LOREAL: Mitigating Low-Resolution Challenges in Vision-Language Models with Attribute-driven Prompt Self-Distillation

Anonymous CVPR submission

Abstract

*Prompt Learning (PL) has emerged as a parameter-efficient technique for adapting Vision-Language Models (VLMs) to downstream tasks. However, almost all existing PL methods are primarily designed and evaluated on well-curated datasets, overlooking a critical post-deployment phenomenon, i.e., the intrinsic connection between input resolution and storage-memory consumption. Specifically, to satisfy the stringent storage-memory constraints on edge devices, models are often limited to low-resolution inputs (e.g., $\leq 224\times224$ for CLIP-ViT/B-16) and generate fewer tokens (with the position embedding resized), which poses a unique challenge in performance robustness. To tackle this issue, we propose LOREAL, an efficient prompt self-distillation framework that learns resolution-invariant representations by excavating attribute semantics. At the heart of LOREAL is a dual-student architecture, i.e., two student models fed with inputs at different resolutions synergistically learn from each other. Building upon this, we contextualize the students' prompt with resolution-invariant attributes queried from the LLM, then leverage cross-modality meta-nets to generate attribute semantics. These meta-nets are bridged between the different encoders of two students, wherein we introduce Low-Level Distillation (LLD) and High-Level Distillation (HLD) to facilitate the learning of more cross-resolution representations. Extensive experiments show that LOREAL significantly improves VLMs' performance and robustness under varied resolution settings, **underscoring significant practical utilities**. Code is in the **Supplementary Material**.*

1. Introduction

The research community has witnessed a remarkable advancement of Vision-Language Models (VLMs) [24, 35, 36, 40, 53], which demonstrates promising zero-shot generalization ability enabled by their joint modeling of multi-modal semantics. As a typical representative, CLIP [53] employs contrastive pre-training over separate modality encoders to learn a unified manifold space. Its prominent cross-modal alignment capabilities have facilitated a wide range of downstream applications [3, 56, 67].

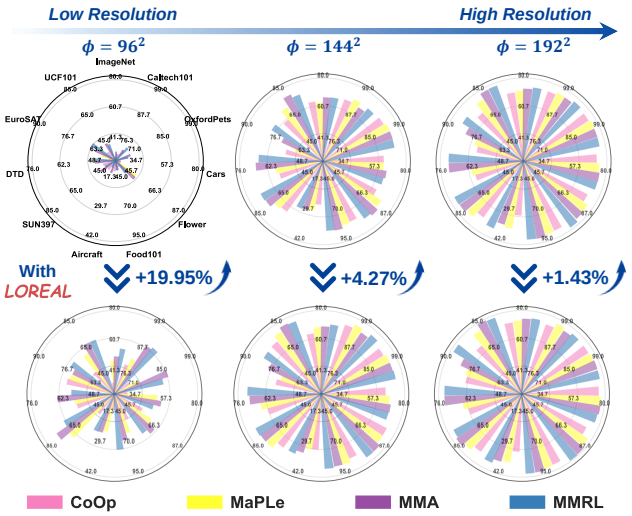

Figure 1. The Harmonic Mean (HM) of four SOTAs (CoOp [83], MaPLe [27], MMA [75], MMRL [15]) w/ or w/o our LOREAL on 11 datasets under varied resolution settings ϕ. LOREAL substantially mitigates their performance degradations under LR settings.

To better adapt VLMs for downstream tasks, an emerging trend focuses on conducting Parameter-Efficient Fine-Tuning (PEFT) [42, 78, 83] over VLMs. Among these techniques, Prompt Learning (PL) [25, 27, 43, 48, 82, 83] has gained significant traction: PL freezes the entire pre-trained VLM and fine-tunes only a small set of learnable prompts inserted into its layers, demonstrating remarkable parameter efficiency and strong compatibility. As a brief revisit, we categorize existing PL methods over VLMs along two dimensions: ❶ ***according to the prompt types***, there are **(a):** uni-modality methods, which employs either pure textual prompts [34, 77, 82, 83] as that for Large Language Models (LLMs), or pure visual prompts [73] for vision encoders; **(b):** bi-modality methods [15, 16, 27, 75], in which prompts are inserted into both encoders with explicit inter-

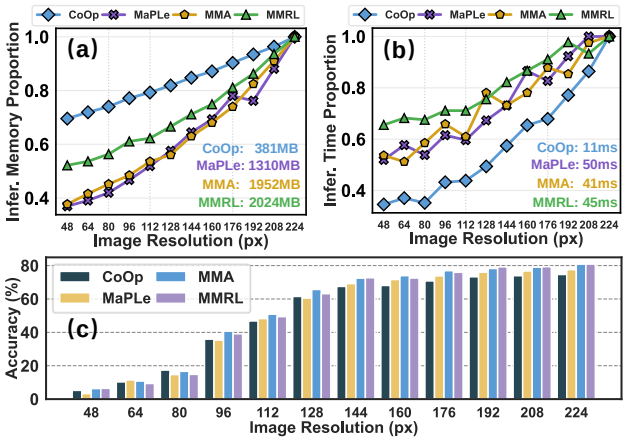

Figure 2. The **(a)** proportion of inference memory respective to that when $\phi = 224^2$ (marked in **Lower-right**), **(b)** proportion of inference time respective to that when $\phi = 224^2$(marked in **Lower-right**), and **(c)** Harmonic Mean (HM) of four existing SOTAs under different resolution settings averaged over 11 datasets.

actions. ❷ ***according to the learning scheme***, there are **(a):** consistency-based methods [15, 47], which employs extra regularization loss to mitigate forgetting or overfitting; **(b):** procedure-based methods [34, 38, 39], which introduces refined tuning procedures like dual-prompt collaboration [34] or prompt-based knowledge distillation [38].

Despite their promising advancements, these methods typically maintain a fixed input resolution (typically, 224×224 for CLIP-ViT-B/16), contradicting real-world applications where edge-deployed VLMs are required to process low-resolution inputs [14, 37, 66, 81] and generate fewer tokens due to constrained computational resources. We name this setting as the Low-Resolution (LR) setting. ***The rationality of LR can be explained from two perspectives***: ❶ the capturing, transmission, and storage overhead required for low-resolution images is nearly orders of magnitude lower than that for standard-resolution images, *thus naturally adapting to the realities of edge devices.* ❷ since the complexity of the visual encoder accounts for a considerable portion of the overall complexity and polynomially increases with token number, reducing visual tokens for low-resolution images significantly cuts off the inference memory, *thereby naturally meeting the requirements of real-time tasks.* Furthermore, we collect the inference cost of four SOTAs under varied resolutions ϕ in Figure 2 **(a,b)**; the results unravel that reducing ϕ can achieve up to 62% memory savings and 64% speed improvement during inference.

Given that LR closely mirrors real-world requirements, how do existing models perform under it? As shown in Figure 2 **(c)**, we observe an overall degradation in existing SOTAs' performance as ϕ decreases; We also find that when ϕ is reduced by approximately half, the model's accuracy shows a dramatic plunge since discriminative visual features are largely blurred, posing unique challenges.

As the first attempt to handle this problem, this paper introduces **LO**w-**RE**solution **A**ttribute-guided prompt **L**earning (LOREAL), a prompt self-distillation scheme designed to equip VLMs with enhanced robustness against resolution shifts. Specifically, inspired by recent progress of automatic attribute-searching [39, 60], we first leverage the Large-Language Models (LLMs) to generate generic attributes which remain salient under varied resolutions. These attributes serve to refine the model's awareness to a portion of resolution-invariant visual features. Then, we construct a contextualized prompt with attributes incorporated: *'A photo of a [CLASS] with S_1 [Attr1] S_2 [Attr2], ...'*, where S are the learnable attribute contents. Furthermore, we introduce cross-modality meta-nets, where each one is dedicated to one attribute and transfers the output visual embeddings into attribute contents. To enable the model to capture resolution-invariant semantics, we further introduce a self-distillation scheme, where two student networks that pin on different input resolutions share the meta-nets to generate attribute contents for each other. These two students further learn from each other via a Low-Level Distillation (LLD) that aligns the prompt attributes and a High-Level Distillation (HLD) that aligns the output logits. Extensive experiments show the superiority of our model (primarily shown in Figure 1). An architecture comparison between our LOREAL and other distillation methods is in Figure 3. Our contribution can be summarized as follows:

- We first study the Low-Resolution (LR) challenges in prompt learning of VLMs, and unveil the vulnerability of existing methods to this realistic setting.
- We propose LOREAL, a prompt self-distillation framework, as the first endeavor towards this problem. LOREAL is guided by resolution-invariant attributes and shared meta-nets to refine the resolution robustness.
- Extensive experiments on various datasets and benchmarks demonstrate the superiority and compatibility of our LOREAL to the realistic LR challenges.

2. Related Work

Prompt Learning of Vision-Language Models. Prompt Learning (PL) originates in natural language processing to steer the Language Model toward producing optimal responses. The pioneer work CoOp first leverages PL to adapt Vision-Language Models (VLMs) with a learnable template combined with classnames. Further advancement focuses on ① strengthening the modality interactions by generating text prompts conditioned on visual priors [73, 77, 82]; ② incorporating multi-modal prompts in intermediate layers [15, 27, 28, 73, 75], typically represented by MaPLe [27] which leverages layer-wise entangled prompts for both encoders; ③ employing distribution alignment to mitigate bias and avoid forgetting in finetuning [15, 28, 73, 84]. For example, PromptSRC [28] and MMRL [28] introduce the

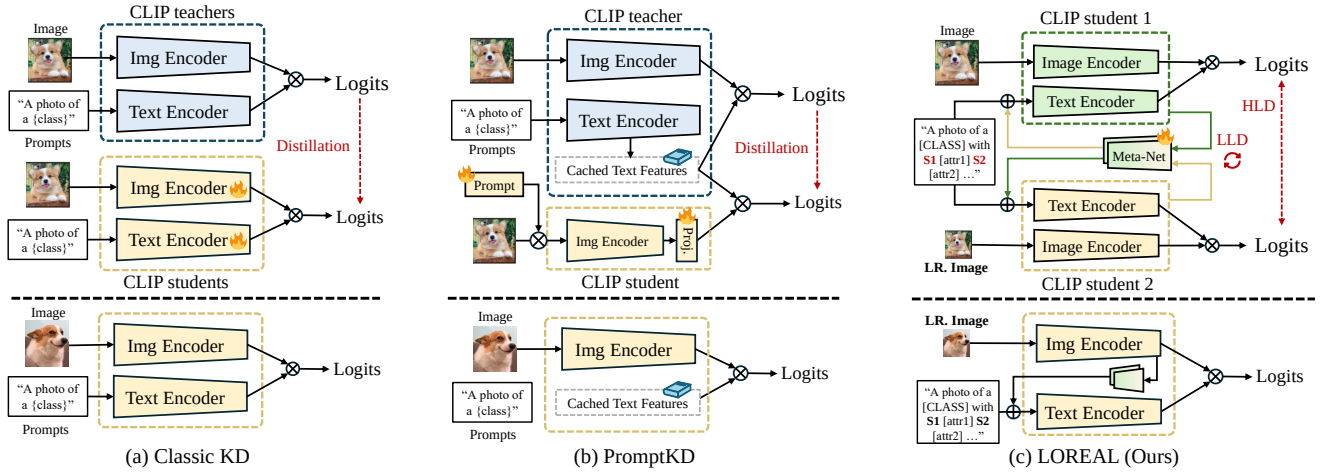

Figure 3. Comparisons of Classic KD [74], PromptKD [38] and our LOREAL. The upper/lower part is the training/inference stage. TE means the Text Encoder. LR means Low-Resolution. **(a)** Classic KD of VLMs, where students are fully-tuned. **(b)** PromptKD, which leverages prompts to learning from teachers. Both (a) and (b) are designed for non-LR inference. **(c)** The proposed LOREAL, a prompt self-distillation scheme to solve the LR challenges. Here, two students are the same models but fed with different inputs. LOREAL leverages fine-grained attribute guidance and simultaneously distills from two levels to boost the model's robustness to data resolutions.

prediction from zero-shot VLM to guide each training step. Recently, Skip-Tuning [71] caches intermediate visual features and drops gradients of unrelated classes for more efficient finetuning. PL for VLMs has also shown wide applications in knowledge distillation [38, 70], federated learning [52], and interpretability analysis [10, 11]. *Notably, although both Skip-Tuning [71] and our work aim to explore lightweighting, it fails to discuss inference memory savings induced by LR settings during inference.*

Knowledge Distillation (KD). Knowledge Distillation (KD) [21] aims to learn a lightweight student network under the guidance of the bulky teacher, thereby achieving comparable performance with greater efficiency and empowering the latent-sensitive applications. Based on the type of distilling sources, KD can be categorized into Logit Distillation (LD) [26, 30, 45, 80](i.e, only to align the final logits), and Feature Distillation (FD) [4, 5, 17, 49, 55, 61–63, 76] which employ extra intermediate features for alignment. Self-Distillation (SD) is a special variant of KD, where the students act as the teachers of themselves and learn from each other. Some recent KD methods [29, 32, 38, 74] turn their gaze to the VLMs, for example, Clip-KD [74] systematically studies the adaptability of distillation techniques to VLMs; PromptKD [38] only leverages prompts to adapt knowledge from teachers. COSMOS [29] introduces a cross-modality self-distillation for VLM pretraining.

Low-Resolution Recognition (LRR). Low-Resolution Recognition [69] (LRR) aims at maintaining the model's performance when only Low-Resolution (LR) images are available in inference. Early LR methods are designed for face recognition [57, 68, 85]; Subsequent works [13, 51, 65] have explored the integration of Knowledge Distillation (KD) to obtain compact yet high-performing models suitable for LR inference: for example, PixelDistillation [14] introduces a cost-flexible heterogeneous framework to distill small convolution models for fast inference; When it comes to LRR on large pre-trained models, existing studies [1, 8, 33, 41, 59] mostly focus on designing position embeddings that enables flexible input resolution for ViTs; for example, ResFormer [59] proposes a multi-resolution training and global-local embedding strategy to learn resolution-invariant semantics; MSPE [41] designs a plug-and-play hierarchical position embedding to adapt different patch sizes. *Despite these advancements, addressing the LR challenge through prompt learning for VLMs remains unexplored.*

3. Methodology

3.1. Preliminaries

Prompt Learning (PL) of Vision-Language (VLMs). Consider a typical VLM, i.e. CLIP; it consists of a visual encoder $\mathcal{E}_v$ and text encoder $\mathcal{E}_t$ where each comprises L layers, i.e $\mathcal{E}_v = \{\mathcal{E}_{v,i}\}_{i=1}^{L}$, $\mathcal{E}_t = \{\mathcal{E}_{t,i}\}_{i=1}^{L}$; For image branch $\mathcal{V}$, denote $\mathcal{B}$ as an image batch; for one image x from $\mathcal{B}$, a patch embedding module `emb` first project it into M-length visual embeddings $\boldsymbol{v}_0$ (m is the fixed token number), then, the process of each layer $\mathcal{E}_{v,i}$ can be formalized as:

$$[\texttt{cls}_i, \boldsymbol{v}_i] = \mathcal{E}_{v,i}([\texttt{cls}_{i-1}, \boldsymbol{v}_{i-1}]), \quad i \in \{1, 2, ..., L\}, \quad (1)$$

where $\texttt{cls}_i \in \mathbb{R}^{D_v}$ means the class embeddings of layer i. D_v means the visual hidden dimension. We apply another projector M_v over $\texttt{cls}_L$ to generate the final visual embeddings $\boldsymbol{f}_v$, i.e: $\boldsymbol{f}_v = \texttt{cls}_L M_v$. For the text branch, we fill all classnames into the prompt P that is initialized from either

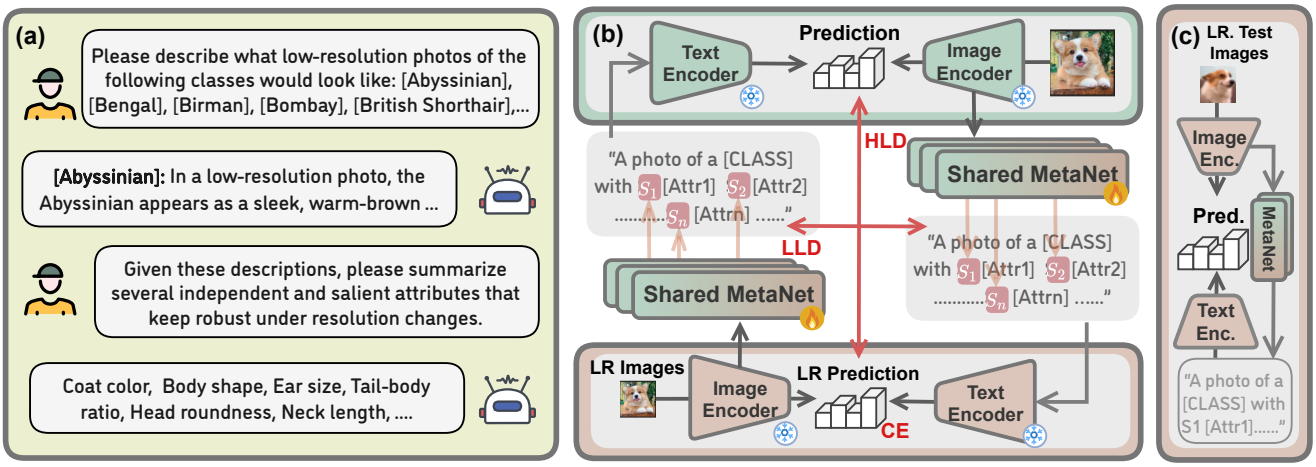

Figure 4. Our LOREAL framework. **(a):** We leverage the LLM to generate several resolution-invariant attributes. **(b):** Self-distillation framework. We utilize the visual embeddings to fill the prompt attributes via meta-nets, then leverage Low-Level Distillation (LLD) and High-Level Distillation (HLD) for self-distillation. Only the meta-nets are learnable, and the parameters of two illustrated meta-nets are shared. LR represents Low-Resolution. **(c):** Inference stage. The model takes LR images and contextualizes prompts with the meta-nets.

pre-defined (i.e "A photo of a [CLASS]") [82, 83] or random values $\mathcal{P}^0$ to generate the text input. Denote $\mathcal{C}$ as all classes, and C as the number of classes. The text input is then tokenized and converted into text embeddings $\boldsymbol{t}_0$. Finally, $\boldsymbol{t}_0$, together with a Begin-Of-Text (BoT) token `bot` and End-Of-Text (EOT) token `eot`, is fed into each $\mathcal{E}_{t,i}$:

$$[\texttt{bot}_i, \boldsymbol{t}_i, \texttt{eot}_i] = \mathcal{E}_{t,i}([\texttt{bot}_{i-1}, \boldsymbol{t}_{i-1}, \texttt{eot}_{i-1}]). \quad (2)$$

We pick up the EOT token $\texttt{eot}_L$ from the last layer L, and project it to the modality-shared space with projector M_t, i.e $\boldsymbol{f}_t = \texttt{eot}_L M_t$. We may also denote $\boldsymbol{f}_t(P)$ as the textual outputs conditioned on prompt P. For the classification stage, we calculate the cosine similarity $\texttt{Sim}(\cdot)$ between $\boldsymbol{f}_v$ and each entry in $\boldsymbol{f}_t \in \mathbb{R}^{C\times d}$ as the raw prediction, i.e:

$$\hat{\boldsymbol{y}}_c = \frac{\exp(\texttt{Sim}(\boldsymbol{f}_v, \boldsymbol{f}_{t,c})/\tau)}{\sum_{c'} \exp(\texttt{Sim}(\boldsymbol{f}_v, \boldsymbol{f}_{t,c'})/\tau)}, \quad (3)$$

where $\boldsymbol{f}_{t,c}$ is the c-th entry of $\boldsymbol{f}_t$. τ is the temperature. We also denote $\hat{\boldsymbol{y}}(\boldsymbol{f}_v, \boldsymbol{f}_{t,c})$ as $\hat{\boldsymbol{y}}$ generated from $\boldsymbol{f}_v$ $\boldsymbol{f}_{t,c}$.

Low-Resolution Recognition (LRR) with VLMs. As we introduced above, LRR generally represents that models are directly fed with low-resolution inputs and generate fewer tokens accordingly. For LRR on VLMs, considering that the length J of visual position embedding $\mathcal{P} \in \mathbb{R}^{J\times D_v}$ is fixed, we follow baseline implementations [23] and reshape it to dynamically meet the shape requirements (ϕ, ϕ) (assume ϕ is the resolution setting of the input image):

$$\mathcal{P} = \texttt{Cat}(\mathcal{P}[:1], \texttt{Intp}(\mathcal{P}[1:], \phi, \phi)), \quad (4)$$

where $[a:b]$ means slicing, and $[:1]$ separates the location of cls-token apart. `Cat` means concatenation, `Intp` is the interpolation function (bicubic mode by default).

3.2. Generating Resolution-invariant Attributes

While the Low-Resolution (LR) settings may induce considerate vague in class distinctions, a natural idea is diving into attribute-level semantics in the image and focusing on those that tend to remain robust under varied resolution. Inspired by recent progress in attribute searching [6, 39, 79], we design an LLM-based pipeline with explicit Chain-of-Thought (CoT) to search resolution-invariant attributes.

Concretely, as shown in Fig. 4, we first instruct the LLM to generate extensive descriptions of how LR photos would typically appear for all [CLASSES]. Leveraging these descriptions as contextual background, we then instruct the LLM to carefully summarize a set of generic and invariant attributes that remain perceptible across varying resolutions. This process favors macro-level attributes (e.g., `coat-color` and `body-shape` for OxfordPets [50]; `window` and `roofline` for StandfordCars [31]) while filtering out more subtle ones (e.g., `texture`, `eye-color` for OxfordPets; `logo` or `badge` for StandfordCars). Note that both macro and subtle attributes are considered generic in standard non-LR settings [39], and *our key insight is to identify and utilize the subset of these generic features that demonstrably retain their saliency under LR.*

Subsequently, with the LLM-generated raw attributes $\{A_k\}_{k=1}^K$, we build contextualized prompts $\boldsymbol{p} = \{\boldsymbol{p}_c\}_{c=1}^C$ by filling them into the attribute slot [A.] as the following to actively refine model's awareness:

$$\text{A photo of a [CLS] with } \underline{\text{S}_1}\ [\text{A}_1]\ \underline{\text{S}_2}\ [\text{A}_2] \cdots \underline{\text{S}_K}\ [\text{A}_K],$$

where the learnable tokens are underlined, $S_i \in \mathbb{R}^{M\times D_t}$ represents the learnable contents of attributes, and M is the learnable token number in each attribute. Notably, the order of attributes is not taken into specific consideration as it yields minor effects on the performance. Details of the

attributes and attribute order are in **Sup. Mat. C.**. To incorporate visual semantics into the prompt, we contextualize S_i using cross-modality meta-nets in the following.

3.3. Cross-modality Meta-Nets

In correspondence with the attributes, we introduce K separate meta-nets, denoted as $\{\mathcal{M}_k\}_{k=1}^{K}$ to dynamically contextualize prompts with visual semantics. Specifically, the meta-nets are built upon a LoRA-like architecture, and the output visual embeddings $\boldsymbol{f}_v$ are fed into each meta-net to generate their respective $\{S_k\}_{k=1}^{K}$:

$$S_k = M_k(\boldsymbol{f}_v) = W_{\uparrow,k}(W_{\downarrow,k}(\boldsymbol{f}_v)), \tag{5}$$

where $W_{\downarrow,k} \in \mathbb{R}^{D_v \times D_s}$, $W_{\uparrow,k} \in \mathbb{R}^{D_s \times D_t}$ are LoRA projection matrices, D_s represents LoRA intermediate dimension. We denote $\boldsymbol{p}(S_k)$ as prompts conditioned on S_k.

3.4. Attribute-driven Self-Distillation

Despite the straightforward idea of finetuning attribute-contextualized prompts to focus on invariant properties, the model's exposure solely to LR images during inference necessitates a more sophisticated approach. We therefore introduce a self-distillation method that guides the model in reconciling semantics across different resolutions.

Our self-distillation incorporates two student VLMs $\{\mathcal{E}_t^{\alpha}, \mathcal{E}_v^{\alpha}\}$ $\{\mathcal{E}_t^{\beta}, \mathcal{E}_v^{\beta}\}$ with all backbone frozen and only the meta-nets learnable. One student $\{\mathcal{E}_t^{\alpha}, \mathcal{E}_v^{\alpha}\}$ is fed with standard image $\boldsymbol{x}$ and the other one $\{\mathcal{E}_t^{\beta}, \mathcal{E}_v^{\beta}\}$ is fed with low-resolution image $\boldsymbol{x}'$ (when combined with other methods besides zero-shot CLIP, the students are pre-tuned with images of their separate resolution). Recall the cross-modality meta-nets in the last subsection, we bridge these meta-nets between the cross-modal encoders of different students:

$$\begin{aligned} \boldsymbol{f}_v^{\alpha} &= \mathcal{E}_v^{\alpha}(\boldsymbol{x}); \;\; \boldsymbol{f}_t^{\beta} = \mathcal{E}_t^{\beta}(\boldsymbol{p}(S(\boldsymbol{f}_v^{\alpha}))) \\ \boldsymbol{f}_v^{\beta} &= \mathcal{E}_v^{\beta}(\boldsymbol{x}'); \;\; \boldsymbol{f}_t^{\alpha} = \mathcal{E}_t^{\alpha}(\boldsymbol{p}(S(\boldsymbol{f}_v^{\beta}))) \end{aligned} \tag{6}$$

Then, to enable inputs of different resolution to generate similar attribute contexts, we introduce Low-Level Distillation (LLD), where we employ the contrastive learning between the prompt contexts $S(\boldsymbol{f}_v^{\alpha})$ and $S(\boldsymbol{f}_v^{\beta}) \in \mathbb{R}^{K \times D_t}$:

$$\mathcal{L}_{\text{LLD}} = -\sum_{k=1}^{K} \log \frac{\exp(\texttt{Sim}(S_k(\boldsymbol{f}_v^{\alpha}), S_k(\boldsymbol{f}_v^{\beta}))/\tau)}{\sum_{k'}^{K} \exp(\texttt{Sim}(S_k(\boldsymbol{f}_v^{\alpha}), S_{k'}(\boldsymbol{f}_v^{\beta}))/\tau)} \tag{7}$$

Meanwhile, we generate the predictions $\hat{\boldsymbol{y}}^{\alpha}$ $\hat{\boldsymbol{y}}^{\beta}$ for both students, i.e: $\hat{\boldsymbol{y}}^{\alpha} = \hat{\boldsymbol{y}}(\boldsymbol{f}_v^{\alpha}, \boldsymbol{f}_{t,c}^{\alpha})$, $\hat{\boldsymbol{y}}^{\beta} = \hat{\boldsymbol{y}}(\boldsymbol{f}_v^{\beta}, \boldsymbol{f}_{t,c}^{\beta})$. These separate predictions are further aligned with Kullback-Leibler (KL)-Divergence as our High-Level Distillation (HLD):

$$\mathcal{L}_{\text{HLD}} = \sum_{c=1}^{C} (\hat{y}_c^{\alpha} \log \hat{y}_c^{\alpha} - \hat{y}_c^{\alpha} \log \hat{y}_c^{\beta}) \tag{8}$$

These designs enable the textual prompts to be dynamically contextualized by visual features from another resolution, thereby progressively aligning them in the multi-modal manifold space. Finally, our two distillation losses are harmonized with the task loss $\mathcal{L}_{\text{CE}} = -\sum_c y_c \log \hat{y}_c^{\beta}$ via two coefficients λ_1 λ_2, i.e, $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{HLD}} + \lambda_2 \cdot {}^1\!/_K \cdot \mathcal{L}_{\text{LLD}}$. In the inference stage, we use the meta-nets to transfer visual embeddings into attribute contents, then calculate the predictions leveraging the attribute-structured prompt.

4. Experiments

4.1. Settings

Datasets. We evaluate our LOREAL on three benchmarks: ❶ ***Low-Resolution Base-to-New (LR-B2N) benchmark*****,** which integrates LR with Base-to-New tasks to further evaluate the models' generalization robustness. The datasets are first split into disjoint base and new classes, then models are trained on base classes and evaluated on both base and new classes. The test-sets are down-sampled to the resolution $\phi \in \{96^2, 144^2, 192^2\}$, and datasets employed for this task are: ImageNet [9], OxfordPets [50], SUN397 [72], EuroSAT [18], Caltech101 [12], StanfordCars [31], DTD [7], UCF101 [58], Flowers102 [46], Food101 [2] and FGVC-Aircraft [44]. ❷ ***Low-Resolution Cross-dataset Evaluation (LR-CE) benchmark*****,** which integrates LR with a cross-dataset task to evaluate models' robustness to LR test-sets of new datasets. We train the models on ImageNet and evaluate them on the above ten datasets down-sampled to ϕ. ❸ ***Low-Resolution Domain Generalization (LR-DG) benchmark*****,** which integrates LR with domain-generalization to evaluate models' robustness to LR test-sets of new domains. We train models on ImageNet and evaluate them on four variants (ImageNet-A [20], ImageNet-V2 [54], ImageNet-R [19], and ImageNet-S [64]) down-sampled to ϕ. All experiments adopt a 16-shot setting, i.e., 16 samples per class.
Implementation Details. Following previous studies, we adopt pretrained CLIP-VIT-B/16 as the backbone. $\tau = 1$. We use the SGD optimizer with a learning rate of 0.002. According to the ablation studies, the intermediate hidden dimension D_s of meta-nets is 32; $M = 2$,$\lambda_1 = 1$,$\lambda_2 = 2$. The employed LLM is GPT-4o [22], and we generate 5 attributes for each class. Notably, when combining our model with methods that cannot perform separate visual or text embeddings [15, 27], we proactively store their visual embeddings offline before distillation and leverage the stored embeddings to fill in the prompts. The training epoch and ϕ are unique for each method. More implementation specifics and experiment results are in **Sup. Mat. A / B.**

4.2. Experiment Results

Results on the LR-B2N Benchmark. As demonstrated in Table 1, integrating our LOREAL with four existing SOTAs under different ϕ settings consistently yields a substantial performance gains. Specifically, we have three key find-

Table 1. Results on the LR-B2N Benchmark over three resolution settings $\phi \in \{96^2, 144^2, 192^2\}$ w/ or w/o our LOREAL. HM means the harmonic mean. η is the ratio of patch tokens relative to the 224^2 input. All results are averaged over 3 runs. **Bold** marks the best results.

$\phi : (96 \times 96), \eta \approx 18.37\%$

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	38.40	33.74	35.54	24.92	23.30	24.08	70.95	65.16	67.93	58.72	57.45	58.08	17.37	14.19	15.62	43.97	21.42	28.81
+LOREAL	**54.81**	**42.77**	**47.48**	**35.96**	**40.84**	**38.24**	**83.21**	**72.60**	**77.54**	**66.92**	**64.20**	**65.53**	**27.61**	**18.16**	**21.91**	**76.73**	**35.04**	**48.11**
MaPLe	37.17	33.71	34.85	21.24	26.74	23.67	69.34	62.88	65.95	53.32	66.44	59.16	12.42	12.82	12.62	50.90	37.52	43.20
+LOREAL	**65.02**	**52.18**	**57.25**	**46.07**	**42.76**	**44.35**	**85.48**	**76.14**	**80.54**	**69.17**	**70.43**	**69.79**	**52.22**	**41.89**	**46.49**	**86.32**	**43.40**	**57.76**
MMA	41.55	39.91	40.50	30.99	25.88	28.21	73.98	66.92	70.27	60.77	69.30	64.76	15.57	16.17	15.86	51.66	38.58	44.17
+LOREAL	**67.81**	**59.74**	**63.14**	**51.80**	**44.29**	**47.75**	**92.32**	**84.61**	**88.30**	**83.47**	**87.47**	**85.42**	**54.10**	**46.01**	**49.73**	**84.90**	**51.49**	**64.10**
MMRL	41.64	36.92	38.90	29.57	23.44	26.15	75.98	68.36	71.97	58.75	68.67	63.32	14.02	11.54	12.66	40.36	33.83	36.81
+LOREAL	**66.97**	**57.80**	**61.71**	**52.54**	**45.34**	**48.68**	**93.29**	**85.26**	**89.09**	**81.61**	**82.27**	**81.94**	**46.53**	**42.63**	**44.49**	**79.36**	**48.72**	**60.38**

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	23.03	30.37	26.20	10.20	9.30	9.73	35.24	30.86	32.90	47.45	36.47	41.24	54.36	52.05	53.18	36.14	30.56	33.12
+LOREAL	**50.92**	**38.96**	**44.14**	**13.87**	**10.14**	**11.72**	**52.83**	**36.03**	**42.84**	**62.85**	**47.45**	**54.07**	**79.71**	**67.41**	**73.05**	**52.33**	**39.64**	**45.11**
MaPLe	19.67	22.83	21.13	7.02	7.86	7.42	26.33	28.03	27.15	50.20	29.71	37.33	64.60	51.13	57.08	33.87	24.82	28.65
+LOREAL	**56.91**	**55.35**	**56.12**	**23.65**	**19.98**	**21.66**	**68.60**	**65.71**	**67.12**	**73.84**	**43.24**	**54.54**	**82.52**	**64.23**	**72.24**	**70.48**	**50.89**	**59.10**
MMA	27.36	30.29	28.75	10.38	11.10	10.73	37.57	38.27	37.92	49.88	40.94	44.97	57.48	64.56	60.81	41.37	36.99	39.06
+LOREAL	**61.90**	**61.55**	**61.72**	**22.63**	**22.80**	**22.71**	**74.50**	**75.39**	**74.94**	**72.80**	**54.83**	**62.55**	**74.90**	**68.92**	**71.79**	**72.54**	**59.76**	**65.53**
MMRL	30.47	27.06	28.66	7.56	9.06	8.24	34.49	32.66	33.55	57.25	37.44	45.27	70.05	58.23	63.60	39.55	35.88	37.63
+LOREAL	**70.46**	**69.43**	**69.94**	**22.93**	**22.62**	**22.77**	**68.55**	**63.46**	**65.91**	**69.56**	**46.74**	**55.91**	**81.74**	**68.08**	**74.29**	**70.10**	**61.28**	**65.39**

$\phi : (144 \times 144), \eta \approx 41.5\%$

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	74.21	61.88	67.00	67.71	59.10	63.11	94.13	89.25	91.63	91.44	93.06	92.24	65.39	57.04	60.93	89.36	56.74	69.41
+LOREAL	**77.30**	**67.37**	**71.73**	**69.22**	**63.92**	**66.46**	**97.32**	**91.66**	**94.41**	**92.77**	**93.96**	**93.36**	**67.89**	**59.98**	**63.69**	**94.40**	**68.73**	**79.55**
MaPLe	74.59	64.54	68.80	67.79	57.55	62.25	94.25	91.48	92.84	92.40	**94.74**	93.56	63.39	59.10	61.17	93.45	64.96	76.64
+LOREAL	**79.46**	**69.32**	**73.72**	**72.79**	**63.26**	**67.69**	**97.81**	**94.21**	**95.98**	**94.26**	94.69	**94.47**	**70.89**	**63.53**	**67.01**	**95.25**	**69.03**	**80.05**
MMA	75.21	69.90	72.30	70.13	62.90	66.32	94.77	87.81	91.16	91.23	95.64	93.38	65.39	63.21	64.28	90.88	69.43	78.72
+LOREAL	**78.68**	**73.08**	**75.60**	**72.01**	**64.78**	**68.20**	**97.35**	**91.59**	**94.38**	**92.93**	**95.97**	**94.43**	**70.21**	**67.18**	**68.66**	**95.54**	**72.36**	**82.35**
MMRL	76.92	68.94	72.43	70.67	62.93	66.58	95.35	90.67	92.95	92.24	94.07	93.15	63.92	60.58	62.21	93.45	65.60	77.09
+LOREAL	**80.93**	**73.03**	**76.56**	**73.74**	**66.38**	**69.87**	**97.81**	**94.23**	**95.99**	**94.21**	**95.53**	**94.87**	**69.74**	**66.28**	**67.97**	**97.06**	**71.99**	**82.67**

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	78.22	79.14	78.68	26.17	22.02	23.92	73.90	64.17	68.69	72.80	46.86	57.02	77.29	60.67	67.98	79.89	52.60	63.43
+LOREAL	**82.25**	**82.57**	**82.41**	**31.75**	**27.17**	**29.28**	**76.59**	**67.66**	**71.85**	**76.39**	**52.12**	**61.96**	**81.24**	**64.97**	**72.20**	**80.46**	**68.35**	**73.91**
MaPLe	79.86	79.34	79.60	27.55	26.15	26.83	72.77	68.84	70.75	73.73	47.71	57.93	76.79	55.08	64.15	78.49	64.95	71.08
+LOREAL	**85.73**	**86.68**	**86.20**	**31.27**	**28.07**	**29.58**	**79.21**	**74.88**	**76.98**	**79.51**	**54.83**	**64.90**	**84.24**	**60.41**	**70.36**	**83.09**	**72.96**	**77.70**
MMA	81.91	83.40	82.65	28.03	26.81	27.41	75.24	73.22	74.22	76.16	60.87	67.66	72.26	70.92	71.58	81.28	74.74	77.87
+LOREAL	**85.27**	**86.68**	**85.97**	**34.33**	**29.87**	**31.95**	**79.03**	**75.77**	**77.37**	**78.59**	**62.92**	**69.89**	**76.93**	**78.54**	**77.73**	**83.25**	**78.21**	**80.65**
MMRL	81.83	82.27	82.05	27.13	28.55	27.82	75.51	73.80	74.65	79.28	57.00	66.32	86.40	71.69	78.36	80.35	71.23	75.52
+LOREAL	**85.65**	**87.15**	**86.39**	**35.65**	**31.75**	**33.59**	**79.02**	**76.94**	**77.97**	**82.05**	**62.44**	**70.91**	**90.71**	**74.97**	**82.09**	**84.61**	**75.66**	**79.89**

$\phi : (192 \times 192), \eta \approx 73.47\%$

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	81.36	66.68	73.29	75.20	65.24	69.87	98.19	91.48	94.72	94.42	96.14	95.27	76.06	66.50	70.96	97.72	64.54	77.74
+LOREAL	**82.55**	**68.60**	**74.93**	**77.10**	**66.18**	**71.22**	**98.32**	**91.70**	**94.89**	**95.84**	**96.42**	**96.13**	**78.81**	**69.52**	**73.87**	**97.15**	**65.62**	**78.33**
MaPLe	82.19	70.75	76.04	75.41	66.28	70.55	98.05	93.56	95.75	95.25	97.04	96.14	74.71	71.30	72.97	97.34	70.99	82.10
+LOREAL	**83.32**	**72.60**	**77.59**	**76.95**	**67.98**	**72.19**	**98.46**	**94.76**	**96.57**	**94.95**	**97.48**	**96.20**	**76.39**	**74.53**	**75.45**	**97.87**	**72.44**	**83.26**
MMA	81.28	75.47	78.27	76.25	68.45	72.14	98.39	92.90	95.57	94.36	97.54	95.92	75.89	72.25	74.03	**97.63**	71.56	82.59
+LOREAL	**82.61**	**76.98**	**79.70**	**78.32**	**69.28**	**73.52**	**98.44**	**93.01**	**95.65**	**95.52**	**98.06**	**96.77**	**77.41**	**75.12**	**76.25**	97.24	**76.24**	**85.47**
MMRL	83.32	75.55	79.25	76.18	69.89	72.90	98.84	94.32	96.53	95.22	96.39	95.80	76.26	71.75	73.94	98.39	74.18	84.59
+LOREAL	**84.39**	**76.67**	**80.34**	**77.60**	**70.85**	**74.07**	**99.10**	**94.43**	**96.71**	**96.01**	**97.09**	**96.55**	**76.89**	**73.00**	**74.89**	**98.48**	**76.11**	**85.86**

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CoOp	87.99	87.15	87.57	34.99	24.30	28.68	80.29	68.20	73.75	78.01	47.83	59.30	87.19	62.13	72.56	84.95	**59.98**	70.31
+LOREAL	**88.25**	**88.67**	**88.46**	**36.61**	**25.91**	**30.34**	**81.28**	**71.01**	**75.80**	**79.94**	**50.75**	**62.09**	**89.71**	**68.26**	**77.53**	**85.04**	59.56	**70.06**
MaPLe	88.50	88.49	88.49	38.06	**32.34**	34.97	80.41	75.47	77.86	78.82	49.20	60.58	92.40	61.85	74.10	85.11	71.77	77.87
+LOREAL	**88.89**	**89.83**	**89.36**	**39.44**	32.13	**35.41**	**82.01**	**76.26**	**79.03**	**81.02**	**53.26**	**64.27**	**94.19**	**67.05**	**78.34**	**86.30**	**72.85**	**79.01**
MMA	89.11	90.42	89.76	36.55	31.73	33.97	81.58	78.63	80.08	80.67	64.61	71.75	78.62	82.51	80.52	85.06	79.61	82.24
+LOREAL	**91.18**	**90.74**	**90.96**	**39.26**	**33.53**	**36.17**	**82.24**	**80.61**	**81.42**	**82.06**	**65.46**	**72.83**	**80.21**	**83.97**	**82.05**	**86.83**	**80.77**	**83.69**
MMRL	89.41	90.79	90.09	37.64	36.95	37.29	81.88	**78.83**	**80.33**	82.99	62.92	71.57	93.71	76.46	84.21	85.99	78.58	82.12
+LOREAL	**90.75**	**91.90**	**91.32**	**38.78**	**37.65**	**38.21**	**81.59**	**78.93**	80.24	**85.06**	**64.01**	**73.05**	**97.38**	**80.38**	**88.07**	**86.61**	**79.02**	**82.64**

ings: ① As ϕ decreases, all methods suffer from a gradually accelerated performance degradation. For CoOp, the degradation of $144^2 \rightarrow 96^2$ (-31.46% HM) is about five times greater than that of $192^2 \rightarrow 144^2$ (-6.29% HM). This trend remains approximately consistent across other methods. This accelerated decline can be attributed to the progressively severe loss of semantic information as ϕ is reduced. ② Our LOREAL demonstrates significant improvements when combined with existing methods, with its effectiveness becoming more pronounced under more challenging (lower ϕ) settings. In MMA, our method brings a surging improvement of 22.64%, 3.30%, and 1.43% under three settings. Notably, for nearly standard ϕ settings (ϕ=192^2), our method essentially functions as cross-modal attribute prompting and also delivers considerable gains for all methods by an average of 1.4%. ③ We also find our

Table 2. Results on the LR-CE Benchmark over three resolution settings $\phi \in \{96^2, 144^2, 192^2\}$ w/ or w/o our LOREAL. η is the ratio of patch tokens relative to the 224^2 input. All results are averaged over 3 runs. **Bold** marks the best results.

Method	ImageNet (Source)			Caltech101			OxfordPets			StanfordCars			Flowers102			Food101		
	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2
CoOp	22.70	57.97	67.64	47.63	93.73	94.49	48.74	91.33	93.08	13.73	58.23	66.20	24.12	63.87	71.00	28.86	80.16	87.75
+LOREAL	**37.65**	**63.28**	**70.92**	**62.41**	**96.90**	**95.51**	**55.13**	**93.22**	**93.88**	**22.40**	**61.52**	**69.64**	**42.06**	**74.51**	**78.30**	**45.23**	**83.87**	**88.08**
MaPLe	20.28	57.30	68.92	52.82	93.07	78.24	49.69	93.92	95.40	14.81	59.52	64.44	30.94	66.77	72.59	33.09	77.26	88.12
+LOREAL	**43.79**	**64.76**	**71.60**	**79.59**	**94.56**	**95.38**	**60.37**	**94.09**	**95.33**	**39.95**	**67.87**	**68.25**	**52.28**	**72.37**	**83.17**	**44.74**	**87.46**	**90.05**
MMA	24.15	58.33	70.89	47.55	93.69	92.78	58.79	94.36	95.26	10.82	61.60	64.79	35.35	60.78	78.68	35.53	73.81	88.57
+LOREAL	**46.53**	**65.91**	**73.26**	**79.38**	**96.68**	**94.80**	**67.58**	**95.20**	**95.65**	**36.67**	**69.83**	**71.60**	**55.84**	**74.89**	**79.12**	**48.00**	**80.83**	**90.02**
MMRL	24.35	62.06	71.87	69.98	93.44	94.89	60.75	94.92	**95.87**	12.11	63.36	70.79	42.05	69.36	**78.94**	37.35	80.25	89.96
+LOREAL	**47.50**	**68.04**	**73.71**	**78.86**	**94.94**	**96.36**	**70.10**	**96.03**	95.60	**37.58**	**65.64**	**73.84**	**58.28**	**75.18**	78.73	**50.65**	**85.95**	**90.48**

Method	FGVCAircraft			SUN397			DTD			EuroSAT			UCF101			Target Avg		
	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2
CoOp	8.25	13.42	22.02	31.50	62.30	70.87	30.74	41.97	51.08	50.23	63.22	63.95	33.40	62.77	71.61	31.72	63.10	69.20
+LOREAL	**13.96**	**20.59**	**28.15**	**46.14**	**69.35**	**76.63**	**54.55**	**61.80**	**66.02**	**69.15**	**67.15**	**77.89**	**46.74**	**68.89**	**72.59**	**45.77**	**69.78**	**74.67**
MaPLe	10.59	13.72	25.54	44.21	64.05	70.95	33.46	47.28	61.73	54.36	62.46	68.49	36.71	62.13	70.61	36.07	64.02	69.61
+LOREAL	**14.26**	**23.11**	**29.27**	**48.02**	**70.76**	**78.03**	**60.58**	**64.00**	**64.25**	**68.49**	**75.16**	**88.72**	**59.38**	**76.72**	**79.50**	**52.77**	**72.61**	**77.19**
MMA	11.52	16.95	26.51	44.65	71.89	74.41	34.32	46.86	62.69	55.97	63.26	65.05	38.23	70.04	**74.04**	37.27	65.32	72.28
+LOREAL	**15.04**	**24.84**	**32.59**	**50.51**	**73.15**	**78.08**	**64.35**	**62.17**	**65.10**	**75.38**	**79.49**	**87.60**	**53.80**	**72.85**	73.55	**54.66**	**72.99**	**76.81**
MMRL	11.82	23.55	32.54	48.22	71.40	76.99	37.85	54.83	54.09	54.67	72.74	81.23	32.94	73.23	77.72	40.77	69.71	75.30
+LOREAL	**17.88**	**29.81**	**37.97**	**52.34**	**74.10**	**78.35**	**62.15**	**67.49**	**72.68**	**70.77**	**76.69**	**89.31**	**52.17**	**75.34**	**78.04**	**55.08**	**74.12**	**79.14**

Table 3. Results on the LR-DG Benchmark over three resolution settings $\phi \in \{96^2, 144^2, 192^2\}$ w/ or w/o our LOREAL. η is the ratio of patch tokens relative to the 224^2 input. All results are averaged over 3 runs. **Bold** marks the best results.

Method	ImageNet (Source)			ImageNet-V2			ImageNet-S			ImageNet-A			ImageNet-R			Target Avg.		
	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2	96^2	144^2	192^2
CoOp	22.70	57.97	67.64	17.68	53.38	60.12	10.27	36.74	42.48	10.51	22.29	38.71	21.63	63.06	71.32	15.02	43.87	53.16
+LOREAL	**37.65**	**63.28**	**70.92**	**31.32**	**56.55**	**63.84**	**18.84**	**38.47**	**42.88**	**17.29**	**28.51**	**43.24**	**25.63**	**69.74**	**73.77**	**23.27**	**48.32**	**55.93**
MaPLe	20.28	57.30	68.92	12.08	50.61	61.28	9.58	35.66	42.15	7.29	20.84	37.18	13.29	59.13	69.75	10.56	41.56	52.59
+LOREAL	**43.79**	**64.76**	**71.60**	**36.83**	**57.25**	**64.16**	**21.02**	**40.46**	**42.93**	**15.91**	**28.97**	**39.19**	**33.68**	**66.98**	**72.85**	**26.86**	**48.42**	**54.78**
MMA	24.15	58.33	70.89	17.98	53.87	61.81	10.24	37.21	43.46	9.53	23.32	38.25	20.11	60.37	69.21	14.47	43.69	53.18
+LOREAL	**46.53**	**65.91**	**73.26**	**35.10**	**57.00**	**62.59**	**19.98**	**38.31**	**43.98**	**14.22**	**27.69**	**39.06**	**32.77**	**66.41**	**71.00**	**25.52**	**47.35**	**54.16**
MMRL	24.35	62.06	71.87	18.10	55.29	63.61	10.84	39.57	46.19	9.80	28.53	44.92	22.94	65.98	74.75	15.42	47.34	57.37
+LOREAL	**47.50**	**68.04**	**73.71**	**42.56**	**58.65**	**65.50**	**21.95**	**41.59**	**46.55**	**12.89**	**32.55**	**45.95**	**45.15**	**69.32**	**75.89**	**30.64**	**50.53**	**58.47**

LOREAL shows more pronounced effectiveness for multi-modal approaches like MaPLe, MMA and MMRL (ϕ=96^2: an average of $+22.81\%$ HM) compared to CoOp (ϕ=96^2: 11.3% HM), possibly attributed to their cross-modal design which facilitates the meta-network's convergence to local minima during resolution-aware semantic refinement.

Results on the LR-CE Benchmark. Table 2 presents the results of models (trained on the standard ImageNet training set) being evaluated on LR test-sets of ImageNet and 10 additional datasets. First, when augmented with LOREAL, all methods achieve average improvements of (21.0%, 6.58%, and 2.54%) on the source ImageNet; Furthermore, LOREAL substantially enhances performance on the diverse set of target datasets, especially under low ϕ settings, underscoring its remarkable effectiveness to cross-dataset tasks.

Results on the LR-DG Benchmark. Table 3 summarizes the results on the domain generalization benchmark, where models trained on ImageNet are evaluated on LR test-sets of four ImageNet variants. LOREAL confers consistent improvements to all backbone methods across varied ϕ: Specifically, at $\phi = 96^2$, all methods exhibit a sharp average performance gain of 12.71%, further validating the effectiveness of LOREAL for domain generalization.

Efficiency Study. Integrating our LOREAL with existing approaches only requires additional meta-nets, resulting in high efficiency. In Table 4, we report the tunable parameters, training, and inference cost of MaPLe and MMRL with/without LOREAL under two ϕ. As observed, the additional computation cost is minimal or even negligible; In contrast, LOREAL brings substantial gains to the HM performance, demonstrating remarkable efficiency.

Table 4. Efficiency study. Tra./Infer. time is tested for a sample.

ϕ	Method	Tunable Para.	Tra. Time	Infer. Time	Infer. Mem.	HM ↑
96^2	MaPLe	3555K	107ms	32ms	612MB	34.85
	+LOREAL	+104.5K	+4ms	+1ms	+33MB	57.25
	MMRL	4992K	113ms	29ms	1236MB	38.90
	+LOREAL	+104.5K	+5ms	+1ms	+33MB	61.71
144^2	MaPLe	3555K	126ms	38ms	844	68.80
	+LOREAL	+104.5K	+4ms	+2ms	+33MB	73.72
	MMRL	4992K	145ms	37ms	1440MB	72.43
	+LOREAL	+104.5K	+5ms	+1ms	+33MB	76.56

4.3. Ablation Studies and Discussions

The Proposed Modules. We ablate each proposed module of LOREAL and report the results in Table 6. "LR→St." represents *using LR image embeddings, to contextualize prompt attributes of the student which accepts standard resolution.* Our key findings are: **(a):** both "LR→St" and "St.→LR" contribute to the performance, where the former exhibits a marginally superior effect (0.84% HM) compared to the latter. This indicates that achieving image-text alignment for LR inputs in training is more conducive to generalization in LR inference. **(b):** Removing either LLD or HLD leads to a considerable performance drop across all

Table 5. Ablation studies on D_s **(left)** and τ **(right)**.

D_s	Base	New	HM
8	73.19	62.18	66.98
16	76.07	63.75	69.11
32	77.30	67.37	71.73
64	76.54	66.96	71.17
128	76.35	67.58	71.44
256	75.78	67.49	71.14
512	75.97	61.20	67.53
1024	75.21	64.53	69.20
2048	73.88	62.74	67.60

τ	Base	New	HM
0.5	75.99	66.34	70.58
1	76.38	67.06	71.16
2	76.55	67.63	71.55
4	77.30	67.37	71.73
6	76.80	66.85	71.22
8	76.10	66.18	70.53
12	75.99	66.11	70.45
16	76.07	66.03	70.44
20	75.95	65.72	70.21

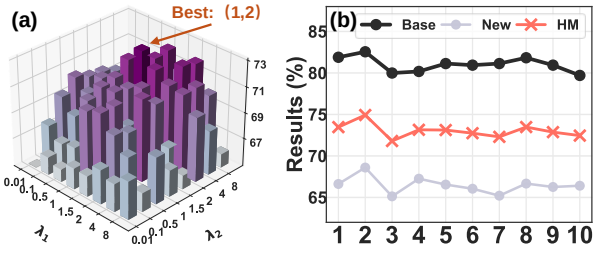

Figure 5. Ablation studies on **(a)** λ_1, λ_2 and **(b)** M.

metrics, confirming that both distillation losses are essential for bridging the semantic gap across resolutions. Between them, LLD contributes more significantly than HLD, likely because its direct attribute-level alignment is more readily optimized compared to the logit-level alignment of HLD.

Balance Coefficients λ_1 and λ_2. We perform a grid-search over these parameters in Figure 5 **(a)**. Although increasing either hyperparameter generally improves HM performance initially, a dramatic degradation occurs when they continually grow. This is likely because an excessively large distillation loss overwhelms the task objective, thereby hindering overall convergence. The optimal value for λ_2 is higher than that for λ_1, further indicating the effectiveness of LLD.

Length of Attribute Prompts M. We vary M from 1 to 10, with results presented in Figure 5**(b)**. We find that increasing M generally leads to a slight performance degradation ($\sim 2\%$), potentially because excessive tokens introduce additional meta-networks and increase optimization difficulty. We fix M to 2 as indicated by the results.

Intermediate Dimension D_s of Meta-nets. The results are shown in Table 5 **(left)**. Results suggest that while increasing D_s brings significant improvement initially, this trend rapidly reverses as D_s undergoes excessive expansion. This may be attributed to the expanded optimization difficulties of meta-nets. According to the results, the optimal D_s is 32.

The Temperature τ. We vary τ from 0.5 to 20 and report the results in Table 5 **(right)**. The results indicate that performance degrades with either excessively small or large values, likely because such values over-sharpen or over-smooth the output distributions, impairing the knowledge transfer. We keep $\tau = 4$ as suggested by the results.

Further Analysis. In this part, we discover the semantic meanings of attribute contents we learned for interpretability analysis. Concretely, we first construct an attribute vocabulary (Details are in **Sup. Mat. D**) by querying an LLM. We then compute the similarity between each vocabulary word's embedding and the learned output text embeddings from the student. In Figure 6 **(lower-left)**, we showcase the word cloud of several most similar attributes in the vocabulary. Our results reveal that the learned attribute contents accurately correspond to visual semantics, demonstrating the validity of our approach. Furthermore, we plot the attention heatmaps with or without LOREAL in Figure 6 **(right)**. With LOREAL, the model maintains focus on semantically central regions even at very low resolutions. Furthermore, it shifts attention toward holistic, resolution-robust features (e.g., the `havanese`'s entirety) rather than details. This demonstrates that our approach guides the model to prioritize invariant semantic concepts.

Table 6. Ablation studies on proposed modules.

LR → St.	St. → LR	LLD	HLD	Base	New	HM
✓				75.45	63.42	68.91
	✓			74.78	62.46	68.07
✓	✓			75.70	64.72	69.78
✓	✓	✓		76.10	66.35	70.89
✓	✓	✓	✓	77.30	67.37	71.73

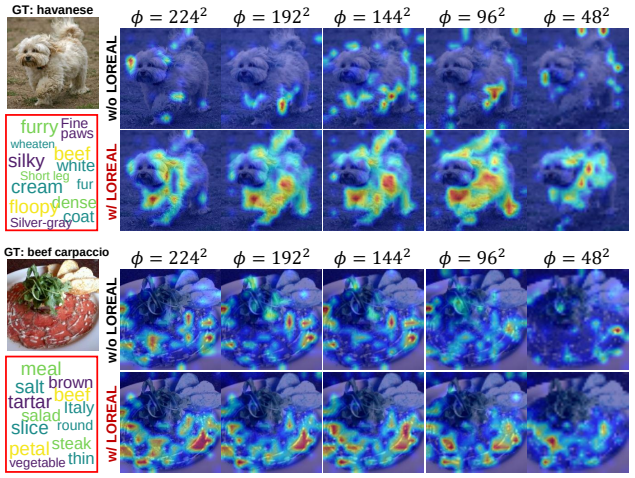

Figure 6. Visualizations of the learned attribute tokens **(Lower-Left)**, and attention heatmaps **(Right)** across varied resolution settings with or without our LOREAL on zero-shot CLIP. Note that here attention maps of different ϕ are resized for better illustration.

5. Conclusion

Edge-deployed VLMs are often constrained by computational resources, necessitating inference on Low-Resolution (LR) inputs and presenting a significant challenge to model robustness. To this end, this paper introduces LOREAL, a self-distillation framework that endows models with robustness to LR via resolution-invariant attributes. LOREAL comprises LLM-based attribute generation and attribute-driven self-distillation. Experiments under various LR settings demonstrate the superiority of our model.

References

[1] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023. 3

[2] 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. 5

[3] Haoxing Chen, Zizheng Huang, Yan Hong, Yanshuo Wang, Zhongcai Lyu, Zhuoer Xu, Jun Lan, and Zhangxuan Gu. Efficient transfer learning for video-language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29129–29138, 2025. 1

[4] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16296–16305, 2021. 3

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5008–5017, 2021. 3

[6] Shiming Chen, Bowen Duan, Salman Khan, and Fahad Shahbaz Khan. Interpretable zero-shot learning with locally-aligned vision-language model. *arXiv preprint arXiv:2506.23822*, 2025. 4

[7] 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. 5

[8] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 3

[9] 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. 5

[10] Tong Ding, Wanhua Li, Zhongqi Miao, and Hanspeter Pfister. Tree of attributes prompt learning for vision-language models. *arXiv preprint arXiv:2410.11201*, 2024. 3

[11] Yingjun Du, Wenfang Sun, and Cees Snoek. Ipo: Interpretable prompt optimization for vision-language models. *Advances in Neural Information Processing Systems*, 37: 126725–126766, 2024. 3

[12] 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. 5

[13] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2018. 3

[14] Guangyu Guo, Dingwen Zhang, Longfei Han, Nian Liu, Ming-Ming Cheng, and Junwei Han. Pixel distillation: Cost-flexible distillation across image sizes and heterogeneous networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9536–9550, 2024. 2, 3

[15] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. *arXiv preprint arXiv:2503.08497*, 2025. 1, 2, 5

[16] Yuncheng Guo and Xiaodong Gu. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *arXiv preprint arXiv:2505.10088*, 2025. 1

[17] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] 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. 5

[19] 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. 5

[20] 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. 5

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[23] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 4

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[25] 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. 1

[26] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023. 3

[27] 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. 1, 2, 5

[28] 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. 2

[29] Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. Cosmos: Cross-modality self-distillation for vision language pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14690–14700, 2025. 3

[30] Youmin Kim, Jinbae Park, YounHo Jang, Muhammad Ali, Tae-Hyun Oh, and Sung-Ho Bae. Distilling global and local logits with densely connected relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6290–6300, 2021. 3

[31] 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. 4, 5

[32] Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25176–25186, 2025. 3

[33] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 3

[34] Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25623–25632, 2025. 1, 2

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[37] Runze Li, Dahun Kim, Bir Bhanu, and Weicheng Kuo. Reclip: Resource-efficient clip by training with small images. *arXiv preprint arXiv:2304.06028*, 2023. 2

[38] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 2, 3

[39] Zheng Li, Yibing Song, Penghai Zhao, Ming-Ming Cheng, Xiang Li, and Jian Yang. Atprompt: Textual prompt learning with embedded attributes. *arXiv preprint arXiv:2412.09442*, 2024. 2, 4

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[41] Wenzhuo Liu, Fei Zhu, Shijie Ma, and Cheng-Lin Liu. Mspe: multi-scale patch embedding prompts vision transformers to any resolution. *Advances in Neural Information Processing Systems*, 37:29191–29212, 2024. 3

[42] Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[43] 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. 1

[44] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[45] Roy Miles, Ismail Elezi, and Jiankang Deng. Vkd: Improving knowledge distillation using orthogonal projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15720–15730, 2024. 3

[46] 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. 5

[47] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 2

[48] Bikang Pan, Qun Li, Xiaoying Tang, Wei Huang, Zhen Fang, Feng Liu, Jingya Wang, Jingyi Yu, and Ye Shi. Nlprompt: Noise-label prompt learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19963–19973, 2025. 1

[49] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 3

[50] 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. 4, 5

[51] Lu Qi, Jason Kuen, Jiuxiang Gu, Zhe Lin, Yi Wang, Yukang Chen, Yanwei Li, and Jiaya Jia. Multi-scale aligned distillation for low-resolution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14443–14453, 2021. 3

[52] Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[53] 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. 1

[54] 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. 5

[55] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[56] Leqi Shen, Guoqiang Gong, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Discovla: Discrepancy reduction in vision, language, and alignment for parameter-efficient video-text retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19702–19712, 2025. 1

[57] Sungho Shin, Joosoon Lee, Junseok Lee, Yeonguk Yu, and Kyoobin Lee. Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. In *European Conference on Computer Vision*, pages 631–647. Springer, 2022. 3

[58] 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. 5

[59] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. Resformer: Scaling vits with multi-resolution training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22721–22731, 2023. 3

[60] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024. 2

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 3

[62] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[63] Can Wang, Defang Chen, Jian-Ping Mei, Yuan Zhang, Yan Feng, and Chun Chen. Semckd: Semantic calibration for cross-layer knowledge distillation. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6305–6319, 2022. 3

[64] 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. 5

[65] Mengjiao Wang, Rujie Liu, Nada Hajime, Abe Narishige, Hidetsugu Uchida, and Tomoaki Matsunami. Improved knowledge distillation for training fast low resolution face recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[66] Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, and Linlin Li. Efficient image captioning for edge devices. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2608–2616, 2023. 2

[67] Xuan Wang, Xitong Gao, Dongping Liao, Tianrui Qin, Yuliang Lu, and Cheng-zhong Xu. A3: Few-shot prompt learning of unlearnable examples with cross-modal adversarial feature alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9507–9516, 2025. 1

[68] Zhifei Wang, Zhenjiang Miao, QM Jonathan Wu, Yanli Wan, and Zhen Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014. 3

[69] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4792–4800, 2016. 3

[70] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *European Conference on Computer Vision*, pages 304–321. Springer, 2024. 3

[71] Shihan Wu, Ji Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14723–14732, 2025. 3

[72] 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. 5

[73] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526, 2025. 1, 2

[74] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 3

[75] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024. 1, 2

[76] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European conference on computer vision*, pages 53–69. Springer, 2022. 3

[77] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 1, 2

[78] 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. 1

[79] Yi Zhang, Ke Yu, Siqi Wu, and Zhihai He. Conceptual codebook learning for vision-language models. In *European Conference on Computer Vision*, pages 235–251. Springer, 2024. 4

[80] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 3

[81] Li Zhong, Ahmed Ghazal, Jun-Jun Wan, Frederik Zilly, Patrick Mackens, Joachim Vollrath, and Bogdan Coseriu. Clip4retrofit: Enabling real-time image labeling on edge devices via cross-architecture clip distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3829–3837, 2025. 2

[82] 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. 1, 2, 4

[83] 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. 1, 4

[84] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023. 2

[85] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011. 3