HISF: Hierarchical Interactive Semantic Fusion for Multi-Modal Prompt Learning

Haohan Feng * and Chen Li *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0717.v1

Keywords: multi-modal prompt learning; semantic alignment; representation learning; hierarchical feature fusion

 doi:10.20944/preprints202511.0717.v1

Article

HISF: Hierarchical Interactive Semantic Fusion for Multi-Modal Prompt Learning

Haohan Feng [1,2,*] and Chen Li [1,2,*]

[1] State Key Laboratory of Media Convergence and Communication, School of Information and Communication Engineering, China

[2] Communication University of China, 100024 Beijing, China

* Correspondence: fenghaohan@cuc.edu.cn (H.F.); lichengood@163.com (C.L.)

Abstract

Recent vision-language pre-training models, like CLIP, have been shown to generalize well across a variety of multitask modalities. Nonetheless, their generalization for downstream tasks is limited. As a light-weight adaptation approach, prompt learning could allow task transfer by optimizing only several learnable vectors, and thus is more flexible for pre-trained models. However, current methods mainly concentrate on the design of unimodal prompts and ignore effective means for multimodal semantic fusion and label alignment, which limits their representation power. To tackle these problems, this paper designs a Hierarchical Interactive Semantic Fusion (HISF) framework for multimodal prompt learning. On top of frozen CLIP backbones, HISF injects visual and textual signals simultaneously in intermediate layers of a Transformer through a cross-attention mechanism as well as fitting category embeddings. This architecture realizes the hierarchical semantic fusion at the modality level with structural consistency kept at each layer. In addition, a Label Embedding Constraint and a Semantic Alignment Loss are proposed to promote category consistency while alleviating semantic drift in training. Extensive experiments across 11 few-shot image classification benchmarks show that HISF improves the average accuracy by around 0.7% compared to state-of-the-art methods and has remarkable robustness in cross-domain transfer tasks. Ablation studies also verify the effectiveness of each proposed part and their combination: hierarchical structure, cross-modal attention, and semantic alignment collaborate to enrich representational capacity. In conclusion, the proposed HISF is a new hierarchical view for multimodal prompt learning and provides a more lightweight and generalizable paradigm for adapting vision-language pre-trained models.

Keywords: multi-modal prompt learning; semantic alignment; representation learning; hierarchical feature fusion

1. Introduction

Massive vision-language pre-trained models like CLIP 2 have recently made great progress in multimodal understanding. Trained from hundreds of millions of image-text pairs 2, these models learn to align visual and textual information into a shared semantics space, and thus provide exciting zero-shot generalization performance across many different recognition tasks. Large-scale models need to be adapted further down-stream, and fine-tuning such large models is expensive and data-consuming, which implies the exploration of parameter-efficient adaptation techniques.

Swift learning has recently become a popular option for knowledge transfer while preserving small differences in the model, which allows only limited parameter updating 3. With the addition of a few learnable prompt vectors, these methods condition the frozen backbone to execute new tasks without changing its base structure. Previous methods, such as CoOp 4 (Context Optimization) and Co-CoOp 5 (Conditional CoOp), have worked successfully by exposing to task-specific textual

 doi:10.20944/preprints202511.0717.v1

prompts. However, these methods are bound to single-modality optimization, and they often do not generalize well to unseen categories or cross-domain scenarios due to the insufficient multi-level multimodal interaction and semantic alignment.

To overcome these limitations, the most recent studies have extended prompt learning to a multimodal space. Representative methods MaPLe 6 and LAMM 7 have provided visual as well as text prompts, which led to better semantical correspondence between the modalities. However, these methods still suffer from shallow semantic fusion–prompts are injected at most to the input level and their cross-modal associations are only partially captured in the Transformer framework 8,9. Accordingly, the method has difficulty in modeling hierarchical semantic relations and deep (inter)action between visual and textual representations.

In this paper, we present a Hierarchical Interactive Semantic Fusion (HISF) framework for fine-grained hierarchical multimodal prompt learning. Based on a frozen CLIP backbone, HISF extends it with a dual-branch prompt structure that enables the joint optimization of visual and textual prompts across multiple Transformer layers using cross-attention to fuse category embeddings and prompt tokens. This architecture can allow for the holistic semantic representation to be effectively propagated across the network in a top-down fashion, improving both intra-class correlation and inter-class discrimination of feature maps. There is also a Label Embedding Constraint and Semantic Alignment Loss to project the label embedding space into the representations learned from the prompt while training to prevent semantic drift as well as facilitating cross-domain generalization.

The main contributions of this work are summarized as follows:

1. Hierarchical Semantic Fusion Framework:

We propose a novel hierarchical prompt learning paradigm that injects multimodal prompts into multiple Transformer layers, allowing progressive semantic interaction between visual and textual modalities.

2. Label-Guided Cross-Attention Mechanism:

A category embedding-guided cross-attention module is designed to dynamically align label semantics with multimodal prompts, achieving deep semantic binding between modalities.

3. Semantic Alignment and Label Constraints:

We introduce a joint optimization scheme consisting of a Label Embedding Constraint and a Semantic Alignment Loss to maintain consistent semantic distribution between the label and prompt spaces.

4. Parameter-Efficient and Robust Adaptation:

HISF requires updating only a small number of parameters while achieving superior performance on few-shot and cross-domain benchmarks, demonstrating both efficiency and generalization.

Extensive experiments on 11 benchmark datasets confirm that HISF consistently outperforms existing prompt learning methods, achieving approximately 0.7% improvement in average accuracy and stronger robustness under domain shifts. These results verify that hierarchical semantic fusion and cross-modal interaction are crucial for efficient and generalizable multimodal adaptation.

2. Related Work

2.1. Vision-Language Pre-Training

Recent progress has been made in this direction with the introduction of large-scale vision-language pre-training (VLP) 21212, such as OSCAR 12, BLIP-2 12, LLaVA 12, that leads to notable improvements for multimodal grounding and transfer learning efficiency through object-semantic alignment and vision-language bootstrapping. In particular, bottom-up and top-down attention mechanisms 12 play a crucial role in visual reasoning tasks such as image captioning and VQA, providing fine-grained contextual grounding for subsequent multimodal adaptation. In early works, including CLIP 2 (Contrastive Language-Image Pre-training), ALIGN 18, and LiT 19, large-scale image-text pairs are used to jointly train visual and text encoders contrastively. The goal is to bring

 doi:10.20944/preprints202511.0717.v1

similar image-text pairs closer together in a shared semantic space and dissimilar pairs further apart, matching disparate modalities. These models have shown strong zero-shot generalization capacity, which can directly perform open-vocabulary recognition 2 based on the input without explicit fine-tuning.

However, those models are highly dependent on large-scale data 3 and computational capabilities that make them unsuitable in domain-specific or low-data settings. Finetuning these models on downstream tasks directly is inefficient and often leads to overfitting. As a result, parameter-efficient adaptation methods (in particular prompt learning) have received significant attention 20, which allow adapting to a new task by updating only a small set of additional parameters while keeping the pre-trained backbone fixed.

2.2. Prompt Learning

Prompt learning originates from natural language processing (NLP), where learnable textual tokens are inserted into the input sequence to guide the model toward specific tasks. Methods such as PET, P-Tuning, and Prefix-Tuning have proven that optimizing only a few prompt vectors can achieve performance comparable to full fine-tuning, dramatically reducing computational cost.

In the vision language domain, this idea was first introduced to CLIP by CoOp 23 (Context Optimization), where hand-crafted textual templates were replaced by learnable contextual vectors (CoOp). As a result, this approach was able to achieve an effective enhancement of task-specific adaptation with efficiency. Nevertheless, CoOp's learned prompts generalize poorly to new categories, as the optimization is only guided by base-class semantics. In order to address this limitation, Co-CoOp 23 (Conditional CoOp) introduced an image-conditioned prompt generator so that the learned prompts would be able to adjust on a per-sample basis. Subsequent works (e.g., PromptStyler 21, ProDA 22, and DualCoOp 23) further explored distributional or style-aware prompt tuning. Other approaches, such as TextRefiner 23, CPT 23, SPGFF 12, focused on fine-grained textual refinement and color-based grounding to enhance visual-text correspondence. Visual Prompting 23 further revealed that lightweight visual perturbations can serve as effective task prompts for frozen models, offering a complementary direction to textual prompt learning. More recently, CoPrompt 23 introduced a consistency-guided regularization strategy to improve generalization under few-shot conditions.

Despite these advances, most prompt learning methods focus exclusively on the textual modality. The visual encoder remains unchanged, and cross-modal semantic interactions are largely ignored. As a result, the learned prompts capture shallow contextual semantics but fail to fully exploit the complementary relationship between vision and language.

2.3. Multi-Modal Prompt Learning

To address the limitations of single-modality prompt learning, multi-modal prompt learning (MMPL) 31 has emerged as a new paradigm that incorporates learnable prompts into both visual and textual branches. This design enables deeper semantic fusion and more balanced representation learning between modalities.

MaPLe (Multi-modal Prompt Learning) 6 injected prompts into intermediate Transformer layers from both visual and textual encoders and captured them via hierarchical semantic connections. It proposed a single initialization sequence and modality-specific projections that guarantee consistency (between joint modalities) while maintaining flexibility between them. LAMM (Label Alignment for Multi-modal Prompt Learning) 7: LAMM further augmented this pipeline with label embeddings and introduced a label-prompt alignment loss that explicitly aligns class semantics to the multimodal prompt using cross-attention. These techniques showed that alignment of semantics across modalities and class labels is a critical factor to enhance generalization in few-shot and cross-domain tasks.

Subsequent studies, such as MMRL 8 and DualPrompt 30, and BiMMPL 23, have further explored multimodal fusion strategies, including bidirectional projection between visual and textual

 doi:10.20944/preprints202511.0717.v1

prompts. Other efficient fusion frameworks, such as PMF 30 and MMA 30, focus on improving cross-modal communication and balancing generalization with discriminability across transformer layers. In addition, dynamic and efficient architectures, such as PCETL 8, demonstrate the necessity of balancing parameter efficiency and computational cost during model adaptation, highlighting the scalability issues in multimodal fusion. In addition, Multi-modal Alignment Prompt (MmAP) 8 extends this idea to a multi-task setting, aligning visual and textual modalities of CLIP through shared and task-specific prompts. Furthermore, question-driven prompt generation 23 introduces a language-conditioned mechanism that utilizes large language models to enhance prompt informativeness and cross-modal alignment, offering new insights for integrating textual reasoning into visual prompt design. Building on these advances, MaPLe 23 and LAMM 23 established hierarchical multimodal prompting frameworks, which jointly align visual and textual semantics to enhance downstream adaptability.

2.4. Semantic Alignment and Label Embedding

Semantic alignment refers to the common objective across all multiple modalities learning tasks where visual and textual features correspond to matching semantics. Common prompt-instantiating methods based on contrastive objectives may generate prompts that drift semantically 33 (prompt embeddings become different from the true label semantics during training). To address this, recent methods have incorporated label embeddings as semantic anchors.

LAMM 7 introduced a label alignment mechanism to regularize the distribution of label embeddings and text prompts in one unified semantic space, leading to stabilized prompt optimization. Another branch of study proposed hierarchical semantic injection, e.g., DeepPrompt29, where category embeddings are gradually integrated through Transformer layers. This not only helps the model capture high-level semantic abstraction but also fine-grained visual detail.

Building upon these ideas, our proposed HISF framework integrates label embeddings and cross-modal prompts through hierarchical attention layers. Similar hierarchical designs have also been discussed in recent works such as Hierarchical Prompt Tuning (HPT) 30, Learning with Enriched Inductive Biases (LwEIB) 30, and Slim Prompt-Averaged Consistency (SPAC) 30, which emphasize structured and consistent semantic learning across multiple levels of representation. Recent studies also emphasize the role of visual tokenization 30 and unsupervised prompt distillation 30 in strengthening semantic transfer, allowing models to better align class-level textual priors with image features. More recently, adaptive frameworks such as MetaPrompt [42] and ProVP [43] aimed to improve prompt robustness under domain shift by progressively refining category embeddings, providing complementary perspectives to our HISF approach.

By embedding category semantics into multiple levels of the Transformer, HISF achieves fine-grained semantic fusion and alleviates inconsistencies between visual and textual spaces.

2.5. Summary

In summary, vision-language pre-training provides a strong foundation for multimodal representation learning, while prompt learning offers a parameter-efficient path for task adaptation. Research has evolved from single-modal prompt optimization (e.g., CoOp, Co-CoOp) to multimodal prompt fusion (e.g., MaPLe, LAMM, MMRL), focusing increasingly on semantic alignment and hierarchical modeling.

Despite these advances, existing methods still face challenges in maintaining consistent cross-modal semantics throughout the Transformer architecture. Our proposed HISF framework addresses these limitations through hierarchical semantic fusion, label-guided cross-attention, and semantic alignment constraints, achieving more stable and generalized multimodal representation learning.

3. Method

 doi:10.20944/preprints202511.0717.v1

In this section, we describe our proposed framework called Hierarchical Interactive Semantic Fusion (HISF) for multimodal prompt learning. HISF can improve the semantic consistency and hierarchical interaction between visual and textual modality in vision-language pre-trained models, such as CLIP. The framework is composed of three key modules including: 1) Prompt Initialization to learn both visual and textual prompts, 2) Semantic Fusion to combine hierarchical semantics via cross-modal attention modulated between prompt tokens and category embeddings; 3) Loss Optimization by leveraging Label Embedding Constraint (LEC) and Semantic Alignment Loss (SAL) for semantic consistency between label space and prompt space.

The overall architecture of HISF is illustrated in Figure 1.

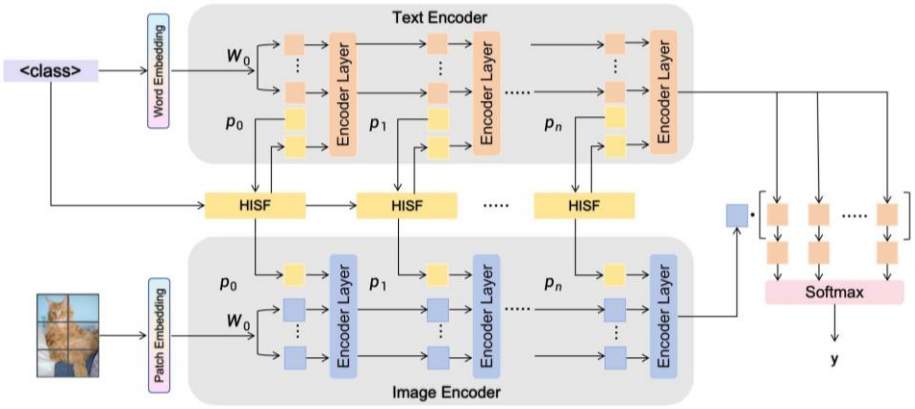

Figure 1. Overall architecture of the proposed HISF model. The model trains prompt vectors while keeping the text and visual encoders frozen. Prompt vectors are generated based on the input label and image embeddings, and then concatenated with the corresponding modality representations before being fed into the CLIP encoders.

3.1. Overall Framework

Given an input image I and its associated category label y, the aim of HISF is to learn a well-mapped multimodal representation so that the visual features can be well coordinated with their textual counterparts in a shared semantic space. Specifically, the visual encoder $f_v(\cdot)$ and text encoder $f_t(\cdot)$ of CLIP are frozen, then we introduce learnable prompts to guide the model's adaptation. The encoded image and text features can be denoted as:

$$z_v = f_v(I, P_v), z_t = f_t(T, P_t) \quad (1)$$

where P_v and P_t denote the visual and textual prompts, respectively, and T denotes the tokenized text corresponding to category y. The HISF framework introduces hierarchical semantic fusion across the Transformer layers of both encoders. Category embeddings are injected via cross-attention modules to enhance inter-modal communication, ensuring consistent semantic representation across hierarchical levels.

3.2. Prompt Initialization

The prompt initialization module serves as the foundation for efficient adaptation. Instead of relying on hand-crafted textual templates (e.g., "a photo of a [class]"), we employ learnable context tokens initialized from a shared semantic space. These learnable tokens are divided into two branches:

·Textual Prompts (P_t), which are appended to the text encoder's input sequence;

·Visual Prompts (P_v), which are inserted into the visual encoder's token sequence.

Following MaPLe, the prompts are initialized using a shared embedding matrix $P_s \in \mathbb{R}^{L_p * d}$, where L_p is the number of prompt tokens and d is the embedding dimension. The shared matrix ensures initial semantic coherence between modalities.

To accommodate modality-specific characteristics, we apply two linear projections:

 doi:10.20944/preprints202511.0717.v1

$$P_v = W_v P_s, P_t = W_t P_s \tag{2}$$

where W_v and W_t are learnable projection matrices for the visual and textual branches. This design allows the prompts to share a common initialization while learning modality-specific representations during optimization. Through this initialization, the HISF model gains a stable starting point for cross-modal alignment and reduces training instability typically observed in multimodal prompt learning.

3.3. Hierarchical Semantic Fusion

While previous multimodal prompt methods (e.g., MaPLe, LAMM) introduce prompts only at the input layer or shallow Transformer layers, HISF adopts a hierarchical injection strategy, enabling semantic fusion across multiple depths of the Transformer architecture.

Specifically, prompts are inserted into the visual and textual Transformers at layer $l \in \{L_1, L_2, L_3, \ldots\ldots, L_k\}$ where k is the number of fusion layers. This allows progressive semantic interaction between visual tokens, textual prompts, and category embeddings.

To strengthen inter-modal communication, we introduce a cross-attention mechanism between prompts and label embedding.

For a given layer, the semantic fusion process is formulated as:

$$Q_p = LayerNorm(P) \cdot W_q \tag{3}$$

$$K_l = E_i \cdot W_k \ (E_i = [e_1, e_2, \ldots, e_d]^T) \in \mathbb{R}^{k*d} \tag{4}$$

$$V_l = E_i \cdot W_v \tag{5}$$

where W_q, W_k, and W_v are learnable projection matrices for the prompt vectors and Q_p, K_l, and V_l correspond to the query, key, and value matrices. In HISF, the prompt tokens serve as queries Q_p, while the label embedding act as keys and values K_l and V_l.

The updated prompt representation at layer l is thus:

$$Attn = Softmax\left(\frac{Q_p \cdot {K_l}^T}{\sqrt{d}}\right) V_l \tag{6}$$

After computing cross-attention, construct and configure a multi-layer perceptron with a bottleneck structure to perform residual semantic augmentation on the weighted prompt vector, thereby accelerating the convergence speed of prompt training. The calculation formula is as follows:

$$P_{fused} = P + Attn \tag{7}$$

$$P_{final} = P_{fused} + MLP\left(P_{fused} \oplus E_i\right) \tag{8}$$

The prompt vector after incorporating semantic information is denoted by P_{fused}, where the operation symbol $\oplus$ represents concatenating the blended prompt vector with label information. This is then fed into a multilayer perceptron, and the final result P_{final} undergoes residual connection to yield the final prompt vector.

This operation allows category semantics to directly guide prompt optimization, aligning the representation of prompts with their corresponding class meanings. By performing this operation hierarchically across Transformer layers, HISF achieves multi-level semantic fusion, where shallow layers capture local appearance information and deeper layers encode abstract semantics.

This hierarchical fusion mechanism not only enhances intra-class feature compactness but also promotes inter-class separability, which is crucial for few-shot recognition and cross-domain generalization.

 doi:10.20944/preprints202511.0717.v1

3.4. Label Embedding and Semantic Alignment

A key innovation of HISF is the introduction of Label Embedding (LE) and Semantic Alignment (SA) mechanisms, which jointly constrain the relationship between prompts and category semantics.

Let $E_y \in \mathbb{R}^d$ denote the label embedding corresponding to category y. During training, we encourage the fused prompt representation P_{final} to remain semantically close to the corresponding label embedding E_i.

To achieve this, we define two complementary objectives:

1. Label Embedding Constraint (L_{le})—encourages the prompt embeddings to align with the label embedding in semantic space:

$$L_{Le} = \frac{1}{N}\sum_{i=1}^{N} ||P_{final} - E_i||_2 \tag{9}$$

2. Semantic Alignment Loss (L_{sa})—ensures consistency between the multimodal representations derived from prompts and the textual embeddings:

$$L_{sa} = \frac{1}{N}\sum_{i=1}^{N} \big(1 - cos(z_v, z_t)\big) \tag{10}$$

where z_v and z_t are the fused visual and textual features obtained after hierarchical semantic injection.

3. Class Prediction Loss (L_{cls})—dominates the recognition accuracy of classification tasks, enabling the representation features obtained through prompt vector concatenation to better approximate the actual content.

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} CE\big(\tau \cdot sim\big(I_x, f(P_{final} + E_i)\big), y_i\big) \tag{11}$$

where τ is a parameter learned from CLIP, f is the CLIP encoder and I_x is the presentation of image, while applies the cross-entropy (CE) loss for the similarity score.

We use the CLIP encoder $f(\cdot)$ and the representation of the image I_x, along with a parameter τ learned from CLIP. A cross-entropy (CE) loss is applied to the resulting similarity score.

The combined objective function is then expressed as:

$$L_{total} = L_{cls} + \lambda_1 \cdot L_{Le} + \lambda_1 \cdot L_{sa} \tag{12}$$

where L_{cls} denotes the original CLIP contrastive loss, and λ_1, λ_2 are balancing hyperparameters.

This formulation jointly optimizes the multimodal alignment and the prompt-label semantic consistency, ensuring that the learned prompts faithfully capture category-level meaning while preserving cross-modal coherence.

3.5. Optimization and Training Strategy

Meanwhile, we fix all the parameters of the CLIP backbone and only update the prompt tokens, projection matrices, and cross-attention parameters in training. We apply an episodic training strategy similar to few-shot learning, where each episode includes a small portion of the base categories to form tasks. The model generalizes to new classes by transferring the hierarchical semantics to them from category embeddings. To stabilize optimization, the learning rate of prompt parameters is greater than that of the projection layers, since prompts should adapt faster. The optimizer we use is AdamW, where the weight decay is only applied to projection matrices. We employ mixed-precision training for all experiments due to efficiency reasons. To summarize, the model is evaluated out-of-the-box in a few-shot and cross-domain setting without any further fine-tuning and shows high generalization capability.

 doi:10.20944/preprints202511.0717.v1

3.6. Discussion

The design motivation for HISF is to model hierarchical semantic relationships in the Transformer structure explicitly. As opposed to existing approaches that treat multimodal prompts as shallow tokens, HISF integrates category-conditioned cross-attention across multiple layers of the model and therefore enables it to iteratively refine its understanding of multimodality. The hierarchical fusion structure not only strengthens the expressiveness of intra-modal features, but also preserves the alignment of semantics between two modalities. Moreover, with the novel Label Embedding Constraint and Semantic Alignment Loss optimization process, we can successively regularize the learning procedure without overfitting on these tasks and achieve strong generalization to both few-shot and cross-domain settings.

In summary, HISF establishes a general and extensible framework for multimodal prompt learning, where hierarchical semantic fusion and label-guided attention serve as the core mechanisms driving efficient and interpretable model adaptation.

4. Experiments

We conduct extensive experiments to verify the effectiveness of our HISF method on 11 widely adopted image classification benchmarks. The datasets are Caltech101, ImageNet, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, UCF101 and DTD, EuroSAT. All evaluations are conducted under the few-shot learning setup: we take 16 samples (called a 16-shot setting) for each class as the training set and report performance on the remaining samples. In cross-dataset transfer experiments, models learned from ImageNet are directly tested on the other datasets without any further fine-tuning to evaluate generalization stability.

4.1. Experimental Hyperparameters

We use the pre-trained ViT-B/16 model as our CLIP backbone. The dimension of the text modality embedding is 512, while the image modality embedding is 768. Prompt tokens are placed from a particular Transformer layer (the index of this layer and the number of prompt tokens are described in the source settings). All models are trained for 50 epochs by setting the batch size to 4. We use SGD as the optimizer with a learning rate of 0.0035. We perform experiments with the NVIDIA A800 GPU.

4.2. Base-to-Novel Generalization Evaluation

For base-to-novel generalization, we adopt the same data split as in MaPLe, where the dataset classes are split into base and novel classes with 16 shots per class. The entire evaluation is done in two steps. At the first stage, a subset of classes is chosen as (base) classes, and the model is trained on a few-shot sample belonging to those base classes. For stage two, the classes outside of state one are regarded as novel classes; we adapt on K-shot support samples from these novel classes and test on the remaining novel-class samples. We report accuracy for the base class, novel class, and their mean (H.M). The specific comparison results over 11 datasets are shown in Table 1.

Table 1. Comparison of the proposed HISF method with recent state-of-the-art approaches across 11 benchmark datasets.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOP	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoOpOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43

 doi:10.20944/preprints202511.0717.v1

ProDA	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC	84.26	76.10	79.97	**77.60**	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ProVP	85.20	73.22	78.76	75.82	69.21	72.36	**98.92**	94.21	**96.51**	**95.87**	97.65	**96.75**
MetaPrompt	83.65	75.48	79.09	77.52	70.83	**74.02**	98.13	94.58	96.32	95.53	97.00	96.26
TCP	84.13	75.36	79.51	77.27	69.87	73.38	98.23	**94.67**	96.42	94.67	97.20	95.92
MMA	83.20	**76.80**	79.87	77.31	71.00	**74.02**	98.40	94.00	96.15	95.40	**98.07**	96.72
HISF	**85.39**	76.12	**80.75**	76.8	70.1	73.4	98.20	94.6	96.4	95.7	96.9	96.3

Method	StanfordCars			Flower102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoOpOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
KgCoOp	71.76	**75.04**	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe	72.94	74.00	73.47	95.92	72.46	82.56	90.71	**92.05**	**91.38**	37.44	35.61	36.50
PromptSRC	78.27	74.97	76.58	98.07	**76.50**	85.95	90.67	91.53	91.10	42.73	**37.87**	40.15
ProVP	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	47.08	29.87	36.55
MetaPrompt	76.34	75.01	75.48	97.66	74.49	84.52	**90.74**	91.85	91.29	40.14	36.51	38.24
TCP	80.80	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
HISF	**81.9**	74.3	**78.1**	**98.6**	74.2	**86.4**	89.5	90.7	90.1	**47.6**	33.8	**40.7**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoOpOP	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
KgCoOp	80.2	76.53	78.3	77.5	54.99	64.3	85.6	64.34	73.4	82.8	76.67	79.6

 doi:10.20944/preprints202511.0717.v1

	9		6	5		5	4		8	9		5
MaPLe	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ProVP	80.67	76.11	78.32	83.95	59.06	69.34	97.12	72.91	83.29	**88.56**	75.55	81.54
MetaPrompt	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
TCP	82.63	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	80.77	83.83
MMA	82.27	78.57	80.38	83.20	**65.63**	73.38	85.46	**82.34**	83.87	86.23	80.03	82.20
HISF	**83.5**	**79.3**	**81.4**	**83.6**	64.8	**74.2**	**97.3**	77.5	**87.4**	86.6	**81.2**	**83.9**

4.3. Cross-Dataset Transfer Evaluation

For the cross-dataset transfer evaluation, we adopt these 11 datasets. We first conduct a few-shot protocol training for the models on ImageNet, and then directly measure the generalization performance of the trained models across datasets on the remaining ten datasets. The transfer results are shown in Table 2. HISF achieves the best or competitive results on most target datasets, and significantly outperforms competing methods, especially on DTD and EuroSAT with distributional shift from ImageNet. This observation reflects that the hierarchical semantic injection mechanism is instrumental in improving cross-domain robustness, and makes pre-trained features more transferable to out-of-distribution tasks of interest.

Table 2. Comparison of HISF with existing methods in a cross-dataset setting.

Method	Source	Target									
	ImageNet	Caltech	OxfordPets	Stanford Cars	Flowers 101	Food101	FGVC Aircraft	SUN397	DTD	EuroSAT	UCF101
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	68.17	41.92	46.39	66.55
CoOpOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
TCP	71.40	93.97	91.25	64.49	71.21	86.69	23.45	67.15	44.35	51.45	68.73
MMA	71.00	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32
HISF	**72.20**	93.49	90.32	65.40	**72.41**	85.41	24.31	67.57	45.91	**54.70**	68.41

4.4. Domain Generalization Evaluation

Domain generalization: We test generalized performance on ImageNet, and directly train on the source domain (ImageNet), without fine-tuning the model with target data. These test sets have the same class labels as ImageNet but involve distributional shifts. Table 3 shows the domain transfer performance of our method and other comparisons, including CLIP, CoOp, Co-CoOp, MaPLe, PromptSRC, and HISF. HISF is competitive across domain-shifted benchmarks as reflected in Table 3 with their average numbers (e.g., on ImageNet, HISF reaches 72.20% (source)) and improves over baselines for all the other domain shifts, like ImageNet-Sketch, ImageNet-A, ImageNet-R, and on six of eight subsets from Imagenet-V2. This highlights the relatively higher robustness of HISF to distributional variations.

 doi:10.20944/preprints202511.0717.v1

Table 3. Comparison of HISF with existing methods in the domain generalization setting.

Method	Source	Target			
	ImageNet	-S	-A	-R	-V2
CLIP	66.73	46.15	47.77	73.96	60.83
CoOp	71.51	47.99	49.71	75.21	64.20
CoOpOp	71.02	48.75	50.63	76.18	64.07
Maple	70.72	49.15	50.90	76.98	64.07
PromptSRC	71.27	49.55	50.90	77.80	64.35
HISF	72.20	49.51	**51.20**	**78.21**	**64.43**

4.5. Ablation Studies

We perform a series of ablation experiments to analyze the contributions of individual design choices and loss components. The ablation studies include: (1) prompt insertion depth; (2) HISF module effectiveness (Prompt Initialization PI and Semantic Fusion SF); and (3) loss term influences.

4.5.1. Prompt Insertion Depth

We explore how the prompt token insertion depth can influence average accuracy. Results show that when the prompts are fed into increasingly deeper layers, the performance of the model typically increases. Performance drops when the prompt tokens are injected into encoder layers that are low; however, deeper layers often store more generalized features, while shallower layers instead encode dataset-specific discriminative patterns. As a result, prompt insertion too low hurts performance on the base class, and intermediate levels of layer insertion achieve better accuracy. But note that going too deep for insertion will also deteriorate the performance, most likely it is because we are restricting how many learnable parameters interact with CLIP's representation, or our control over backbone networks. The curve is given in Figure 2 (prompt depth vs average accuracy).

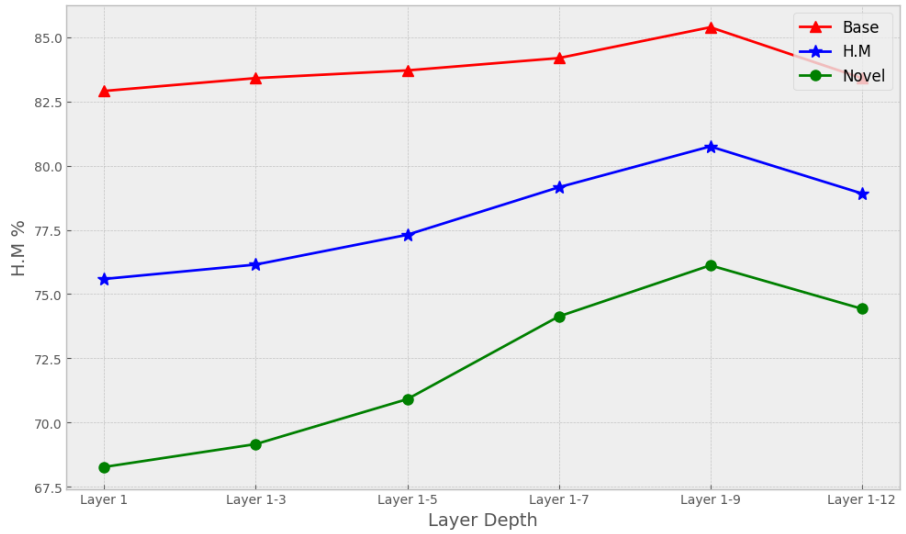

Figure 2. Influence of prompt depth on the model's average accuracy.

4.5.2. HISF Component Effectiveness

We validate the effort of HISF components by performing module-wise ablation. PI represents the Prompt initialization, and SF is short for Semantic Fusion. Results of HISF on the MaPLe experimental setting and its ablated variants are shown in Table 4. The ablative studies confirm the significance of each module towards HISF's performance. Taking out the Prompt Initialization module (w/o PI), there are negligible shifts in base and novel accuracies under the MaPLe setting, indicating that shared initialization has a stabilizing effect at the beginning of training. Without the Semantic Fusion module (w/o SF), we obtain another remarkable performance drop, verifying that the hierarchical approach plays a critical role in aligning modalities. The model achieves the best

 doi:10.20944/preprints202511.0717.v1

performance when both PI and SF are exploited (PI + SF) (Base 85.39, Novel 76.12, H.M 80.75), showing that prompt initialization and semantic fusion are complementary: The former establishes a consistent semantic entry point to each task, while the latter imposes deep cross-modal interplays for enhanced discrimination ability and generalizability.

Table 4. Evaluation of the impact of different HISF modules on model performance.

Method	Base	Novel	H.M
HISF (Maple setting)	82.28	75.14	78.55
HISF (w/o PI)	83.76	75.57	79.66
HISF (w/o SF)	82.45	74.96	78.71
HISF (PI + SF)	85.39	76.12	80.75

4.5.3. Loss Component Influence

We test to what extent the three loss terms (label embedding loss, semantic alignment loss, and classification loss) contribute to overall performance by conducting experiments to remove each part independently. The performance degrades significantly if we remove the label embedding loss, which suggests that this loss is crucial for learning discriminative class-specific representations. The degradation of cross-dataset transfer performance and domain generalization if the semantic alignment loss is disabled: The model's inference accuracy on the target datasets decreases, which shows that it can help in forming more effective semantic correspondence across different datasets for distributional change robustness. The quantitative results in Table 5 collectively confirm that the three losses facilitate together to make a contribution to the base-to-novel generalization, cross-dataset transfer, and domain robustness; the combinational effect from them is crucial to obtain the performance in experiments of HISF.

Table 5. Evaluation of the effect of different loss functions on model performance.

L_{CE}	L_S	L_{LP}	H.M
		√	79.41
√	√	√	80.75
	√	√	80.67
√		√	80.35

4.6. Summary

The above experiments show the effectiveness of HISF under several evaluation protocols. HISF outperforms existing methods on base-to-novel generalization, cross-dataset transfer, and domain generalization. Ablation studies confirm the efficacy of timely insertion depth, prompt initialization, semantic fusion, and the joint loss design. These findings empirically support our claim that the hierarchical semantic injection and label-constrained alignment work together to enhance visual identification with language-based meta-training for prompt-based adaptation of vision-language pre-trained models.

5. Conclusion

In this work, we present a new multimodal prompt learning framework called HISF (Hierarchical Interleaved Semantic Fusion). In this paper, we propose a hierarchical semantic fusion (HISF) to make bidirectional information flow between visual and textual modality, thus substantially improving fine-grained semantic alignment. With hierarchical fusion and label-guided learning, richer multimodal interaction is captured, and better generalization in base as well as novel categories is obtained.

 doi:10.20944/preprints202511.0717.v1

In particular, the PI module gives a uniform initialization, which stabilizes training and prevents overfitting in few-shot cases; while the SF module learns multi-level semantics to facilitate deeper interaction between modalities. Moreover, a label embedding and semantic alignment-based joint optimization approach can prevent information loss, and meanwhile discriminate the learned features effectively.

Extensive experiments across eleven benchmark datasets validate the superiority of HISF over existing prompt learning methods. Ablation results further confirm that each proposed component contributes meaningfully to performance improvement, particularly in maintaining semantic coherence and enhancing model adaptability under limited data conditions.

In future work, we plan to extend HISF to broader multimodal tasks such as video-language understanding and visual question answering, and to explore adaptive fusion strategies for dynamic prompt interaction. We believe that the proposed framework offers a promising direction for developing more generalizable and interpretable multimodal prompt learning systems.

Author Contributions: H.F. (Haohan Feng), and C.L. (Chen Li) contributed equally to the conception, development, writing, editing, and analysis of this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data relevant to the study are included in the article.

Acknowledgments: During the preparation of this manuscript, the author(s) used ChatGPT-4 for the purposes of structuring the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HISF	Hierarchical Interactive Semantic Fusion
PI	Prompt Initialization
LE	Label Embedding
LA	Label Alignment

References

1. Liu, Y.; Deng, Y.; Liu, A.; Liu, Y.; Li, S. Fine-grained multi-modal prompt learning for vision–language models. *Neurocomputing* **2025**, *636*, 130028. https://doi.org/10.1016/j.neucom.2025.130028.
2. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
3. Gu, Y.; Han, X.; Liu, Z.; Huang, M. PPT: Pre-trained prompt tuning for few-shot learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 8410–8423. https://doi.org/10.18653/v1/2022.acl-long.576.
4. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. https://doi.org/10.1007/s11263-022-01653-1.
5. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16795–16804. https://doi.org/10.1109/CVPR52688.2022.01631.
6. Khattak, M.U.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F.S. MaPLe: Multi-modal prompt learning. In

 doi:10.20944/preprints202511.0717.v1

Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19113–19122. https://doi.org/10.1109/CVPR52729.2023.01832.

7. Gao, J.; Ruan, J.; Xiang, S.; Yu, Z.; Ji, K.; Xie, M.; Liu, T.; Fu, Y. LAMM: Label alignment for multi-modal prompt learning. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 1815–1823. https://doi.org/10.1609/aaai.v38i3.27950.
8. Guo, Y.; Gu, X. MMRL: Multi-modal representation learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 25015–25025.
9. Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting vision transformers for scalable visual recognition. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November 2022–9 December 2022; pp. 16664–16678. https://doi.org/10.5555/3600270.3601482.
10. Du, Y.; Liu, Z.; Li, J.; Zhao, W.X. A survey of vision-language pre-trained models. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022), Vienna, Austria, 23–29 July 2022; pp. 5436–5443. https://doi.org/10.24963/ijcai.2022/762.
11. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Jitsev, J.; Komatsuzaki, A. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. In Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual Event, 6–14 December 2021; pp. 1–5.
12. Chen, F.; Zhang, D.; Han, M.; Chen, X.; Shi, J.; Xu, S.; Xu, B. VLP: A survey on vision-language pre-training. *Mach. Intell. Res.* **2023**, *20*, 38–56. https://doi.org/10.1007/s11633-022-1369-5.
13. Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, J.; Zhou, J.; Zhou, C. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *arXiv* **2022**, arXiv:2211.01335. https://doi.org/10.48550/arXiv.2211.01335.
14. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the 16th European Conference on Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 121–137. https://doi.org/10.1007/978-3-030-58577-8_8.
15. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA, 23–29 July, 2023; pp. 19730–19742. https://doi.org/10.5555/3618408.3619222.
16. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 34892–34916. https://doi.org/10.5555/3666122.3667638.
17. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086. https://doi.org/10.1109/CVPR.2018.00636.
18. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 4904–4916.
19. Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; Beyer, L. LIT: Zero-shot transfer with locked-image text tuning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP), New Orleans, LA, USA, 18–24 June 2022; pp. 18102–18112. https://doi.org/10.1109/CVPR52688.2022.01759.
20. Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; Qiao, Y. CLIP-Adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.* **2024**, *132*, 581–595. https://doi.org/10.1007/s11263-023-01891-x.
21. Cho, J.; Nam, G.; Kim, S.; Yang, H.; Kwak, S. PromptStyler: Prompt-driven style generation for source-free domain generalization. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 15656–15666. https://doi.org/10.1109/ICCV51070.2023.01439.
22. Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; Tian, X. Prompt distribution learning. In Proceedings of the 2022 IEEE/CVF

 doi:10.20944/preprints202511.0717.v1

Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5196–5205. https://doi.org/10.1109/CVPR52688.2022.00514.

23. Sun, X.; Hu, P.; Saenko, K. DualCoOp: Fast adaptation to multi-label recognition with limited annotations. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November 2022–9 December 2022; pp. 30569–30582. https://doi.org/10.5555/3600270.3602486.
24. Xie, J.; Zhang, Y.; Peng, J.; Huang, Z.; Cao, L. TextRefiner: Internal visual feature as efficient refiner for vision-language models prompt tuning. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 8718–8726. https://doi.org/10.1609/aaai.v39i8.32942.
25. Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.S.; Sun, M. CPT: Colorful prompt tuning for pre-trained vision-language models. *AI Open* **2024**, *5*, 30–38. https://doi.org/10.1016/j.aiopen.2024.01.004.
26. Zhang, J.; Wang, S. Text-guided visual prompt learning with semantic prompt generation and feature fusion. *Neurocomputing* **2025**, *654*, 131253. https://doi.org/10.1016/j.neucom.2025.131253.
27. Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; Isola, P. Exploring visual prompts for adapting large-scale models. *arXiv* **2022**, arXiv:2203.17274. https://doi.org/10.48550/arXiv.2203.17274.
28. Roy, S.; Etemad, A. Consistency-guided prompt learning for vision-language models. *arXiv* **2023**, arXiv:2306.01195. https://doi.org/10.48550/arXiv.2306.01195.
29. Zhu, J.; Ruan, Y.; Chang, J.; Sun, W.; Wan, H.; Long, J.; Luo, C. Deep prompt multi-task network for abuse language detection. In Proceedings of the 27th International Conference on Pattern Recognition, Kolkata, India, 1–5 December 2024; pp. 249–263. https://doi.org/10.1007/978-3-031-78107-0_16.
30. Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.Y.; Ren X.; Su, G.; Perot, V.; Dy, J.; et al. DualPrompt: Complementary prompting for rehearsal-free continual learning. In Proceedings of the 17th European Conference on Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 631–648. https://doi.org/10.1007/978-3-031-19809-0_36.
31. Yin, H.; Zhao, Y. Multi-modal prompt learning with bidirectional layer-wise prompt fusion. *Inf. Fusion* **2025**, *117*, 102919. https://doi.org/10.1016/j.inffus.2024.102919.
32. Li, Y.; Quan, R.; Zhu, L.; Yang, Y. Efficient multimodal fusion via interactive prompting. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2604–2613. https://doi.org/10.1109/CVPR52729.2023.00256.
33. Yang, L.; Zhang, R.; Wang, Y.; Xie, X. MMA: Multi-modal adapter for vision-language models. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 23826–23837. https://doi.org/10.1109/CVPR52733.2024.02249.
34. Wu, Q.; Yu, W.; Zhou, Y.; Huang, S.; Sun, X.; Ji, R. Parameter and computation efficient transfer learning for vision-language pre-trained models. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 41034–41050. https://doi.org/10.5555/3666122.3667908.
35. Xin, Y.; Du, J.; Wang, Q.; Yan, K.; Ding, S. MmAP: Multi-modal alignment prompt for cross-domain multi-task learning. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 16076–16084. https://doi.org/10.1609/aaai.v38i14.29540.
36. Özdemir, Ö.; Akagündüz, E. Enhancing visual question answering through question-driven image captions as prompts. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 1562–1571. https://doi.org/10.1109/CVPRW63382.2024.00163.
37. Wang, Y.; Jiang, X.; Cheng, D.; Li, D.; Zhao, C. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 5749–5757. https://doi.org/10.1609/aaai.v38i6.28387.
38. Yang, L.; Zhang, R.; Chen, Q.; Xie, X. Learning with enriched inductive biases for vision-language models. *Int. J. Comput. Vis.* **2025**, *133*, 3746–3761. https://doi.org/10.1007/s11263-025-02354-1.
39. He, S.; Wang, S.; Long, S. A slim prompt-averaged consistency prompt learning for vision–language model. *Knowl. Based Syst.* **2025**, *310*, 113011. https://doi.org/10.1016/j.knosys.2025.113011.
40. Wang, G.; Ge, Y.; Ding, X.; Kankanhalli, M.; Shan, Y. What makes for good visual tokenizers for large language models?. *arXiv* **2023**, arXiv:2305.12223. https://doi.org/10.48550/arXiv.2305.12223.

 doi:10.20944/preprints202511.0717.v1

41. Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; Yang, J. PromptKD: Unsupervised prompt distillation for vision-language models. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16 –22 June 2024; pp. 26607–26616. https://doi.org/10.1109/CVPR52733.2024.02513.
42. Zhao, C.; Wang, Y.; Jiang, X.; Shen, Y.; Song, K.; Li, D.; Miao, A. Learning domain invariant prompt for vision-language models. *IEEE Trans. Image Process.* **2024**, *33*, 1348–1360. https://doi.org/10.1109/TIP.2024.3362062.
43. Xu, C.; Zhu, Y.; Shen, H.; Chen, B.; Liao, Y.; Chen, X.; Wang, L. Progressive visual prompt learning with contrastive feature re-formation. *Int. J. Comput. Vis.* **2025**, *133*, 511–526. https://doi.org/10.1007/s11263-024-02172-x.