Dropout Prompt Learning: Towards Robust and Adaptive Vision-Language Models

Biao Chen, Lin Zuo*, Mengmeng Jing, Kunbin He, Yuchen Wang

School of Information and Software Engineering,
University of Electronic Science and Technology of China
chenbiao@std.uestc.edu.cn, linzuo@uestc.edu.cn, {jingmeng1992, hekunbin19, indramoe472}@gmail.com

Abstract

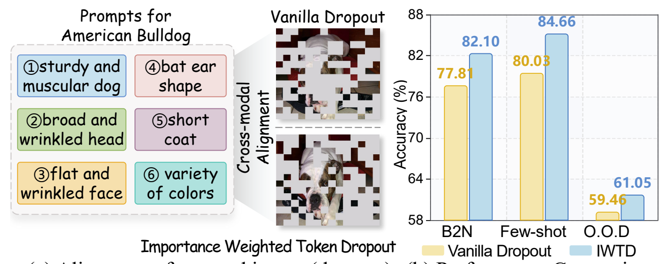
Dropout is a widely used regularization technique which improves the generalization ability of a model by randomly dropping neurons. In light of this, we propose Dropout Prompt Learning, which aims for applying dropout to improve the robustness of the vision-language models. Different from the vanilla dropout, we apply dropout on the tokens of the textual and visual branches, where we evaluate the token significance considering both intra-modal context and inter-modal alignment, enabling flexible dropout probabilities for each token. Moreover, to maintain semantic alignment for general knowledge transfer while encouraging the diverse representations that dropout introduces, we further propose residual entropy regularization. Experiments on 15 benchmarks show our method's effectiveness in challenging scenarios like low-shot learning, long-tail classification, and out-of-distribution generalization. Notably, our method surpasses regularization-based methods including KgCoOp by 5.10% and PromptSRC by 2.13% in performance on base-to-novel generalization.

Introduction

Vision-Language Models (VLMs) such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) have achieved remarkable advantages in zero-shot scenarios. While prompt learning (Zhou et al. 2022b; Khattak et al. 2023a) offers a parameter-efficient approach for adapting pre-trained VLMs to downstream tasks, its generalization capability remains limited by overfitting issues, particularly in low-data scenarios (Park, Ko, and Kim 2024; Khattak et al. 2023b).

In the past decade, dropout is applied as an effective regularization technique in deep neural networks, significantly mitigating overfitting and improving generalization by randomly dropping neurons during training (Srivastava et al. 2014). Dropout prevents complex co-adaptations among feature detectors and implicitly averages over an exponential number of thinned network architectures, a critical factor in the success of models such as AlexNet (Krizhevsky, Sutskever, and Hinton 2012). While dropout has shown remarkable success across various deep learning architectures, its potential in prompt learning for VLMs remains unexplored. Motivated by the effectiveness of dropout in learning

*Corresponding author.



(a) Alignment of text and image (dropout) (b) Performance Comparison

Figure 1: (a) Vanilla dropout randomly removes visual tokens, disrupting image-text alignment (top). Importance Weighted Token Dropout (IWTD) preserves semantically relevant tokens for alignment (bottom). (b) Comparison on base-to-novel, out-of-distribution and few-shot image classification.

robust models, we propose to incorporate dropout mechanisms into VLM prompt learning to enhance model generalization, particularly in low-data regimes (Zhu et al. 2023a; Yu et al. 2025; Chen et al. 2025).

However, VLM prompt learning presents distinct challenges compared to traditional deep learning, raising three critical questions regarding dropout implementation: (1) *Where to drop*: VLMs rely on tokens as fundamental semantic units to facilitate fine-grained semantic alignment across modalities (e.g., the contrastive learning mechanism of CLIP (Radford et al. 2021)), while vanilla dropout would destroy this alignment. As shown in Fig. 1(a), randomly dropping critical visual tokens impairs their matching with textual descriptions, leading to degraded performance (Fig. 1(b)). While existing works in unimodal tasks (Ke et al. 2020; Zhai and Wang 2018) improve the vanilla dropout through adaptive probabilities, these approaches are unsuitable for VLMs requiring cross-modal dependencies. (2) *What degree to drop*: Unlike traditional neural networks where high parameter redundancy enables effective dropout without performance loss, VLMs process semantically dense tokens with limited token-level redundancy. The inherent token sparsity means that high dropout ratios on semantically rich tokens could severely degrade performance, while low ratios on less informative tokens provide insufficient regularization. This creates a challenge in determining optimal dropout scheduling that balances feature preservation with regularization. (3) *How to* *learn from dropout*: To prevent semantic drift between learnable and frozen branches, existing approaches (Yao, Zhang, and Xu 2023; Khattak et al. 2023b) often enforce strict $\mathcal{L}_1$ or $\mathcal{L}_2$ regularization. This operation, however, is overly strict for dropout prompt learning. Such strict constraints limit the benefits from dropout-induced variations, suggesting the need for a mechanism that balances consistency and diversity.

Towards more robust and general prompt learning, we propose **Dropout Prompt Learning**, a principled framework that incorporates dropout mechanisms into vision-language prompt learning by regularizing through token dropout. Based on the framework of dropout prompt learning, we present the *Importance Weighted Token Dropout*, termed as IWTD, which formulates dropout as a token importance estimation problem in the multimodal space.

Importance weighted token dropout is carefully designed to handle the three challenges. For the first challenge of *where to drop*, we leverage a comprehensive importance metric to jointly model intra-modal context, inter-modal alignment, and task-specific relevance through a unified attention mechanism. This enables the identification of semantically critical tokens that maintain cross-modal alignment. For the second challenge of *what degree to drop*, different samples exhibit varying semantic densities in their tokens. Tokens carrying minimal semantic information can tolerate higher dropout rates for enhancing generalization, while samples with high semantic density require lower dropout rates to preserve crucial tokens for cross-modal alignment. This motivates flexible dropout probability assignment according to token significance. For the third challenge of *how to learn from dropout*, we propose residual entropy regularization, which computes residuals between pre- and post-dropout feature representations, and maximizes the predictive entropy on these residuals, simultaneously maintaining alignment with general knowledge transfer while encouraging representational diversity. Our main contributions are as follows:

- We propose Dropout Prompt Learning, a novel learning paradigm that extends dropout regularization to vision-language model adaptation. By introducing token-level dropout strategies, this framework enhances model generalization ability while maintaining cross-modal alignment.
- We present importance weighted token dropout, an effective implementation of dropout prompt learning. It dynamically adjusts dropout probabilities by jointly considering intra-modal context and cross-modal alignment. The residual entropy regularization is further adopted to maintain semantic alignment for general knowledge transfer while encouraging diverse feature representations.
- Extensive experiments on 15 benchmark datasets comprehensively validate the robustness and superior performance of the proposed method under various challenging settings, including low-shot learning, long-tail classification, and out-of-distribution generalization.

Related Work

Prompt Learning in VLMs. Prompt learning has evolved from NLP (Li and Liang 2021) to VLMs, with CoOp (Zhou et al. 2022b) introducing soft prompts for CLIP. Recent advances explore multimodal prompting (Khattak et al. 2023a; Guo et al. 2025). However, limited data often leads to overfitting (Khattak et al. 2023b; Park, Ko, and Kim 2024), inspiring various regularization methods: ProGrad (Zhu et al. 2023a) aligns prompt gradients with general knowledge. KgCoOp (Yao, Zhang, and Xu 2023) minimizes distance between learned and hand-crafted embeddings. PSRC (Khattak et al. 2023b) uses self-regularization through mutual agreement, and ProMetaR (Park, Ko, and Kim 2024) employs meta-learning with task augmentation. GalLoP (Lafon et al. 2024) applies dropout on multiple complete candidate texts to enhance diversity, yet this coarse-grained operation differs from dropout's principle of fine-grained dropping of individual units. Despite these approaches, the potential of dropout in the prompt learning remains underexplored. To address this limitation, we propose a token-level adaptive dropout framework for robust prompt learning in VLMs.

Dropout regularization. Vanilla Dropout (Srivastava et al. 2014) and its variants like DropConnect (Wan et al. 2013), DropBlock (Ghiasi, Lin, and Le 2018), and Curriculum Dropout (Morerio et al. 2017) apply fixed dropout probabilities during training. However, static dropout rates cannot adapt to varying feature importance across inputs and layers. This limitation motivates adaptive dropout methods that dynamically adjust probabilities. StandOut (Ba and Frey 2013) pioneered input-dependent rate learning via auxiliary networks, while subsequent work leveraged Rademacher complexity (Zhai and Wang 2018) and feature distributions (Ke et al. 2020). Recent works include attention mechanisms (Yang et al. 2022) and GFlowNet (Liu et al. 2023) which learns data-dependent dropout masks via posterior inference. However, adaptive dropout remains underexplored in multimodal settings such as VLMs. Thus, we propose a multimodal importance metric that jointly considers intra-modal context and inter-modal alignment for adaptive dropout.

Consistency regularization. Consistency regularization preserves model generalization by minimizing discrepancies between learned and reference features. Common approaches maintain consistency via feature alignment using $\mathcal{L}_1$ or $\mathcal{L}_2$ norms (Laine and Aila 2016; Tarvainen and Valpola 2017; Sajjadi, Javanmardi, and Tasdizen 2016) or cosine similarity (Hoe et al. 2021), or through distribution matching with KL divergence (Li et al. 2018). These methods have proven effective in enhancing model robustness (Wang et al. 2021). Consistency regularization also effectively mitigates overfitting and knowledge forgetting in VLM prompt learning. Methods like KgCoOp (Yao, Zhang, and Xu 2023), PSRC (Khattak et al. 2023b), and CoPrompt (Roy and Etemad 2024) work by constraining learnable prompts using references such as frozen CLIP features. While beneficial for generalization, enforcing strict consistency between multi-source features often limits model flexibility. Our residual entropy regularization relaxes this strict constraint, enabling the model to balance semantic alignment with the diverse representations introduced by adaptive dropout.

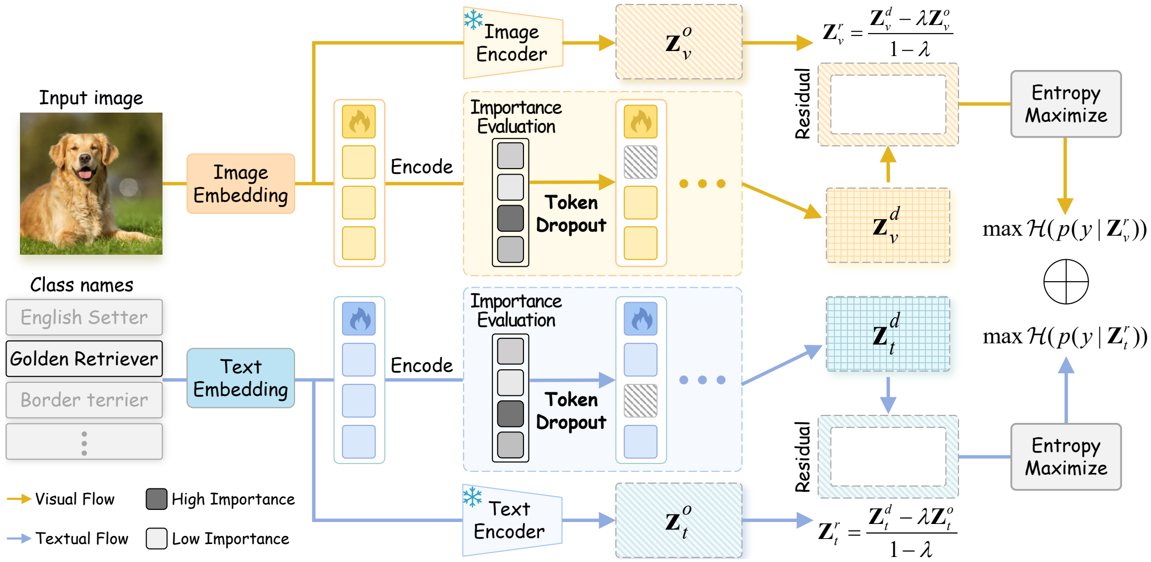


Figure 2: **Method overview of Importance Weighted Token Dropout.** Textual and visual modalities are processed by parallel encoding pathways, a frozen branch and a learnable branch. In the learnable branch, we compute an intra-/inter-modal importance metric for tokens at each layer, which guides adaptive token dropout. Then, residual features derive from learnable and frozen branch differences. Finally, maximizing entropy constrains dropout for both visual and textual residuals.

Methodology

Preliminaries

CLIP. Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) uses image ($\mathcal{F}$) and text ($\mathcal{G}$) encoders to produce aligned features ($\mathbf{z}_v = \mathcal{F}(x), \mathbf{z}_t = \mathcal{G}(t)$) via contrastive loss. For K-way zero-shot classification, image features $\mathbf{z}_v$ are compared against K class text features $\{\mathbf{z}_{t,k}\}_{k=1}^K$ using cosine similarity and a softmax function with temperature τ to compute probabilities $p(y = k|x)$:

$$p(y = k|x) = \frac{\exp(\text{sim}(\mathbf{z}_v, \mathbf{z}_{t,k})/\tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{z}_v, \mathbf{z}_{t,j})/\tau)}. \quad (1)$$

Prompt learning. Prompt learning optimizes VLMs by incorporating learnable prompts instead of full fine-tuning (Zhou et al. 2022b; Khattak et al. 2023a). The text and visual input sequences at layer i are defined as: $T_{input}^{(i)} = \{t_{bos}, P_t^{(i)}, T_{embed}, t_{eos}\}$ and $V_{input}^{(i)} = \{v_{cls}, E_{patch}, P_v^{(i)}\}$, where $P_t^{(i)} = \{p_t^1, p_t^2, ..., p_t^\eta\}$ and $P_v^{(i)} = \{p_v^1, p_v^2, ..., p_v^M\}$ are learnable prompts with dimensions $\mathbb{R}^\eta$ and $\mathbb{R}^M$.

Adaptive Dropout. Adaptive dropout dynamically adjusts dropout probabilities based on feature importance rather than fixed rates. For unit i with importance score I_i, the dropout probability is $p_i = f(I_i)$, where $f(\cdot)$ maps importance scores to dropout probabilities. The dropout operation then follows:

$$\boldsymbol{\mu}_i \sim \text{Bernoulli}(1 - p_i), \quad \boldsymbol{\phi}_i = (\mathbb{W}_i \boldsymbol{\theta}) \odot \boldsymbol{\mu}_i / (1 - p_i), \quad (2)$$

where $1/(1 - p_i)$ maintains expected output during training.

Dropout Prompt Learning

Dropout prompt learning extends dropout principles to VLMs by applying dropout at the token level to enhance robustness and generalization. Let $U^{(i)} \in \{V_{\text{input}}^{(i)}, T_{\text{input}}^{(i)}\}$ denote the visual or textual token sequence at layer i, where $V_{\text{input}}^{(i)} = \{v_{\text{cls}}, E_{\text{patch}}, P_v^{(i)}\}$ and $T_{\text{input}}^{(i)} = \{t_{\text{bos}}, P_t^{(i)}, T_{\text{embed}}, t_{\text{eos}}\}$. Here, $P_v^{(i)}$ and $P_t^{(i)}$ are learnable prompts. For training phase, Dropout prompt learning applies the dropout operation $\mathcal{D}_{\text{token}}$ through the following formulation:

$$U_{\text{dropped}}^{(i)} = f_{\text{CLIP}}(\mathcal{D}_{\text{token}}(\text{Enc}(U^{(i)}; \theta_e) \odot \mathcal{B}(p)_{1\times n}); \theta_{\text{CLIP}}), \quad (3)$$

where $\text{Enc}(\cdot; \theta_e)$ is the modality-specific encoder with parameters θ_e, $\mathcal{B}(p)_{1\times n}$ denotes an n-dimensional vector of independent Bernoulli random variables with dropout rate p, and $f_{\text{CLIP}}(\cdot; \theta_{\text{CLIP}})$ denotes the CLIP model with parameters θ_{CLIP}. During inference, following the standard dropout protocol (Srivastava et al. 2014), dropout is disabled and we directly apply the trained model:

$$U_{\text{infer}}^{(i)} = f_{\text{CLIP}}(\text{Enc}(U^{(i)}; \theta_e); \theta_{\text{CLIP}}). \quad (4)$$

Nevertheless, vanilla token dropout can disrupt cross-modal semantic alignment in VLMs, potentially degrading performance by randomly dropping critical visual or textual tokens from $U^{(i)}$. To address this, dropout prompt learning requires dropout strategies $\mathcal{D}_{\text{token}}$ designed explicitly for VLMs. We propose importance weighted token dropout (IWTD), which evaluates token significance from multiple perspectives to guide adaptive dropout. As shown in Fig. 2, IWTD measures token importance from various aspects and employs these to assign flexible dropout probabilities. Furthermore, it incorporates residual entropy regularization to constrain the adaptive token dropout process.

Importance Weighted Token Dropout (IWTD)

Based on the dropout prompt learning framework, we propose importance weighted token dropout. This assigns dropout probabilities via a multimodal importance metric instead of uniform randomness, preserving tokens critical for cross-modal alignment while enabling effective regularization.

Multimodal Importance Metric. The metric is termed as $I(\mathbf{x}_j^{(i)})$, which quantifies the significance of a token $\mathbf{x}_j^{(i)}$ at layer i from multiple sources:

$$I(\mathbf{x}_j^{(i)}) = f\left(S_{cls}^{(i)}(j), S_{self}^{(i)}(j), S_{cross}^{(i)}(j)\right), \quad (5)$$

where j indexes the tokens in the sequence of length L at layer i, and $f(\cdot)$ is an averaging function. S_{cls}, S_{self}, and S_{cross} denote the class attention, self-attention and cross-modal attention score respectively, each defined as follows.

First, to capture intra-modal relationships, we introduce the Self-Attention Score (S_{self}), which quantifies token interactions within its modality. Given the self-attention tensor $\mathbf{A}_{self}^{(i)} \in \mathbb{R}^{B\times H\times L\times L}$ at layer i, where B and H denote batch size and number of attention heads, we compute:

$$S_{self}^{(i)}(j) = \frac{1}{H}\sum_{h=1}^{H}\max_{k\neq j}\left((\mathbf{A}_{self}^{(i)})_{b,h,j,k}\right), \quad (6)$$

where token k excludes global tokens $[v_{cls}]$ and $[t_{eos}]$.

Second, to capture intra-modal task-specific importance of tokens, we leverage the attention patterns of the primary task-specific token (i.e., $[v_{cls}]$ for vision and $[t_{eos}]$ for text). The Class Attention Score (S_{cls}) is defined as:

$$S_{cls}^{(i)}(j) = \frac{1}{H}\sum_{h=1}^{H}\left((\mathbf{A}_{self}^{(i)})_{b,h,cls,j}\right). \quad (7)$$

Since task-specific tokens aggregate intra-modal information, their attention reflects each token's task relevance.

Third, in vision-language tasks where cross-modal semantic alignment is essential for understanding inter-modality interactions, we propose the Cross-modal Attention Score (S_{cross}) to quantify token importance via cross-modal alignment. Specifically, as shown in Fig. 3, given visual tokens $V_{input}^{(i)} \in \mathbb{R}^{N'\times D_v}$ and textual tokens $T_{input}^{(i)} \in \mathbb{R}^{M'\times D_t}$ at the i-th layer, we employ linear projections to map these tokens into a shared d-dimensional semantic space, yielding $\mathbf{V}'^{(i)} \in \mathbb{R}^{N'\times d}$ and $\mathbf{T}'^{(i)} \in \mathbb{R}^{M'\times d}$ respectively. To facilitate cross-modal interaction, we introduce ξ learnable bridge tokens $\mathbf{E} \in \mathbb{R}^{\xi\times d}$ as semantic anchors, reducing complexity from direct cross-modal attention $\mathcal{O}(N'\times M')$ to $\mathcal{O}(\xi\times(N'+M'))$. We then compute attention maps $\mathbf{A}_{cross}^{\mathcal{M}}$ between these bridge tokens and projected modality-specific features. For each modality $\mathcal{M}\in\{v,t\}$ with projected features $\mathbf{X}'^{(i)}$ (denoting either $\mathbf{V}'^{(i)}$ or $\mathbf{T}'^{(i)}$),

$$\mathbf{A}_{cross}^{\mathcal{M}} = \text{softmax}\left(\frac{\mathbf{E}(\mathbf{X}'^{(i)})^\top}{\sqrt{d}}\right), \quad \text{for } \mathcal{M}\in\{v,t\}. \quad (8)$$

Based on the computed attention maps, we define the cross-modal importance score $S_{cross}^{\mathcal{M}}(j)$ for each token j in modality $\mathcal{M}$ as the maximum attention weight received from any bridge token κ. This score quantifies each token's engagement with the cross-modal bridge tokens:

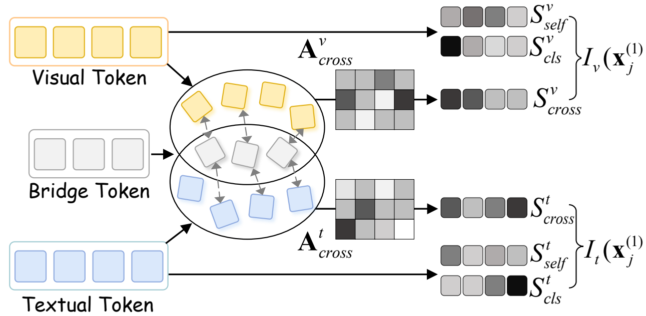


Figure 3: **Multimodal Importance Metric**, which simultaneously considers both intra-modal attention S_{self}, S_{cls} and inter-modal attention S_{cross}.

$$S_{cross}^{\mathcal{M}}(j) = \max_{\kappa\in\{1,\ldots,\xi\}}\left((\mathbf{A}_{cross}^{\mathcal{M}})_{\kappa,j}\right). \quad (9)$$

For each modality $\mathcal{M}$, we denote $S_{cross}^{\mathcal{M}}(j)$ as $S_{cross}^{(i)}(j)$ at layer i, which combines with other importance measures to form our final token importance metric $I(\mathbf{x}_j^{(i)})$. This unified metric guides the adaptive token dropout process.

Token Dropout. Let $I(\mathbf{x}_j^{(i)})$ be the importance score for token $\mathbf{x}_j^{(i)}$ at layer i, and $\mathcal{J}_{target}^{(i)}$ denote the indices of token excluding $[v_{cls}]$ and $[t_{eos}]$ which are essential for cross-modal similarity computation. Our objective is to adaptively adjust token dropout, where the dropout probability p_j of each token is inversely mapped from its normalized importance score:

$$p_j = p_{max} - \hat{I}(\mathbf{x}_j^{(i)})(p_{max}-p_{min}), \quad j\in\mathcal{J}_{target}^{(i)}, \quad (10)$$

where $\hat{I}(\mathbf{x}_j^{(i)})$ is the normalized importance score, and p_{max}, p_{min} specify the probability bounds. For each token representation, we apply the dropout operation $\mathcal{D}(\cdot)$ with its corresponding probability:

$$\mathbf{x}_{out,j}^{(i)} = \begin{cases}\mathcal{D}(\mathbf{x}_j^{(i)}) & \text{if } j\in\mathcal{J}_{target}^{(i)} \text{ and rand}() < p_j \\ \mathbf{x}_j^{(i)} & \text{otherwise}\end{cases}. \quad (11)$$

Unlike vanilla dropout that applies uniform randomness globally, our method restricts randomness to importance-determined probability ranges while preserving stochastic sampling within these ranges. This strategy preserves critical tokens for cross-modal alignment while providing effective regularization.

Residual Entropy Regularization. Prompt learning methods typically employ consistency regularization to align learnable prompts with pre-trained representations for general knowledge transfer (Yao, Zhang, and Xu 2023; Khattak et al. 2023b). Conventional $\mathcal{L}_1$ or $\mathcal{L}_2$ consistency regularization, while mitigating semantic drift between variable and fixed branch, is overly restrictive against beneficial dropout-induced variations. Therefore, we propose residual entropy regularization to permit such advantageous diversity. Specifically, we denote the output embedding features from the VLM branch processed by IWTD as $\mathbf{z}^d$, and the original VLM output embedding features as $\mathbf{z}^o$. We define the residual component $\mathbf{z}^r$ as the variation introduced by IWTD. To isolate this residual component, an intuitive method is to subtract $\mathbf{z}^d$ and $\mathbf{z}^o$. For more flexible control over the residual component, we consider the following linear relationship:

$$\mathbf{z}^d = \lambda \mathbf{z}^o + (1-\lambda)\mathbf{z}^r, \tag{12}$$

where $\mathbf{z}^r = \frac{\mathbf{z}^d - \lambda \mathbf{z}^o}{1-\lambda}$ is the residual component and $\lambda \in (0,1)$. We employ an annealing strategy (Jing et al. 2023): $\lambda = \lambda_0[1-(1+10t/T)^{3/4}]$, where t, T are current and total iterations. This gradually increases λ to balance original and residual components. Based on Occam's razor principle, linearity is a good inductive bias (Zhang et al. 2018; Jing et al. 2023), and Eq. (12) is an invertible operation that can easily infer $\mathbf{z}^r$ given $\mathbf{z}^d$ and $\mathbf{z}^o$. Besides, assuming IWTD primarily targets on non-critical information, the residual component $\mathbf{z}^r$ should ideally contain minimal class-discriminative features. To encourage this, we aim to maximize the uncertainty associated with the class prediction based on $\mathbf{z}^r$.

Taking the visual modality as an example, we compute the cosine similarity $\mathrm{sim}(\cdot)$, between the visual residual component $\mathbf{z}_v^r$ and the original textual class embeddings $\mathcal{W}_t^o = \{\mathbf{z}_{t,k}^o\}_{k=1}^K$. $\mathbf{z}_{t,k}^o$ are obtained by feeding text descriptions corresponding to each class k into the VLM's text encoder $\mathcal{G}$. The probability distribution over classes given the visual residual is:

$$p(y=k|\mathbf{z}_v^r) = \frac{\exp(\mathrm{sim}(\mathbf{z}_v^r, \mathbf{z}_{t,k}^o)/\tau)}{\sum_{j=1}^K \exp(\mathrm{sim}(\mathbf{z}_v^r, \mathbf{z}_{t,j}^o)/\tau)}. \tag{13}$$

Next, we maximize the conditional entropy $\mathcal{H}(p(y|\mathbf{z}_v^r))$ to enlarge the uncertainty of $\mathbf{z}_v^r$'s prediction. Therefore, our objective for the visual modality is as follows:

$$\mathcal{L}_{RE}^v = -\mathcal{H}(p(y|\mathbf{z}_v^r)) = \sum_{k=1}^K p(y=k|\mathbf{z}_v^r) \log p(y=k|\mathbf{z}_v^r). \tag{14}$$

By minimizing $\mathcal{L}_{RE}^v$, we regularize $\mathbf{z}_v^r$ to have an approximately equal probability of being associated with any category, thereby ensuring $\mathbf{z}_v^r$ does not contain class-discriminative information.

An analogous procedure is applied to the textual modality, yielding a residual entropy loss $\mathcal{L}_{RE}^t$. The total residual entropy regularization: $\mathcal{L}_{RE} = \mathcal{L}_{RE}^v + \mathcal{L}_{RE}^t$. This constrains IWTD by ensuring dropout-altered information does not carry significant class-specific signals, maintaining semantic alignment for knowledge transfer while allowing beneficial robustness variations. During inference, class probabilities are computed using the learned prompts without dropout:

$$p(y=k|x) = \frac{\exp(\mathrm{sim}(\mathcal{F}(V_{input}), \mathcal{G}(T_{input,k}))/\tau)}{\sum_{j=1}^K \exp(\mathrm{sim}(\mathcal{F}(V_{input}), \mathcal{G}(T_{input,j}))/\tau)}. \tag{15}$$

The learned representations from our dropout-based training directly contribute to robust inference.

Experiments

In this section, we conduct extensive experiments on widely-used benchmarks to evaluate our proposed method. We assess its performance in base-to-novel generalization, cross-dataset evaluation, few-shot classification, and out-of-distribution generalization, comparing it against competitive vision-language prompt learning baselines.

Datasets. Following the previous work (Zhou et al. 2022b), our experiments utilize a diverse array of 11 image classification datasets: UCF101 (Soomro, Zamir, and Shah 2012) (action recognition), DTD (Cimpoi et al. 2014) (texture analysis), SUN397 (Xiao et al. 2016) (scene recognition), EuroSAT (Helber et al. 2019) (satellite imagery), five fine-grained datasets (Flowers102 (Nilsback and Zisserman 2008), FGVCAircraft (Maji et al. 2013), Food101 (Bossard, Guillaumin, and Van Gool 2014), OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013)), and two generic object datasets (Caltech101 (Fei-Fei, Fergus, and Perona 2004), ImageNet (Deng et al. 2009)). For evaluating out-of-distribution generalization, ImageNet serves as the source dataset, while its variants (ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), ImageNet-Sketch (Wang et al. 2019), ImageNet-V2 (Recht et al. 2019)) are target datasets.

Implementation Details. Our implementation is based on CLIP-B/16 (Radford et al. 2021). The number of bridge tokens ξ is set to 64. The textual and visual prompt learning layers are set to 9 and 6, with corresponding adaptive dropout layers are 6. Learnable prompt length is 4. The textual token space is augmented with 16 learnable tokens for expanded dropout operation, with $T_{\text{input}}^{(i)} = \{t_{\text{bos}}, P_t^{(i)}, T_{\text{embed}}, t_{\text{eos}}\}$, where $P_t^{(i)}$ denotes supplementary tokens. The default dropout probability ranges from 10% to 50%. λ_0 is set to 0.1. Experiments are conducted on 2 NVIDIA 4090 GPUs.

Base-to-Novel Generalization

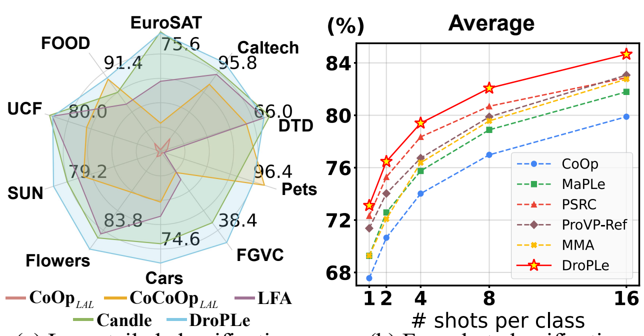
In the base-to-novel generalization, we evaluate our dropout prompt learning method, dubbed as **DroPLe**, against recent approaches specifically designed to enhance prompt learning generalization via regularization techniques (KgCoOp (Yao, Zhang, and Xu 2023), PSRC (Khattak et al. 2023b), DeKgTCP (Li et al. 2025)), the baseline CoOp (Zhou et al. 2022b), the LLM-enhanced method TAP (Ding et al. 2024) and TAC (Hao et al. 2025) which incorporates textual-visual consistency regularization. As shown in Table 1, with 16-shot training on base classes and zero-shot evaluation on novel classes, DroPLe achieves 86.12% base accuracy and 78.44% novel accuracy, with 82.10% HM significantly surpassing TAP (81.04%) and DeKgTCP (80.44%). Notably, DroPLe achieves the best overall HM across all 11 datasets and demonstrates superior performance on fine-grained recognition, achieving 43.75% HM on FGVCAircraft (+2.89% over TAC). Furthermore, to verify DroPLe's generalization in long-tailed scenarios, we follow Candles (Shi et al. 2024), replacing loss functions in CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) with LA Loss (Ren et al. 2020). As shown in Fig. 4 (a), under an imbalance ratio of 10, DroPLe outperforms LFA (Ouali et al. 2023) and

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp$_{(\text{IJCV'22})}$	82.69	63.22	71.66	76.47	67.88	71.92	96.00	89.81	93.73	93.67	95.29	94.47
KgCoOp$_{(\text{CVPR'23})}$	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
PSRC$_{(\text{ICCV'23})}$	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
DeKgTCP$_{(\text{ICLR'25})}$	84.96	76.38	80.44	77.40	69.20	73.07	98.64	95.20	96.89	94.47	97.76	96.09
TAP$_{(\text{ICLR'25})}$	84.75	77.63	81.04	77.97	70.40	73.99	**98.90**	95.50	97.17	95.80	97.73	96.76
TAC$_{(\text{CVPR'25})}$	85.24	77.60	81.24	**78.57**	71.03	74.61	98.57	95.27	96.89	95.93	**98.17**	97.04
DroPLe$_{(\text{Ours})}$	**86.12**	**78.44**	**82.10**	78.24	**71.38**	**74.65**	98.72	**96.06**	**97.37**	**96.38**	98.13	**97.25**

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp$_{(\text{IJCV'22})}$	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
KgCoOp$_{(\text{CVPR'23})}$	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
PSRC$_{(\text{ICCV'23})}$	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
DeKgTCP$_{(\text{ICLR'25})}$	81.18	74.75	77.83	98.58	75.18	85.30	90.73	91.55	91.14	45.20	35.09	39.51
TAP$_{(\text{ICLR'25})}$	80.70	74.27	77.35	97.90	75.57	85.30	90.97	91.83	91.40	44.40	36.50	40.06
TAC$_{(\text{CVPR'25})}$	81.63	74.17	77.72	97.97	76.87	86.15	90.87	91.87	91.37	44.60	37.70	40.86
DroPLe$_{(\text{Ours})}$	**82.87**	**75.04**	**78.76**	**98.61**	**78.29**	**87.28**	**91.18**	**92.20**	**91.69**	**49.26**	**39.35**	**43.75**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp$_{(\text{IJCV'22})}$	80.60	65.89	72.51	79.44	41.18	54.24	93.19	54.74	68.69	84.69	56.05	67.46
KgCoOp$_{(\text{CVPR'23})}$	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
PSRC$_{(\text{ICCV'23})}$	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
DeKgTCP$_{(\text{ICLR'25})}$	82.52	78.30	80.35	83.80	59.66	69.70	94.02	81.69	87.42	88.06	81.77	84.80
TAP$_{(\text{ICLR'25})}$	82.87	79.53	81.17	84.20	**68.00**	75.24	90.70	82.17	86.22	87.90	82.43	85.08
TAC$_{(\text{CVPR'25})}$	83.70	80.03	81.82	83.37	64.27	72.58	94.37	82.60	88.10	88.07	81.67	84.75
DroPLe$_{(\text{Ours})}$	**83.82**	**80.07**	**81.90**	**85.43**	67.32	**75.30**	**94.73**	**82.70**	**88.31**	**88.16**	**82.73**	**85.39**

Table 1: **Base-to-novel generalization.** Comparison with CoOp and the methods mainly focusing on regularization techniques to improve generalization across 11 image recognition datasets. Bold values indicate the best results. HM: Harmonic Mean.



(a) Long-tailed classification (b) Few-shot classification

Figure 4: (a) Base-to-novel generalization with **imbalance ratio 10**. (b) **Few-shot classification** on 11 datasets.

other long-tail classification methods in terms of harmonic mean accuracy. On EuroSAT, DroPLe achieves +4.6% gain over the 75.6% baseline, outperforming GLA's +3.8% (Zhu et al. 2023b). Notably, DroPLe is orthogonal to such post-hoc methods, with GLA+DroPLe yielding +5.4% improvement. This further demonstrates its effectiveness and robustness in handling imbalanced data distributions.

Few-shot Classification

Fig. 4(b) presents the comprehensive few-shot classification results averaged across all 11 datasets. DroPLe demonstrates strong few-shot learning capabilities, outperforming baseline methods across different shot settings. The method's superior performance stems from our importance weighted token dropout, which provides effective regularization in data-limited scenarios while maintaining crucial semantic information. These results validate that our adaptive dropout strategy successfully introduces beneficial diversity for improved generalization in few-shot learning.

Method	Source	Target				
	ImgNet	-V2	-S	-A	-R	*OOD*
KgCoOp	71.20	64.10	48.97	50.69	76.70	60.11
PSRC	71.27	64.35	49.55	50.90	77.80	60.65
CoPrompt	70.80	64.25	49.43	50.50	77.51	60.42
ProMetaR	71.29	64.39	49.55	51.25	77.89	60.77
GalLoP*	71.14	64.32	49.56	50.83	77.42	60.53
SPTR	70.05	64.40	48.78	51.30	**77.90**	60.59
DroPLe	**71.94**	**64.90**	**50.36**	**51.33**	77.59	**61.05**

Table 2: **Out-of-distribution generalization.** '*' means reproduced results. Best results highlighted in **first** , second .

Out-of-distribution Generalization

We further evaluate DroPLe's robustness on out-of-distribution (OOD) generalization. As shown in Table 2, DroPLe attains the highest average OOD accuracy of 61.05%, surpassing SPTR (Cui et al. 2025) by 0.46%. Notably, com-

IWTD	$\mathcal{L}_{RE}$	Component	Method	Base	Novel	HM
✗	✗	*Dropout*	Baseline	83.82	75.50	79.44
			+Dropout$_{0.5}$	82.71	73.45	77.81
			+Dropout$_{0.3}$	83.44	74.53	78.73
			+Dropblock	82.63	73.75	77.94
✓	✗		+IWTD (Ours)	85.34	77.38	81.17
✓	✗	*Consistency Reg.*	$\mathcal{D}_{cos}$	85.53	77.60	81.37
			$\mathcal{D}_{KL}$+$\mathcal{L}_1$	85.80	77.53	81.46
✓	✓		$\mathcal{L}_{RE}$ (Ours)	86.12	78.44	82.10
✓	✓	*Importance Scores*	DroPLe$^{\blacktriangle}_{S_{self}}$	84.78	76.92	80.66
			DroPLe$^{\blacklozenge}_{S_{self,cls}}$	85.24	77.41	81.14
			DroPLe$^{\bigstar}_{S_{self,cls,cross}}$	**86.12**	**78.44**	**82.10**

Table 3: **Ablation studies of components** on base-to-novel task. $\mathcal{D}_{cos}$ is cosine similarity. $\mathcal{D}_{KL}$ is KL divergence. DroPLe$^{\blacktriangle}$ only uses the score S_{self}, DroPLe$^{\blacklozenge}$ adds S_{cls} on $^{\blacktriangle}$, and DroPLe$^{\bigstar}$ further adds S_{cross} on $^{\bigstar}$.

pared with GalLoP (Lafon et al. 2024) that applies text-level dropout, our token-level adaptive dropout strategy achieves better performance on ImageNet-S (+0.80%) and ImageNet-V2 (+0.58%). This indicates that token-level dropout enables more precise regularization than text-level dropout, leading to better generalization across distribution shifts.

Ablation Study

Component Ablation. Table 3 shows the ablation analysis on base-to-novel task. The baseline implements a similar architecture to Fig. 2 but removes IWTD and $\mathcal{L}_{RE}$, using $\mathcal{L}_2$ norm to constrain dual-branch embeddings as in (Yao, Zhang, and Xu 2023). While vanilla dropout (randomly dropping individual tokens, dropout probability is 50% or 30%) and dropblock (dropping consecutive 3 tokens) show limited gains, IWTD achieves substantial improvement (81.17% HM) by preserving semantically meaningful tokens. Upon IWTD, we examine consistency regularization approaches. Although CoPrompt's $\mathcal{D}_{cos}$ and PSRC's $\mathcal{D}_{KL}$+$\mathcal{L}_1$ improve performance, our proposed $\mathcal{L}_{RE}$ further boosts HM to 82.10% by encouraging representation diversity. Finally, ablating scores in Eq. (5) shows the cross-modal score S_{cross} contributes the largest gain, validating its importance for VLM alignment.

Configuration Analysis. Results of hyperparameter and baseline analysis are shown in Table 4. Vanilla dropout typically employs a 50% dropout probability, as higher rates risk eliminating critical features. Our method performs targeted dropout based on token importance, maintaining robustness even with higher probability thresholds, 70%. Empirically, the 10-50% range yields optimal results and serves as our default configuration. The number of shared tokens ξ is set to 64, balancing cross-modal semantic representation with computational cost. For the initial value λ_0 of λ in Eq. (12), we set $\lambda_0 = 0.1$ to emphasize IWTD residual features during early training and enhance model robustness. As demonstrated in Table 4(d), DroPLe consistently improves generalization on challenging OOD tasks across various prompt learning baselines, confirming our method's effectiveness.

Visualization. Grad-CAM visualizations reveal that our importance weighted token dropout produces concentrated attention on semantically relevant regions, while vanilla dropout exhibits scattered patterns (Fig. 5(a)), showing our method effectively maintains cross-modal semantic alignment.

(a) Range of **dropout probability**.

Range (%)	Base	Novel	HM
0-40	86.08	78.21	81.96
5-45	86.04	78.35	82.02
10-50	86.12	78.44	82.10
20-70	85.85	78.13	81.81

(b) Number of **shared tokens** ξ.

ξ	Base	Novel	HM
16	85.74	78.39	81.90
32	85.97	78.36	81.99
64	86.12	78.44	82.10
128	86.20	78.41	82.12

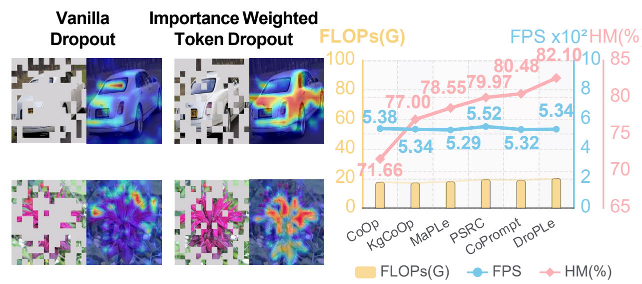
(c) Initial **balance parameter** λ_0.

λ_0	Base	Novel	HM
0.05	86.02	78.48	82.08
0.10	86.12	78.44	82.10
0.20	85.97	78.28	81.94
0.40	85.93	78.26	81.92

(d) OOD generalization of **other approaches**.

Method	ImgNet	-V2	-S	-A	-R
CoOp	71.51	64.20	47.99	49.71	75.21
+DroPLe	71.32	64.23	49.06	50.93	77.04
MaPLe	70.72	64.07	49.15	50.90	76.98
+DroPLe	71.26	64.38	49.56	51.28	77.84

Table 4: **Ablation analysis of different settings.** The default configuration is colored blue .



(a) CAM visualization (b) Computational cost

Figure 5: (a) **Grad-CAM visualizations for different dropout methods.** Redder colors indicate higher feature attention. (b) **Computational cost and performance** of different prompt learning methods.

Computational Efficiency. Our method maintains comparable FLOPs and inference speed while achieving superior performance (HM: 82.10%) over CoPrompt (80.48%) and PSRC (79.97%), demonstrating efficient enhancement without computational overhead, as shown in Fig. 5(b).

Conclusion

This paper introduces Dropout Prompt Learning, a novel paradigm that enhances vision-language model generalization through token-level dropout. We propose importance weighted token dropout, which dynamically assesses token significance by considering intra-modal context, inter-modal alignment, and task-specific relevance. Additionally, we introduce residual entropy regularization to maintain general knowledge and promote representational diversity. Extensive experiments across diverse benchmarks demonstrate our method's effectiveness in low-shot learning, long-tail classification, and out-of-distribution generalization, showing promise for robust VLM adaptation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62276054 and 62406060; in part by the Sichuan Science and Technology Program under Grant No. 2025ZNSFSC1500; and in part by the Xiaomi Young Scholar Program.

References

Ba, J.; and Frey, B. 2013. Adaptive dropout for training deep neural networks. *Advances in neural information processing systems*, 26.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Chen, B.; He, K.; Zheng, Z.; Jing, M.; and Zuo, L. 2025. Chain-of-Thought Guided Semantic Debiasing for Low-Shot Vision-Language Tasks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 4600–4609.

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

Cui, F.; Fong, J.; Zeng, R.; Tian, X.; and Yu, J. 2025. A Similarity Paradigm Through Textual Regularization Without Forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16100–16108.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, T.; Li, W.; Miao, Z.; and Pfister, H. 2024. Tree of Attributes Prompt Learning for Vision-Language Models. *arXiv preprint arXiv:2410.11201*.

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

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2018. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31.

Guo, F.; Chen, B.; Chen, Z.; Zheng, Z.; Zuo, L.; and Li, W. 2025. DviT: Debiased variational inference for multi-modal mutual prompt tuning. *Knowledge-Based Systems*, 113798.

Hao, F.; He, F.; Wu, F.; Wang, T.; Song, C.; and Cheng, J. 2025. Task-Aware Clustering for Prompting Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14745–14755.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Hoe, J. T.; Ng, K. W.; Zhang, T.; Chan, C. S.; Song, Y.-Z.; and Xiang, T. 2021. One loss for all: Deep hashing with a single cosine similarity based learning objective. *Advances in Neural Information Processing Systems*, 34: 24286–24298.

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

Jing, M.; Zhen, X.; Li, J.; and Snoek, C. G. 2023. Order-preserving consistency regularization for domain adaptation and generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18916–18927.

Ke, Z.; Wen, Z.; Xie, W.; Wang, Y.; and Shen, L. 2020. Group-wise dynamic dropout based on latent semantic variations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11229–11236.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023b. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15190–15200.

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

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lafon, M.; Ramzi, E.; Rambour, C.; Audebert, N.; and Thome, N. 2024. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, 264–282. Springer.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Li, J.; Luo, P.; Lin, F.; and Chen, B. 2018. Conversational model adaptation via KL divergence regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Association for Computational Linguistics*.

Li, Y.; Cheng, M.; Han, X.; and Song, W. 2025. Divergence-enhanced knowledge-guided context optimization for visual-language prompt tuning. In *The Thirteenth International Conference on Learning Representations*.

Liu, D.; Jain, M.; Dossou, B. F.; Shen, Q.; Lahlou, S.; Goyal, A.; Malkin, N.; Emezue, C. C.; Zhang, D.; Hassen, N.; et al. 2023. Gflowout: Dropout with generative flow networks. In *International Conference on Machine Learning*, 21715–21729. PMLR.

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

Morerio, P.; Cavazza, J.; Volpi, R.; Vidal, R.; and Murino, V. 2017. Curriculum dropout. In *Proceedings of the IEEE international conference on computer vision*, 3544–3552.

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

Ouali, Y.; Bulat, A.; Matinez, B.; and Tzimiropoulos, G. 2023. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15534–15546.

Park, J.; Ko, J.; and Kim, H. J. 2024. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26940–26950.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

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

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

Roy, S.; and Etemad, A. 2024. Consistency-guided Prompt Learning for Vision-Language Models. In *International Conference on Learning Representations*.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29.

Shi, J.-X.; Zhang, C.; Wei, T.; and Li, Y.-F. 2024. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2663–2673.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; and Fergus, R. 2013. Regularization of neural networks using dropconnect. In *International conference on machine learning*, 1058–1066. PMLR.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

Wang, Y.; Huang, G.; Song, S.; Pan, X.; Xia, Y.; and Wu, C. 2021. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3733–3748.

Xiao, J.; Ehinger, K. A.; Hays, J.; Torralba, A.; and Oliva, A. 2016. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119: 3–22.

Yang, T.; Deng, J.; Quan, X.; Wang, Q.; and Nie, S. 2022. Ad-drop: Attribution-driven dropout for robust language model fine-tuning. *Advances in Neural Information Processing Systems*, 35: 12310–12324.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6757–6767.

Yu, Y.; Guo, F.; Tian, X.; Chen, B.; Jing, M.; and Zuo, L. 2025. Visual residual aggregation network for visual-language prompt tuning. *Applied Intelligence*, 55(15): 988.

Zhai, K.; and Wang, H. 2018. Adaptive dropout with rademacher complexity regularization. In *International conference on learning representations*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023a. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15659–15669.

Zhu, B.; Tang, K.; Sun, Q.; and Zhang, H. 2023b. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. *Advances in Neural Information Processing Systems*, 36: 64663–64680.