

Article

Bidirectional Multi-Modal Spatial-Consistent Prompt Tuning for Enhanced Vision-Language Alignment

Sikai Bai [†], Haoxi Li [†] and Jie Zhang ^{1,*}

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China

* Correspondence: csejzhang@cse.ust.hk

[†] These authors contributed equally to this work.

How To Cite: Bai, S.; Li, H.; Zhang, J. Bidirectional Multi-modal Spatial-Consistent Prompt Tuning for Enhanced Vision-Language Alignment. *Edge Intelligence and Systems* **2026**, *1*(1), 6.

Received: 7 May 2026

Revised: 30 July 2026

Accepted: 6 August 2026

Published: 24 August 2026

Abstract: Multi-modal prompt tuning in vision-language models like CLIP has significantly enhanced machine comprehension and interaction with multi-modal content. However, prior works have primarily focused on unidirectional semantic alignment across modalities, limiting their effectiveness in dynamic and open-world visual concept comprehension due to loosely constrained mappings. To address this limitation, we introduce a novel Bidirectional Multi-modal Spatial-consistent Prompt Tuning (BMSPT) framework to ensure a dynamic and fine-grained alignment between text and image modalities. Specifically, our proposed framework facilitates bidirectional interactions between modalities through two novel Vision-Language (V-L) and Language-Vision (L-V) transfer functions. This bidirectional alignment is supported by a modality-alignment loss using the Wasserstein distance to enhance semantic coherence and a content-alignment loss to ensure information integrity across transformations. Extensive experiments on 11 diverse image recognition benchmarks demonstrate that BMSPT can significantly improve adaptability and accuracy and achieve an adaptive, multi-modal learning systems.

Keywords: multiple modality; prompt tuning; Vision-Language model

1. Introduction

Pre-trained vision-language models such as CLIP [1] have demonstrated remarkable zero-shot capabilities across various downstream tasks by encoding large-scale datasets into a unified representational space through contrastive learning [2,3]. At the inference stage, CLIP traditionally relies on manually crafted text prompts (e.g., ‘a photo of [class]’) to align output classes with pre-trained knowledge. However, this reliance on static prompts limits the system’s flexibility and adaptability due to their sensitivity to semantic variations. Recent advancements in Natural Language Processing (NLP) [4–7] have inspired the development of dynamic, learnable prompts to maintain the semantic integrity between sentences and labels [8–10]. These learnable prompts allow the pre-trained model to adapt dynamically to new tasks without the need for re-training [11,12].

When it comes to computer vision tasks [13,14] and multimodal learning tasks [15,16], early research efforts concentrated on exploring learnable prompts within vision-language models (VLMs), especially in CLIP, to improve their generalization capabilities and achieve better performance in zero-shot and few-shot learning tasks [17,18]. Language prompting involves constructing trainable language contexts that serve as inputs to the text encoder, with these tokens being optimized during cross-modal contrastive learning.

As a pioneering framework in prompt learning for visual tasks, CoOp [9] introduced learnable prompt contexts into CLIP. It retains the pre-trained CLIP text encoder in a fixed state while optimizing the learnable contexts, thereby minimizing additional training overhead. Subsequently, CoCoOp [10] was introduced as an



advanced framework that dynamically conditions context learning on latent visual features. It is achieved through a lightweight meta-network and bridges visual representations with adaptable textual contexts. Different from CoOp and CoCoOp, TCP [19] introduced Textual Knowledge Embedding (TKE) to map class-level textual knowledge into class-aware prompts, enhancing generalization and discriminability for unseen classes by integrating these prompts into the text encoder. Instead of using fixed text templates like TCP, CuTCP [20] further utilized large language models to create detailed, category-specific prompts and provided richer semantic information. However, these language prompting predominantly rely on language models to conduct vision-language alignment. They fail to ensure spatial consistency between the captured visual information and the learnable language contexts, while also struggling to fully unleash the optimization potential of prompt tuning paradigms [21,22].

Furthermore, Visual Prompt Tuning (VPT) has gained prominence by applying vision prompts to the vision transformers and tuning very few learnable parameters [23]. For instance, by directly introducing task-specific learnable parameters into the input sentences of the transformer layers in the vision encoder (left part of Figure 1) [24,25], VPT strategies can make full use of the potential of image encoders to model the adaptation. However, this patch-level focus might restrict the visual modality's natural expressiveness due to its coarse-grained pattern. Figure 1 vividly demonstrates how traditional methods might misclassify isolated patches of a zebra's coat as belonging to a hyena, due to the similarity in patch patterns. Therefore, it is critical for the desired methods to respect the holistic nature of visual data and enhance its inherent qualities without compromising on task-relevant feature extraction.

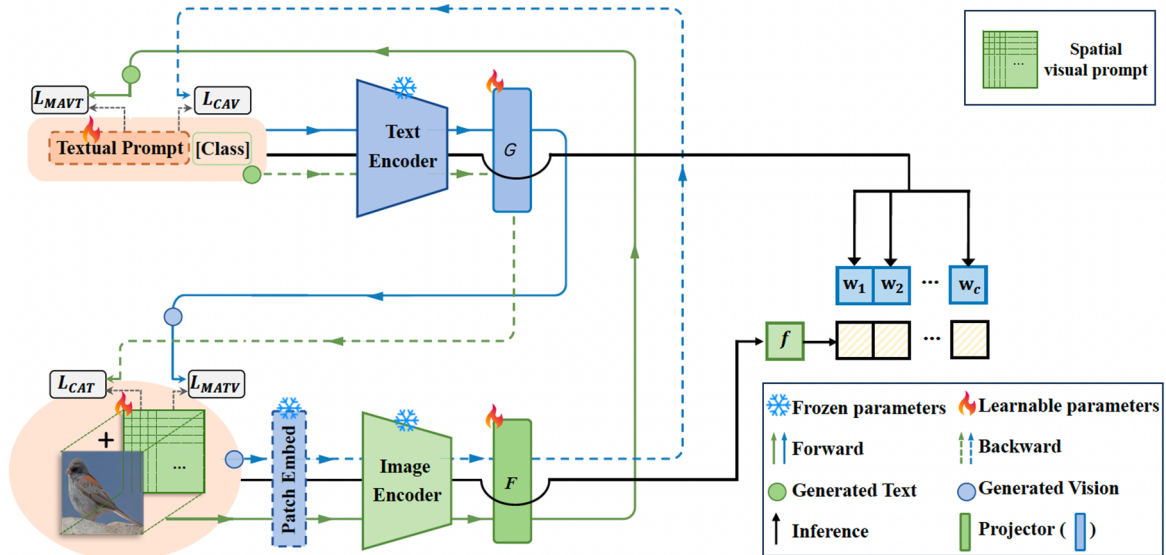


Figure 1. Overview of the proposed BMSPT framework. BMSPT first aligns modalities by minimizing the W-distance between the distributions of transformed and original data (modality-alignment loss), which are applied to both L-V and V-L directions. Then we propose content-alignment loss on each side to preserve content integrity across modalities, specifically visual-to-text-to-visual and text-to-visual-to-text.

Another trend sprouting up at prompt tuning is the integration of Textual Prompt Tuning (TPT) and VPT, striving to enhance multi-modal prompts across both domains synergistically [18,21,22]. In this case, a unidirectional alignment strategy is usually employed during the tuning process, where textual descriptions guide the visual recognition process [26–28]. However, these methods often lead to significant communication challenges between modalities alignment. As depicted in right upper part of Figure 1, a ‘dog’—symbolizing textual data—transmits its characteristics from textual into a visual space, is fraught with uncertainties. It is unclear whether the data maintains its integrity, remains undistorted, or is interpreted correctly by the visual system. Consequently, a ‘dog’ might be mistakenly recognized as a ‘bear’ due to distortions and misinterpretations. This example underscores the risks involved in one-way communication between modalities and highlights the necessity for a robust, bidirectional multi-modal prompt tuning framework.

Inspired by the above observations, we introduce a novel prompt tuning framework, named Bidirectional Multi-modal Spatial-consistent Prompt Tuning (BMSPT), which employs specific Language-Vision (L-V) and Vision-Language (V-L) transfer functions to facilitate dynamic interactions between modalities while preserving the original spatial dimensions of visual data. Specifically, the L-V transfer function maps the original image

aligned with our proposed spatial visual prompt to the textual semantic space, ensuring spatial and semantic integrity. Conversely, the V-L transfer function accurately reflects textual insights back into the visual domain. To ensure robust alignment and semantic coherence across the bidirectional transformations, we further incorporate two-pronged alignment loss functions: a modality-alignment loss utilizing the Wasserstein distance for optimal distribution matching, and a content-alignment loss that maintains semantic fidelity throughout the transformation cycles. Extensive experiments on several popular image recognition benchmarks demonstrate the superiority and effectiveness of our proposed BMSPT framework. The contributions of the paper are summarized as follows:

- To the best of our knowledge, this is the first work that introduces Bidirectional Multi-modal Spatial-consistent Prompt Tuning (BMSPT) to align image and textual prompts within the original image dimension.
- We design an innovative two-pronged alignment mechanism to harmonize textual and visual prompts, ensure the semantic integrity of transformations across multiple modalities.
- We apply our methodology across diverse scenarios, including few-shot classification, dataset transfer learning, domain generalization, image editing, video understanding, object detection and segmentation. The extensive experimental results demonstrate the superiority of our proposed method.

2. Materials and Methods

In this section, we introduce the Bidirectional Multi-modal Spatial-Consistent Tuning (BMSPT) framework, which is meticulously designed to ensure fine-grained alignment across modalities. The methodology is structured into three primary components: initialization, the overview of BMSPT framework, and the bidirectional multi-modal prompt tuning.

2.1. The Overview of BMSPT Framework

Figure 1 presents an overview of our BMSPT framework, which integrates the foundational architectures of CLIP's pre-trained encoders to support bidirectional interactions between textual and visual modalities. The framework employs innovative transfer functions and a two-pronged training approach, enabling multi-modal prompts to achieve semantic-coherent and information-integral representations across transformations, thereby ensuring effective multi-modal alignment.

The core of the BMSPT framework is a bidirectional alignment strategy that meticulously preserves the spatial dimensions during cross-modal interactions. The Vision-to-Language (V-L) projector, denoted as F , adapts visual data to the textual embedding space while maintaining spatial coherence. Conversely, the Language-to-Vision (L-V) projector, represented by G , converts textual insights back into visual representations, ensuring the retention of spatial integrity.

To foster the development of semantic-coherent and distribution-consistent representations, we employ a two-pronged strategy that fine-tunes prompts within both the V-L and L-V projections. This approach significantly enhances the framework's adaptability, particularly in scenarios characterized by unbalanced datasets and limited sample sizes. By aligning visual prompts P_v in the visual space V with textual prompts P_t in the textual space T , we optimize the cross-modal interaction.

Utilizing CLIP's encoder in conjunction with our V-L projector, we establish a mapping $F \circ f: V \rightarrow T$ that aligns the outputs $\hat{y} = \hat{f}(f(\alpha P_v \oplus x))$ with the target text $[P_t, y] \in T$, consistent with the Wasserstein distance. This setup theoretically induces an output distribution for $\hat{y}$ that closely mirrors the empirical distribution $p_{data}(y)$, ensuring the stochastic nature of F . Consequently, the function $F \circ f$ transforms visual inputs into a domain $\tilde{T}$ that closely matches T .

Furthermore, we incorporate a content-alignment loss that reinforces the preservation of content integrity across modalities, expressed as $(G \circ g)(F \circ f)(x) \approx x$ and $(F \circ f)(G \circ g)(y) \approx y$. This loss, combined with modality-alignment loss for both visual and textual domains, forms a comprehensive objective for our bidirectional multi-modal spatial-consistent alignment, further enhancing semantic integrity and modal harmony.

2.2. Bidirectional Multi-Modal Prompt Tuning

Let x denote an input image with class label c , and let s_c denote the textual description of class c . We use E_v and E_t to denote the frozen CLIP visual and textual encoders, respectively. The prompted visual and textual representations are defined as

$$z_v = E_v(x \oplus P_v), \quad z_t = E_t([P_t; s_c]). \quad (1)$$

Here, P_v is the learnable spatial visual prompt, P_t denotes the learnable textual context tokens, $\oplus$ represents the spatial prompt-composition operation, and $[\cdot; \cdot]$ denotes concatenation.

BMSPT contains two learnable transfer functions. The Vision-to-Language projector $F_{V \rightarrow T}: \mathcal{V} \rightarrow \mathcal{T}$ maps a visual representation to the textual representation space, whereas the Language-to-Vision projector $G_{T \rightarrow V}: \mathcal{T} \rightarrow \mathcal{V}$ maps a textual representation to the visual representation space. The transformed representations are

$$\hat{z}_t = F_{V \rightarrow T}(z_v), \quad \hat{z}_v = G_{T \rightarrow V}(z_t). \quad (2)$$

(a) Modality-Alignment Loss

The modality-alignment objective reduces the distribution discrepancy between the transformed representations and the representations in the corresponding target modality. We formulate the two transfer directions separately to avoid ambiguity. Let p_V and p_T denote the distributions of visual and textual representations, respectively. Let D_T and D_V denote the critic functions defined in the textual and visual representation spaces. For the Vision-to-Language direction, the modality-alignment loss is defined as

$$\mathcal{L}_{MA}^{V \rightarrow T}(F_{V \rightarrow T}, D_T) = \mathbb{E}_{z_t \sim p_T}[D_T(z_t)] - \mathbb{E}_{z_v \sim p_V}[D_T(F_{V \rightarrow T}(z_v))]. \quad (3)$$

This objective measures the Wasserstein distance between the distribution of real textual representations z_t and that of the transformed visual representations $\hat{z}_t = F_{V \rightarrow T}(z_v)$. For the Language-to-Vision direction, the corresponding modality-alignment loss is

$$\mathcal{L}_{MA}^{T \rightarrow V}(G_{T \rightarrow V}, D_V) = \mathbb{E}_{z_v \sim p_V}[D_V(z_v)] - \mathbb{E}_{z_t \sim p_T}[D_V(G_{T \rightarrow V}(z_t))]. \quad (4)$$

This objective measures the Wasserstein distance between the distribution of real visual representations z_v and that of the transformed textual representations $\hat{z}_v = G_{T \rightarrow V}(z_t)$.

Both D_T and D_V are constrained to be 1-Lipschitz functions. Following the weight-clipping strategy, after each critic update, their parameters w are clipped to the bounded interval $w \leftarrow \text{clip}(w, -c, c)$, where $c > 0$ is the clipping threshold specified in the implementation details. During optimization, the critics D_T and D_V maximize Equations (3) and (4), respectively, to estimate the corresponding Wasserstein distances. In contrast, the transfer functions $F_{V \rightarrow T}$ and $G_{T \rightarrow V}$ minimize these distances so that the transformed representations approach the distributions of their target modalities.

(b) Content-Alignment Loss

The modality-alignment losses constrain the distributions of the generated representations to match their corresponding target modalities. However, distribution-level alignment alone does not guarantee that the semantic content of each individual representation is preserved during transformation.

For textual content preservation, a textual representation z_t is first mapped into the visual space and then transformed back into the textual space:

$$\tilde{z}_t = F_{V \rightarrow T}(G_{T \rightarrow V}(z_t)).$$

The textual content-alignment loss is defined as

$$\mathcal{L}_{CA}^T(F_{V \rightarrow T}, G_{T \rightarrow V}) = \mathbb{E}_{z_t \sim p_T}[\|\tilde{z}_t - z_t\|_1]. \quad (5)$$

This loss encourages the text-to-visual-to-text transformation to preserve the original textual content. For visual content preservation, a visual representation z_v is first mapped into the textual space and then transformed back into the visual space:

$$\tilde{z}_v = G_{T \rightarrow V}(F_{V \rightarrow T}(z_v)).$$

Because visual samples belonging to the same category may exhibit natural appearance variations, exact reconstruction is not required. Instead, no reconstruction penalty is imposed when the recovered visual representation lies within an ϵ -neighborhood of the original representation. The visual content-alignment loss is therefore defined as

$$\mathcal{L}_{CA}^V(F_{V \rightarrow T}, G_{T \rightarrow V}) = \mathbb{E}_{z_v \sim p_V}[\mathbb{I}(\|\tilde{z}_v - z_v\|_2 > \epsilon) \|\tilde{z}_v - z_v\|_1], \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function:

$$\mathbb{I}(\|\tilde{z}_v - z_v\|_2 > \epsilon) = \begin{cases} 1, & \|\tilde{z}_v - z_v\|_2 > \epsilon, \\ 0, & \|\tilde{z}_v - z_v\|_2 \leq \epsilon. \end{cases}$$

Here, ϵ controls the tolerated visual variation. Equation (8) applies an L_1 reconstruction penalty only when the distance between the reconstructed and original visual representations exceeds this tolerance.

Together, the textual and visual content-alignment losses encourage the bidirectional transformations to preserve sample-specific semantic information while allowing limited visual variation.

(c) Full Objective

In addition to the bidirectional alignment losses, the model is optimized using the task-specific classification loss

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}_{(x,c)}[\log p(c | x)],$$

where $p(c | x)$ is computed from the similarity between the prompted image representation and the class-specific textual representations.

The complete BMSPT objective is formulated as $\min_{F_{V \rightarrow T}, G_{T \rightarrow V}, P_v, P_t, D_T \in \text{Lip}_1, D_V \in \text{Lip}_1} \max \mathcal{L}_{\text{BMSPT}}$ with

$$\mathcal{L}_{\text{BMSPT}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{VT} \mathcal{L}_{\text{MA}}^{V \rightarrow T} + \lambda_{TV} \mathcal{L}_{\text{MA}}^{T \rightarrow V} + \lambda_T \mathcal{L}_{\text{CA}}^T + \lambda_V \mathcal{L}_{\text{CA}}^V. \quad (7)$$

where Lip_1 denotes the set of 1-Lipschitz functions. The coefficients λ_{cls} , λ_{VT} , λ_{TV} , λ_T , and λ_V control the relative contributions of the classification, modality-alignment, and content-alignment objectives.

The critic parameters are optimized to maximize the two Wasserstein objectives, whereas the prompts and bidirectional transfer functions are optimized to minimize the complete objective. The CLIP visual and textual encoders remain frozen throughout training.

Through these mechanisms, BMSPT combines task-specific prompt learning, bidirectional distribution alignment, and cycle-based content preservation to improve semantic consistency between the textual and visual modalities.

3. Results

To make a comprehensive evaluation, we perform extensive experiments on 4 task settings, such as base-to-novel generalization and cross-dataset transfer learning. Those experiments are conducted on 15 widely used image datasets, varying in scale and domains, including ImageNet [29], Caltech101 [30], OxfordPets [31], StanfordCars [32], Flowers102 [33], Food101 [34], FGVCAircraft [35], EuroSAT [36], UCF101 [37], DTD [38], SUN397 [39], ImageNetV2 [40], ImageNet-Sketch [41], ImageNet-A [42] and ImageNetR [43].

3.1. Implementation Details

We use the official OpenAI CLIP ViT-B/16 checkpoint as the vision-language backbone. All input images are converted to RGB and processed at a resolution of 224×224 pixels. During training, we apply random resized cropping to 224×224 using bicubic interpolation, followed by random horizontal flipping, conversion to a tensor, and normalization using the CLIP-specific channel means (0.48145466, 0.4578275, 0.40821073) and standard deviations (0.26862954, 0.26130258, 0.27577711). During evaluation, the shorter edge of each image is resized to 224 pixels using bicubic interpolation, followed by a 224×224 center crop, tensor conversion, and the same CLIP normalization. Both the CLIP visual encoder and text encoder remain frozen throughout training. The textual prompt contains 1024 learnable context tokens. The Language-to-Vision transfer module first projects each 512-dimensional textual feature into a 128-dimensional representation using a fully connected layer, batch normalization, and LeakyReLU. The projected representation is reshaped to $C \times 128 \times 1 \times 1$ and concatenated with 96-dimensional Gaussian noise, yielding a latent tensor of shape $C \times 224 \times 1 \times 1$. The class-conditioned outputs are averaged to obtain a spatial visual prompt of shape $3 \times 224 \times 224$. We optimize the learnable prompts and transfer functions using SGD with a learning rate of 0.0025, momentum 0.9, and weight decay 5×10^{-4} . A constant warm-up learning rate of 1.2×10^{-5} is used for the first two epochs, followed by cosine-annealing learning-rate scheduling. The batch size is 4. Few-shot experiments are trained for 25 epochs, while base-to-novel, cross-dataset, and domain-generalization experiments are trained for 30 epochs. We report the mean and standard deviation over the three independent runs. All experiments are conducted using an NVIDIA A100 GPU with 80 GB memory.

3.2. Experimental Protocol

For the base-to-novel experiments, we follow the class partitions provided by CoOp. Each model is trained using only the base-class training samples and evaluated separately on the base-class and novel-class test sets. The harmonic mean is calculated as

$$\text{HM} = \frac{2A_{\text{base}}A_{\text{novel}}}{A_{\text{base}} + A_{\text{novel}}}. \quad (8)$$

For few-shot learning, we sample $K \in \{1, 2, 4, 8, 16\}$ training examples per class using the same sampled subsets for all compared methods. The subsets are generated independently for each random seed, and their image indices are included in the released split files.

For cross-dataset transfer, the model is trained on ImageNet and directly evaluated on the ten target datasets without target-dataset adaptation. For domain-shift evaluation, the ImageNet-trained model is directly evaluated on ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R.

The 11 standard image-recognition datasets are used for base-to-novel and few-shot evaluation. Together with the four ImageNet distribution-shift datasets, the complete study contains 15 image datasets.

3.3. Performance Comparison

We can find that BMSPT demonstrates strong zero-shot generalization in Table 1, improving novel category performance by 1.49% over TCP and 0.75% over CuTCP. Unlike many methods that struggle to maintain zero-shot performance while learning new tasks, BMSPT achieves substantial gains in both zero-shot and few-shot scenarios. Specifically, on the SUN 397 dataset, BMSPT shows a 0.41% improvement in the base categories and 1.53% improvement in the novel categories over TCP. In terms of the harmonic mean on EuroSAT dataset, BMSPT shows a 1.56% improvement over CuTCP and a 2.68% improvement over TCP. When evaluating the harmonic mean for individual datasets, BMSPT outperforms other state-of-the-art methods across 9 out of 11 datasets, further validating its efficacy and versatility.

Table 1. Performance comparison of our proposed BMSPT with state-of-the-art methods on base-to-novel generalization. The best accuracies are bolded. HM indicates the harmonic mean.

Methods	Base	Average Novel	HM	Base	ImageNet Novel	HM	Base	Caltech 101 Novel	HM	Base	Oxford Pets Novel	HM
CLIP	69.34	74.22	71.70	72.34	68.14	70.18	96.84	94.00	95.39	91.17	97.26	94.11
CoOp	82.69	63.22	71.66	76.14	67.88	71.77	98.00	89.81	93.72	93.67	95.29	94.47
CoCoOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
QCoOp	80.68	74.44	77.43	76.17	70.73	73.35	95.57	74.67	83.84	95.17	97.60	96.37
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
QMaPLe	83.02	75.57	79.12	76.93	70.73	73.70	97.97	95.00	96.46	95.67	97.63	96.64
PromptSR C	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.67	97.93	96.79
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
CuTCP	84.21	76.10	79.95	77.73	70.50	73.94	98.47	95.27	96.84	95.07	97.83	96.43
BMSPT	84.89 ± 0.14	76.85 ± 0.21	80.46 ± 0.17	78.22 ± 0.12	71.12 ± 0.18	74.50 ± 0.15	98.50 ± 0.08	94.82 ± 0.15	96.62 ± 0.11	95.69 ± 0.10	98.10 ± 0.07	96.88 ± 0.08
Stanford Cars			Flowers 102			Food 101			FGVC Aircraft			
	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	68.12	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
QCoOp	73.73	72.90	73.31	95.57	74.67	83.84	90.87	91.90	91.38	37.50	34.03	35.68
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
QMaPLe	75.00	73.67	74.33	96.43	74.33	83.95	90.63	92.10	91.36	39.10	34.90	36.88
PromptSR C	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
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
CuTCP	80.23	74.27	77.13	98.10	75.58	85.38	90.47	91.77	91.11	42.43	36.37	39.17
BMSPT	79.40 ± 0.25	75.12 ± 0.31	77.2 ± 0.27	98.11 ± 0.09	76.92 ± 0.28	86.23 ± 0.18	90.79 ± 0.11	91.85 ± 0.10	91.32 ± 0.09	43.20 ± 0.35	37.90 ± 0.42	40.38 ± 0.37
SUN 397			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
CoCoOp	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
QCoOp	79.20	77.93	78.56	74.97	58.37	65.63	83.53	69.80	76.05	81.87	75.93	78.79
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
QMaPLe	81.33	78.27	79.77	80.77	57.63	67.27	94.30	79.47	86.25	85.10	77.50	81.12
PromptSR C	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
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
CuTCP	83.00	78.23	80.55	83.00	59.40	69.24	90.87	77.13	83.44	86.87	80.80	83.72
BMSPT	83.04 ± 0.16	79.76 ± 0.20	81.37 ± 0.17	83.43 ± 0.28	63.01 ± 0.39	71.80 ± 0.31	95.90 ± 0.22	76.33 ± 0.45	85.00 ± 0.29	87.43 ± 0.18	80.42 ± 0.24	83.78 ± 0.20

Few-Shot Learning. we evaluate our model on few-shot scenario, where models are trained on 1, 2, 4, 8, and 16 shots and then applied to the test sets. We report the average accuracy scores of all models across 11 datasets in

Figure 2a. Overall, our proposed BMSPT outperforms other methods in most cases, demonstrating BMSPT’s ability to distill cross-modal knowledge and efficiently adapt a pre-trained vision- language model to downstream tasks.

Transfer Learning. To investigate the generalizability of our BMSPT framework across datasets and domains, we first train our model on ImageNet, utilizing all 1000 classes, and then apply the model to ten other datasets for cross-dataset evaluation. The results are presented in Table 2. It shows BMSPT outperforms state-of-the-art methods on 7 out of 11 cross-dataset benchmarks, achieving the highest average accuracy 67.18%. Meanwhile, BMSPT demonstrates consistent performance superiority in some challenging datasets (e.g., Stanford Cars, Aircraft, Sun397, DTD, EuroSAT, and UCF101), outperforming TCP and PromptSRC. These results indicate BMSPT’s robustness in knowledge transfer across various datasets with different characteristics.

Table 2. Cross-dataset benchmark evaluation.

	Source				Target							
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	Sun397	DTD	EuroSAT	UCF101	Average
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
QCoOp	70.63	94.07	90.53	65.97	71.33	86.23	22.73	66.80	44.20	48.23	69.17	65.93
PromptS												
RC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
TCP	71.40	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73	66.29
BMSPT	71.81 ± 0.15	94.20 ± 0.09	90.66 ± 0.11	65.99 ± 0.27	71.81 ± 0.25	86.52 ± 0.12	24.98 ± 0.36	67.41 ± 0.20	46.88 ± 0.32	48.95 ± 0.40	68.96 ± 0.23	67.18 ± 0.17

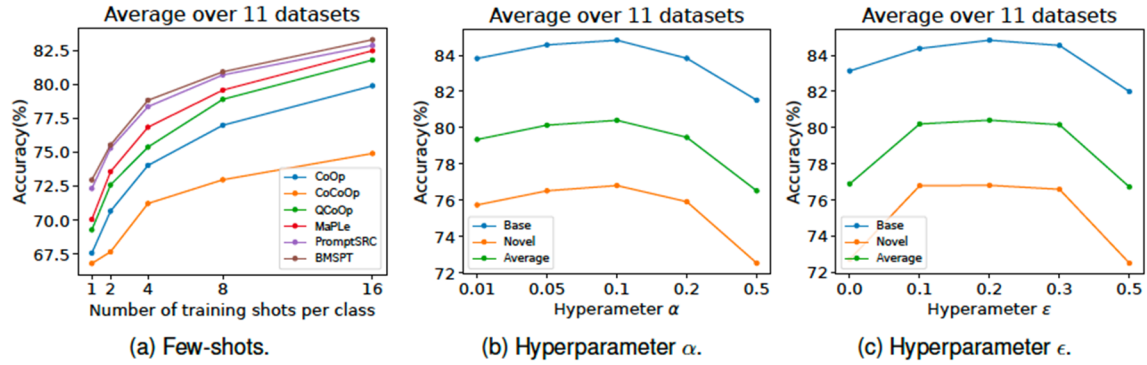


Figure 2. Quantitative analysis of BMSPT averaged over 11 image-recognition datasets: (a) classification accuracy under the 1-, 2-, 4-, 8-, and 16-shot settings; (b) sensitivity to the spatial visual-prompt coefficient α ; and (c) sensitivity to the visual content-alignment tolerance ϵ . Error bars denote the standard deviation over three independent runs.

3.4. Ablation Study

Components Analysis. To further provide an in-depth demonstration of the effectiveness of the bidirectional architecture in our BMSPT framework, we validate the role of each module within the BMSPT framework. As shown in Table 3, the analysis of different components reveals the contributions of each module within BMSPT. The baseline IVLP achieves a harmonic mean (HM) of 77.90%, while IVLP + Init improves this to 79.48%, demonstrating the benefit of the initialized stage. IVLP + L-V and IVLP + V-L further enhance performance, achieving HMs of 79.87% and 79.59%, respectively, highlighting the importance of Language-to-Vision and Vision-to-Language prompts projection. Semi-BMSPT uses a one-way modality alignment loss, which isolates cycle-consistency alignment and results in poor performance. BMSPT with L1 and L2 regularization achieves HMs of 80.25% and 80.07%, respectively, indicating the role of regularization in refining the model. The complete BMSPT framework incorporates bidirectional alignment and Wasserstein distance, and achieves the highest HM of 80.46%, underscoring the effectiveness of each module in contributing to the overall performance. These results validate the importance of bidirectional cycle-consistency alignment, initialization, and regularization in the BMSPT framework, ensuring a robust and efficient model.

Hyperparameter Analysis. We first explore the influence of the spatial visual prompt ratio (α) on the performance of our BMSPT framework, with results depicted in Figure 2b. Our analysis across 11 datasets indicates that tuning α is crucial, as setting it to 0.1 maximizes accuracy, yielding the highest average results across base and novel classes. This emphasizes the importance of precisely adjusting the prompt ratio to enhance model

performance. Concurrently, we assess the effects of the boundary support coefficient (ϵ), which dictates the tolerance for visual variability within the same class, shown in Figure 2c. Adjusting ϵ also significantly impacts performance, with optimal settings leading to peak accuracy. These findings highlight the necessity of finely tuning both α and ϵ to ensure accurate content translation and to accommodate variability across modalities, thus optimizing the overall effectiveness of the BMSPT framework.

Qualitative Analysis. Besides the extensive qualitative results, we explore the visual performance of our BMSPT. Specifically, we compare the bidirectional BMSPT model with a semi-BMSPT variant to understand the impact of our proposed bidirectional alignment mechanism. As illustrated in Figure 3, our BMSPT focuses on distinct visual features that are semantically relevant to the class label. To quantitatively evaluate the precision of these attention maps, we compute the Energy-based Pointing Game metric, which measures the fraction of attention energy captured within the ground-truth bounding boxes of the target objects. Across 500 randomly sampled images from the evaluation pool, BMSPT achieves an average localization accuracy of 84.3%, significantly surpassing Semi-BMSPT's 71.2%. This empirically validates that bidirectional alignment forces the model to focus on tighter and more semantically rigorous core features rather than dispersed local background noise. Although BMSPT exhibits comprehensive improvements across various multi-modal benchmarks, it is not without limitations. In the spirit of scientific candor, we observe that our framework faces boundary degradation under extreme low-light conditions or when dealing with highly occluded objects. In these severe scenarios, the spatial visual information integrity is drastically compromised. Consequently, the Vision-to-Language projector F receives heavily corrupted image tokens and passes ambiguous embeddings to the text space. This error propagates through the cycle-consistency content-alignment loop, occasionally causing the model to over-correct or hallucinate wrong contextual features. Future work will explore integrating robust low-light visual restorers or robust multi-modal representations to address these extreme OOD sensory anomalies.

Table 3. Analysis of different components.

	Base	Novel	HM
IVLP	82.15 $\pm$ 0.20	74.07 $\pm$ 0.27	77.90 $\pm$ 0.22
IVLP-Init	83.76 $\pm$ 0.17	75.94 $\pm$ 0.24	79.48 $\pm$ 0.19
IVLP + L-V	84.35 $\pm$ 0.15	76.21 $\pm$ 0.21	79.87 $\pm$ 0.19
IVLP + V-L	84.36 $\pm$ 0.14	75.72 $\pm$ 0.23	79.59 $\pm$ 0.18
Semi-BMSPT	84.23 $\pm$ 0.14	76.20 $\pm$ 0.31	80.01 $\pm$ 0.16
BMSPT(w L1)	84.69 $\pm$ 0.12	76.25 $\pm$ 0.18	80.25 $\pm$ 0.14
BMSPT(w L2)	84.28 $\pm$ 0.13	76.26 $\pm$ 0.19	80.07 $\pm$ 0.15
BMSPT	84.89 $\pm$ 0.14	76.85 $\pm$ 0.21	80.46 $\pm$ 0.17

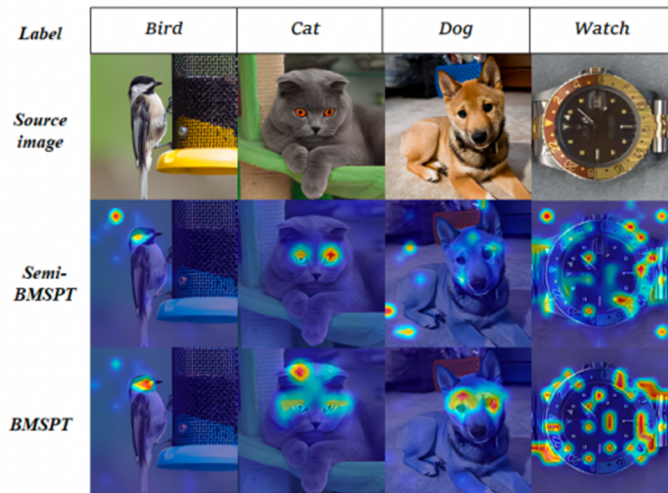


Figure 3. Visualization analysis for the effectiveness of bidirectional prompt tuning.

Efficiency Analysis. To investigate the trade-off between the performance and efficiency of our BMSPT, we conduct an analysis of parameters and computational complexity. We observe that multi-modal prompt tuning (MPT) methods, such as CoCoOp and MaPLe, generally require more trainable parameters and longer inference times compared to single-modal methods. As illustrated in Table 4, BMSPT, which employs a bidirectional alignment strategy with a two-pronged approach to learn multimodal prompts, achieves a superior balance. It attains the highest accuracy of 80.46% with fewer parameters (3,347,810) than MaPLe (7,165,440) and maintains

a competitive inference speed of 102 FPS, significantly faster than CoCoOp (37 FPS) and MaPLe (62 FPS) due to the design of spatial visual prompts. While our bidirectional training regularizations introduce a slight overhead in training time (485 s/epoch) and peak memory (16.1 GB) compared to CoCoOp, it remains significantly lighter than MaPLe during training. Crucially, it maintains a high inference speed of 102 FPS and the highest accuracy of 80.46%. These findings highlight BMSPT’s effectiveness in harmonizing parameter efficiency, computational complexity, and accuracy.

Table 4. Efficient analysis of parameters and computational complexity.

	#Param	FPS	Accuracy
CoOp	2048	645	71.66
CoCoOp	35,360	37	75.83
MaPLe	7,165,440	62	78.55
BMSPT	3,347,810	102	80.46

Analysis of Generalization Ability. (1) Video Understanding: we report the comparisons on the accuracy with state-of-the-art method A5 [44] in terms of video understanding in Table 5. The results demonstrate that BMSPT achieves superior performance across all metrics. Specifically, BMSPT attains a base accuracy of 71.73%, and a novel accuracy of 52.64%, higher than A5’s 70.4% and 51.7%, respectively. Furthermore, BMSPT’s harmonic mean (HM) of 60.50% outperforms A5’s HM of 59.6%. These results highlight the effectiveness of BMSPT in video understanding tasks, showcasing it generalizes well across both base and novel categories on the HMDB51 dataset [45]. (2) Object Detection and Segmentation: We conduct an exploratory evaluation on the VisA anomaly-analysis benchmark. For image-level anomaly detection, BMSPT achieves an AUROC of 90.8% and an AP of 91.9%, compared with 78.1% and 81.2% for WinCLIP [46]. These results indicate that BMSPT is effective for image-level anomaly discrimination. As presented in the Table 6, these results on VisA dataset [47] highlight the superior performance of BMSPT in object detection, achieving an AUROC of 90.8% and an AP of 91.9%, significantly outperforming WinCLIP’s AUROC of 78.1% and AP of 81.2%. For pixel-level anomaly segmentation, BMSPT achieves an AUROC of 64.6% and an AUPRO of 52.9%. Although BMSPT outperforms the standard CLIP and CoOp baselines, its segmentation performance remains below WinCLIP, which achieves an AUROC of 79.6% and an AUPRO of 56.8%. WinCLIP is specifically designed for dense patch-level anomaly localization, whereas BMSPT primarily performs global cross-modal prompt alignment. We therefore do not claim superiority over WinCLIP in pixel-level segmentation. These results demonstrate strong image-level anomaly-detection performance while also revealing the limitation of BMSPT in pixel-level anomaly localization.

Table 5. Additional experiments in terms of video understanding on HMDB51 dataset.

	Base	Novel	HM
CLIP	53.3	46.8	49.8
A5	70.4	51.7	59.6
BMSPT	71.73	52.64	60.50

Table 6. Additional experiments in terms of object detection and semantic segmentation on VisA dataset.

	Detection		Segmentation	
	AUROC	AP	AUROC	AUPRO
CLIP	66.4	71.5	46.6	14.8
CoOp	62.8	68.1	24.2	3.8
WinCLIP	78.1	81.2	79.6	56.8
BMSPT	90.8	91.9	64.6	52.9

Diffusion Model Transfer. To further investigate the potential of our BMSPT, we design a simple yet highly effective approach for the image manipulation task by combining BMSPT with Stable Diffusion. Specifically, instead of finetuning the diffusion model, we construct an auxiliary alignment loss between noisy images and learnable prompts to preserve some personalized information (e.g., the identity of the subject) based on prompt tuning, the detailed architecture is illustrated in Figure 4. Meanwhile, as shown in Figure 5, our method enables better modification for background, texture, and motion while preserving the underlying structure from visualized perspective. For example, when compared with other methods, our BMSPT demonstrates the similar texture of and background for two bananas to source image. Moreover, we conducted additional quantitative experiments to strengthen its progressiveness, comparing our work against the state-of-the-art using three metrics: CLIP score,

GPT-4-based image quality assessment, and manual image quality assessment. The results demonstrate that BMSPT significantly outperforms the other methods across all metrics. To validate the visual quality and fidelity quantified in Section 3.3, we formalize our assessment protocols. For the GPT-4-based image quality assessment, we input randomized pairs of manipulated images into the GPT-4 Vision API alongside the specific system prompt: “Analyze the source image and evaluate which generated image best preserves the core subject’s identity while successfully manipulating the background/motion according to the target text.” The model outputs a definitive win/loss choice across an evaluation pool of 100 sample images. For the manual image quality assessment, three independent annotators with expertise in computer vision conduct a double-blind forced-choice preference test over the identical image pool. Final preference rates are aggregated utilizing a standard majority-voting scheme to minimize subjective bias. We aspire that this novel exploration can offer a fresh perspective toward cutting-edge challenges in text-to-image generation and multi-modal learning.

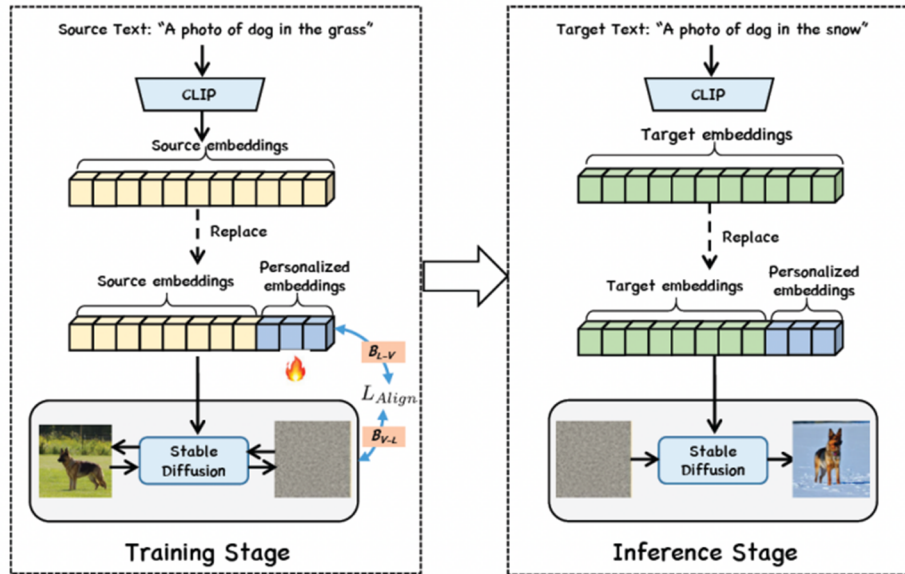


Figure 4. The overall framework of diffusion model with our BMSPT.



Figure 5. Visualization analysis in terms of image editing.

4. Related Work

Multimodal fusion and confidence aggregation. In addition to representation-level alignment, multimodal learning has also been studied from the perspectives of decision-level fusion, adaptive modality weighting, and uncertainty-aware confidence aggregation [48–52]. For example, two-level fusion combines modality-specific predictions at the decision stage [48], while discounted belief fusion explicitly models uncertainty and conflicts among modalities [49]. More recently, fuzzy-logic-driven confidence aggregation has been introduced to combine textual and visual evidence while accounting for annotator uncertainty and potentially contradictory modality-specific predictions [50]. Dynamic attention and hierarchical fusion methods further adjust modality contributions or progressively integrate multimodal representations [51,52]. In contrast, BMSPT focuses on representation-level alignment by learning bidirectional mappings between visual and textual prompt spaces. These directions are complementary but operate at different stages of multimodal learning.

5. Discussion

The experimental results indicate that bidirectional alignment can improve the adaptation of pretrained vision-language models in the evaluated classification and transfer settings. In contrast to unidirectional prompt-transfer methods, BMSPT constrains both the Vision-to-Language and Language-to-Vision transformations. The Wasserstein objectives encourage distribution-level correspondence, while the cycle-based losses reduce the risk that cross-modal transformations discard sample-specific semantic information.

The spatial visual prompt is designed to preserve the organization of the visual input rather than representing the prompt only as a small set of independent transformer tokens. The qualitative attention analysis suggests that this design can encourage the model to focus on regions that are more closely associated with the target category. However, the attention visualization should be interpreted as supporting evidence rather than a direct explanation of the model’s decision process.

The exploratory downstream results reveal that the effectiveness of BMSPT depends on the task. The method performs favourably in image-level anomaly detection, but remains below the specialized WinCLIP method in pixel-level anomaly segmentation. Similarly, the video-understanding and diffusion-editing results demonstrate possible extensions of the framework but do not constitute exhaustive task-specific evaluations.

Several limitations remain. First, the use of two transfer functions and two critics introduces additional training complexity relative to single-direction prompt tuning. Second, the method requires several alignment hyperparameters, including α , ϵ , and the loss weights. Third, performance may degrade under severe low-light conditions or heavy occlusion because corrupted visual representations can propagate through the bidirectional cycle. Finally, the current exploratory evaluations do not include sufficient task-specific baselines or ablations to establish general superiority in video understanding, anomaly segmentation, or image generation.

6. Conclusions

In this paper, we introduce the Bidirectional Multi-modal Spatial-consistent Prompt Tuning (BMSPT) framework, a novel strategy designed to significantly enhance modality alignment in vision-language models. This framework employs a dual approach to ensure both semantic coherence and information integrity across modalities within their original spatial dimensions. Our method notably improves model generalization across novel categories, facilitates cross-dataset transfer, and effectively addresses challenges in datasets with domain shifts. Furthermore, by integrating BMSPT with the diffusion model, we demonstrate the framework’s potential to offer innovative solutions to cutting-edge challenges in text-to- image generation and multi-modal learning, paving the way for future advancements in the field.

Author Contributions

S.B.: methodology design, coding, paper writing; H.L.: conceptualization, methodology design, software, strategy investigation, data curation; J.Z.: supervision, conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

The research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The benchmark datasets used in this study are publicly available: ImageNet (<https://image-net.org/index.php>, accessed on 9 June 2026); Caltech101 (http://www.vision.caltech.edu/Image_Datasets/Caltech101/101_ObjectCategories.tar.gz accessed on 9 June 2026); OxfordPets (<https://www.robots.ox.ac.uk/~vgg/data/pets/data/>, accessed on 9 June 2026); StanfordCars (<http://ai.stanford.edu/~jkrause/car196/>); Flowers102 (<https://www.robots.ox.ac.uk/~vgg/data/flowers/102/>); Food101 (https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/, accessed on 9 June 2026); FGVC Aircraft (<https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/archives/fgvc-aircraft-2013b.tar.gz>, accessed on 9 June 2026); SUN397 (<http://vision.princeton.edu/projects/2010/SUN/>, accessed on 9 June 2026); DTD (<https://www.robots.ox.ac.uk/~vgg/data/dtd/download/dtd-r1.0.1.tar.gz>, accessed on 9 June 2026); EuroSAT (<http://madm.dfki.de/files/sentinel/EuroSAT.zip>, accessed on 9 June 2026); UCF101 (<https://drive.google.com/file/d/10Jqome3vtUA2keJkNanAiFpgbyC9Hc2O/view?usp=sharing>, accessed on 9 June 2026); ImageNetV2 (https://s3-us-west-2.amazonaws.com/imagenetv2_public/, accessed on 9 June 2026); ImageNet-Sketch (<https://github.com/HaohanWang/ImageNet-Sketch>, accessed on 9 June 2026); ImageNet-A (<https://github.com/hendrycks/natural-adv-examples>, accessed on 9 June 2026); ImageNet-R (<https://github.com/hendrycks/imagenet-r>, accessed on 9 June 2026).

Acknowledgments

The authors thank colleagues and collaborators who provided feedback on earlier drafts of this manuscript.

Conflicts of Interest

Given the role as the Editorial Board Member, Jie Zhang had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used OpenAI Codex to assist with language editing and proof correction. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Radford, A.; Kim, J.W.; Hallacy, C.; et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2021; pp. 8748–8763.
2. Liu, P.; Yuan, W.; Fu, J.; et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35.
3. Wang, Z.; Zhang, Z.; Lee, C.-Y.; et al. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 139–149.
4. Brown, T.; Mann, B.; Ryder, N.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
5. Jiang, Z.; Xu, F.F.; Araki, J.; et al. How can we know what language models know? *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 423–438.
6. Bai, S.; Li, H.; Zhang, J.; et al. Diep: Adaptive mixture-of-experts compression through differentiable expert pruning. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 56090–56115.
7. Bai, S.; Li, H.; Zhang, J.; et al. Ttvs: Boosting self-exploring reinforcement learning via test-time variational synthesis. In *Findings of the Association for Computational Linguistics*; ACL: San Diego, CA, USA, 2026; pp. 32651–32665.
8. Xiong, S.; Zhao, Y.; Zhang, J.; et al. Dual prompt tuning based contrastive learning for hierarchical text classification. In *Findings of the Association for Computational Linguistics*; ACL: Bangkok, Thailand, 2024; pp. 12146–12158.

9. Zhou, K.; Yang, J.; Loy, C.C.; et al. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348.
10. Zhou, K.; Yang, J.; Loy, C.C.; et al. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16816–16825.
11. Long, Y.; Han, J.; Huang, R.; et al. Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 16277–16287.
12. Zhao, C.; Wang, Y.; Jiang, X.; et al. Learning domain invariant prompt for vision-language models. *IEEE Trans. Image Process.* **2024**, *33*, 1348–1360.
13. Jia, M.; Tang, L.; Chen, B.-C.; et al. Visual prompt tuning. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 709–727.
14. Bai, S.; Li, S.; Zhuang, W.; et al. Combating data imbalances in federated semi-supervised learning with dual regulators. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 10989–10997.
15. Yang, J.; Li, Z.; Zheng, F.; et al. Prompting for multi-modal tracking. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3492–3500.
16. Tsimpoukelli, M.; Menick, J.L.; Cabi, S.; et al. Multimodal few-shot learning with frozen language models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 200–212.
17. Yao, H.; Zhang, R. and Xu, C., Visual-language prompt tuning with knowledge-guided context optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6757–6767.
18. Wang, D.; Li, M.; Liu, X.; et al. Tuning multi-mode token-level prompt alignment across modalities. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 52792–52810.
19. Yao, H.; Zhang, R.; Xu, C. Tc: Textual-based class-aware prompt tuning for visual-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 23438–23448.
20. Huang, M.; Yang, C.; Yu, X. Cutcp: Custom text generation-based class-aware prompt tuning for visual-language models. *Sci. Rep.* **2025**, *15*, 2681.
21. Khattak, M.U.; Rasheed, H.; Maaz, M.; et al. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–19 June 2023.
22. Khattak, M.U.; Wasim, S.T.; Naseer, M.; et al. Self-regulating prompts: Foundational model adaptation without forgetting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 15190–15200.
23. Cai, R.; Yang, G.; Averbuch-Elor, H.; et al. Learning gradient fields for shape generation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 364–381.
24. Gao, Y.; Shi, X.; Zhu, Y.; et al. Visual prompt tuning for test-time domain adaptation. *arXiv* **2022**, arXiv:2210.04831.
25. Sohn, K.; Chang, H.; Lezama, J.; et al. Visual prompt tuning for generative transfer learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19840–19851.
26. Cao, G.; Shi, K.; Fu, H.; et al. Aple: Token-wise adaptive for multi-modal prompt learning. *arXiv* **2024**, arXiv:2401.06827.
27. Bai, S.; Zhang, J.; Guo, S.; et al. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27274–27283.
28. Xin, Y.; Du, J.; Wang, Q.; et al. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 16076–16084.
29. Deng, J.; Dong, W.; Socher, R.; et al. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
30. Fei-Fei, L.; Fergus, R.; Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In Proceedings of the Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004.
31. Parkhi, O.M.; Vedaldi, A.; Zisserman, A.; et al. Cats and dogs. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3498–3505.
32. Krause, J.; Stark, M.; Deng, J.; et al. 3d object representations for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 554–561.

33. Nilsback, M.-E.; Zisserman, A. Automated flower classification over a large number of classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
34. Bossard, L.; Guillaumin, M.; Van Gool, L. Food-101—Mining discriminative components with random forests. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VI 13. Springer: Berlin/Heidelberg, Germany, 2014; pp. 446–461.
35. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv* **2013**, arXiv:1306.5151.
36. Helber, P.; Bischke, B.; Dengel, A.; et al. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226.
37. Soomro, K.; Zamir, A.R.; Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
38. Cimpoi, M.; Maji, S.; Kokkinos, I.; et al. Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3606–3613.
39. Xiao, J.; Hays, J.; Ehinger, K.A.; et al. Sun database: Large-scale scene recognition from abbey to zoo. In Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
40. Recht, B.; Roelofs, R.; Schmidt, L.; et al. Do imagenet classifiers generalize to imagenet? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 5389–5400.
41. Wang, H.; Ge, S.; Lipton, Z.; et al. Learning robust global representations by penalizing local predictive power. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 10506–10518.
42. Hendrycks, D.; Zhao, K.; Basart, S.; et al. Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15262–15271.
43. Hendrycks, D.; Basart, S.; Mu, N.; et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 8340–8349.
44. Ju, C.; Han, T.; Zheng, K.; et al. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 105–124.
45. Kuehne, H.; Jhuang, H.; Garrote, E.; et al. Hmdb: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
46. Jeong, J.; Zou, Y.; Kim, T.; et al. Winclip: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616.
47. Zou, Y.; Jeong, J.; Pemula, L.; et al. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 392–408.
48. Sun, J.; Yin, H.; Tian, Y.; et al. Two-level multimodal fusion for sentiment analysis in public security. *Secur. Commun. Netw.* **2021**, *2021*, 6662337.
49. Bezirganyan, G.; Sellami, S.; Berti-Equille, L.; et al. Multimodal learning with uncertainty quantification based on discounted belief fusion. *arXiv* **2024**, arXiv:2412.18024.
50. Balderas-Díaz, S.; Guerrero-Contreras, G.; Bueno-Crespo, A.; et al. Fuzzy logic-driven confidence aggregation for multimodal sentiment classification. *Multimed. Tools Appl.* **2026**, *85*, 289.
51. Feng, X.; Lin, Y.; He, L.; et al. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics; EMNLP: Miami, FL, USA, 2024*; pp. 14755–14766.
52. Majumder, N.; Hazarika, D.; Gelbukh, A.; et al. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. -Based Syst.* **2018**, *161*, 124–133.