An Effective Levelling Paradigm for Unlabeled Scenarios

Fangming Cui[1,2] **Zhou Yu**[3] **Di Yang**[4] **Yuqiang Ren**[4]
Liang Xiao[1*] **Xinmei Tian**[5*]
[1]Defense Innovation Institute
[2]Shanghai Jiao Tong University [3]The Key Laboratory of Complex Systems Modeling and Simulation, the School of Computer Science, Hangzhou Dianzi University, China
[4]ByteDance Inc. [5]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
cuifangming@sjtu.edu.cn yuz@hdu.edu.cn
xiaoliang@nudt.edu.cn xinmei@ustc.edu.cn

Abstract

Advancements in direct-integration fine-tuning frameworks have underscored their potential to enhance the performance of labeled scenarios and tasks. To enhance the generalization of different categories in the same dataset, some methods have added visual loss to these frameworks for unlabeled scenarios. However, the performance of these methods through visual loss does not improve significantly in domain generalization and cross-dataset generalization tasks. This may be attributed to the uncoordinated learning of the two-modalities alignment and visual loss. To mitigate this issue of uncoordinated learning, we propose a novel method called Levelling Paradigm (LePa) to improve performance for unlabeled tasks or scenarios. The proposed LePa, designed as a plug-in module, dynamically constrains and coordinates multiple objective functions, thereby improving the generalization of these baseline methods. Comprehensive experiments have shown that our design can effectively address generalized scenarios and tasks.

1 Introduction

Vision-Language Models (VLMs), such as the frozen CLIP [1], are garnering increasing attention for their exceptional generalization capabilities. These VLMs have been trained to align textual and visual components by utilizing large datasets [2]. As an example, CLIP undergoes training on a massive dataset containing 400 million text-image pairs, enabling VLMs to encode a wide range of concepts in a unified embedding space. By aligning and integrating textual and visual information, VLMs can bridge the divide between language and visual representations, thereby improving their comprehension of context within the shared embedding space [3, 4]. VLMs and large language models [5, 6, 7] can exhibit strong performance across a diverse range of downstream tasks [8, 9, 10, 11]. One remarkable aspect of CLIP is its exceptional zero-shot generalization capability. This ability is achieved by employing predefined text inputs, like "a photo of a [class]," known as prompts, to generate recognition weights during the inference stage. Through the use of prompts, CLIP can effectively transfer its learned knowledge to achieve precise recognition, even for classes that have not been encountered before. The zero-shot capability enhances the adaptability and versatility of CLIP, allowing it to tackle various tasks [12, 13, 14] and domains without the need for extensive fine-tuning and adaptation [15].

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent advancements involve keeping the weights of CLIP fixed while learning a set of textual parameters [16] for prompts to fine-tune CLIP for downstream image recognition tasks. Researchers have achieved notable advancements through investigations into the synchronization of images and prompts, surpassing the boundaries of earlier research endeavors. These analyses have acknowledged the promise of learning prompts for individual images, outstripping the efficacy of hand-crafted prompts. A pivotal revelation has been the importance of preserving the class name as inherent knowledge to guarantee that the acquired prompts can adeptly construct a classifier. Moreover, the word embeddings of prompts, serving as contextual indicators, are regarded as adjustable parameters. Remarkably, the trainable terms within the prompts are initialized using the "a photo of a" [1].

Recent research (Vision-Language Prompting, VLP) has found that initializing and fine-tuning visual prompt [17] directly based on learnable textual prompt can better improve the performance of supervised learning [18] and base classes of the same datasets [19, 20, 21]. In order to improve novel classes (unlabeled samples) within the same dataset, PromptSRC [22] incorporates traditional constraints to avoid forgetting general knowledge for enhancing the base-to-novel generalization through hand-crafted prompts (gray 'T' of Figure 1) and visual loss. Recently, CoPrompt [23] employs two learning adapters to enhance the base-to-novel generalization through the visual loss. ProMetaR [24] meta-learns both the visual regularizer and the soft prompts to harness the task-specific knowledge from the downstream tasks and task-agnostic general knowledge. However, the performance of these direct-integration VLP through visual loss [22, 23] does not improve significantly in higher difficulty generalization tasks, shown in Table 3 and Table 1 for unlabeled scenarios. This may be attributed to the uncoordinated learning [25, 26, 27] of the CE loss and visual loss [22, 23]. The theoretical basis is that the purpose of CE loss is to fine-tune and enhance few-shot learning ability [16], while the purpose of visual loss is to avoid forgetting pre-trained features and enhance generalization ability [28, 23, 24, 22]. There is a compromise and tradeoff between these two losses, a natural conflict where one goes up and the other goes down. To alleviate the uncoordinated learning, we design a flexible way to shape a dynamic constraint objective, improving the domain generalization and cross-dataset generalization. Our contributions can be summarized as follows:

- We propose a plug-and-play design called Levelling Paradigm (LePa) that is compatible with representative VLP baselines.
- The LePa dynamically coordinates multiple functions, thereby improving the generalization of direct-integration VLP with visual regularization.
- Representative tasks across 11 real datasets on generalization from base-to-novel, cross-dataset generalization, and domain generalization demonstrate that our design can effectively address generalized scenarios and tasks.

2 Methods

2.1 Frozen CLIP

Our method is founded on frozen CLIP [1] (as shown in Figure 1), a pioneering pre-trained vision-language model backbone known for its effectiveness in zero-shot learning applications. The pre-trained CLIP comprises two encoders: text encoder $\mathcal{G}_t(\cdot)$ and image encoder $\mathcal{G}_v(\cdot)$, which separately map textual input embeddings $\mathbf{p}$ and a visual training images, i.e., an image, $\mathbf{x}$ into a feature space through ViT blocks. The output features of two encoders of pre-trained CLIP are denoted as $\mathcal{G}_t(\mathbf{p})$ and $\mathcal{G}_v(\mathbf{x})$. Within the CLIP framework, the image encoder plays a crucial role in converting raw input images into feature embeddings, capturing intricate visual nuances, and deriving meaningful representations. Meanwhile, the text encoder is meticulously designed to generate representations for sequences of word embeddings, empowering the model to understand and process textual information effectively. Throughout the pre-training phase of CLIP, both the image and text encoders undergo simultaneous training on expansive datasets comprising text-image pairs. This alignment aligns seamlessly with the overarching goal of zero-shot recognition, a concept that can be formally delineated as follows,

$$p(z \mid \mathbf{x}) = \frac{\exp\left(\text{sim}\left(\mathcal{G}_t\left(\mathbf{p}_z\right), \mathcal{G}_v(\mathbf{x})\right)/\tau\right)}{\sum_{\mathbf{p}_i \in \mathcal{P}} \exp\left(\text{sim}\left(\mathcal{G}_t\left(\mathbf{p}_i\right), \mathcal{G}_v(\mathbf{x})\right)/\tau\right)}, \tag{1}$$

where z is the label of training image $\mathbf{x}$, $\mathbf{p}_i$ denotes the pre-defined prompts, $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, and τ is temperature.

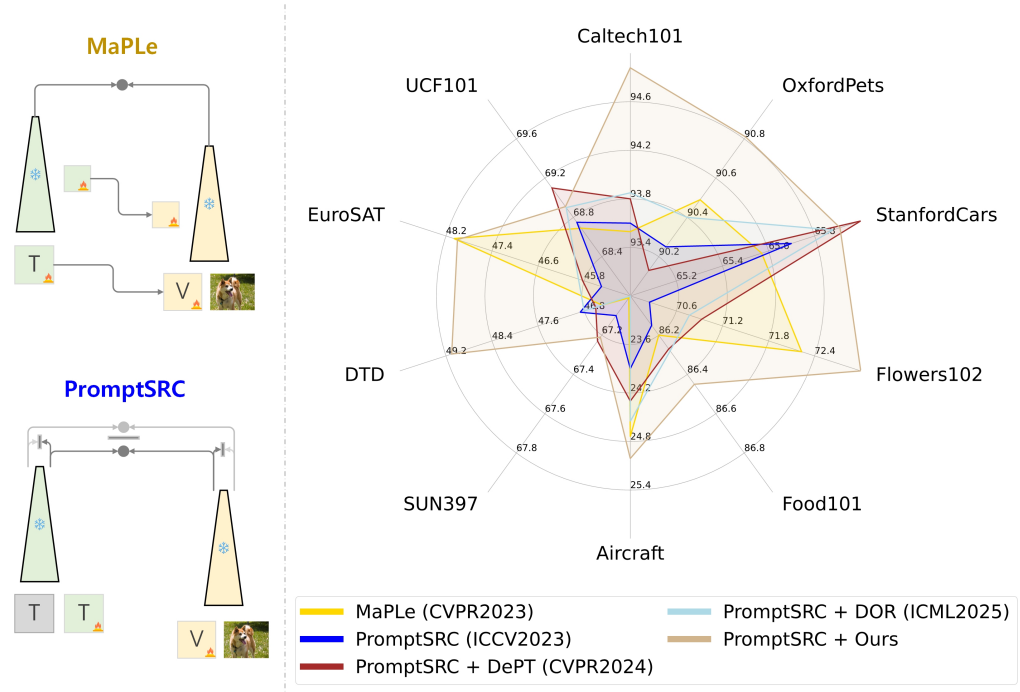

Figure 1: Performance comparison of Ours with prompting plug-in methods under the cross-dataset generalization. For cross-dataset generalization, the MaPLe (without visual loss) is higher than PromptSRC (with visual loss), as shown in Table 1. To improve the cross-dataset generalization of PromptSRC (with visual loss), we propose a plug-in design, improving the performance, surpassing the plug-and-play DePT [29] (CVPR2024) and DOR [30] (ICML2025). The green 'T' and yellow 'V' represent learnable textual prompts and visual prompts, respectively.

2.2 Language Prompt Learning

In recent studies, the weights of CLIP are kept unchanged [16], while a set of textual parameters are learned for prompts—these parameters help fine-tune CLIP for downstream image recognition. In this regard, textual prompting [31, 16, 32] has been proposed to improve base performance and few-shot supervised learning tasks. For class c, the tuning feature of the text encoder is denoted as t_c in a dataset with total C classes. The cross-entropy concept of the textual prompting method is computed as:

$$p(z \mid \mathbf{x}) = \frac{\exp\left(\operatorname{sim}\left(t_c, \mathcal{G}_v(\mathbf{x})\right)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(\operatorname{sim}\left(t_{c'}, \mathcal{G}_v(\mathbf{x})\right)/\tau\right)}. \tag{2}$$

2.3 Vision-Language Prompt Learning for Non-generalizable Fine-tuning

Recent research has found that initializing and learning visual prompt learning [17] directly based on textual prompt learning can better improve the performance of base classes of the same datasets and non-generalizable few-shot learning task [18]. Specifically, the output features of the text encoder in this Vison-Language Prompting (VLP) are denoted as $\boldsymbol{t}_p$, and the output features of image encoder is denoted as $\boldsymbol{v}_p$, where $\mathcal{L}_{\mathrm{CE}}(\cdot,\cdot)$ is the cross-entropy loss with vision and language tuned prompts $\mathbf{p}$ for downstream training samples $\mathcal{N}$:

$$\mathcal{L}_{\mathrm{CE}} = \arg\min_{\mathbf{p}} \mathbb{E}_{(\mathbf{x},z)\sim\mathcal{N}} \mathcal{L}\left(\operatorname{sim}\left(\boldsymbol{t}_p, \boldsymbol{v}_p\right), z\right). \tag{3}$$

Although VLP methods excel in base classes of base-to novel generalization, it still performs poorly in novel class recognition tasks (Avg. 11 datasets). This is because learnable prompts can overfit downstream data [22, 16].

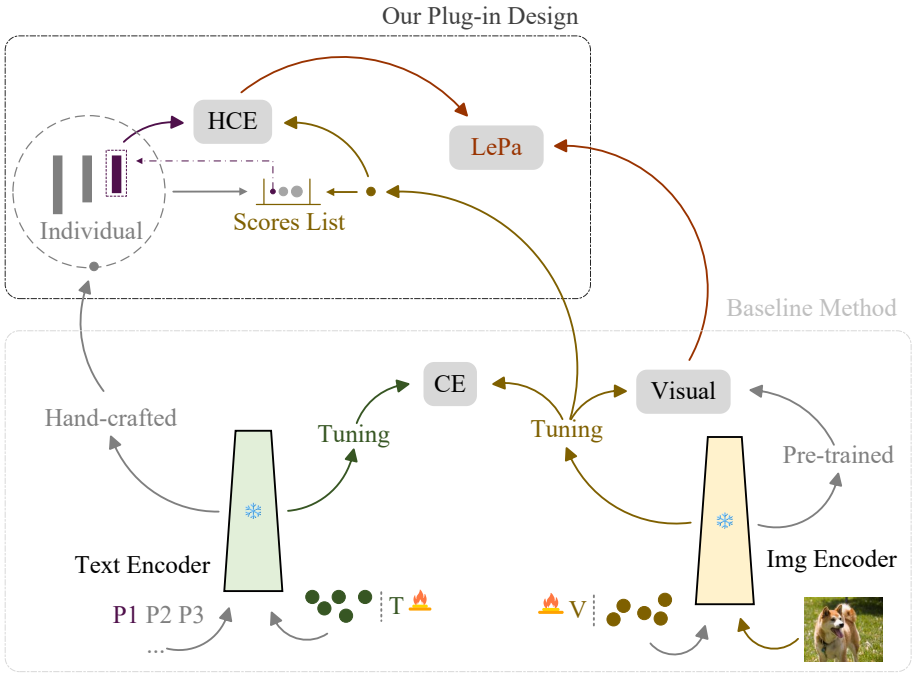

Figure 2: An overview of our plug-in design. We propose a plug-in design called Levelling Paradigm (LePa) in Figure 2 (Top). LePa regularizes the objective functions of two-modal alignment and visual regularization, thereby improving the robustness of coordinated fine-tuning and alleviating uncoordinated learning, thereby enhancing generalization tasks across 11 real datasets. Visual loss exists in baseline works [24, 23, 28, 22]. Three representative generalization experiments demonstrate that our method can effectively address generalized scenarios and tasks.

2.4 Visual Loss for Improving Generalizable Novel Classes

To improve this novel classes within the same dataset, some representative VLP works [22, 23, 28, 24] incorporates the constraint with $L1$ loss [33] on task-specific feature $\boldsymbol{v}_p$ and pre-trained $\boldsymbol{v}$ for visual branch, avoiding the forgetting of frozen CLIP's original generalization capability [34, 1, 22] and improving the novel classes of base-to-novel generalization task, as shown in Table 2. Following the $L1$ loss, the $\mathcal{L}_{\text{Visual}}(\cdot,\cdot)$ represents the visual loss:

$$\mathcal{L}_{\text{Visual}} = |\boldsymbol{v}_p - \boldsymbol{v}|\,. \tag{4}$$

2.5 Challenge and Motivation

However, the performance of these direct-integration VLP through visual loss does not improve significantly in domain generalization and cross-dataset generalization. For novel classes (Table 2), the MaPLe (without visual loss) is lower than PromptSRC (with visual loss). However, for cross-dataset generalization (Table 1), the MaPLe (without visual loss) is higher than PromptSRC (with visual loss). It seems that the visual loss performs poorly when faced with higher difficulty generalization tasks. This may be attributed to the uncoordinated learning [25, 26, 27] of the CE loss and visual loss. The theoretical basis is that the purpose of CE loss is to fine-tune and enhance few-shot learning ability [16], while the purpose of visual loss is to avoid forgetting pre-trained features and enhance generalization ability [22]. There is a compromise and tradeoff between these two losses, a natural conflict where one goes up and the other goes down.

2.6 Proposed Levelling Paradigm (LePa)

To improve the domain and cross-dataset generalizable performance and prevent visual loss convergence instability, we propose a novel plug-in method called Levelling Paradigm (LePa) that

normalizes Hand-crafted CE (HCE) and visual loss, using this method to maintain the stability of visual loss. HCE is composed of pre-trained manual text features, having generalization ability. Specifically, the LePa dynamically constrains the objective functions between vision-language alignment and visual regularization, thereby improving the robustness of coordinated fine-tuning. In Figure 2, these two objectives of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Visual}}$ may exhibit uncoordinated learning.

Specifically, uncoordinated learning can be defined as the scenario where the objective of the vision-language alignment performs well ($\mathcal{L}_{\text{CE}} \approx 0$) while the objective of visual regularization does not ($\mathcal{L}_{\text{Visual}} \gg 0$), or ($\mathcal{L}_{\text{CE}} \gg 0$) with ($\mathcal{L}_{\text{Visual}} \approx 0$). The difference in the value of these two objectives can serve as a measure of uncoordinated learning of direct-integration VLP with visual regularization. The uncoordinated learning not only diminishes the generalization learning of individual objectives but also disrupts the balance of objectives across the entire direct-integration VLP models.

Thus, we propose the $\mathcal{L}_{\text{LePa}}$ to balance the overall framework. We employ N hand-crafted prompts (P1: a picture of a, P2: a photo of a, P3: a drawing of a, etc) to obtain individual N hand-crafted features, as shown in Figure 2. We calculate cosine similarity as the scores between these hand-crafted generated features and the visual features generated by learnable visual embeddings of baseline methods. Afterwards, we will receive a score list.

Note that CLIP's general pre-trained features through hand-crafted prompts have a strong generalization ability [1, 22, 33]. In order to improve generalization, we fully capture the individual semantics of manual prompts in depth, and we obtained the hand-crafted prompt template with the w-worst score based on the scores list, and automatically recorded it. Further, we vectorized the recorded prompt word template text, and we average these w manually output features to represent the overall situation of the bad case for robustness perspective. Further, we align the frozen hand-crafted textual features $\boldsymbol{t}$ with the learnable visual features $\boldsymbol{v}_p$ using cross-entropy loss to obtain $\mathcal{L}_{\text{HCE}}$.

In this regard, we employ the $\mathcal{L}_{\text{LePa}}$ to control the $\mathcal{L}_{\text{HCE}}$ and the visual loss $\mathcal{L}_{\text{Visual}}$ generated by pre-trained visual features $\boldsymbol{v}$ to alleviate uncoordinated learning of direct-integration VLP with visual regularization. Specifically, the $\mathcal{L}_{\text{LePa}}$ can be realized with the following form,

$$\left| \mathcal{L}_{\text{HCE}}\left(\boldsymbol{t}, \boldsymbol{v}_p\right) - \mathcal{L}_{\text{Visual}}\left(\boldsymbol{v}_p, \boldsymbol{v}\right) \right|. \tag{5}$$

Here, model parameters are omitted for simplicity, $\boldsymbol{v}_p$ represents the task-specific visual embeddings. According to the equation, minimizing the objective function $\mathcal{L}_{\text{LePa}}$ can make the model calibrate the uncoordinated learning of entire framework, reflecting in the performance improvement of higher difficulty generalization experiments (Table 1). Based on the above design, we can alleviate the situation of hindering model tuning and alignment: V-L alignment performs well, with the visual branch does not. The training objective of our method involves hyperparameters γ_1 and γ_2. The total loss $\mathcal{L}_{\text{total}}$ can be formulated as follows,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \gamma_1 \mathcal{L}_{\text{LePa}} + \gamma_2 \mathcal{L}_{\text{Visual}}. \tag{6}$$

The proposed design highlights the necessity for generating features that ensure consistent performance across various objectives. Recognizing and addressing this uncoordinated learning would aid in promoting uniformity through the optimization of embeddings and models, aligning with the ethos of bridging the divide between language and visual representations of CLIP. Consequently, LePa regularizes the interaction between two-modal alignment and visual regularization to enhance the performance of the existing direct-integration approach, thereby addressing the uncoordinated learning in these frameworks [22, 23, 28, 24]. It ensures continuous robustness of the overall direct-integration VLP with visual regularization. Note that the proposed LePa is a plug-in design that is compatible with existing direct-integration VLP frameworks, as shown in (Table 3, Table 1, and Table 2).

3 Main Results

We introduce compare methods, implementation details, 11 datasets, and the key experiments (cross-dataset generalization, domain generalization, and base-to-novel generalization).

3.1 Compared Prompting Methods

We conduct performance analysis based on various direct-integration VLP baselines, including MaPLe [19] (CVPR2023), PromptSRC (PSRC) [22] (ICCV2023), and CoPrompt (CoP) [23]

(ICLR2024). The MaPLe employs the learnable hidden multi-layer for the fusion of two encoders without visual loss. The PromptSRC is independent vision and language prompts which employs traditional KL loss and L1 loss to avoid forgetting pre-trained knowledge through visual loss. The CoPrompt exists learnable hidden multi-layer which employs two learning adapters and two perturbed inputs to enhance the base-to-novel generalization through visual loss. The DePT [29] (CVPR2024) and DOR [30] (ICML2025) is plug-in methods. For the compared prompting methods, we adopt the optimal settings. We use a ViT-B/16-based CLIP model with compared methods for fair comparison.

Table 1: Cross-dataset generalization. * indicates our reproduced results.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
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
+ DePT	73.27	92.55	90.22	64.58	70.09	85.33	23.78	66.26	45.11	40.25	67.88	64.60
+ DOR	71.50	93.10	90.00	64.11	73.23	86.64	25.13	67.04	46.25	49.30	68.70	66.35
+ Ours	72.33	94.11	90.85	65.88	73.81	86.55	24.81	68.89	46.73	49.22	68.11	**66.89**
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
+ DePT	71.60	93.80	90.13	66.00	70.93	86.27	24.30	67.23	46.60	45.83	69.10	66.02
+ DOR	71.55	93.85	90.40	65.88	70.77	86.28	24.55	67.00	46.80	45.90	68.90	66.03
+ Ours	71.50	94.88	90.81	65.91	73.00	86.45	25.01	67.21	49.10	48.00	68.91	**66.93**
CoPrompt*	70.70	93.40	90.15	65.38	72.15	86.00	24.10	66.08	47.25	51.15	69.00	65.46
+ DePT	71.01	93.50	90.18	66.33	70.71	86.00	24.10	67.08	46.01	45.06	69.40	65.84
+ DOR	71.00	93.11	90.03	66.04	70.32	86.10	24.23	67.13	45.71	45.20	69.00	65.69
+ Ours	71.08	93.15	90.30	66.33	73.01	87.88	24.62	67.00	46.96	48.05	69.10	**66.64**

3.2 Implementation Details

After the ablation experiment, we obtained the implementation details. We set textual and visual embeddings to 4 based on all VLP methods. We use deep prompting with multi-modal encoders and an SGD optimizer with a learning rate of 0.0026 on a single A5000 GPU. Training for 30 epochs for base-to-novel generalization by 16-shot, 20 epochs for domain generalization and cross-dataset evaluation setting. We train the ImageNet source model on all classes with 16-shot in the first 3 transformer layers for domain generalization and cross-dataset evaluation. We set $\gamma_1 = \gamma_2 = 5$ for multi-modal regularization in total loss. We set $w = 4$ for w-worst cases. For the base-to-novel generalization, we set the learning depth to 9. We fix $N = 60$ hand-crafted prompts [1, 22], following CLIP. The hand-crafted prompt template used in this paper is: "a photo of a .", "a bad photo of a .", "a photo of many .", "a sculpture of a .", "a photo of the hard to see .", etc.

3.3 Datasets

The open-source real datasets cover multiple recognition tasks. We conducted base-to-novel generalization experiments and cross-dataset generalization experiments on 11 datasets. We conduct domain generalization experiments on four variants of ImageNet [35]. The datasets encompass various recognition tasks, including ImageNet [35], Caltech101 [36] for generic objects, OxfordPets [37], StanfordCars [38], Flowers102 [39], Food101 [40], FGVCAircraft [41] for fine-grained classification, SUN397 [42] for scene recognition, UCF101 [43] for action recognition, DTD [44] for texture classification, and EuroSAT [45] for satellite images. For the domain generalization benchmark, we use ImageNetA [46], ImageNet-R [47], ImageNet-Sketch [48] and ImageNetV2 [49].

3.4 Cross-Dataset Generalization Task

In Table 1, training on the source dataset (first column) is a non-generalization task only for ImageNet. The model trained on ImageNet will directly obtain an accuracy of 50,000 images, which is the source data in the first column of the Table 1. We test our ImageNet-trained model directly on the other 10 datasets to validate the potential of our method in cross-dataset transfer [50]. This experimental

Table 2: Base-to-novel generalization. * indicates our reproduced results.

(a) **Avg. 11 datasets**

	Base	Novel	HM
MaPLe	82.28	75.14	78.55
+ DePT	84.85	74.82	79.52
+ DOR	84.13	76.15	79.94
+ Ours	84.48	76.46	**80.27**
PSRC	84.26	76.10	79.97
+ DePT	85.19	76.17	80.43
+ DOR	84.34	76.99	80.48
+ Ours	85.61	77.46	**81.33**
CoP*	83.89	76.75	80.13
+ DePT	84.03	76.77	80.23
+ DOR	84.10	77.69	80.76
+ Ours	84.86	77.79	**81.16**

(b) ImageNet

	Base	Novel	HM
MaPLe	76.66	70.54	73.47
+ DePT	77.87	70.23	73.85
+ DOR	76.57	72.00	74.23
+ Ours	78.03	71.05	74.37
PSRC	77.60	70.73	74.01
+ DePT	78.20	70.27	74.02
+ DOR	77.71	71.54	74.49
+ Ours	78.81	71.44	74.94
CoP*	77.00	71.10	74.02
+ DePT	77.01	71.05	74.00
+ DOR	77.21	72.03	74.58
+ Ours	78.15	70.88	74.46

(c) Caltech101

	Base	Novel	HM
MaPLe	97.74	94.36	96.02
+ DePT	98.53	95.03	96.75
+ DOR	97.55	94.90	96.22
+ Ours	98.28	96.00	97.12
PSRC	98.10	94.03	96.02
+ DePT	98.57	94.10	96.28
+ DOR	98.25	94.80	96.51
+ Ours	98.55	95.91	97.21
CoP*	97.30	95.05	96.17
+ DePT	97.81	95.11	96.46
+ DOR	97.44	96.00	96.71
+ Ours	98.00	95.10	96.54

(d) OxfordPets

	Base	Novel	HM
MaPLe	95.43	97.76	96.58
+ DePT	95.03	97.83	96.41
+ DOR	95.55	98.50	97.02
+ Ours	95.76	98.50	97.11
PSRC	95.33	97.30	96.30
+ DePT	95.43	97.33	96.37
+ DOR	95.66	98.40	97.03
+ Ours	95.61	97.80	96.69
CoP*	95.10	97.13	96.11
+ DePT	94.60	97.35	95.97
+ DOR	95.35	98.00	96.67
+ Ours	95.01	98.13	96.56

(e) EuroSAT

	Base	Novel	HM
MaPLe	94.07	73.23	82.35
+ DePT	94.43	76.23	84.36
+ DOR	94.11	74.25	82.84
+ Ours	94.19	75.71	83.94
PSRC	92.90	73.90	82.32
+ DePT	92.23	77.90	84.88
+ DOR	92.80	74.81	82.71
+ Ours	95.41	76.90	85.16
CoP*	94.24	75.88	84.12
+ DePT	94.75	76.00	84.59
+ DOR	94.56	76.60	84.94
+ Ours	95.11	76.32	84.97

(f) UCF101

	Base	Novel	HM
MaPLe	83.00	78.66	80.77
+ DePT	86.87	78.10	82.25
+ DOR	83.22	79.60	81.38
+ Ours	87.00	79.31	82.97
PSRC	87.10	78.80	82.74
+ DePT	87.73	77.70	82.46
+ DOR	87.21	79.60	83.32
+ Ours	88.74	79.11	83.65
CoP*	86.00	79.63	82.76
+ DePT	86.10	79.80	82.93
+ DOR	86.44	80.77	83.56
+ Ours	86.81	80.00	83.37

(g) StanfordCars

	Base	Novel	HM
MaPLe	72.94	74.00	73.47
+ DePT	80.93	71.73	76.06
+ DOR	92.87	75.10	83.02
+ Ours	78.51	75.05	76.74
PSRC	78.27	74.97	76.58
+ DePT	80.80	75.00	77.79
+ DOR	78.50	75.90	77.19
+ Ours	80.80	76.70	78.69
CoP*	80.80	74.33	77.50
+ DePT	81.13	74.00	77.49
+ DOR	80.95	75.50	78.19
+ Ours	81.12	76.05	78.56

(h) Flowers102

	Base	Novel	HM
MaPLe	95.92	72.46	82.56
+ DePT	98.03	73.17	83.79
+ DOR	95.80	74.00	83.63
+ Ours	98.18	74.88	84.96
PSRC	98.07	76.50	85.95
+ DePT	98.40	77.10	86.46
+ DOR	98.10	77.40	86.89
+ Ours	98.70	78.36	87.36
CoP*	96.10	77.25	86.04
+ DePT	96.81	77.05	86.33
+ DOR	96.65	77.94	86.59
+ Ours	97.88	78.02	87.27

(i) Food101

	Base	Novel	HM
MaPLe	90.71	92.05	91.38
+ DePT	90.33	91.53	90.93
+ DOR	90.88	92.90	91.88
+ Ours	90.28	92.95	91.59
PSRC	90.67	91.53	91.10
+ DePT	90.87	91.57	91.22
+ DOR	90.90	92.63	91.76
+ Ours	90.58	92.51	91.53
CoP*	90.65	92.03	91.34
+ DePT	91.00	92.72	91.85
+ DOR	90.45	93.00	91.70
+ Ours	91.10	92.57	91.83

(j) FGVCAircraft

	Base	Novel	HM
MaPLe	37.44	35.61	36.50
+ DePT	44.53	32.80	37.78
+ DOR	37.85	36.70	37.32
+ Ours	43.00	37.35	39.97
PSRC	42.73	37.87	40.15
+ DePT	45.70	36.73	40.73
+ DOR	42.80	38.71	40.65
+ Ours	45.78	39.87	42.62
CoP*	41.40	39.80	40.58
+ DePT	41.01	38.75	39.85
+ DOR	41.51	40.60	41.05
+ Ours	45.12	44.20	44.65

(k) SUN397

	Base	Novel	HM
MaPLe	80.82	78.70	79.75
+ DePT	82.90	76.40	79.52
+ DOR	80.76	79.56	80.16
+ Ours	82.13	79.65	80.87
PSRC	82.67	78.47	80.52
+ DePT	83.27	78.97	81.06
+ DOR	82.80	79.35	81.04
+ Ours	83.88	80.77	82.29
CoP*	82.49	78.70	80.56
+ DePT	82.16	79.10	80.59
+ DOR	82.68	80.01	81.32
+ Ours	83.01	80.11	81.54

(l) DTD

	Base	Novel	HM
MaPLe	80.36	59.18	68.16
+ DePT	83.87	59.93	69.91
+ DOR	80.35	60.20	68.84
+ Ours	83.92	60.65	70.41
PSRC	83.37	62.97	71.75
+ DePT	84.80	61.20	71.09
+ DOR	83.11	63.81	72.19
+ Ours	84.90	62.70	72.13
CoP*	81.78	63.40	71.45
+ DePT	82.06	63.55	71.63
+ DOR	81.90	64.22	71.99
+ Ours	82.25	64.39	72.26

setup is similar to the domain generalization experimental setup. Compared with DePT, our method shows improved performance in average precision. For cross-dataset generalization, the MaPLe (without visual loss) is higher than PromptSRC (with visual loss). Through our plugin integration, the cross-dataset performance of PromptSRC has been improved. In Figure 1, it suggests that our method demonstrates a clear advantage in terms of generalization across most classification scenes.

3.5 Base-to-Novel Generalization Task

This experiment is to test whether the model can handle small-scale generalization. We follow a setting where the same datasets are split into base and novel classes. So the distribution of the base classes is similar to that of the novel classes. Dividing all classes of a dataset into two parts is a process of random division. Please note that this is divided into two equally sized parts. The model is trained only on the base classes in a 16-shot setting and tested on base (non-generalization task) and novel classes (generalization task). Finally, we use the HM (Harmonic Mean) [51] to evaluate the model's ability to balance generalization and non-generalization. The HM task aims to observe the model's fine-tuning ability and similar classes classification ability, expressed as:

$$HM = \frac{2 \times Acc_{\text{base}} \times Acc_{\text{novel}}}{Acc_{\text{base}} + Acc_{\text{novel}}}. \quad (7)$$

Table 3: Domain generalization Task. * indicates our reproduced results.

	Source	Target			
	ImageNet	-V2	-S	-A	-R
MaPLe	70.72	64.07	49.15	50.90	76.98
+ DePT	73.27	65.00	49.05	51.15	77.30
+ DOR	71.50	64.94	48.56	52.00	77.10
+ Ours	72.33	65.41	50.40	52.15	78.15
PSRC	71.27	64.35	49.55	50.90	77.80
+ DePT	71.60	64.51	50.15	51.88	77.18
+ DOR	71.55	64.00	50.20	52.15	77.65
+ Ours	71.50	65.15	51.05	52.33	78.75
CoP*	70.70	64.15	49.48	50.60	77.40
+ DePT	71.01	64.60	50.05	51.05	77.45
+ DOR	71.00	63.90	50.13	52.00	77.51
+ Ours	71.08	65.60	51.08	52.00	78.20

As shown in Table 2 (left-top), our plugin design is based on these representative VLP architectures, which have all improved novel performance. Moreover, previous methods have compromised their generalization capabilities when it comes to handling more specialized datasets. For the specialized EuroSAT [45], our method provides the highest novel accuracy. In Table 2, our method provides the best-averaged results on the novel classes. Overall, our method demonstrates significant improvements on an average of 11 datasets for HM.

3.6 Domain Generalization Task

In Table 3, we train our method on the ImageNet dataset with 1,000 classes and then test it on the domain shift datasets to evaluate the robustness [52, 53, 54] of our approach. In domain generalization experiments, training on the source dataset, ImageNet, is a non-generalization task only for this dataset. The remaining 4 ImageNet variant datasets are tested directly using ImageNet-trained models to determine the model's generalization ability. Our method consistently outperforms all existing baselines (MaPLe, PromptSRC, CoPrompt) on target datasets. When compared to plug-in DePT [29] and DOR [30], our method demonstrates improved performance across variants of the ImageNet dataset. Our method is purposefully crafted to bolster generalization capabilities when faced with domain shifts. This experiment is commonly set up in real-life scenarios, such as whether the model can recognize objects with minor tampering.

4 Further Analysis

In this section, we conduct crucial ablation experiments. The first experiment is cost training, and the second experiment is cross-dataset generalization experiment with different semantics.

4.1 Computational Cost

In Table 4a and Table 4b, the computational cost analysis is performed using the SUN397 over 10 epochs on a single GPU. Our method (row 3) may exhibit more training time due to the necessity of performing multiple cosine similarity calculations. In future work, we will consider how to alleviate

this time consumption. Compared to PromptSRC and MaPLe, our plug-and-play method does not increase in terms of parameter count. The DePT method increases the number of learning parameters, which are related to the parameters of the base class. Our component only adds the type of loss.

Table 4: The cost analysis is performed using the SUN397 [42] over 10 epochs. 'N': the number of classes in the base task [29].

(a) The cost analysis is based on PromptSRC [22].

Method	Train time	Learnable para.	HM
PromptSRC	13.13 min	5120	79.97
+ DePT	13.88 min	+ (2+N/2) K	80.13
+ Ours	14.80 min	+ 0 K	81.01

(b) The cost analysis is based on MaPLe [19].

Method	Train time	Learnable para.	HM
MaPLe	10.55 min	3.55M	79.59
+ DePT	10.66 min	+ (2+N/2)K	79.69
+ Ours	13.56 min	+ 0 K	79.98

4.2 Analysis of Semantic Misalignment

In Table 5, we design and conduct a new experimental setup. In this experiment, we trained 10 datasets (Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, UCF101, DTD, EuroSAT) with with 16-shot learning and 5 epochs, and then tested each dataset directly on the remaining 9 datasets. We did not use the ImageNet dataset because it has too many categories and would include other datasets. This setting is to simulate the situation of semantic mismatch. For example, after the model is trained on the SUN397 dataset, the current embeddings are fitted to the SUN397 dataset. Then when we use this model to test the Food101 dataset, the semantics of the text branch embeddings and the visual branch test dataset are not aligned. The cross-dataset generalization of this pattern is a difficult generalization setting, and in real-world scenarios, the distribution gap between the trained dataset and the generalized dataset is often very large. We use this experiment to observe the generalization effect of this pattern in different unlabeled scenarios.

Table 5: Analysis of Semantic Misalignment.

Source	Target									
	Caltech101 [36]	Food101 [40]	DTD [44]	UCF101 [43]	Flowers102 [39]	OxfordPets [37]	Aircraft [41]	StanfordCars [38]	SUN397 [42]	EuroSAT [45]
Caltech101 (94.7)		86.0	45.6	65.2	67.2	88.3	23.8	65.5	65.7	44.6
Food101 (87.1)	92.7		43.3	64.9	67.9	88.3	23.5	65.5	64.8	47.5
DTD (58.3)	93.1	85.6		64.8	69.8	87.5	22.4	64.9	62.0	43.7
UCF101 (76.5)	91.0	85.4	43.7		65.9	83.1	20.9	63.7	62.8	42.8
Flowers102 (87.5)	89.5	80.5	39.1	58.2		82.9	19.9	44.6	52.6	51.9
OxfordPets (93.1)	93.1	86.2	42.1	63.5	66.4		21.6	44.6	52.6	48.7
Aircraft (31.0)	81.3	84.1	40.7	63.3	62.4	83.2		44.6	58.2	45.0
StanfordCars (69.2)	90.3	85.2	40.7	62.8	59.4	85.0	22.3		61.0	47.7
SUN397 (70.0)	91.9	85.7	38.4	63.9	63.7	83.2	21.7	62.4		49.1
EuroSAT (76.7)	84.3	68.7	31.2	54.2	32.0	43.6	17.5	39.1	53.6	

5 Related Work

5.1 Vision-Language Models (VLMs)

VLMs have revolutionized the field of artificial intelligence by seamlessly blending visual and textual data to enhance comprehension of multimodal information [55, 56, 57]. This innovative technology serves as a robust solution for analyzing and producing content that merges images and text, resulting in major advancements across diverse domains like image captioning [4, 58, 59] and other tasks [60, 61, 62]. Central to the architecture of VLMs is the amalgamation of visual and textual modalities, enabling the system to acquire intricate representations that encapsulate the semantic

connections between images and their corresponding textual descriptions [63, 64]. Through the concurrent processing of visual and textual data, VLMs effectively bridge the divide between disparate modalities, generating coherent outputs that capitalize on the synergistic information inherent in each domain [65, 66, 67]. The VLMs can also demonstrate value in other fields [68, 69, 70].

5.2 Prompt Learning

Prompt Learning serves as a prevalent method in the field of NLP, aiding in the acquisition of skills for various subsequent tasks [71, 72]. Employing textual prompts [73], serving as directives for the linguistic component of a VLMs, replicating this practice is a prevalent method to augment comprehension of tasks. The adapter [74, 75] approaches can achieve competitive performance to adapt VLMs to downstream tasks [76, 77, 78]. Nevertheless, the limitations of these approaches have spurred the investigation of novel techniques inspired by prompt tuning in Natural Language Processing (NLP). The image-conditional prompt [31] significantly contributes to enhancing generalization to unseen classes (unlabeled samples). This conditioning aids in improving the generalization to unseen examples. Some methods [32, 33] constrain the learnable prompts to prior distribution learning. TAP [79] first instructs large language models to generate a tree of attributes with a "concept - attribute - description" structure for each category. In addition to single-modal prompt tuning and two-modal prompt tuning, FAP [80] introduces robust attack [81, 82, 83] for prompt learning based on VLP. Moreover, prompt learning can also demonstrate value in robot fields [84, 85] and innovative research areas [86, 87].

6 Limitations

Our design has more training time, we will explore how to reduce this overhead. In Table 3 (column-1) and Table 1 (column-1), it seems that our method has a lower performance compared to DePT on the source dataset of training labeled ImageNet. The LePa aims to enhance cross-dataset generalization and domain generalization experiments, but there is a compromise and tradeoff with specifying ImageNet dataset fine-tuning. In real-world industrial settings, obtaining labeled training data can often be challenging or impractical. In future work, we will try to improve our performance on few-shot learning and generalization, such as 1-shot and 2-shot scenarios with little training data.

7 Conclusion

It has recently been discovered that VLP frameworks have underscored their potential to improve the fine-tuning performance of labeled scenarios for pre-trained CLIP. In order to improve the generalization of novel classes in the same dataset, some methods have added visual regularization to VLP frameworks. However, the performance of these baseline methods does not improve significantly in domain generalization and cross-dataset generalization tasks. This may be attributed to the uncoordinated learning of the fine-tuning CE loss and generalizable visual loss. To address this problem of uncoordinated learning, we propose an effective design called Levelling Paradigm (LePa) to improve performance for unlabeled tasks or scenarios. The proposed LePa, designed as a plug-and-play method, dynamically constrains and coordinates multiple objective functions, thereby enhancing the generalization of these baseline methods. Representative tasks across 11 real datasets on generalization from base-to-novel, cross-dataset generalization, and domain generalization demonstrate that our design can effectively address generalized scenarios and tasks.

Broader Impacts. This paper presents work whose goal is to advance the field of Machine Learning. None of these points we feel must be specifically highlighted here. This finding carries important implications for the deployment of VLMs across diverse real-world applications. By enhancing zero-shot recognition capabilities, our approach offers substantial benefits to industries that depend on large-scale image analysis, including the improvement of visual search systems, the refinement of automated image annotation pipelines, and the advancement of tools in imaging. The societal implications of our research involve democratizing the availability of potent AI resources, as our approach can achieve impressive performance even in the absence of extensive labeled data, thereby enhancing the accessibility and utility of advanced VLMs in environments with limited resources. Furthermore, our method promotes ethical AI practices by minimizing the necessity for extensive model training and adaptation, thereby supporting sustainability and efficiency objectives.

Acknowledgement

We extend our heartfelt gratitude to the anonymous reviewers for their insightful comments, which greatly improved the quality of this paper. We sincerely express our gratitude to the anonymous AC for the responsible coordination throughout the entire process. This work was supported in part by NSFC No. 62222117. This work was supported in part by the National Natural Science Foundation of China under Grants 62422204, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LDT23F02025F02, and in part by the Key Research and Development Program of Zhejiang Province under Grant 2025C01026.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, and Marco Fumero. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv preprint arXiv:2110.02624*, 2021.

[3] Djamahl Etchegaray, Zi Huang, Tatsuya Harada, and Yadan Luo. Find n'propagate: Open-vocabulary 3d object detection in urban environments. *arXiv preprint arXiv:2403.13556*, 2024.

[4] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European Conference on Computer Vision*, pages 512–531. Springer, 2022.

[5] Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. Unsupervised post-training for multi-modal llm reasoning via grpo. *arXiv preprint arXiv:2505.22453*, 2025.

[6] Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:39501–39521, 2024.

[7] Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

[8] Yiyuan Pan, Yunzhe Xu, Zhe Liu, and Hesheng Wang. Planning from imagination: Episodic simulation and episodic memory for vision-and-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6345–6353, 2025.

[9] Kefan Jin, Zhe Liu, Jian Wang, and Hesheng Wang. Unmanned surface vehicle navigation under disturbances: World model enhanced reinforcement learning. *IEEE/ASME Transactions on Mechatronics*, 2025.

[10] Haoang Li, Yixin Mai, Ming Gao, Junjie He, Zhe Liu, and Hesheng Wang. Large-scale lidar-based loop closing via combination of equivariance and invariance on se (3). *IEEE/ASME Transactions on Mechatronics*, 2025.

[11] Jiuming Liu, Dong Zhuo, Zhiheng Feng, Siting Zhu, Chensheng Peng, Zhe Liu, and Hesheng Wang. Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment. In *European Conference on Computer Vision*, pages 475–493. Springer, 2024.

[12] Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang, Le Song, and Yu Li. Progress and opportunities of foundation models in bioinformatics. *Briefings in Bioinformatics*, 25(6):bbae548, 2024.

[13] Hang Yu, Ruilin Li, Shaorong Xie, and Jiayan Qiu. Shadow-enlightened image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7860, 2024.

[14] Jinkun Hao, Naifu Liang, Zhen Luo, Xudong Xu, Weipeng Zhong, Ran Yi, Yichen Jin, Zhaoyang Lyu, Feng Zheng, Lizhuang Ma, et al. Mesatask: Towards task-driven tabletop scene generation via 3d spatial reasoning. *arXiv preprint arXiv:2509.22281*, 2025.

[15] Ross Greer, Bjørk Antoniussen, Mathias V Andersen, Andreas Møgelmose, and Mohan M Trivedi. The why, when, and how to use active learning in large-data-driven 3d object detection for safe autonomous driving: An empirical exploration. *arXiv preprint arXiv:2401.16634*, 2024.

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[18] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022.

[19] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[20] Jinhao Li, Haopeng Li, Sarah Monazam Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models. In *Forty-first International Conference on Machine Learning*.

[21] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 52792–52810. Curran Associates, Inc., 2023.

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.

[23] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023.

[24] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024.

[25] Jianqi Gao, Hao Wu, Yiu-ming Cheung, Jian Cao, Hang Yu, and Yonggang Zhang. Mitigating forgetting in adapting pre-trained language models to text processing tasks via consistency alignment. In *Proceedings of the ACM on Web Conference 2025*, pages 3492–3504, 2025.

[26] Jianqi Gao, Hang Yu, Yiu-ming Cheung, Jian Cao, Raymond Chi-Wing Wong, and Yonggang Zhang. Shaping pre-trained language models for task-specific embedding generation via consistency calibration. *Neural Networks*, page 107754, 2025.

[27] Jianqi Gao, Jian Cao, Hang Yu, Yonggang Zhang, and Zhen Fang. Sena: Leveraging set-level consistency adversarial learning for robust pre-trained language model adaptation. *Knowledge-Based Systems*, page 113831, 2025.

[28] Yonggang Zhang and Xinmei Tian. Consistent prompt learning for vision-language models. *Knowledge-Based Systems*, 310:112974, 2025.

[29] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024.

[30] Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. In *Forty-second International Conference on Machine Learning*.

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[32] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[33] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[34] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6757–6767, June 2023.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[36] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[39] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[40] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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*, pages 446–461. Springer, 2014.

[41] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[42] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[44] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[45] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[46] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[47] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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*, pages 8340–8349, 2021.

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[50] Yonggang Zhang, Jie Lu, Bo Peng, Zhen Fang, and Yiu-ming Cheung. Learning to shape in-distribution feature space for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:49384–49402, 2024.

[51] Joseph Cavanaugh. Encyclopedia of statistical sciences, 2007.

[52] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 32(10):4309–4322, 2020.

[53] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *International conference on machine learning*, pages 3122–3132. PMLR, 2021.

[54] Shijia Liu, Zhenghua Chen, Min Wu, Hao Wang, Bin Xing, and Liangyin Chen. Generalizing wireless cross-multiple-factor gesture recognition to unseen domains. *IEEE Transactions on Mobile Computing*, 23(5):5083–5096, 2023.

[55] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[56] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 19–27, 2020.

[57] Yadan Luo, Zhuoxiao Chen, Zhen Fang, Zheng Zhang, Mahsa Baktashmotlagh, and Zi Huang. Kecor: Kernel coding rate maximization for active 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18279–18290, 2023.

[58] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

[59] Yadan Luo, Zi Huang, Yang Li, Fumin Shen, Yang Yang, and Peng Cui. Collaborative learning for extremely low bit asymmetric hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(12):3675–3685, 2021.

[60] Bo Du, Lixiang Ru, Chen Wu, and Liangpei Zhang. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9976–9992, 2019.

[61] Bo Du and Liangpei Zhang. Random-selection-based anomaly detector for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 49(5):1578–1589, 2011.

[62] Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen, and Dacheng Tao. Exploring representativeness and informativeness for active learning. *IEEE Transactions on Cybernetics*, 47(1):14–26, 2017.

[63] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[64] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[65] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[66] Yadan Luo, Ziwei Wang, Zi Huang, Yang Yang, and Cong Zhao. Coarse-to-fine annotation enrichment for semantic segmentation learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 237–246, 2018.

[67] Yadan Luo, Zijian Wang, Zhuoxiao Chen, Zi Huang, and Mahsa Baktashmotlagh. Source-free progressive graph learning for open-set domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[68] Mingkui Tan, Peihao Chen, Hongyan Zhi, Jiajie Mai, Benjamin Rosman, Dongyu Ji, and Runhao Zeng. Source-free elastic model adaptation for vision-and-language navigation. *IEEE Transactions on Multimedia*, pages 1–13, 2025.

[69] Mingkui Tan, Gengqin Ni, Xu Liu, Shiliang Zhang, Xiangmiao Wu, Yaowei Wang, and Runhao Zeng. Bidirectional posture-appearance interaction network for driver behavior recognition. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13242–13254, 2022.

[70] Mingkui Tan, Zhuangwei Zhuang, Sitao Chen, Rong Li, Kui Jia, Qicheng Wang, and Yuanqing Li. Epmf: Efficient perception-aware multi-sensor fusion for 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8258–8273, 2024.

[71] Fangming Cui, Yonggang Zhang, Xuan Wang, Xinmei Tian, and Jun Yu. Enhancing target-unspecific tasks through a features matrix. In *Forty-second International Conference on Machine Learning*, 2025.

[72] Fangming Cui, Xun Yang, Chao Wu, Liang Xiao, and Xinmei Tian. Advancing prompt learning through an external layer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 807–816, 2024.

[73] Mengjia Wang, Fang Liu, Licheng Jiao, Shuo Li, Lingling Li, Puhua Chen, Xu Liu, and Wenping Ma. Vcgprompt: Visual concept graph-aware prompt learning for vision-language models. *Pattern Recognition*, page 112012, 2025.

[74] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[75] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

[76] Wenxi Li, Yuchen Guo, Jilai Zheng, Haozhe Lin, Chao Ma, Lu Fang, and Xiaokang Yang. Sparseformer: Detecting objects in hrw shots via sparse vision transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4851–4860, 2024.

[77] Zhe Liu, Yu Zhai, Jiaming Li, Guangming Wang, Yanzi Miao, and Hesheng Wang. Graph relational reinforcement learning for mobile robot navigation in large-scale crowded environments. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8776–8787, 2023.

[78] Bo Du, Yuxiang Zhang, Liangpei Zhang, and Dacheng Tao. Beyond the sparsity-based target detector: A hybrid sparsity and statistics-based detector for hyperspectral images. *IEEE Transactions on Image Processing*, 25(11):5345–5357, 2016.

[79] Tong Ding, Wanhua Li, Zhongqi Miao, and Hanspeter Pfister. Tree of attributes prompt learning for vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[80] Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. *Advances in Neural Information Processing Systems*, 37:3122–3156, 2024.

[81] Yonggang Zhang, Ya Li, Tongliang Liu, and Xinmei Tian. Dual-path distillation: A unified framework to improve black-box attacks. In *International Conference on Machine Learning*, pages 11163–11172. PMLR, 2020.

[82] Yonggang Zhang, Xinmei Tian, Ya Li, Xinchao Wang, and Dacheng Tao. Principal component adversarial example. *IEEE Transactions on Image Processing*, 29:4804–4815, 2020.

[83] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021.

[84] Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. Flame: Learning to navigate with multimodal llm in urban environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9005–9013, 2025.

[85] Zhe Liu, Kai Li, Tian Hao, and Hesheng Wang. Visual servoing of rigid-link flexible-joint manipulators in the presence of unknown camera parameters and boundary output. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(8):5096–5105, 2023.

[86] Wenyao Zhang, Shipeng Lyu, Feng Xue, Chen Yao, Zheng Zhu, and Zhenzhong Jia. Predict the rover mobility over soft terrain using articulated wheeled bevameter. *IEEE Robotics and Automation Letters*, 7(4):12062–12069, 2022.

[87] Yuanze Wang, Yichao Yan, Dianxi Shi, Wenhan Zhu, Jianqiang Xia, Tan Jeff, Songchang Jin, Ke Gao, Xiaobo Li, and Xiaokang Yang. Nerf-ibvs: visual servo based on nerf for visual localization and navigation. *Advances in Neural Information Processing Systems*, 36:8292–8304, 2023.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Please refer to the main text.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please refer to the main text.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Please participate in the main text. Our article covers the reasoning details and experimental setup.

   Guidelines:

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

5. **Open access to data and code**

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

Answer: [No]

Justification: We are organizing the code.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Our paper includes ablation experiments with parameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Our experiment is an average obtained through three runs, which is convincing.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Our paper describes computing resources and machines.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discussed it.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

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

13. **New assets**

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

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

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

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

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

Answer: [NA]

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

Guidelines:

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

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

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

Answer: [NA]

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

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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