the shared visual patterns from support samples, we propose Cross-modal Iterative Prompting (CIP), which jointly leverages the class label and K -shot support images to generate precise and visually grounded class descriptions. CIP follows a structured reasoning process inspired by Chain-of-Thought (CoT), which improves the semantic interpretability and generation accuracy of Large Language Models (LLMs).

CIP decomposes the generation process into four structured reasoning stages, including strategy, perception, refinement, and conclusion, as shown in Fig. 3. Every stage guided by the prompting design corresponds to outlining the problem, interpreting relevant information from the image, proceeding with a step-by-step reasoning process, and ultimately reaching a well-supported conclusion, respectively. Each stage is marked using structured tags (e.g., <STRATEGY>...</STRATEGY>), which allows the LLM to maintain reasoning boundaries throughout a single inference pass. This design removes the need for multi-turn interactions or external filtering, reducing manual effort and latency. The entire structural reasoning process and more cases are detailed in the Appendix.

As shown in Fig. 4, the generated description is then fed into a text-to-image generative model to produce synthetic images with semantic consistency in a zero-shot manner. To ensure semantic fidelity, we incorporate an LLM-based pairwise comparison strategy, which selects the top- K images per class by ranking them against the textual description. An enriched support set is obtained, containing both real and synthetic samples without compromising the low-data regime. Formally, an N -way $(K + K)$ -shot support set can be constructed from N textual descriptions T_N as:

$$\hat{\mathcal{S}}_{N \times (K+K)} = \{\text{Synthesize}(T_N), \mathcal{S}_{N \times K}\}. \quad (3)$$

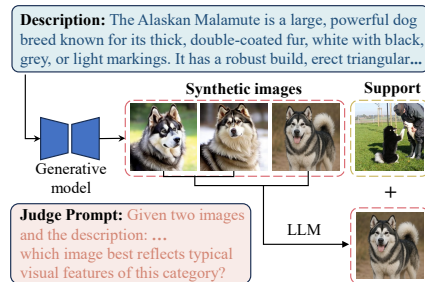


Figure 4: Illustration of visual synthetic images in the 1-shot task.

3.3 Cross-modal Geometric Alignment

Cross-modal Fusion. First, we leverage textual prompts to adaptively adjust support features extraction based on rich semantics. The textual features Z_t encoded from the CLIP are linearly mapped to the same embedding space, followed by integrating with support features z_s by a lightweight and effective two-layer network, as follows:

$$\beta = \sigma(W_2 \sigma(W_1 [Z_t; \text{Avg}(z_s)])), \quad (4)$$

where W_i and W_2 are parameters of the network, σ is sigmoid activation function. $\text{Avg}(\cdot)$ denotes the average of all patch tokens. The vector β is injected into each token of support features, modulating support features along the channel dimension.

Z_t and z_s are then concatenated along the spatial dimension to capture the relevance between tokens based on rich semantics via Multi-head Self-Attention (MSA) of the Transformer block:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \quad (5)$$

$$Z_s = (AV)W, \quad (6)$$

where Q , K , and V are linearly mapped by each token. d is the dimension of each head. The attention matrix A is used to aggregate information from the value V . The outputs from all attention heads are concatenated and projected linearly by the weight matrix W , obtaining the enhanced Z_s .

Kernelized Volume-based Contrastive Learning. To achieve comprehensive consistency alignment, we propose a kernelized volume-based contrastive objective to improve global and nonlinear interactions among all features. Unlike traditional pairwise contrastive objectives, we measure the alignment among multiple vectors via the volume of the k -dimensional parallelotope they span in a shared embedding space [59]. A smaller volume indicates closer alignment. Specifically, K normalized embeddings $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$ construct the matrix $\mathbf{A} = [\mathbf{v}_1, \dots, \mathbf{v}_k]$ and their extremities lie on the surface of the unit hypersphere. The Gram matrix $\mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) \in \mathbb{R}^{k \times k}$ is first defined, reflecting the square of the corresponding volume Vol as follows:

$$\mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \mathbf{A}^\top \mathbf{A} = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_1, \mathbf{v}_k \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_2, \mathbf{v}_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_k, \mathbf{v}_1 \rangle & \langle \mathbf{v}_k, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_k, \mathbf{v}_k \rangle \end{bmatrix}. \quad (7)$$

$$\text{Vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \sqrt{\det(\mathbf{G})}, \quad \mathbf{G}_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle. \quad (8)$$

To further model nonlinear interactions, this metric is extended into a high-dimensional Reproducing Kernel Hilbert Space (RKHS) using a Radial Basis Function (RBF) kernel mapping $\kappa(\cdot, \cdot)$. The volume is computed from the kernel Gram matrix $\mathbf{K}$:

$$\text{Vol}_{\mathcal{H}}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \sqrt{\det(\mathbf{K})}, \quad \mathbf{K}_{ij} = \kappa(\mathbf{v}_i, \mathbf{v}_j). \quad (9)$$

Next, Textual, synthetic, and enhanced support features are transformed into normalized triplets $(\mathbf{Z}_t, \mathbf{Z}_v, \mathbf{Z}_s)$. Selecting textual modality as the anchor focuses the alignment around it. Formal derivations of relevant theory and ablations on anchor and kernel function choices are detailed in the Appendix. Minimizing the kernelized volume via the contrastive loss is defined as:

$$\mathcal{L}_{\text{D2A}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(-\text{Vol}_{\mathcal{H}}(Z_t^i, Z_s^i, Z_v^i)/\tau)}{\sum_{j=1}^K \exp(-\text{Vol}_{\mathcal{H}}(Z_t^i, Z_s^j, Z_v^i)/\tau)}, \quad (10)$$

$$\mathcal{L}_{\text{A2D}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(-\text{Vol}_{\mathcal{H}}(Z_t^i, Z_s^i, Z_v^i)/\tau)}{\sum_{j=1}^K \exp(-\text{Vol}_{\mathcal{H}}(Z_t^i, Z_s^j, Z_v^j)/\tau)}. \quad (11)$$

Finally, the overall loss also includes the Cross-Entropy loss between the probability p_i of the query sample q to the i -th class in Eq. (2) and the corresponding ground-truth label, as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^M \text{CrossEntropy}(p_i, y_i) + \frac{1}{2}(\mathcal{L}_{\text{D2A}} + \mathcal{L}_{\text{A2D}}). \quad (12)$$

Table 1: Results (%) on *miniImageNet* [12] and *tieredImageNet* [60]. The average accuracy with 95% confidence interval is reported. Bold and blue font indicates the best and suboptimal results.

Model	Venue	Backbone	$\approx$ # Params	<i>miniImageNet</i>		<i>tieredImageNet</i>	
				1-shot	5-shot	1-shot	5-shot
MatchNet [12]	NeurIPS'16	ResNet-12	12.4M	65.64 $\pm$ 0.20	78.72 $\pm$ 0.15	68.50 $\pm$ 0.92	80.60 $\pm$ 0.71
ProtoNet [11]	NeurIPS'17	ResNet-12	12.4M	62.39 $\pm$ 0.21	80.53 $\pm$ 0.14	68.23 $\pm$ 0.23	84.03 $\pm$ 0.16
AM3 [25]	NeurIPS'19	ResNet-12	12.4M	65.30 $\pm$ 0.49	78.10 $\pm$ 0.36	69.08 $\pm$ 0.47	82.58 $\pm$ 0.31
DeepEMD [14]	CVPR'20	ResNet-12	12.4M	65.91 $\pm$ 0.82	82.41 $\pm$ 0.56	71.16 $\pm$ 0.87	86.03 $\pm$ 0.58
PCPK [29]	CVPR'21	ResNet-12	12.4M	73.13 $\pm$ 0.85	82.06 $\pm$ 0.54	81.04 $\pm$ 0.89	87.42 $\pm$ 0.57
SUN [17]	ECCV'22	Visformer-S	12.3M	67.80 $\pm$ 0.45	83.25 $\pm$ 0.30	72.99 $\pm$ 0.50	86.74 $\pm$ 0.33
FewTure [47]	NeurIPS'22	ViT-S/16	22.0M	68.02 $\pm$ 0.88	84.51 $\pm$ 0.53	72.96 $\pm$ 0.92	86.43 $\pm$ 0.67
SVAE [27]	CVPR'22	ResNet-12	12.4M	74.84 $\pm$ 0.23	83.28 $\pm$ 0.40	76.98 $\pm$ 0.65	85.77 $\pm$ 0.50
Meta-AdaM [61]	NeurIPS'23	ResNet-12	12.4M	59.89 $\pm$ 0.49	77.92 $\pm$ 0.43	65.31 $\pm$ 0.48	85.24 $\pm$ 0.35
ProtoDiff [62]	NeurIPS'23	ResNet-12	12.4M	66.63 $\pm$ 0.21	83.48 $\pm$ 0.15	72.95 $\pm$ 0.24	85.15 $\pm$ 0.18
CPEA [63]	ICCV'23	ViT-S/16	22.0M	71.97 $\pm$ 0.65	87.06 $\pm$ 0.38	76.93 $\pm$ 0.70	90.12 $\pm$ 0.45
SP [23]	CVPR'23	Visformer-T	10.0M	72.31 $\pm$ 0.40	83.42 $\pm$ 0.30	78.03 $\pm$ 0.46	88.55 $\pm$ 0.32
NIW-Meta [64]	ICLR'24	WRN28-10	36.5M	68.54 $\pm$ 0.26	84.81 $\pm$ 0.28	74.59 $\pm$ 0.33	89.76 $\pm$ 0.23
FeatWalk [16]	AAAI'24	ResNet-12	12.4M	70.21 $\pm$ 0.44	87.38 $\pm$ 0.27	75.25 $\pm$ 0.48	89.92 $\pm$ 0.29
BECLR [65]	ICLR'24	ResNet-18	11.7M	75.74 $\pm$ 0.62	84.93 $\pm$ 0.33	76.44 $\pm$ 0.66	84.85 $\pm$ 0.37
SIFT [49]	IJCV'24	WRN-28-10	36.5M	77.31 $\pm$ 0.67	86.95 $\pm$ 0.53	77.86 $\pm$ 0.77	89.89 $\pm$ 0.52
SemFew [31]	CVPR'24	Swin-T	29.0M	78.94 $\pm$ 0.66	86.49 $\pm$ 0.50	82.37 $\pm$ 0.77	89.89 $\pm$ 0.52
UAP [66]	NeurIPS'24	ResNet-12	12.4M	81.63 $\pm$ 0.28	79.05 $\pm$ 0.19	79.68 $\pm$ 0.30	76.78 $\pm$ 0.21
ECER [30]	AAAI'25	Visformer-T	10.0M	81.14 $\pm$ 0.15	-	81.81 $\pm$ 0.51	-
VT-FSL	ours	Visformer-T	10.0M	83.66 $\pm$ 0.31	88.38 $\pm$ 0.25	88.02 $\pm$ 0.34	91.71 $\pm$ 0.27

We detail the algorithm of the entire training process in the Appendix. During testing, the prototype with textual prompts c_t is obtained by averaging the enhanced support features. The prototype with visual prompts c_v is computed by averaging the expanding support set in Eq. (3). The final classification prototype C is then obtained by integrating c_v and c_t in a convex combination manner [25]:

$$C = uc_t + (1 - u)c_v, \quad (13)$$

where $u \in [0, 1]$ is a manually controlled fusion factor, determined in the validation set.

4 Experiments

4.1 Experimental Details

Datasets. Extensive experiments are conducted in three distinct few-shot learning scenarios. (1) Four datasets in standard FSL: *miniImageNet* [12], *tieredImageNet* [60], CIFAR-FS [67], and FC100 [19]. (2) Three datasets in fine-grained FSL: CUB [68], Cars [69], and Dogs [70]. (3) Three datasets in cross-domain FSL: CUB, Places [71], and Plantae [72]. The detailed dataset statistics regarding the number of categories and images are introduced in the Appendix.

Implementation Details. Following recent few-shot studies [24, 23, 30], we adopt Visformer-Tiny [3] as the feature extractor and apply the text encoder from ViT-B/16 CLIP [37] where the output dimension is 512. Qwen2.5-VL-32B [34] and Janus-Pro [36] models are utilized to generate textual and visual prompts by default. More architecture comparisons are detailed in the Appendix. Our training framework adopts a two-stage framework [17], consisting of pre-training and meta-tuning stages. Input images are resized into 224 $\times$ 224 [31]. The AdamW optimizer [73] is used with a learning rate of 5e-4 and a cosine scheduler. Pre-training runs for 300 epochs in *tieredImageNet* and 800 epochs in other datasets with a batch size of 512, followed by meta-tuning for 100 epochs via an episodic training strategy. The hyperparameter τ is set as 0.2 according to validation accuracy. All experiments are performed with an NVIDIA RTX 6000 Ada.

Evaluation Protocol. For evaluation, we adopt the widely used episodic protocol [23–28] in FSL. Specifically, 2000 classification tasks are uniformly sampled from the novel classes that do not overlap with training categories. Each task follows the standard N -way K -shot setting, where 15 query samples per class are included for evaluation. The final performance is reported as the mean classification accuracy across all sampled tasks, along with the 95% confidence interval.

4.2 Main Results

Standard Few-Shot Classification. Table 1 and Table 2 show the comparative results on four benchmarks. It is worth noting that (1) Semantic-based methods (AM3, PCPK, SVAE, SP, SIFT,

SemFew, and ECER) generally outperform others, highlighting the effectiveness of semantic information in learning more generalizable feature representations. (2) The proposed VT-FSL uses a lightweight Visformer-T backbone yet outperforms all methods with larger backbones (e.g., ViT-S/16, Swin-T, and WRN28-10), demonstrating that VT-FSL fully utilizes the feature extraction potential of the backbone. (3) VT-FSL outperforms previous methods by a large margin. Compared with the competing methods on four benchmarks, VT-FSL surpasses SemFew [31] by 3.7%-5.7% and 1.0%-2.7% in the 1-shot and 5-shot settings. Moreover, compared to complex fusion mechanisms [25–27], VT-FSL achieves such improvements using a simple two-layer network with only 0.7M parameters, significantly fewer than the 4.3M parameters used in SemFew. These results highlight the effectiveness of constructing complementary multimodal prompts with global alignment across all features for full semantic integration.

Table 2: Results (%) with the average accuracy is reported on CIFAR-FS [67] and FC100 [19].

Method	Venue	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [11]	NeurIPS'17	72.20	83.50	41.54	57.08
MABAS [74]	ECCV'20	73.51	85.65	42.31	58.16
FewTURE [47]	NeurIPS'22	77.76	88.90	47.68	63.81
MAdaM [61]	NeurIPS'23	—	—	41.12	56.14
SP [23]	CVPR'23	82.18	88.24	48.53	61.55
ALFA [39]	TPAMI'24	76.32	86.73	44.54	58.44
LastShot [40]	TPAMI'24	76.76	87.49	44.08	59.14
SemFew [31]	CVPR'24	84.34	89.11	54.27	65.02
ECER [30]	AAAI'25	86.01	—	57.34	—
VT-FSL	ours	88.67	91.45	57.99	67.68

Table 3: Results (%) on CUB [68], Dogs [70], and Cars [69]. The average accuracy with 95% confidence interval is reported. Bold and blue font indicates the best and suboptimal results.

Method	Venue	CUB-200-2011		Stanford-Dogs		Stanford-Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [62]	NeurIPS'17	63.44±0.56	83.17±0.35	41.61±0.50	76.78±0.36	45.01±0.49	87.19±0.31
FRN [75]	CVPR'21	83.55±0.19	92.92±0.10	49.37±0.20	67.13±0.17	58.90±0.22	79.65±0.15
DAN [76]	AAAI'22	72.89±0.50	86.60±0.31	59.81±0.50	77.19±0.35	70.21±0.50	85.55±0.31
MFGN [77]	IJCAI'22	84.01±0.39	84.01±0.39	74.81±0.44	86.52±0.26	—	—
TDM [78]	CVPR'22	84.36±0.19	93.37±0.10	57.64±0.22	75.03±0.16	68.36±0.22	86.14±0.13
BSFA [79]	TCSVT'23	86.00±0.41	92.53±0.23	69.58±0.50	82.59±0.33	88.93±0.38	95.20±0.20
MLI [80]	TIP'24	85.94±0.42	93.50±0.29	76.32±0.47	88.25±0.27	—	—
C2-Net [81]	AAAI'24	83.31±0.41	92.18±0.23	75.50±0.49	87.65±0.28	88.96±0.37	95.16±0.20
SUITED [82]	AAAI'25	86.02±0.47	94.13±0.24	76.55±0.47	88.86±0.27	89.97±0.36	96.53±0.16
VT-FSL	ours	91.08±0.28	94.63±0.19	86.58±0.30	90.69±0.25	92.95±0.24	96.62±0.15

Fine-grained Few-Shot Classification. The classification results on three fine-grained datasets are presented in Table 3. It can be observed that the proposed VT-FSL method obtains the best classification results, outperforming the second-best SUITED [82] method by a significant margin of 3.0%-10.3% in the challenging 1-shot task across three benchmarks, and is far superior to other competing methods. This shows that our VT-FSL can also be effective on the fine-grained few-shot image classification tasks to capture subtle inter-class differences and preserve intra-class consistency by bridging the cross-modal semantic gap and enhancing multimodal integration.

Table 4: The average accuracy (%) is reported on cross-domain *mini*ImageNet [12] → CUB [68], Places [68], and Plantae [72]. Bold and blue font indicates the best and suboptimal results.

Method	Venue	CUB		Places		Plantae	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
GNN [83]	ICLR'18	44.40±0.68	62.87±0.65	52.42±0.80	70.91±0.65	33.60±0.56	48.51±0.59
FWT [84]	ICLR'20	45.50±0.75	64.97±0.68	53.44±0.79	70.70±0.67	32.56±0.58	49.66±0.62
AFA [85]	ECCV'22	46.86±0.70	68.25±0.65	54.04±0.75	76.21±0.60	36.76±0.65	54.26±0.68
UCD [86]	NeurIPS'22	40.65±0.68	58.54±0.70	51.84±0.72	72.19±0.60	37.28±0.67	54.15±0.66
ATA [87]	AIJ'23	45.00±0.50	66.22±0.50	53.57±0.50	75.48±0.40	34.42±0.40	52.69±0.40
LDP-net [88]	CVPR'23	49.82±0.70	70.39±0.66	53.82±0.71	72.90±0.63	39.84±0.68	58.49±0.69
StyleAdv [89]	CVPR'23	48.49±0.72	68.72±0.67	58.58±0.83	77.73±0.62	41.13±0.67	61.52±0.68
FAP [90]	IJCAI'24	50.56±0.73	64.17±0.69	57.34±0.72	72.05±0.60	37.44±0.64	53.58±0.66
FLOR [91]	CVPR'24	49.99±0.68	70.39±0.67	53.18±0.70	72.31±0.62	40.10±0.65	55.80±0.66
MEFP [92]	NeurIPS'24	51.55±0.70	73.61±0.66	52.06±0.69	73.78±0.61	41.55±0.65	61.39±0.67
SVasP [93]	AAAI'25	49.49±0.72	68.95±0.66	59.07±0.81	77.78±0.62	41.22±0.62	60.63±0.64
VT-FSL	ours	66.86±0.47	81.02±0.36	73.68±0.41	81.52±0.33	45.90±0.40	61.54±0.38

Cross-Domain Few-Shot Classification. Following the setup [85, 87], the model is trained on the training set of *mini*ImageNet and evaluated on three novel datasets: *mini*ImageNet → CUB [68], Places [68], and Plantae [72], which is more challenging due to significant domain shifts. As shown in

Table 5: Ablation study on three datasets under the 1-shot and 5-shot settings. P_{text} means textual prompts, and P_{vision} means visual prompts. $\mathcal{L}_{align}$ is kernelized volume-based contrastive loss.

P_{text}	P_{vision}	$\mathcal{L}_{align}$	<i>miniImageNet</i>		CIFAR-FS		<i>tieredImageNet</i>	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
			68.47 $\pm$ 0.43	82.63 $\pm$ 0.30	76.43 $\pm$ 0.45	86.60 $\pm$ 0.31	75.88 $\pm$ 0.37	87.39 $\pm$ 0.29
✓			78.82 $\pm$ 0.36	86.01 $\pm$ 0.27	84.76 $\pm$ 0.36	89.23 $\pm$ 0.29	86.15 $\pm$ 0.38	89.65 $\pm$ 0.31
✓		✓	78.96 $\pm$ 0.36	86.35 $\pm$ 0.29	86.54 $\pm$ 0.34	89.74 $\pm$ 0.29	86.91 $\pm$ 0.37	90.19 $\pm$ 0.29
	✓		79.17 $\pm$ 0.35	86.26 $\pm$ 0.28	86.01 $\pm$ 0.34	89.75 $\pm$ 0.28	85.54 $\pm$ 0.38	90.01 $\pm$ 0.29
	✓	✓	79.76 $\pm$ 0.35	86.72 $\pm$ 0.28	86.64 $\pm$ 0.34	90.13 $\pm$ 0.28	85.93 $\pm$ 0.38	90.30 $\pm$ 0.29
✓	✓		82.08 $\pm$ 0.31	87.06 $\pm$ 0.27	87.72 $\pm$ 0.33	90.14 $\pm$ 0.29	87.13 $\pm$ 0.36	90.78 $\pm$ 0.28
✓	✓	✓	83.66$\pm$0.31	88.38$\pm$0.25	88.67$\pm$0.32	91.45$\pm$0.28	88.02$\pm$0.34	91.71$\pm$0.27

Table 4, the proposed VT-FSL consistently outperforms all baselines by a large margin. Specifically, in the 1-shot task, VT-FSL surpasses the second-best methods by 4.35% (vs. SVasP) to 15.31% (vs. MEFP) across the three datasets. These results demonstrate that VT-FSL effectively learns more transferable representations by establishing a comprehensive interaction with the class-specific cross-modal semantics, generalizing well to novel categories even under distribution shifts.

4.3 Model Analysis

Ablation Study. We conduct an ablation study on three datasets to evaluate the effectiveness of our VT-FSL, as shown in Table 5. It is to be noted that (1) Introducing textual prompts alone improves performance over the baseline across all datasets, showing that precise and visually grounded semantics facilitate the extraction of discriminative features. Adding the kernelized volume-based contrastive loss, i.e., alignment loss, further improves results, especially on *miniImageNet* and *tieredImageNet* 1-shot settings, suggesting better cross-modal consistency. (2) Visual prompts also bring performance gains. The model with only visual prompts outperforms the baseline, and combining them with the alignment loss achieves further improvements from 89.75% to 90.14% in the 5-shot task on CIFAR-FS, indicating the alignment loss enhances intra-modal structure. (3) The best performance is achieved when both prompts and alignment loss are used. The full model consistently outperforms all ablations, demonstrating the complementarity of textual and visual prompts and the importance of jointly aligning them with the proposed geometric objective.

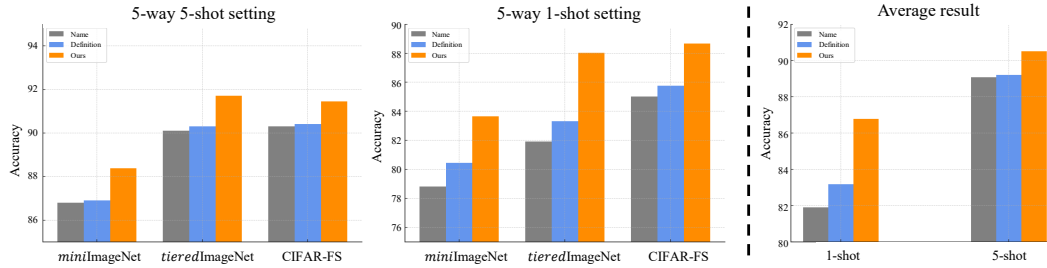


Figure 5: Comparison with textual semantics from name, definition of SemFew [31], and ours.

Comparison with Different Textual Semantics. As shown in Fig. 5, the texts generated by the proposed VT-FSL method achieve the best performance across all settings on three datasets, followed by SemFew [31, 30, 31] definitions and class names. This trend suggests that class names provide minimal information and that SemFew definitions, although generated by the LLM, rely only on class names and naive prompting, limiting the effectiveness. VT-FSL yields richer and more precise textual semantics by combining class names with support images and structured iterative prompting.

Comparison with Contrastive Learning Methods. As shown in Table 6a, we compare three contrastive learning strategies: InfoNCE [94], volume-based loss [58], and our proposed kernelized volume-based loss. The baseline is utilized without contrastive objective. InfoNCE performs the worst, especially in the 1-shot setting, due to its reliance on pairwise similarity, which overlooks interactions among multiple modalities. While the volume-based loss improves upon this, it remains limited to linear space and fails to model complex semantic structures. In contrast, our method

Table 6: Comparison with contrastive learning methods and different prototypes for inference

(a) Contrasting learning methods					(b) Inference prototypes				
Method	<i>miniImageNet</i>		CIFAR-FS		Method	<i>miniImageNet</i>		CIFAR-FS	
	1-shot	5-shot	1-shot	5-shot		1-shot	5-shot	1-shot	5-shot
baseline	82.49	87.35	87.01	90.34	C	71.63	86.13	80.75	90.34
InfoNCE [94]	79.96	86.61	86.47	88.35	$T \Rightarrow C$	80.05	86.92	87.43	90.43
Volume [58]	82.33	87.59	87.20	89.72	$V \Rightarrow C$	82.83	87.55	87.27	90.63
VT-FSL	83.66	88.38	88.67	91.45	$T + V \Rightarrow C$	83.66	88.38	88.67	91.45

introduces a Reproducing Kernel Hilbert Space (RKHS) to enable nonlinear alignment, substantially enhances the ability to integrate cross-modal semantics and achieves the best results across settings.

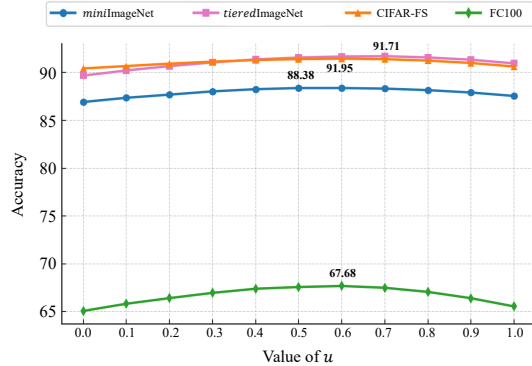
Comparison with Different Inference Prototypes. In Table 6b, using only support features (C) yields the lowest accuracy, indicating limited discriminative power under scarce supervision. Incorporating textual or visual prompts significantly improves performance, demonstrating the value of external semantic guidance. The best results are obtained when both prompts are jointly integrated ($T + V \Rightarrow C$), confirming their complementarity and the effectiveness of cross-modal synergy.

Table 7: Comparison of training and inference times on *miniImageNet* under 5-way 1-shot tasks.

Method	Prompt (h)	Training (min)	Inference (ms)	Acc (%)
SP [23]	-	1.7	78	72.31 $\pm$ 0.40
SemFew [31]	1.5	2.3	105	78.94 $\pm$ 0.66
ECER [30]	0.7	3.0	119	81.14 $\pm$ 0.15
VT-FSL (ours)	0.7	1.1	76	83.66$\pm$0.31

The effect of computational overhead with LLMs. To further investigate the efficiency of VT-FSL, we analyze the impact of computational overhead introduced by large language models (LLMs). The experiment is conducted on the same server configuration, comparing to SP [23] without any large models to generate descriptions/images and SemFew [31] and ECER [30] with these models. SP and ECER employ the same Visformer-T backbone as VT-FSL during training/inference. As shown in the Table 7, VT-FSL achieves both the lowest training time and inference time, while attaining the highest accuracy (83.66%). Compared to ECER, VT-FSL reduces training time from 3.0 min to 1.1 min and inference time from 119 ms to 76 ms, while improving accuracy by 2.5%. These results confirm that the plug-in design of VT-FSL transfers knowledge from large models efficiently in a one-time offline stage to construct cross-modal prompts. Any downstream FSL model can directly use the resulting cross-modal prompts, adding virtually no extra overhead during training or inference. This avoids the extra manual or algorithmic corrections required by SemFew and ECER, leading to a better trade-off between efficiency and performance.

The effect of fusion factor u . u controls the relative weight between textual and visual features to obtain the final inference prototypes in Eq. 13. For each dataset, u is automatically determined by a grid search over $[0, 1]$ on the corresponding validation set, selecting the value that yields the highest validation accuracy. As shown in the Fig. 6, The optimal values are 0.5 (*miniImageNet*), 0.7 (*tieredImageNet*), 0.6 (*CIFAR-FS*), and 0.6 (*FC100*). The results show that performance peaks at intermediate values of u , confirming that textual and visual features are complementary. Using only visual prompts ($u = 0$) or only textual prompts ($u = 1$) consistently leads to inferior results.

Figure 6: The effect of the fusion factor u .

The effect of the Synthetic Image Number. We evaluate how the number of synthetic images per class affects performance in the 1-shot setting, as shown in Fig. 7a. Accuracy improves notably when

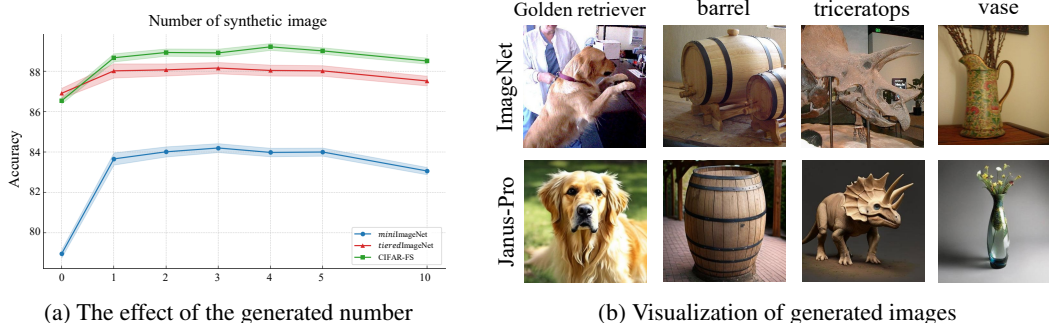


Figure 7: The effect of the generated number and visualization of generated images.

adding 1-3 synthetic images, but plateaus or slightly declines beyond that. This trend is consistent across all datasets. These results validate our design choice of generating K synthetic images for K -shot tasks, which effectively augments the support set without compromising the low-data regime. However, increasing the number further introduces noise, as lower-ranked generations tend to be less discriminative and may degrade representation quality. In Fig 7b, we visualize the synthetic images generated by Janus-Pro on ImageNet [95]. Benefiting from precise descriptions with key visual attributes, synthetic images can well highlight the low-level semantics of the target category.

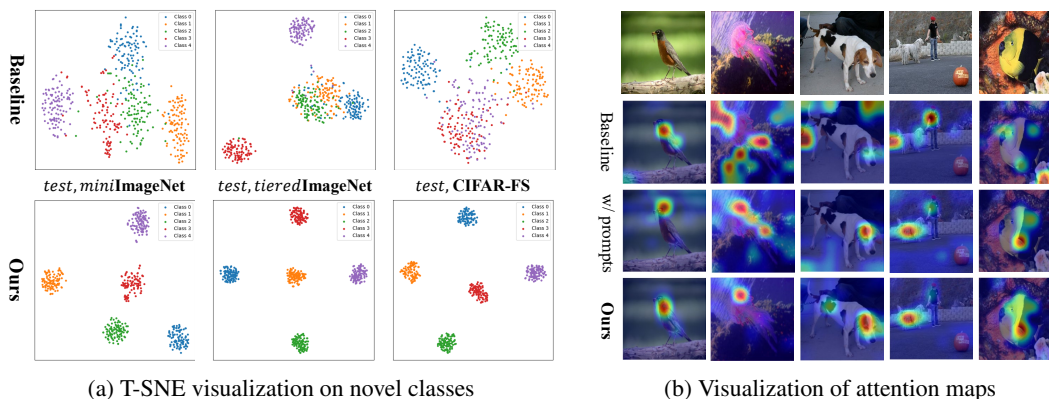


Figure 8: Visualization of T-SNE and attention maps.

Visualization of T-SNE and Attention Maps. Fig. 8a shows that VT-FSL produces well-separated and compact class clusters, in contrast to the scattered distributions of the baseline, indicating improved semantic alignment. As shown in Fig. 8b, the baseline suffers from attention to irrelevant regions, while our method focuses more accurately on key object areas, guided by cross-modal prompts and further enhanced through geometry-aware semantic alignment.

5 Conclusion

In this paper, we propose VT-FSL, a novel framework that bridges vision and text representations with large language models to advance few-shot learning by generating complementary cross-modal prompts and integrating them via a geometry-aware consistency alignment. Specifically, we introduce Cross-modal Iterative Prompting (CIP) to generate precise descriptions with high-level semantics from both class names and support images by a single structure inference pass, enabling the zero-shot synthesis of semantically consistent images with low-level diversity. We further propose Cross-modal Geometric Alignment (CGA) to comprehensively align the fused textual, support, and synthetic features by minimizing the volume in their kernelized parallelotope space, capturing global and nonlinear cross-modal dependencies. Extensive experiments on ten FSL benchmarks demonstrate the effectiveness of VT-FSL, improving the classification accuracy by 4.2% on average.

Acknowledgment: This work was supported in part by the Natural Science Foundation of China under Grant U23A20389, 62176139, and 62406177, in part by the Shandong Excellent Young Scientists Fund (Oversea) under Grant 2024HWYQ-027, in part by the Natural Science Foundation of Shandong province under Grant ZR2023QF124, in part by the Young Scholars Program of Shandong University, and in part by the Fundamental Research Funds of Shandong University.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 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,” in *ICLR*, 2021.
- [3] Z. Chen, L. Xie, J. Niu, X. Liu *et al.*, “Visformer: The vision-friendly transformer,” in *ICCV*, 2021, pp. 589–598.
- [4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [5] P. Zhao, X. Song, X. Xi, X. Nie, X. Meng, Y. Qu, and Y. Yin, “Biomarkers-aware asymmetric bibranch gan with adaptive memory batch normalization for prediction of anti-vegf treatment response in neovascular age-related macular degeneration,” *IEEE Journal of Biomedical and Health Informatics*, 2023.
- [6] S. Kim, S. An, P. Chikontwe, M. Kang, E. Adeli, K. M. Pohl, and S. H. Park, “Few shot part segmentation reveals compositional logic for industrial anomaly detection,” in *AAAI*, vol. 38, no. 8, 2024, pp. 8591–8599.
- [7] Z. Qi, L. Meng, Z. Chen, H. Hu, H. Lin, and X. Meng, “Cross-silo prototypical calibration for federated learning with non-iid data,” in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 3099–3107.
- [8] K. Shi, Z. Qi, J. Zhu, L. Meng, Y. Zhang, H. Huang, and X. Meng, “Protoconnet: Prototypical augmentation and alignment for open-set few-shot image classification,” *arXiv preprint arXiv:2507.11845*, 2025.
- [9] Z. Han, X.-J. Gui, H. Sun, Y. Yin, and S. Li, “Towards accurate and robust domain adaptation under multiple noisy environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6460–6479, 2022.
- [10] Z. Han, H. Sun, and Y. Yin, “Learning transferable parameters for unsupervised domain adaptation,” *IEEE Transactions on Image Processing*, vol. 31, pp. 6424–6439, 2022.
- [11] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *NeurIPS*, 2017, pp. 4077–4087.
- [12] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, “Matching networks for one shot learning,” in *NeurIPS*, 2016, pp. 3630–3638.
- [13] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *CVPR*, 2018, pp. 1199–1208.
- [14] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *CVPR*, 2020, pp. 12 203–12 213.
- [15] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, “Improved few-shot visual classification,” in *CVPR*, 2020, pp. 14 493–14 502.
- [16] D. Chen, J. Zhang, W.-S. Zheng, and R. Wang, “Featwalk: Enhancing few-shot classification through local view leveraging,” in *AAAI*, vol. 38, no. 2, 2024, pp. 1019–1027.
- [17] B. Dong, P. Zhou, S. Yan, and W. Zuo, “Self-promoted supervision for few-shot transformer,” in *ECCV*. Springer, 2022, pp. 329–347.
- [18] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, “Finding task-relevant features for few-shot learning by category traversal,” in *CVPR*, 2019, pp. 1–10.
- [19] B. Oreshkin, P. Rodríguez López, and A. Lacoste, “Tadam: Task dependent adaptive metric for improved few-shot learning,” *NeurIPS*, vol. 31, 2018.

- [20] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, “Rethinking few-shot image classification: a good embedding is all you need?” in *ECCV*. Springer, 2020, pp. 266–282.
- [21] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, “Few-shot learning via embedding adaptation with set-to-set functions,” in *CVPR*, 2020, pp. 8808–8817.
- [22] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, “Meta-baseline: Exploring simple meta-learning for few-shot learning,” in *ICCV*, 2021, pp. 9062–9071.
- [23] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, “Semantic prompt for few-shot image recognition,” in *CVPR*, 2023, pp. 23 581–23 591.
- [24] W. Li, Q. Wang, P. Zhao, and Y. Yin, “Knn transformer with pyramid prompts for few-shot learning,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1082–1091.
- [25] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. O Pinheiro, “Adaptive cross-modal few-shot learning,” in *NeurIPS*, vol. 32, 2019.
- [26] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, “Boosting few-shot learning with adaptive margin loss,” in *CVPR*, 2020, pp. 12 576–12 584.
- [27] J. Xu and H. Le, “Generating representative samples for few-shot classification,” in *CVPR*, 2022, pp. 9003–9013.
- [28] R. Zhang, X. Hu, B. Li, S. Huang, H. Deng, Y. Qiao, P. Gao, and H. Li, “Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners,” in *CVPR*, 2023, pp. 15 211–15 222.
- [29] B. Zhang, X. Li, Y. Ye, Z. Huang, and L. Zhang, “Prototype completion with primitive knowledge for few-shot learning,” in *CVPR*, 2021, pp. 3754–3762.
- [30] M. Liu, F. Wu, B. Li, Z. Lu, Y. Yu, and X. Li, “Envisioning class entity reasoning by large language models for few-shot learning,” in *AAAI*, vol. 39, no. 18, 2025, pp. 18 906–18 914.
- [31] H. Zhang, J. Xu, S. Jiang, and Z. He, “Simple semantic-aided few-shot learning,” in *CVPR*, 2024, pp. 28 588–28 597.
- [32] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [33] OpenAI, A. Hurst, A. Lerer *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [34] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang *et al.*, “Qwen2.5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [35] I. I. G. with Better Captions, “James betker and gabriel goh and li jing and tim brooks and others,” <https://cdn.openai.com/papers/dall-e-3.pdf>, 2024.
- [36] X. Chen, Z. Wu, X. Liu *et al.*, “Janus-pro: Unified multimodal understanding and generation with data and model scaling,” *arXiv preprint arXiv:2501.17811*, 2025.
- [37] 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 *ICML*. PMLR, 2021, pp. 8748–8763.
- [38] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*. JMLR. org, 2017, pp. 1126–1135.
- [39] S. Baik *et al.*, “Learning to learn task-adaptive hyperparameters for few-shot learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1441–1454, 2024.
- [40] H.-J. Ye, L. Ming, D.-C. Zhan, and W.-L. Chao, “Few-shot learning with a strong teacher,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1425–1440, 2024.
- [41] B. Zhang, C. Luo, D. Yu, X. Li, H. Lin, Y. Ye, and B. Zhang, “Metadiff: Meta-learning with conditional diffusion for few-shot learning,” in *AAAI*, vol. 38, no. 15, 2024, pp. 16 687–16 695.
- [42] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, “Fast context adaptation via meta-learning,” in *ICML*. PMLR, 2019, pp. 7693–7702.

- [43] M. A. Jamal and G.-J. Qi, “Task agnostic meta-learning for few-shot learning,” in *CVPR*, 2019, pp. 11 719–11 727.
- [44] T. Elsken, B. Staffler, J. H. Metzen, and F. Hutter, “Meta-learning of neural architectures for few-shot learning,” in *CVPR*, 2020, pp. 12 365–12 375.
- [45] C. Doersch, A. Gupta, and A. Zisserman, “Crosstransformers: spatially-aware few-shot transfer,” in *NrurIPS*, vol. 33, 2020, pp. 21 981–21 993.
- [46] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, “Boosting few-shot visual learning with self-supervision,” in *ICCV*, 2019, pp. 8059–8068.
- [47] M. Hiller, R. Ma *et al.*, “Rethinking generalization in few-shot classification,” *NeurIPS*, vol. 35, pp. 3582–3595, 2022.
- [48] Y. Rong, X. Lu, Z. Sun, Y. Chen, and S. Xiong, “Espt: A self-supervised episodic spatial pretext task for improving few-shot learning,” in *AAAI*, vol. 37, no. 8, 2023, pp. 9596–9605.
- [49] M.-H. Pan, H.-Y. Xin, and H.-B. Shen, “Semantic-based implicit feature transform for few-shot classification,” *IJCV*, vol. 132, no. 11, pp. 5014–5029, 2024.
- [50] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*. PMLR, 2022, pp. 12 888–12 900.
- [51] Z. Yang, J. Wang, and Y. Zhu, “Few-shot classification with contrastive learning,” in *ECCV*, 2022, pp. 293–309.
- [52] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, and L. Zhang, “Learning a few-shot embedding model with contrastive learning,” in *AAAI*, vol. 35, no. 10, 2021, pp. 8635–8643.
- [53] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, “Prompt-aligned gradient for prompt tuning,” in *ICCV*, 2023, pp. 15 659–15 669.
- [54] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *ICCV*, 2023, pp. 11 975–11 986.
- [55] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, “Tip-adaptor: Training-free adaption of clip for few-shot classification,” in *ECCV*. Springer, 2022, pp. 493–510.
- [56] J. Pennington and P. Worah, “Nonlinear random matrix theory for deep learning,” *NeurIPS*, vol. 30, 2017.
- [57] I. Nejjar, Q. Wang, and O. Fink, “Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices,” in *CVPR*, 2023, pp. 11 744–11 754.
- [58] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, “Gramian multimodal representation learning and alignment,” *arXiv preprint arXiv:2412.11959*, 2024.
- [59] F. R. Gantmakher, *The theory of matrices*. American Mathematical Soc., 2000, vol. 131.
- [60] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, “Meta-learning for semi-supervised few-shot classification,” *arXiv preprint arXiv:1803.00676*, 2018.
- [61] S. Sun and H. Gao, “Meta-adam: An meta-learned adaptive optimizer with momentum for few-shot learning,” *NeurIPS*, vol. 36, pp. 65 441–65 455, 2023.
- [62] Y. Du, Z. Xiao, S. Liao, and C. G. M. Snoek, “Protodiff: Learning to learn prototypical networks by task-guided diffusion,” in *NeurIPS*, vol. 36, 2023, pp. 46 304–46 322.
- [63] F. Hao *et al.*, “Class-aware patch embedding adaptation for few-shot image classification,” in *ICCV*, 2023, pp. 18 905–18 915.
- [64] M. Kim and T. Hospedales, “A hierarchical bayesian model for few-shot meta learning,” in *ICLR*, 2024.
- [65] S. Poulakakis-Daktylidis and H. Jamali-Rad, “Beclr: batch enhanced contrastive few-shot learning,” in *ICLR*, 2024.
- [66] Y. Hu, Y. Zou, R. Li, and Y. Li, “Generate universal adversarial perturbations for few-shot learning,” *NeurIPS*, vol. 37, pp. 102 672–102 695, 2024.
- [67] K. Lee, S. Maji *et al.*, “Meta-learning with differentiable convex optimization,” in *CVPR*, 2019, pp. 10 657–10 665.

- [68] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [69] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in *ICCV workshops*, 2013, pp. 554–561.
- [70] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *CVPR workshop*, vol. 2, no. 1, 2011.
- [71] B. Zhou, A. Lapedriza *et al.*, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
- [72] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, “The inaturalist species classification and detection dataset,” in *CVPR*, 2018, pp. 8769–8778.
- [73] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [74] J. Kim, H. Kim, and G. Kim, “Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning,” in *ECCV*. Springer, 2020, pp. 599–617.
- [75] D. Wertheimer, L. Tang, and B. Hariharan, “Few-shot classification with feature map reconstruction networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.
- [76] S.-L. Xu, F. Zhang, X.-S. Wei, and J. Wang, “Dual attention networks for few-shot fine-grained recognition,” in *AAAI*, vol. 36, no. 3, 2022, pp. 2911–2919.
- [77] Y. Yu, D. Zhang, and Z. Ji, “Masked feature generation network for few-shot learning,” in *IJCAI*, 2022, pp. 3695–3701.
- [78] S. Lee, W. Moon, and J.-P. Heo, “Task discrepancy maximization for fine-grained few-shot classification,” in *CVPR*, 2022, pp. 5331–5340.
- [79] Z. Zha, H. Tang, Y. Sun, and J. Tang, “Boosting few-shot fine-grained recognition with background suppression and foreground alignment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3947–3961, 2023.
- [80] L.-J. Zhao, Z.-D. Chen, Z.-X. Ma, X. Luo, and X.-S. Xu, “Angular isotonic loss guided multi-layer integration for few-shot fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 33, pp. 3778–3792, 2024.
- [81] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, X. Luo, and X.-S. Xu, “Cross-layer and cross-sample feature optimization network for few-shot fine-grained image classification,” in *AAAI*, vol. 38, no. 5, 2024, pp. 4136–4144.
- [82] Z.-X. Ma, Z.-D. Chen, T. Zheng, X. Luo, Z. Jia, and X.-S. Xu, “Few-shot fine-grained image classification with progressively feature refinement and continuous relationship modeling,” in *AAAI*, vol. 39, no. 6, 2025, pp. 6036–6044.
- [83] V. Garcia and J. Bruna, “Few-shot learning with graph neural networks,” in *ICLR*, 2018.
- [84] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, “Cross-domain few-shot classification via learned feature-wise transformation,” in *ICLR*, 2020.
- [85] Y. Hu and A. J. Ma, “Adversarial feature augmentation for cross-domain few-shot classification,” in *ECCV*. Springer, 2022, pp. 20–37.
- [86] J. Oh, S. Kim, N. Ho, J.-H. Kim, H. Song, and S.-Y. Yun, “Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty,” *NeurIPS*, vol. 35, pp. 2622–2636, 2022.
- [87] H. Wang, H. Mai, Y. Gong, and Z.-H. Deng, “Towards well-generalizing meta-learning via adversarial task augmentation,” *Artificial Intelligence*, vol. 317, p. 103875, 2023.
- [88] F. Zhou, P. Wang, L. Zhang, W. Wei, and Y. Zhang, “Revisiting prototypical network for cross domain few-shot learning,” in *CVPR*, 2023, pp. 20 061–20 070.
- [89] Y. Fu, Y. Xie, Y. Fu, and Y.-G. Jiang, “Styleadv: Meta style adversarial training for cross-domain few-shot learning,” in *CVPR*, 2023, pp. 24 575–24 584.

- [90] T. Zhang, Q. Cai, F. Gao, L. Qi, and J. Dong, “Exploring cross-domain few-shot classification via frequency-aware prompting,” *IJCAI*, 2024.
- [91] Y. Zou, Y. Liu *et al.*, “Flatten long-range loss landscapes for cross-domain few-shot learning,” in *CVPR*, 2024, pp. 23 575–23 584.
- [92] F. Zhou, P. Wang, L. Zhang, Z. Chen, W. Wei, C. Ding, G. Lin, and Y. Zhang, “Meta-exploiting frequency prior for cross-domain few-shot learning,” *NeurIPS*, vol. 37, pp. 116 783–116 814, 2024.
- [93] W. Li, P. Fang, and H. Xue, “Svasp: Self-versatility adversarial style perturbation for cross-domain few-shot learning,” in *AAAI*, vol. 39, no. 15, 2025, pp. 15 275–15 283.
- [94] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [95] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. Ieee, 2009, pp. 248–255.
- [96] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The key claims in the abstract and introduction accurately reflect the contribution and scope of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations of the work is discussed in the supplemental material.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: A short proof sketch is provided in the main paper and theoretical assumptions and proofs are detailed in the supplemental material

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all the training and testing details necessary to understand the results (e.g., data partitioning, hyperparameters, how to select them, type of optimizer, etc.).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments are averaged over multiple runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides information on the computer resources needed

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work does not contain any negative social impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage are deeply involved in the design of core methodology of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Appendix / supplemental material

A.1 Training Algorithm

Algorithm 1: Training algorithm of the proposed VT-FSL.

Input: Training set $\mathcal{C}_{train}$, feature extractor f_θ , linear projector g_η , two-layer perception g_φ
while not converged do

1. Sample an N -way K -shot task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ from $\mathcal{C}_{train}$ by the episodic strategy;
- for each class y_i in $\mathcal{Q}$ do**
 2. Generate textual descriptions $\mathcal{T}_N$ via LLM with class name and K support images, using four structured reasoning stages: Strategy, Perception, Refinement, Conclusion;
 3. Generate M synthetic images from $\mathcal{T}_N$ using a text-to-image model;
 4. Select top- K images with LLM-based pairwise comparison according to Eq. 3;
5. Extract support features z_s and synthetic visual features Z_v using f_ϕ ;
6. Encode textual features Z_t via CLIP, and project them to visual space with g_η ;
7. Fuse Z_t and z_s via g_φ along channel and spatial dimensions, generating enhanced support features Z_s according to Eq. 4, Eq. 5, and Eq. 6;
8. Calculate the kernelized volume-based contrastive loss $\mathcal{L}_{D2A}$ and $\mathcal{L}_{A2D}$, jointly denoted as $\mathcal{L}_{align}$ with Z_t , Z_s , and Z_v according to Eq. 10 and Eq. 11;
9. Compute class prototypes $C = \{c_i\}_{i=1}^N$ by averaging Z_s per class according to Eq. 1;
- for each query image x_j in $\mathcal{Q}$ do**
 10. Compute the prediction scores p_{x_j} according to Eq. 2;
 11. Calculate the cross-entropy loss between p_{x_j} and the ground-truth label y_j ;
12. Calculate the overall loss $\mathcal{L}_{total}$ according to Eq. 12;
13. Update θ , η , and φ via gradient backpropagation

Output: Final prototypes with text and synthetic images for testing according to Eq. 13.

We describe the entire training procedure of VT-FSL in detail in Algorithm 1. The training process follows an episodic training paradigm and incorporates two key modules: (1) *Cross-modal Iterative Prompting (CIP)*, which generates class-specific textual descriptions based on both class names and support samples through a structured iterative reasoning process; and (2) *Cross-modal Geometric Alignment (CGA)*, which integrates textual, synthetic visual, and support features via a volume-based contrastive loss in a kernelized embedding space. These modules jointly enable the construction of cross-modal prototypes that are both semantically rich and visually grounded. The complete process, including prompt generation, feature extraction, fusion, prototype construction, loss computation, and parameter updates, is formalized in Algorithm 1.

A.2 Text Generation Scheme

Some studies [24–28] extract additional semantic information from class names to compensate for the lack of representative semantics in limited support samples. However, class names typically provide minimal contextual information and are inherently ambiguous. For example, the term “mouse” may refer to either an animal or a computer device, offering little discriminative value without additional context.

To address this, SemFew [31] uses a large language model (LLM) to expand a class name into a detailed definition. However, to ensure textual-visual consistency, it relies on external knowledge bases such as WordNet [32], introducing additional matching constraints and reducing generation efficiency. ECER [30] also employs LLMs to generate a set of visual attributes from the class name, but due to the lack of guidance from support images, it requires a carefully designed attribute-filtering mechanism to remove visually irrelevant semantics.

In contrast, our method incurs no extra matching overhead and fully leverages the LLM’s reasoning and generation capabilities to produce precise and visually grounded class descriptions. We propose a **Cross-modal Iterative Prompting (CIP)** framework, which conditions the LLM on both the class name and K support images to elicit structured and grounded definitions. Specifically, CIP decomposes the generation process into four explicitly defined reasoning stages:

- **Strategy Stage:** Acts as a visual taxonomy expert, interpreting the class name to establish an initial semantic anchor and outlining the reasoning process.
- **Perception Stage:** Analyzes the K support images to identify shared and discriminative visual traits that define the category.
- **Refinement Stage:** Refines the initial interpretation through logical reasoning, aligning class semantics with observed visual evidence, while eliminating hallucinated or irrelevant attributes.
- **Conclusion Stage:** Produces a concise, scientifically accurate, and generalizable definition, following a unified format and concrete example.

We guide the LLM to produce answers at each stage using a carefully designed prompt structure for iterative optimization of the final generated textual descriptions. The input is formatted using four special tags: `<STRATEGY></STRATEGY>`, `<PERCEPTION></PERCEPTION>`, `<REFINEMENT></REFINEMENT>`, and `<CONCLUSION></CONCLUSION>`, corresponding to outlining the problem, interpreting visual information, conducting reasoning, and producing the final definition, respectively.

The complete prompt template is shown below:

Prompt Template for Cross-modal Iterative Prompting (CIP)

You are an expert in computer vision and concept definition.
 Given a class name and a set of representative images, your task is to generate a brief, scientifically accurate, and visually grounded definition. The definition should be primarily guided by the **semantic meaning of the class name**, and refined using the **visual evidence** from the images to resolve ambiguities and enhance relevance.
 Please strictly follow the structured reasoning format below:

`<SUMMARY>`
 Briefly describe your overall approach: (1) Interpret the class name and infer its possible meanings; (2) Analyze the image content to validate or refine the interpretation; (3) Formulate a revised description based on both sources. `</SUMMARY>`

`<CAPTION>`
 Describe **common visual elements** observed across all images that help clarify the concept, such as objects, shapes, textures, colors, or background environments. `</CAPTION>`

`<REASONING>`
 Explain how the visual patterns confirm or adjust the class meaning, especially if it is **ambiguous, polysemous, or abstract**. Use *step-by-step reasoning* to refine the concept. `</REASONING>`

`<CONCLUSION>`
 Rewrite the class definition as a concise, scientifically sound, and visually consistent paragraph. Avoid redundancy; prioritize clarity and relevance.
Example: The American robin is a widely recognized songbird, characterized by a rust-red to orange breast, dark grayish-blue upperparts, a white eye ring, and a slender yellow bill. It is commonly found in open woodlands and suburban gardens across North America. `</CONCLUSION>`

Class Name: {class_label}
Images: [K image inputs attached here]

A.3 Related Theories of Cross-modal Geometric Alignment

A.3.1 Gram Matrix-based Volume

Volume Computation Based on Gram Matrix. Given a set of k vectors $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$, we aim to compute the volume of the k -dimensional parallelotope spanned by them. Let these vectors be arranged as columns in a matrix $\mathbf{A} = [\mathbf{v}_1, \dots, \mathbf{v}_k] \in \mathbb{R}^{n \times k}$. The associated *Gram matrix* is defined as:

$$\mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \mathbf{A}^\top \mathbf{A} = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_1, \mathbf{v}_k \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_2, \mathbf{v}_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_k, \mathbf{v}_1 \rangle & \langle \mathbf{v}_k, \mathbf{v}_2 \rangle & \cdots & \langle \mathbf{v}_k, \mathbf{v}_k \rangle \end{bmatrix}. \quad (14)$$

This matrix captures all geometric information of the set $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, including vector lengths and pairwise angles. The k -volume of the parallelotope is computed by:

$$\text{Vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \sqrt{\det \mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k)}. \quad (15)$$

This formulation generalizes the intuition of Euclidean norm: when $k = 1$, the volume reduces to the norm of a single vector, i.e., $\text{Vol}(\mathbf{v}) = \|\mathbf{v}\|$.

Theoretical Validity. To justify the above definition, we consider three cases:

- **Case I:** $k = n$. Then,

$$\det \mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \det(\mathbf{A}^\top \mathbf{A}) = (\det \mathbf{A})^2 = |\mathbf{A}|^2 \geq 0. \quad (16)$$

Therefore,

$$\text{Vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = |\det \mathbf{A}| = \sqrt{\det \mathbf{A}^\top \mathbf{A}} = \sqrt{\det \mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k)}. \quad (17)$$

- **Case II:** $k < n$. In this case, the volume lies within the k -dimensional subspace spanned by the vectors. Suppose $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ are linearly independent. We can extend this set to an orthonormal basis $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ of $\mathbb{R}^n$, and define an orthogonal transformation $\mathbf{O}$ that maps the basis to a new coordinate frame.

The transformed vectors become:

$$\mathbf{v}'_i = \mathbf{O}(\mathbf{v}_i), \quad \mathbf{v}'_i = \begin{bmatrix} m_{1i} \\ m_{2i} \\ \vdots \\ m_{ki} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (18)$$

which visibly lie in a k -dimensional subspace of $\mathbb{R}^n$. Since orthogonal transformations preserve inner products, we have:

$$\langle \mathbf{v}'_i, \mathbf{v}'_j \rangle = \langle \mathbf{v}_i, \mathbf{v}_j \rangle, \quad (19)$$

and thus,

$$\text{Vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \text{Vol}(\mathbf{v}'_1, \dots, \mathbf{v}'_k) = \sqrt{\det \mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k)}. \quad (20)$$

- **Case III:** $k > n$. In this case, the vectors must be linearly dependent, and hence the k -dimensional volume is zero:

$$\det \mathbf{G}(\mathbf{v}_1, \dots, \mathbf{v}_k) = 0, \quad \text{Vol}(\mathbf{v}_1, \dots, \mathbf{v}_k) = 0. \quad (21)$$

Summary. The determinant of the Gram matrix provides a robust and theoretically grounded measure of the volume of the parallelotope spanned by a set of vectors. This geometric quantity generalizes norm and angle in high-dimensional embedding spaces, and forms the basis of our proposed alignment loss for cross-modal representation matching.

A.3.2 Volume as a Generalized Alignment Metric

We demonstrate that volume computation based on the Gram matrix provides a more comprehensive and expressive alignment mechanism compared to cosine similarity, particularly when multiple types of embeddings are involved.

We first consider the case of two vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^n$ with unit norm, i.e., $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$. The Gram matrix is:

$$\mathbf{G} = \begin{bmatrix} \langle \mathbf{v}_1^\top, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1^\top, \mathbf{v}_2 \rangle \\ \langle \mathbf{v}_2^\top, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2^\top, \mathbf{v}_2 \rangle \end{bmatrix}, \quad (22)$$

with determinant:

$$\det(\mathbf{G}) = \langle \mathbf{v}_1^\top \mathbf{v}_1 \rangle \langle \mathbf{v}_2^\top \mathbf{v}_2 \rangle - \langle \mathbf{v}_1^\top \mathbf{v}_2 \rangle^2. \quad (23)$$

Hence, the volume spanned by these two vectors is:

$$\text{Vol} = \sqrt{\det(\mathbf{G})} = \sqrt{1 - \langle \mathbf{v}_1^\top \mathbf{v}_2 \rangle^2}. \quad (24)$$

Letting $\cos(\theta) = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle$, this becomes:

$$\text{Vol} = \sqrt{1 - \cos^2(\theta)} = \sin(\theta). \quad (25)$$

This reveals that in the two-vector case, the volume is directly proportional to the sine of the angle between them, capturing the degree of geometric misalignment. Unlike cosine similarity, which increases as vectors become more aligned, volume achieves maximum when vectors are orthogonal, offering complementary information and sensitivity to structural differences.

Volume Captures Full Pairwise Interactions. We extend the analysis to our setup involving three types of embeddings: textual features $\mathbf{T}$, support embeddings $\mathbf{S}$, and synthetic visual embeddings $\mathbf{V}$. The corresponding Gram matrix is:

$$\mathbf{G} = \begin{bmatrix} TT & TS & TV \\ ST & SS & SV \\ VT & VS & VV \end{bmatrix}, \quad (26)$$

where $TS = \langle \mathbf{T}, \mathbf{S} \rangle$, $TV = \langle \mathbf{T}, \mathbf{V} \rangle$, and so on. All embeddings are normalized such that $TT = SS = VV = 1$. The determinant of this matrix expands as:

$$\begin{aligned} \det(\mathbf{G}) &= TT \cdot (SS \cdot VV - SV \cdot VS) - TS \cdot (ST \cdot VV - SV \cdot VT) \\ &\quad + TV \cdot (ST \cdot VS - SS \cdot VT), \end{aligned} \quad (27)$$

and simplifies to:

$$\begin{aligned} \det(\mathbf{G}) &= 1 \cdot (1 - SV^2) - TS \cdot (TS - SV \cdot VT) + TV \cdot (TS \cdot SV - TV) \\ &= 1 - SV^2 - TS^2 + TS \cdot SV \cdot VT + TV \cdot TS \cdot SV - TV^2 \\ &= 1 - SV^2 - TS^2 - TV^2 + 2 \cdot TS \cdot SV \cdot VT. \end{aligned} \quad (28)$$

This result shows that volume-based computation naturally incorporates all pairwise interactions among the three embeddings—textual ($\mathbf{T}$), support ($\mathbf{S}$), and synthetic visual ($\mathbf{V}$). In particular, it accounts for relationships like SV , which are not explicitly considered when only using pairwise cosine similarities between the anchor ($\mathbf{T}$) and the other embeddings.

Comparison with Cosine-based Alignment. In many existing methods, only the similarities between the anchor and the remaining embeddings (e.g., TS and TV) are computed, while relationships among the non-anchor embeddings (e.g., SV) are ignored. This may lead to suboptimal alignment, especially when synthetic features or support embeddings deviate semantically.

In contrast, volume-based alignment considers all pairwise inner products in a unified geometric form. The determinant of the Gram matrix thus encodes higher-order consistency among the embeddings, enabling a more robust alignment objective that jointly optimizes their global compatibility.

A.3.3 Kernelized Volume for Nonlinear Alignment

While the volume computed from the standard Gram matrix already captures higher-order geometric relationships among textual, support, and synthetic visual embeddings, it remains a linear measure. To further model complex nonlinear relationships, we extend our formulation to a high-dimensional Reproducing Kernel Hilbert Space (RKHS) via kernel embedding.

Let $\kappa : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a positive definite kernel function. The most common example is the Radial Basis Function (RBF) kernel, defined as:

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right), \quad (29)$$

where σ controls the smoothness of the kernel. Such kernels define an implicit mapping $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ into a high-dimensional Hilbert space $\mathcal{H}$, where

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathcal{H}}. \quad (30)$$

Given a set of k embeddings $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, we construct a kernel Gram matrix $\mathbf{K} \in \mathbb{R}^{k \times k}$ as:

$$\mathbf{K}_{ij} = \kappa(\mathbf{v}_i, \mathbf{v}_j), \quad \forall i, j \in \{1, \dots, k\}. \quad (31)$$

Due to the positive definiteness of κ , the kernel Gram matrix $\mathbf{K}$ is symmetric and positive semi-definite. Thus, we can generalize the definition of volume in the RKHS as:

$$\text{Vol}_{\mathcal{H}}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \sqrt{\det(\mathbf{K})}. \quad (32)$$

This formulation retains the original motivation of measuring the mutual independence or spread among embeddings, but now in a nonlinear feature space where complex dependencies can be captured.

Why Kernelized Volume Matters. The kernelized volume metric $\text{Vol}_{\mathcal{H}}$ has several key advantages over both cosine similarity and its linear volume counterpart:

- **Nonlinear feature interactions:** Through the kernel map ϕ , we effectively transform the original embeddings into a high-dimensional space where nonlinear interactions become linearly separable. This enables detection of subtle structural mismatches between the three embeddings (text $\mathbf{T}$, support $\mathbf{S}$, synthetic vision $\mathbf{V}$) that may not be captured by inner product or cosine measures in the original space.
- **Higher expressive power:** Unlike cosine similarity, which only encodes directional alignment between pairs of vectors, the determinant $\det(\mathbf{K})$ captures how all k embeddings interact geometrically in the kernel space. This includes all mutual pairwise kernel similarities and their higher-order arrangements.
- **Connection to independence and compactness:** When the embeddings $\phi(\mathbf{v}_1), \dots, \phi(\mathbf{v}_k)$ are highly correlated, the volume tends to zero, indicating collapse in representation diversity. Conversely, a larger volume suggests that the embeddings are geometrically well-distributed and carry complementary information. This is especially valuable for few-shot learning, where redundancy among features may severely harm generalization.

Kernel Properties and Geometric Interpretation. From the theory of positive definite kernels, we know that for any set of points $\{x_1, x_2, \dots, x_N\}$, the associated kernel matrix:

$$\mathbf{K} = [\kappa(x_i, x_j)]_{N \times N} \quad (33)$$

is guaranteed to be symmetric and positive semi-definite. This ensures that the square root of its determinant always yields a valid (possibly zero) volume in RKHS.

In essence, $\text{Vol}_{\mathcal{H}}$ defines the volume of a parallelotope formed by $\phi(\mathbf{v}_1), \dots, \phi(\mathbf{v}_k)$ in $\mathcal{H}$:

$$\text{Vol}_{\mathcal{H}}(\mathbf{v}_1, \dots, \mathbf{v}_k) = \|\phi(\mathbf{v}_1) \wedge \dots \wedge \phi(\mathbf{v}_k)\|, \quad (34)$$

where $\wedge$ denotes the exterior product. This geometric interpretation highlights its capacity to quantify not just similarity, but high-dimensional structural diversity.

Table 8: The splits of categories and the number of categories/images in each few-shot dataset.

Dataset	#(Class)			#(Image)	
	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{valid}}$	$\mathcal{D}_{\text{test}}$	Test	Total
<i>mini</i> ImageNet [12]	64	16	20	12,000	60,000
<i>tiered</i> ImageNet [60]	351	97	160	206,209	779,165
CIFAR-FS [67]	64	16	20	12,000	60,000
FC100 [19]	60	20	20	12,000	60,000
CUB-200-2011 [68]	100	50	50	2,958	11,788
Stanford-Cars [69]	130	17	49	4,103	16,185
Stanford-Dogs [70]	70	20	30	5,115	20,580
Places [71]	183	91	91	18,200	73,000
Plantae [72]	100	50	50	17,253	68,461

Summary. By embedding the embeddings into a kernel-induced space and computing their geometric volume, our approach moves beyond simple pairwise similarity, capturing richer nonlinear relationships among text, support, and synthetic visual features. This kernelized extension of the CGA loss offers greater alignment flexibility and improves robustness under complex cross-space discrepancies.

A.4 Experiments

A.4.1 Dataset Details

We evaluate our method on a range of widely used few-shot learning benchmarks across three scenarios: standard few-shot classification, fine-grained few-shot classification, and cross-domain few-shot classification. Detailed statistics of all datasets are summarized in Table 8.

(1) Standard Few-Shot Classification Benchmarks. Following [23, 31, 47], we adopt the following four benchmark datasets:

- ***mini*ImageNet** [12]: A subset of ImageNet [95] consisting of 100 object categories, each with 600 images, totaling 60,000 images. The dataset is split into 64 classes for training, 16 for validation, and 20 for testing.
- ***tiered*ImageNet** [60]: A larger subset of ImageNet [95] comprising 779,165 images from 608 categories, grouped into broader semantic superclasses. The training/validation/test split contains 351, 97, and 160 categories, respectively.
- **CIFAR-FS** [67]: Derived from CIFAR-100 [96] by randomly splitting the 100 classes into 64 for training, 16 for validation, and 20 for testing. Each class contains 600 images, resulting in 60,000 images in total.
- **FC100** [19]: Also based on CIFAR-100 [96], but categories are split according to their semantic superclasses. It uses 60 classes from 12 superclasses for training, 20 classes from 4 superclasses for validation, and 20 classes from 4 different superclasses for testing. The large semantic gap across splits makes FC100 more challenging.

(2) Fine-Grained Few-Shot Classification Benchmarks. Following [80–82], we evaluate on three datasets focused on fine-grained visual categories:

- **CUB-200-2011 (CUB)** [68]: Comprises 11,788 images from 200 bird species. The classes are divided into 100 for training, 50 for validation, and 50 for testing. Each image is cropped to the human-annotated bounding box of the bird.
- **Stanford-Cars** [69]: Contains 16,185 images from classes of cars. The split includes 130 training classes, 17 validation classes, and 49 testing classes.
- **Stanford-Dogs** [70]: Consists of 20,580 images of 120 dog breeds. The dataset is split into 70 training, 20 validation, and 30 testing classes.

Table 9: The performance of different backbone architectures. *indicates our implementation.

Method	Backbone	<i>miniImageNet</i>		<i>tieredImageNet</i>	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [11]	Resnet-12	62.39 $\pm$ 0.21	80.53 $\pm$ 0.14	68.23 $\pm$ 0.23	84.03 $\pm$ 0.16
ProtoNet*	Visformer-T	62.48 $\pm$ 0.35	79.78 $\pm$ 0.26	68.85 $\pm$ 0.37	83.65 $\pm$ 0.28
Meta-Baseline [22]	Resnet-12	63.17 $\pm$ 0.23	79.26 $\pm$ 0.17	68.62 $\pm$ 0.27	83.29 $\pm$ 0.18
Meta-Baseline*	Visformer-T	62.59 $\pm$ 0.34	79.88 $\pm$ 0.27	68.01 $\pm$ 0.35	82.75 $\pm$ 0.29
ours	Visformer-T	83.66$\pm$0.31	88.38$\pm$0.25	88.02$\pm$0.34	91.71$\pm$0.27

(3) Cross-Domain Few-Shot Classification Benchmarks. Following [84, 88, 91], we train the model on the *miniImageNet* training set and evaluate it on domain-shifted datasets:

- **CUB** [68]: Same as above.
- **Places** [71]: contains 73,000 images with 365 scene categories. The classes are split into 183 base classes, 91 validation classes, and 91 test classes with 18,200 images.
- **Plantae** [72]: A subset of the iNaturalist [72] dataset focusing on plant species with 68,461 images. We follow the 100/50/50 split for training, validation, and testing classes, with 17,253 images in the test set.

A.4.2 Ablation Study of Backbone Architectures

Table 9 showcases the performance comparison of different backbone architectures on the *miniImageNet* and *tieredImageNet* datasets under 1-shot and 5-shot settings. The results reveal that directly replacing the standard ResNet-12 backbone with the more advanced Visformer-T in existing baseline methods, such as ProtoNet and Meta-Baseline, fails to produce consistent performance gains. For example, in ProtoNet, the 1-shot accuracy only slightly improves from 62.39% to 62.48%, while the 5-shot accuracy even drops from 80.53% to 79.78%. A similar trend is observed for Meta-Baseline, where performance fluctuates around the original ResNet-12 values without notable improvement. This suggests that simply substituting the backbone does not effectively enhance the feature extraction or task-specific generalization ability of these methods.

In contrast, our method, VT-FSL, achieves a substantial performance boost when built upon the Visformer-T backbone. Specifically, VT-FSL attains 83.66% and 88.38% in 1-shot and 5-shot settings on *miniImageNet*, and 88.02% and 91.71% on *tieredImageNet*, outperforming all backbone-based baselines by a large margin. This significant improvement is attributed to the effective design of VT-FSL, which fully unleashes the potential of the backbone through cross-modal prompts and geometric-aware alignment. By leveraging rich cross-modal semantic information and global consistency among all features, VT-FSL dynamically enhances the extracted features and facilitates more discriminative class-specific representations, leading to superior generalization in few-shot scenarios.

A.4.3 The Effect of Different Kernel Functions

Table 10: The effect of different kernel functions from kernelized volume-based contrastive loss

Type	<i>miniImageNet</i>		CIFAR-FS		<i>tieredImageNet</i>	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Linear	82.01 $\pm$ 0.30	87.10 $\pm$ 0.27	87.33 $\pm$ 0.33	90.25 $\pm$ 0.28	87.35 $\pm$ 0.36	90.07 $\pm$ 0.28
Poly	82.76 $\pm$ 0.31	87.80 $\pm$ 0.28	87.82 $\pm$ 0.33	89.87 $\pm$ 0.30	87.51 $\pm$ 0.36	90.70 $\pm$ 0.29
RBF	83.66$\pm$0.31	88.38$\pm$0.25	88.67$\pm$0.32	91.45$\pm$0.28	88.02$\pm$0.34	91.71$\pm$0.27

The Effect of Kernel Functions in CGA . We evaluate the impact of different kernel functions used in our kernelized volume-based contrastive loss. The three kernels considered are:

$$\kappa(\mathbf{x}, \mathbf{y}) = \begin{cases} \mathbf{x}^\top \mathbf{y}, & \text{Linear kernel} \\ (\mathbf{x}^\top \mathbf{y} + c)^d, & \text{Polynomial kernel (Poly)} \\ \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), & \text{Radial Basis Function (RBF)} \end{cases} \quad (35)$$

Table 10 reports the results on *miniImageNet*, CIFAR-FS, and *tieredImageNet* under both 1-shot and 5-shot settings. We observe that the RBF kernel consistently yields the best performance across all datasets. For instance, it achieves 83.66% and 88.38% on *miniImageNet*, 88.67% and 91.45% on CIFAR-FS, outperforming the other two kernels.

The superior performance of the RBF kernel is attributed to its ability to project data into an infinite-dimensional Hilbert space, capturing fine-grained nonlinear relationships among textual, support, and synthetic visual embeddings. In contrast, the polynomial kernel introduces limited nonlinearity and is sensitive to hyperparameters such as degree and offset. The linear kernel performs the worst, as it lacks the expressiveness required to model complex cross-modal structures and effectively reduces to a dot product in the original feature space. These results validate the effectiveness of kernelized geometric alignment and confirm the choice of the RBF kernel as the default setting in our approach.

A.4.4 The Effect of Anchor Selection in Contrastive Learning

Table 11: The effect of different Anchors from kernelized volume-based contrastive loss

Type	<i>miniImageNet</i>		CIFAR-FS		<i>tieredImageNet</i>	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Vision	82.70±0.30	87.60±0.28	87.91±0.35	90.58±0.30	87.30±0.36	90.92±0.29
Text	83.66±0.31	88.38±0.25	88.67±0.32	91.45±0.28	88.02±0.34	91.71±0.27

The Effect of Anchor Selection in Contrastive Learning. We conduct an ablation study to investigate the effect of anchor selection in the kernelized volume-based contrastive loss. Specifically, we compare two anchor choices: (1) **Text**, where the textual feature Z_t generated by the LLM serves as the anchor; and (2) **Vision**, where the synthetic visual feature Z_v , generated by the text-to-image model from the same class description, is used as the anchor. The results are reported in Table 11. As shown, using the textual feature as the anchor consistently leads to better performance across all datasets and settings. For instance, on *miniImageNet*, the accuracy improves from 82.70% to 83.66% in the 1-shot setting and from 87.60% to 88.38% in the 5-shot setting when switching from the Vision to the Text anchor. Similar trends are observed on CIFAR-FS (88.67% vs. 87.91%) and *tieredImageNet* (88.02% vs. 87.30%) under the 1-shot setting.

This result validates our default choice of **textual anchor** in contrastive learning. Although both generated text and synthetic images are used to enrich the support representation Z_s , the textual feature Z_t provides a class-level semantic anchor that is more stable and discriminative. In contrast, the synthetic visual feature Z_v , while visually grounded, is more prone to noise and distributional artifacts due to limitations in generative models. Furthermore, aligning support and visual features around a textual anchor enables the model to centralize its representation around high-level semantics, which is particularly beneficial under few-shot scenarios where visual variability is high and labeled samples are scarce. Therefore, anchoring on the generated text helps better guide the cross-modal geometric alignment and leads to more robust class representations.

A.4.5 The Effect of Different LLMs and Text-to-Image Models

We investigate the effect of large language models (LLMs) and Text-to-Image (T2I) Models on VT-FSL: LLM used for class description generation, and the text-to-image model used for synthetic image generation.

Large language models. As illustrated in Fig 9a, we compare GPT-4o [33] and Qwen2.5-VL-32B [34] under both 1-shot and 5-shot settings on *miniImageNet* and CIFAR-FS. GPT-4o slightly outperforms Qwen2.5-VL-32B in 1-shot settings and maintains comparable or superior performance in 5-shot settings. The consistent results validate the robustness of VT-FSL across different LLMs,

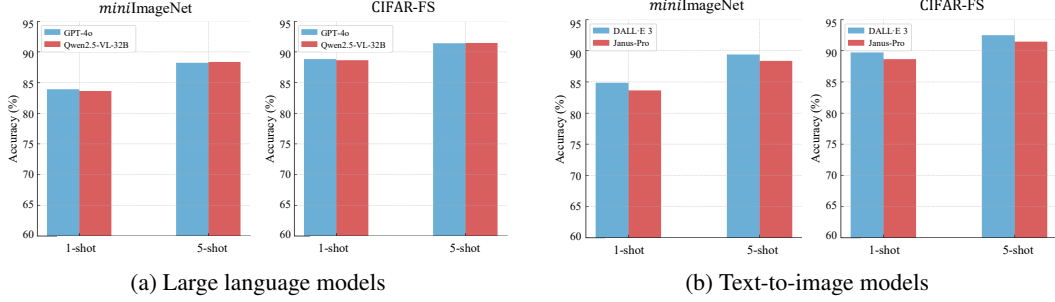


Figure 9: The Effect of different LLMs and text-to-image models

but also suggest that GPT-4o’s structured reasoning and richer world knowledge provide more semantically precise class descriptions, which are particularly beneficial when visual samples are scarce.

Text-to-image models. In Fig 9b, we examine the impact of two popular T2I models: DALL·E 3 [35] and Janus-Pro [36]. Results demonstrate that DALL·E 3 provides stronger visual priors than Janus-Pro, yielding higher accuracy across both datasets and settings. This is likely because DALL·E 3 generates more visually consistent and semantically faithful images from class-level descriptions, improving the quality of visual support features used during training. Despite Janus-Pro being multimodally capable, its generation quality is more variable and less optimized for fine-grained alignment with textual descriptions.

Overall, these results highlight that while VT-FSL is generally robust to backbone choices, the selection of higher-capacity and better-aligned LLMs and T2I models can further boost performance by improving the quality of generated semantic priors.

A.5 Limitations

While VT-FSL demonstrates strong performance across a variety of few-shot classification benchmarks, several limitations remain. First, although we evaluate the method on cross-domain datasets such as Places and Plantae, these settings still exhibit certain similarities to the source domain. The robustness of VT-FSL under more challenging distribution shifts (e.g., medical images) has not been fully assessed. Second, VT-FSL depends on the quality of external generative models to provide textual descriptions and synthetic visual samples. While the proposed Cross-modal Iterative Prompting (CIP) fully leverages the reasoning and generation capabilities of large models to produce semantically rich descriptions, the overall quality of the generated content remains inherently bounded by the capacity of the underlying models. For example, weaker LLMs may produce generic or noisy descriptions, and low-quality image synthesis could introduce misleading visual signals. Finally, although we introduce a kernelized volume-based contrastive loss to enhance multimodal alignment, its theoretical behavior under high-dimensional, noisy, or semantically entangled feature distributions remains underexplored. Further study is needed to rigorously understand its convergence properties and sensitivity to kernel choice. We hope these observations inspire future improvements in robustness, interpretability, and generalization for few-shot multimodal learning.