A DATASET IS WORTH DECOMPOSING INTO MULTIPLE LATENT DOMAINS: REFRAMING AND MITIGATING BASE–NEW TRADE-OFF IN PROMPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt learning provides an efficient way to adapt vision-language models, yet optimizing prompts on seen base classes often compromises generalization to unseen new classes, giving rise to the Base–New Trade-off. We identify the prevalent *monolithic assumption* as a key source of this tension, with a single global prompt expected to model the complex and non-uniform semantic manifold spanning all classes in a dataset. To mitigate this tension, we reframe the class semantic space of a dataset as multiple latent domains and propose **L**atent **D**omain **P**rompt (**LDP**). LDP constructs semantic anchors from the frozen textual geometry of the full predefined class vocabulary and learns a global prompt together with a set of latent domain bases using only base-class visual supervision. For each class, LDP synthesizes a class-adaptive prompt by weighting these bases according to the class's semantic affinities to the anchors. Through this affinity-weighted synthesis, latent domains serve as structural bridges between base and new classes, allowing new classes to inherit domain-level knowledge learned from base classes without requiring new-class images. Theoretically, monolithic prompting contains a first-order global-displacement term, whereas LDP is governed by local fitting and reconstruction errors, with a second-order geometric term under approximate barycentric reconstruction. Extensive experiments spanning 15 datasets show that LDP consistently improves four representative prompt learning baselines across base-to-new, cross-dataset, and cross-domain settings.

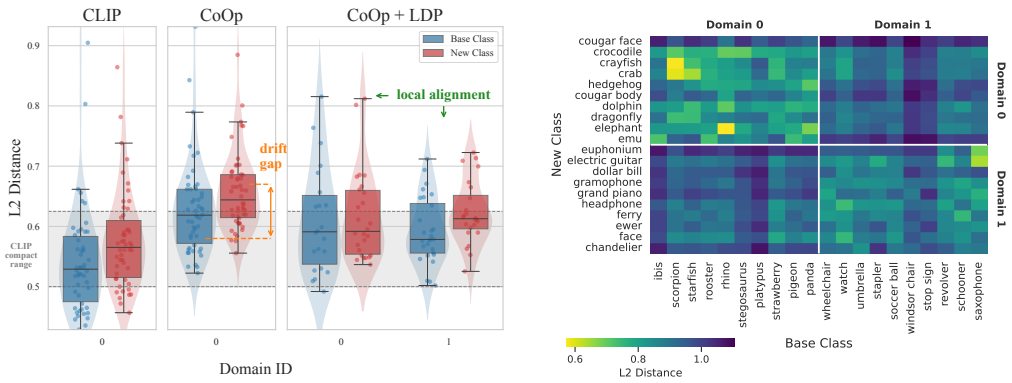
1 INTRODUCTION

Motivation. Vision-language models (VLMs), such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), enable strong zero-shot recognition by aligning visual and textual representations in a shared semantic space. To adapt them to downstream tasks, strategies ranging from full-model fine-tuning to parameter-efficient adaptation have been explored (Gao et al., 2024; Hu et al., 2022). Among them, prompt learning (Zhou et al., 2022a; Khattak et al., 2023a; Xu et al., 2025; Hua et al., 2025) optimizes a small set of prompt vectors while keeping the pretrained encoders frozen. However, optimizing prompts on seen base classes creates a persistent *structural tension*: improved base-class recognition often compromises generalization to unseen new classes, giving rise to the Base–New Trade-off.

A geometric observation. Figure 1(a) compares the distance distributions of prompt-induced textual representations. The dispersed distribution under CoOp (Zhou et al., 2022b) suggests that optimizing a single global prompt on base classes can distort compact semantic structure and weaken new-class generalization. In contrast, LDP produces a more compact distribution, indicating that preserving local semantic neighborhoods can mitigate prompt-induced drift. Figure 1(b) visualizes pairwise distances between selected base and new classes under LDP, where affinity-related categories form visible local blocks. This raises a central question: *Why does a single global prompt tend to disrupt the semantic structure shared by base and new classes?*

The monolithic assumption. A prevailing view attributes the Base–New Trade-off to prompt overfitting on base classes (Zhou et al., 2022a; Yao et al., 2023; Khattak et al., 2023b). We further trace this tendency to a *monolithic assumption*: a single global prompt can adequately represent the en-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Semantic Drift and Local Alignment (b) Latent Structural Alignment

Figure 1: **(a) Semantic Drift and Local Alignment:** We compare the ℓ_2 distances between prompt-induced textual representations and corresponding centroids. CoOp exhibits a more dispersed distribution, suggesting greater semantic drift under a single global prompt. In contrast, LDP yields a more compact distribution by aligning classes with latent domains, thereby better preserving the semantic manifold. **(b) Latent Structural Alignment:** The pairwise-distance heatmap exhibits local blocks among the selected base and new classes, consistent with shared local semantic structure.

tire semantic space. Inspired by continual learning, where distinct prompts model heterogeneous tasks (Wang et al., 2022b; Douillard et al., 2022; Wang et al., 2022c;a), we argue that a prompt trained solely with base-class visual supervision becomes a global compromise shaped primarily by base-supported semantic regions. This forced unification can bias visual–textual alignment toward seen samples, reshaping the prompt-induced geometry and disrupting local relationships between base and new classes. As illustrated in Figure 2(Left), unrelated categories such as the base class *watch* and the new class *elephant* can consequently become spuriously close, weakening new-class transfer.

Latent domains bridge base and new classes. To mitigate this tension, we introduce *latent domains* as a structural bridge between base and new classes. As shown in Figure 2(Right), each latent domain is characterized by a semantic anchor and a learnable base, and LDP synthesizes a class-adaptive prompt by weighting these bases according to class-anchor affinities. Thus, the new class *elephant* and the related base class *rhino* can draw on the same latent domain base, transferring domain-level knowledge from base classes. This affinity-weighted sharing preserves local semantic neighborhoods without requiring new-class images, thereby mitigating the Base–New Trade-off.

Our contributions are threefold:

- We **propose Latent Domain Prompt (LDP)**, which mitigates the Base–New Trade-off by decomposing the class semantic space into multiple latent domains. LDP constructs anchors from textual geometry, learns corresponding latent domain bases using base-class visual supervision, and synthesizes class-adaptive prompts through affinity weighting, establishing a structural bridge between base and new classes.
- We **derive approximation-error expansions** that formalize the global-to-local reframing: monolithic prompting contains a first-order global-displacement term, whereas LDP is governed by local fitting and reconstruction errors, with a second-order geometric term under approximate barycentric reconstruction.
- We **conduct extensive experiments** on 15 datasets across base-to-new, cross-dataset, and cross-domain settings, demonstrating consistent improvements over representative methods.

2 Related Work

Prompt Learning in VLMs. The emergence of VLMs has demonstrated remarkable zero-shot generalization capabilities through contrastive learning on large-scale web image-text pairs. To adapt these frozen pre-trained models to downstream tasks, prompt learning has emerged. CoOp (Zhou et al., 2022b) significantly improved few-shot classification performance by replacing fixed hand-

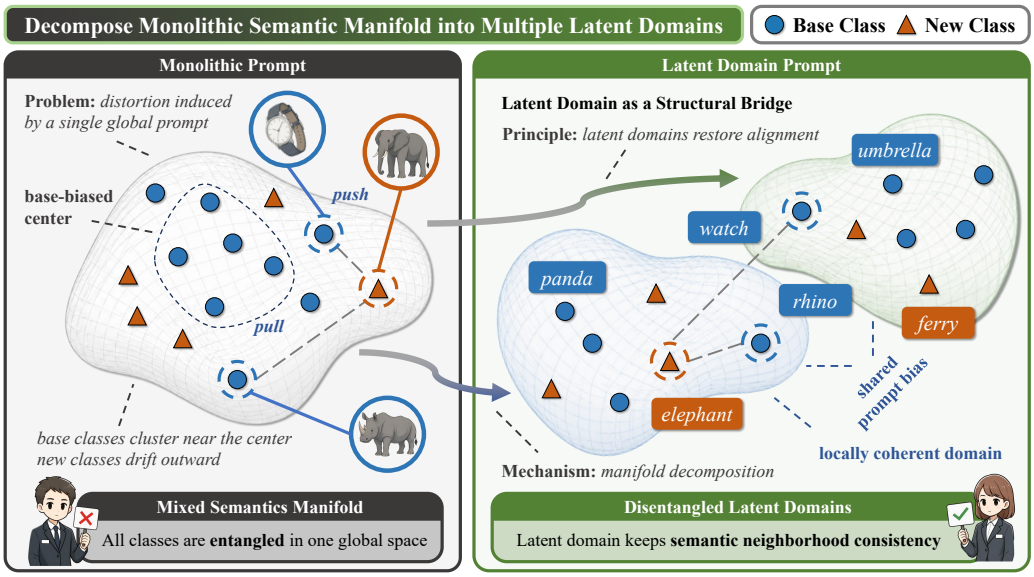


Figure 2: **(Left) Monolithic Prompt:** A single global prompt, optimized on base classes, can bias the prompt-induced semantic geometry toward base-supported regions, distorting local relationships between base and new classes and bringing unrelated categories, such as *watch* and *elephant*, spuriously close. **(Right) Latent Domain Prompt:** our **LDP** models the class semantic manifold through multiple latent domains. This localized structure preserves coherent semantic neighborhoods, allowing the new class *elephant* to draw on domain-level knowledge shared with the semantically related base class *rhino*.

crafted templates with learnable continuous text vectors. Subsequent studies have expanded this paradigm to visual and multi-modal contexts (Zhou et al., 2022a; Khattak et al., 2023a; Jia et al., 2022; Yang et al., 2024), evolving from shallow to deep hierarchical optimization (Khattak et al., 2023b; Yao et al., 2023; Qiao et al., 2024; Wu et al., 2025).

Base–New Trade-off. Existing prompt learning methods often suffer from severe performance degradation on new classes, a phenomenon referred to as the Base–New Trade-off. Regularization-based methods (Yao et al., 2023; Roy & Etemad, 2024; Khattak et al., 2023b; Park et al., 2024) introduce consistency constraints to prevent learnable prompts from deviating excessively from pre-trained knowledge. However, they fail to fundamentally resolve the base-new distributional discrepancy. LLM-guided methods (Tian et al., 2024; Zheng et al., 2024; Wang et al., 2024) leverage the rich external knowledge of LLM to enhance prompts but introduce additional computational overhead. DePT (Zhang et al., 2024) and DPC (Li et al., 2025a) decouple prompts for base and new classes but require task identifiers at inference. ATPrompt (Li et al., 2025b) connects them via fixed attribute vocabularies, yet relies on manual prior filtering per dataset. Prior methods (Li et al., 2026; Hao et al., 2025) cluster either image features or task-level contexts. In contrast, LDP clusters class-level textual representations into anchors to synthesize class-adaptive prompts, restructuring them to implicitly bridge via latent domains.

3 Diagnosis of the Base–New Trade-off

3.1 Preliminaries

CLIP (Radford et al., 2021) comprises an image encoder $\mathcal{E}_v$ and a text encoder $\mathcal{E}_t$ that align visual and textual representations in a shared space. Given an image $\mathbf{x}$, the image encoder produces $\mathbf{v} = \mathcal{E}_v(\mathbf{x})$. For a classification task with class set $\mathcal{C}$, CLIP typically uses a hand-crafted prompt template. Let $\mathbf{w}_c$ denote the class-name token embeddings for class c and $\mathbf{t}_c = \mathcal{E}_t(\mathbf{w}_c)$ its textual representation. The probability of assigning $\mathbf{x}$ to class c is

$$p(y = c|\mathbf{x}) = \frac{\exp(\cos(\mathbf{v}, \mathbf{t}_c)/\tau)}{\sum_{i \in \mathcal{C}} \exp(\cos(\mathbf{v}, \mathbf{t}_i)/\tau)}, \tag{1}$$

where τ is the temperature parameter. To avoid manual prompt engineering, CoOp (Zhou et al., 2022b) replaces hand-crafted context words with M learnable context vectors $\mathbf{P} = \{\mathbf{p}_m\}_{m=1}^{M}$. For class c, the prompt sequence is $\mathbf{s}_c = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M, \mathbf{w}_c]$. For the following analysis, all prompt parameterizations lie in a common finite-dimensional Hilbert space $\mathcal{H}$ equipped with the Frobenius norm; squared distance defines the best class-independent approximation, while approximation errors use the norm itself.

3.2 Approximation Error of Monolithic Prompting

Our analysis formalizes the approximation component of the monolithic-assumption diagnosis. The central distinction is between class-dependent oracle prompts and a single class-independent prompt that is constrained to jointly approximate the prompt requirements of all base classes.

Assumption 1 (Oracle Prompt Map)**.** There exists a function $\Phi^\star : \mathcal{U} \to \mathcal{H}$ that maps a class textual representation to an oracle prompt for that class. The set $\mathcal{U} \subset \mathbb{R}^D$ is an open convex neighborhood containing all textual representations and anchors used in the analysis. The map $\Phi^\star$ is a theoretical reference and need not be available to the algorithm.

Define the base-class semantic centroid and spread as $\mathbf{c}_b = \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\mathbf{t}_c$ and $V_b = \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\|\mathbf{t}_c - \mathbf{c}_b\|_2^2$. The best class-independent (i.e., constant across classes) approximation to the base-class oracle prompts under the squared prompt-space distance surrogate is

$$\mathbf{P}_b^\circ \in \arg\min_{\mathbf{Q}\in\mathcal{H}} \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\|\mathbf{Q} - \Phi^\star(\mathbf{t}_c)\|_F^2. \tag{2}$$

Since $\mathcal{H}$ is a finite-dimensional Hilbert space, $\mathbf{P}_b^\circ = \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\Phi^\star(\mathbf{t}_c)$. We refer to $\mathbf{P}_b^\circ$ as the oracle global compromise induced by base-class supervision.

Assumption 2 (Monolithic Fitting Consistency)**.** The learned monolithic prompt $\mathbf{P}$ satisfies $\|\mathbf{P} - \mathbf{P}_b^\circ\|_F \le \varepsilon_m$, where ε_m measures the discrepancy between the learned prompt and the oracle compromise under the squared prompt-space surrogate.

Assumption 3 (Local Smoothness)**.** For any $\mathbf{u}, \mathbf{v} \in \mathcal{U}$,

$$\Phi^\star(\mathbf{u}) = \Phi^\star(\mathbf{v}) + J_{\Phi^\star}(\mathbf{v})(\mathbf{u} - \mathbf{v}) + \mathbf{R}(\mathbf{u}, \mathbf{v}), \tag{3}$$

where $\|\mathbf{R}(\mathbf{u}, \mathbf{v})\|_F \le C\|\mathbf{u} - \mathbf{v}\|_2^2$ for a constant $C > 0$.

This assumption controls the local Taylor remainder of the oracle prompt map within the semantic neighborhood considered in the analysis. It is motivated by the smoothness and manifold perspectives in representation learning (Bengio et al., 2013).

Theorem 1 (Characterization of Monolithic Prompt Approximation Error)**.** Under Assumptions 1–3, for a textual representation $\mathbf{t} \in \mathcal{U}$, let $\mathbf{\Delta}(\mathbf{t}) = \mathbf{t} - \mathbf{c}_b$. Then

$$\mathbf{P} - \Phi^\star(\mathbf{t}) = -J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t}) + \boldsymbol{\xi}_m(\mathbf{t}), \tag{4}$$

where $\|\boldsymbol{\xi}_m(\mathbf{t})\|_F \le \varepsilon_m + CV_b + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2$. Consequently, with $\mathcal{E}_m(\mathbf{t}) = \|\mathbf{P} - \Phi^\star(\mathbf{t})\|_F$, we have

$$\left|\mathcal{E}_m(\mathbf{t}) - \|J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t})\|_F\right| \le \varepsilon_m + CV_b + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2. \tag{5}$$

In particular,

$$\mathcal{E}_m(\mathbf{t}) \le \varepsilon_m + CV_b + \|J_{\Phi^\star}(\mathbf{c}_b)\|\|\mathbf{\Delta}(\mathbf{t})\|_2 + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2. \tag{6}$$

Theorem 1 thus characterizes the prompt-space approximation induced by monolithic prompting, with the first-order global-displacement term becoming dominant when it exceeds the residual terms.

4 Method

4.1 Latent Domain Prompt

Inspired by our theoretical diagnosis of the approximation error of monolithic prompting, we introduce LDP to relax the monolithic assumption by augmenting the shared global prompt with class-adaptive latent domain prompts. As shown in Figure 3, LDP follows a structure-to-prompt pipeline: Stage I organizes the full vocabulary into local semantic regions, while Stage II transfers this structure into class-adaptive prompts through learned latent domain bases.

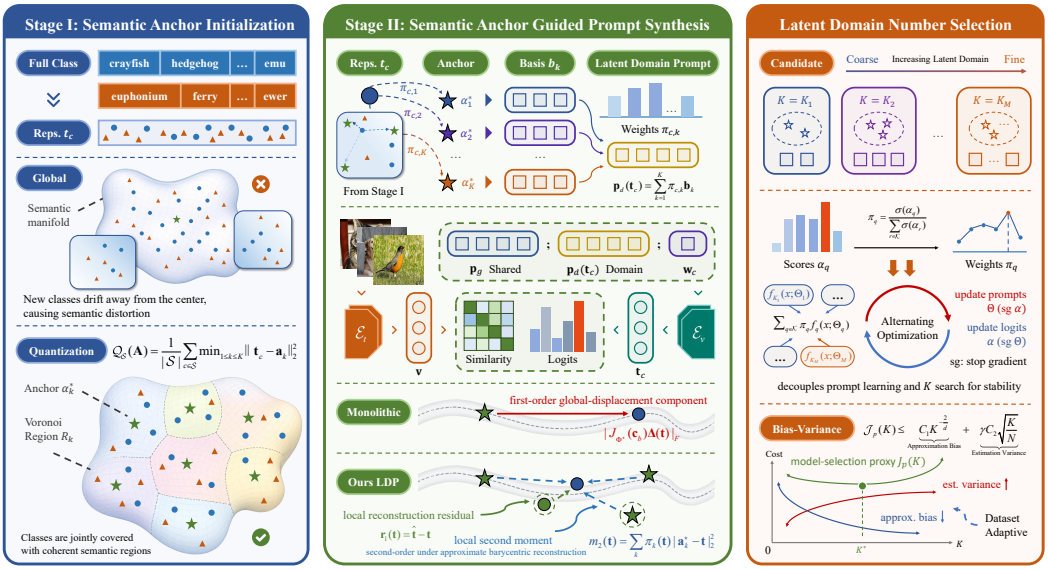


Figure 3: Illustration of LDP. **(Left)** LDP quantizes the semantic manifold to obtain semantic anchors and corresponding Voronoi regions. **(Middle)** LDP learns a set of latent domain bases and combines them according to affinity-weighted prompt synthesis. **(Right)** LDP evaluates candidate latent domain numbers through alternating optimization and selects a dataset-specific K.

4.1.1 Stage I: Semantic Anchor Initialization

Definition 1 (Semantic Quantization)**.** Let $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^K$ be a set of K semantic anchors. For a non-empty class subset $\mathcal{S} \subseteq \mathcal{C}$, we define its empirical semantic quantization risk as

$$\mathcal{Q}_{\mathcal{S}}(\mathbf{A}) = \frac{1}{|\mathcal{S}|} \sum_{c \in \mathcal{S}} \min_{1 \le k \le K} \|\mathbf{t}_c - \mathbf{a}_k\|_2^2. \tag{7}$$

The full-vocabulary semantic anchors are defined by $\mathbf{A}^* \in \arg\min_{\mathbf{A}} \mathcal{Q}_{\mathcal{C}}(\mathbf{A})$.

In practice, it is optimized using K-means (MacQueen, 1967). The optimized anchors induce a Voronoi partition of the full vocabulary:

$$R_k = \left\{ c \in \mathcal{C} : k = \arg\min_{1 \le j \le K} \|\mathbf{t}_c - \mathbf{a}_j^*\|_2^2 \right\} \tag{8}$$

Proposition 1 (Full-Vocabulary Coverage Rebalancing)**.** Let $\mathcal{C} = \mathcal{C}_b \,\dot\cup\, \mathcal{C}_n$, with both subsets nonempty. For the same number of anchors and the same distance metric, let $\mathbf{A}_b^\star \in \arg\min_{\mathbf{A}} \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A})$ and $\mathbf{A}^\star \in \arg\min_{\mathbf{A}} \mathcal{Q}_{\mathcal{C}}(\mathbf{A})$ be global minimizers of the base-vocabulary and full-vocabulary quantization risks, respectively. Then

$$\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}_b^\star) \le \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}^\star), \qquad \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}^\star) \le \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}_b^\star). \tag{9}$$

Remark 1. Proposition 1 concerns average risks at exact global optima, not individual classes or arbitrary local solutions returned by K-means. Stage I is transductive in the predefined vocabulary but uses no new-class images. The use of the full vocabulary is consistent with CLIP-based classification, which requires all class names for textual classifiers.

4.1.2 Stage II: Semantic Anchor Guided Prompt Synthesis

Definition 2 (Global and Latent Domain Prompt Parameterization)**.** We define $\mathbf{p}_g \in \mathbb{R}^{L_g \times D}$ as the global prompt capturing class agnostic knowledge, and establish a set of learnable latent domain bases $\mathcal{B} = \{\mathbf{b}_k\}_{k=1}^K \subset \mathbb{R}^{L_d \times D}$, where each $\mathbf{b}_k$ corresponds to the representation of the local semantic manifold R_k. For class c, we compute the semantic affinity between class c and anchor $\mathbf{a}_k^*$ as

$$\pi_{c,k} = \frac{\exp\left(-\|\mathbf{t}_c - \mathbf{a}_k^*\|_2^2/\lambda\right)}{\sum_{i=1}^K \exp\left(-\|\mathbf{t}_c - \mathbf{a}_i^*\|_2^2/\lambda\right)}, \tag{10}$$

where $\lambda > 0$ is a temperature parameter. Since the affinities are normalized, they satisfy $\pi_{c,k} \geq 0$ and $\sum_{k=1}^{K} \pi_{c,k} = 1$. The latent domain prompt for class c is then synthesized by $\mathbf{p}_d(\mathbf{t}_c) = \sum_{k=1}^{K} \pi_{c,k} \mathbf{b}_k$. The final prompt sequence for class c is $\mathbf{s}_c = [\mathbf{p}_g; \mathbf{p}_d(\mathbf{t}_c); \mathbf{w}_c]$.

Proposition 2 (Affinity Weighted Gradient Sharing)**.** During training, the semantic affinities in Eq. equation 10 are fixed with respect to the parameters. For a mini batch $\mathcal{X}$ sampled from the base class training set, the gradient of the loss with respect to the k-th latent domain prompt base satisfies

$$\frac{\partial \mathcal{L}_{\mathcal{X}}}{\partial \mathbf{b}_k} = \sum_{(\mathbf{x},y)\in\mathcal{X}} \sum_{c\in\mathcal{C}_b} \pi_{c,k} \frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{p}_d(\mathbf{t}_c)}. \tag{11}$$

Thus, soft affinities share gradients across related regions, alleviating sparse updates from hard assignment.

Definition 3 (Latent Domain Number Selection)**.** Let $\mathcal{K} = \{K_1, \ldots, K_M\}$ be a candidate set of latent domain numbers. For each $q \in \mathcal{K}$, we construct semantic anchors $\mathbf{A}_q^*$ and build a branch $f_q(\cdot; \Theta_q)$ with latent domain bases. Let $\boldsymbol{\alpha} = \{\alpha_q\}_{q\in\mathcal{K}}$ be learnable search scores over candidate branches. Inspired by DARTS (Liu et al., 2019) and ATPrompt (Li et al., 2025b), we normalize the branch selection weight as $\pi_q = \sigma(\alpha_q)/\sum_{r\in\mathcal{K}} \sigma(\alpha_r)$. During search, we alternately optimize the prompt parameters $\Theta = \{\Theta_q\}_{q\in\mathcal{K}}$ and the search logits $\boldsymbol{\alpha}$:

$$\hat{\Theta} = \arg\min_{\Theta} \mathcal{L}_{\mathrm{CE}}^{\mathrm{tr}} \left(\sum_{q\in\mathcal{K}} \mathrm{sg}(\pi_q) f_q(\mathbf{x}; \Theta_q), y \right), \tag{12}$$

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_{\mathrm{CE}}^{\mathrm{val}} \left(\sum_{q\in\mathcal{K}} \pi_q f_q(\mathbf{x}; \mathrm{sg}(\Theta_q)), y \right), \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes the stop gradient operation. The latent domain number is selected by $K^* = \arg\max_{q\in\mathcal{K}} \pi_q$.

4.2 Approximation Error of LDP

To characterize the approximation error of LDP, we rewrite its class-adaptive prompt as a mixture of full prompt prototypes. For each latent domain k, let $\mathbf{Q}_k \coloneqq [\mathbf{p}_g; \mathbf{b}_k] \in \mathcal{H}$. The resulting prompt is $\mathbf{P}_d(\mathbf{t}) \coloneqq [\mathbf{p}_g; \mathbf{p}_d(\mathbf{t})] = \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{Q}_k$. We further define the prototype fitting error $\mathbf{e}_k \coloneqq \mathbf{Q}_k - \Phi^\star(\mathbf{a}_k^\star)$ and its magnitude $\delta_k \coloneqq \|\mathbf{e}_k\|_F$.

For a textual representation $\mathbf{t}$, let $\delta(\mathbf{t}) \coloneqq \left\| \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{e}_k \right\|_F$ denote the aggregated fitting error. By the triangle inequality, $\delta(\mathbf{t}) \leq \sum_{k=1}^{K} \pi_k(\mathbf{t}) \delta_k$. We also define the first-order residual $\mathbf{r}_1(\mathbf{t}) \coloneqq \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{a}_k^\star - \mathbf{t}$ and the second-order moment $m_2(\mathbf{t}) \coloneqq \sum_{k=1}^{K} \pi_k(\mathbf{t}) \|\mathbf{a}_k^\star - \mathbf{t}\|_2^2$. Here, $\mathbf{r}_1(\mathbf{t})$ is the first-order anchor-reconstruction residual, and $m_2(\mathbf{t})$ is the affinity-weighted local second moment.

Theorem 2 (Local-Mixture Expansion for LDP)**.** Under Assumptions 1 and 3, for any $\mathbf{t} \in \mathcal{U}$ whose positively weighted anchors also lie in $\mathcal{U}$,

$$\mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t}) = \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{e}_k + J_{\Phi^\star}(\mathbf{t}) \mathbf{r}_1(\mathbf{t}) + \boldsymbol{\rho}_d(\mathbf{t}), \tag{14}$$

where $\|\boldsymbol{\rho}_d(\mathbf{t})\|_F \leq C m_2(\mathbf{t})$. Consequently,

$$\mathcal{E}_d(\mathbf{t}) \coloneqq \|\mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t})\|_F \leq \delta(\mathbf{t}) + \|J_{\Phi^\star}(\mathbf{t})\| \|\mathbf{r}_1(\mathbf{t})\|_2 + C m_2(\mathbf{t}). \tag{15}$$

Assumption 4 (Uniform Approximate Local Barycentric Reconstruction)**.** For the target representations under consideration, there exists a constant $\eta \geq 0$, independent of the local anchor radius, such that

$$\|\mathbf{r}_1(\mathbf{t})\|_2 \leq \eta m_2(\mathbf{t}). \tag{16}$$

Corollary 1 (Second-Order Local Geometry Bound)**.** Under the conditions of Theorem 2 and Assumption 4,

$$\mathcal{E}_d(\mathbf{t}) \leq \delta(\mathbf{t}) + \big(\eta \|J_{\Phi^\star}(\mathbf{t})\| + C\big) m_2(\mathbf{t}). \tag{17}$$

The second-order bound requires Assumption 4; without it, Theorem 2 remains valid, but the first-order reconstruction residual need not be absorbed into the local second moment. Together with Theorem 1, this result formalizes the shift from global to affinity-weighted local approximation.

Proposition 3 (Proxy Approximation-Estimation Trade-off)**.** Let q_K denote the optimal K-anchor semantic quantization risk and e_K the affinity-averaged prototype estimation error. Under the covering-number and effective sample-size conditions detailed in the appendices, we have $q_K \leq C_1 K^{-\frac{2}{d}}$ and $e_K \leq C_2\sqrt{\frac{K}{N}}$. Consequently, for $\mathcal{J}_p(K) = q_K + \gamma e_K$,

$$\mathcal{J}_p(K) \leq C_1 K^{-\frac{2}{d}} + \gamma C_2 \sqrt{\frac{K}{N}}. \tag{18}$$

This proxy provides a theoretical rationale for selecting K, which is determined empirically through validation in practice. *Detailed proofs and clarifications are provided in the appendices.*

Table 1: Comparison of base-to-new generalization on 11 datasets. We report the accuracy on base classes, new classes, and their harmonic mean (H); full results are provided in the appendices. LDP improves the average H score for each baseline.

Method	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
SDPT	85.00	75.20	79.80	78.30	70.70	74.30	98.60	95.50	97.00	95.30	98.20	96.70	80.80	70.40	75.20	97.70	74.70	84.70
MMA++	85.24	76.80	80.81	78.33	71.20	74.56	98.70	94.53	96.57	95.83	97.40	96.61	82.63	74.83	78.54	95.07	77.17	85.19
CASPA-G	85.24	77.18	81.01	77.70	70.85	74.12	98.91	94.31	96.56	95.43	97.16	96.29	81.37	76.41	78.81	98.14	77.74	86.76
CoOp	81.19	70.25	75.32	76.23	69.74	72.84	97.93	93.56	95.69	95.22	97.54	96.36	72.83	73.04	72.93	95.79	72.65	82.63
w/LDP	82.13	73.77	**77.73**	76.86	70.07	**73.31**	97.98	95.38	**96.66**	95.78	97.52	**96.64**	74.17	73.06	**73.61**	97.12	74.96	**84.62**
MaPLe	83.34	73.92	78.35	77.02	69.39	73.00	98.30	94.76	**96.50**	95.09	97.39	96.23	76.52	71.78	74.08	97.34	71.82	82.65
w/LDP	83.56	74.72	**78.89**	77.21	69.29	**73.04**	98.09	94.61	96.32	95.66	97.39	**96.52**	76.69	72.89	**74.74**	96.96	73.19	**83.42**
PromptSRC	83.31	75.58	79.26	77.36	70.66	73.86	97.83	94.43	96.10	95.34	97.22	96.27	75.82	75.15	75.49	97.79	76.00	85.53
w/LDP	83.67	76.65	**80.01**	77.50	70.79	**73.99**	98.06	94.43	**96.21**	95.59	97.18	**96.38**	76.69	75.13	**75.90**	97.53	76.99	**86.05**
MMRL++	85.44	77.78	81.43	77.73	71.37	74.41	98.90	94.40	96.60	95.43	96.97	96.19	81.23	75.23	78.12	98.50	77.47	**86.73**
w/LDP	85.65	78.23	**81.77**	77.90	71.67	**74.65**	99.17	94.80	**96.93**	95.97	97.20	**96.58**	81.73	75.27	**78.37**	98.00	77.70	86.68

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
SDPT	90.80	92.10	91.40	45.40	31.10	36.90	82.80	77.50	80.10	84.10	59.80	69.90	94.20	78.90	85.90	86.50	78.70	82.40
MMA++	90.83	91.63	91.23	48.64	38.05	42.70	82.80	79.57	81.15	82.30	64.30	72.19	94.55	75.97	84.25	88.30	80.17	84.04
CASPA-G	91.86	92.56	92.21	46.98	37.51	41.71	81.94	78.17	80.01	83.09	61.65	70.78	94.03	82.06	87.64	88.24	80.56	84.23
CoOp	90.46	91.88	**91.17**	35.85	29.95	32.64	80.51	75.45	77.90	78.90	46.46	58.48	86.84	50.33	63.73	82.52	72.13	76.98
w/LDP	90.52	91.70	91.11	36.99	33.67	**35.25**	81.42	75.77	**78.49**	79.09	54.47	**64.51**	90.01	69.64	**78.53**	83.47	75.25	**79.15**
MaPLe	90.21	91.68	**90.94**	39.38	34.29	36.66	81.75	77.10	**79.36**	82.37	58.33	**68.30**	94.18	69.76	80.15	84.63	76.78	80.51
w/LDP	90.36	91.32	90.84	40.52	35.43	**37.80**	81.78	76.90	79.27	81.56	57.61	67.52	94.78	75.56	**84.09**	85.52	77.77	**81.46**
PromptSRC	90.68	91.74	**91.21**	39.18	35.87	37.45	82.35	78.85	80.56	82.91	60.95	70.25	91.29	71.56	80.23	85.82	79.00	82.27
w/LDP	90.62	91.79	91.20	39.88	37.63	**38.72**	82.58	78.98	**80.74**	82.71	61.43	**70.50**	92.60	80.13	**85.91**	86.64	78.71	**82.48**
MMRL++	90.50	91.70	91.10	46.47	38.50	**42.11**	82.93	79.57	81.22	85.07	65.83	74.22	95.73	84.17	89.58	87.37	80.43	83.76
w/LDP	90.50	91.87	**91.18**	47.53	36.83	41.50	83.23	79.70	**81.43**	84.87	66.67	**74.67**	95.77	87.33	**91.36**	87.53	81.50	**84.41**

5 EXPERIMENTS

5.1 EXPERIMENT SETTINGS

Datasets. Following CoCoOp (Zhou et al., 2022a), we evaluate our method on a comprehensive benchmark suite of 15 datasets covering base-to-new, cross-dataset, and cross-domain generalization settings. Specifically, 11 datasets are used for base-to-new and cross-dataset generalization, while 4 ImageNet variants are adopted for cross-domain generalization. Detailed information is provided in the appendices.

Implementation Details. We integrate LDP with four representative prompt learning baselines, including CoOp (Zhou et al., 2022b), MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b), and MMRL++ (Guo & Gu, 2025). To ensure a fair comparison, we use the same training configurations as the corresponding baseline methods unless otherwise specified, and report the average results over three seeds. For LDP, semantic anchors are re-initialized for each seed, so the reported averages also reflect its stability under different K-means initializations. We also include several recent methods (Zhou et al., 2026; Yang et al., 2026; Chatterjee et al., 2026) as reference baselines for comparison.

5.2 EXPERIMENTAL RESULTS

Table 2: Comparison of cross-dataset generalization from source dataset to 10 target datasets. LDP consistently improves the average transfer performance across all baselines, demonstrating enhanced robustness against distribution shifts.

Method	Source	Target										Avg.
	ImageNet	Caltech101	Pets	Cars	Flowers	Food	FGVCA	SUN397	DTD	EuroSAT	UCF101	
CoOp	71.39	93.48	89.94	64.31	70.15	85.79	20.35	66.02	43.63	47.24	69.59	65.05 ±0.09
w/LDP	**72.03**	**94.17**	**90.53**	**66.11**	**71.20**	**86.18**	**23.09**	**66.15**	**43.93**	**50.02**	**69.72**	**66.11** ±0.14
MaPLe	72.48	93.00	90.04	63.30	69.85	84.70	21.65	65.73	42.93	39.67	66.76	63.76 ±0.06
w/LDP	**72.67**	**93.67**	**90.66**	**64.36**	**70.25**	**84.99**	**23.42**	**66.31**	**43.89**	**44.80**	**67.27**	**64.96** ±0.41
PromptSRC	72.57	93.06	90.19	65.87	66.04	85.89	23.63	66.80	44.98	38.98	66.04	64.92 ±0.30
w/LDP	**72.74**	**93.83**	**90.94**	**66.42**	**71.13**	**86.14**	**24.75**	**67.12**	**47.01**	**43.40**	**69.70**	**66.04** ±0.17
MMRL++	71.37	94.57	91.23	66.60	73.17	**86.67**	26.17	68.13	45.87	52.07	68.87	67.33 ±0.30
w/LDP	**71.67**	**94.90**	**91.47**	**66.67**	**73.97**	**86.67**	**26.47**	**68.17**	**46.07**	**55.87**	**69.60**	**67.98** ±0.11

Table 3: Comparison of cross-domain generalization from source to 4 distribution-shifted variants. LDP consistently improves transfer performance under visual shifts.

Method	Source	Target				Avg.
	ImageNet	-V2	-S	-A	-R	
CoOp	71.39	64.47	48.76	50.66	76.34	60.06 ±0.13
w/LDP	**72.03**	**65.04**	**48.93**	**50.98**	**76.44**	**60.35** ±0.09
MaPLe	72.48	64.92	47.72	46.55	75.16	58.59 ±0.09
w/LDP	**72.67**	**65.13**	**48.05**	**47.27**	**75.54**	**59.00** ±0.17
PromptSRC	72.57	65.51	49.25	49.24	77.46	60.37 ±0.20
w/LDP	**72.74**	**65.77**	**49.63**	**49.39**	**77.51**	**60.58** ±0.05
MMRL++	71.37	64.50	49.27	51.17	77.57	60.62 ±0.17
w/LDP	**71.67**	**64.77**	**49.60**	**51.53**	**77.93**	**60.96** ±0.05

Base-to-New Generalization. Table 1 reports the base-to-new results on 11 datasets. LDP consistently improves the average H score across all baselines. Specifically, it increases the average H score from 75.32% to 77.73% for CoOp and further improves the strongest baseline MMRL++ from 81.43% to 81.77%. This gain is non-trivial in the saturated few-shot base-to-new setting, where recent methods such as DPC (Li et al., 2025a) and NeRP (Tian et al., 2026) report comparable improvements of +0.32% and +0.48%, respectively. Moreover, LDP achieves these gains with lightweight latent domain bases, without substantial additional parameters or computational overhead. Beyond the numerical improvements, the results suggest that the learned latent domains capture intrinsic local data structures beyond simple semantic clustering. By modeling these structures, LDP provides a semantic bridge between base and new classes, enabling more effective transfer of knowledge to new classes.

Cross-Dataset Generalization. Table 2 reports cross-dataset results, with models trained on ImageNet and evaluated on 10 new target datasets. LDP improves the average accuracy across all baselines, increasing the strongest baseline MMRL++ from 67.33% to 67.98%. These gains suggest that LDP may mitigate overfitting to source-specific patterns by decomposing source knowledge into latent domain prompts. The resulting structured prompt bases enable more adaptive context synthesis under target-domain shifts, leading to more robust transfer across datasets.

Cross-Domain Generalization. Table 3 reports cross-domain results, with models trained on ImageNet and evaluated on four ImageNet variants under visual shifts. LDP consistently improves the robustness of all baselines. In particular, it improves the average accuracy of the strongest baseline MMRL++ from 60.62% to 60.96%, with gains across all target domains. Because ImageNet variants preserve category semantics while altering visual appearance, LDP constrains prompt learning around latent semantic anchors derived from class names, reducing reliance on source-specific visual shortcuts. This text-derived structure provides a stable semantic prior under visual shifts, helping preserve class discrimination as appearance statistics change.

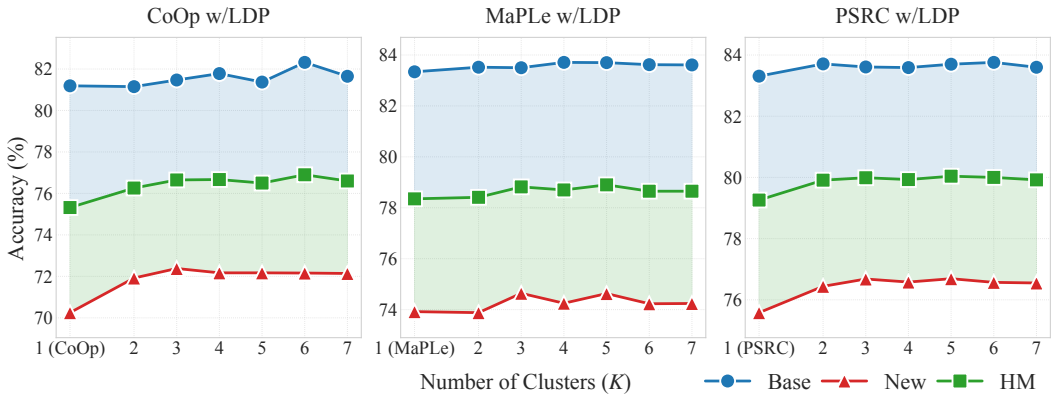


Figure 4: Ablation study on the number of latent domains K. While CoOp and PromptSRC benefit immediately from $K = 2$, MaPLe requires a finer granularity to effectively decouple disparate semantic domains.

5.3 Ablation Studies

Impact of Latent Domain Number K. We vary K to examine its effect on performance. As shown in Figure 4, introducing latent domains consistently improves over monolithic prompting, suggesting that latent domain bases better approximate complex semantic manifolds. CoOp and PromptSRC benefit from a coarse partition with $K = 2$, whereas MaPLe generally requires a finer decomposition because of its deeper cross-modal coupling. As K becomes too large, the gains tend to saturate or decline, indicating a bias–variance trade-off: excessive partitioning fragments the training data across latent domains and destabilizes prompt optimization.

Impact of Semantic Routing. We compare the original baselines with permuted affinity routing, nearest-neighbor matching, and affinity-weighted prompt synthesis. As shown in Table 4, permuting affinity vectors across classes reduces H by 0.30% and 0.52% for CoOp and MaPLe, respectively. PromptSRC is less sensitive to permutation and retains a marginal improvement, possibly because its self-regulating constraints preserve pre-trained knowledge and mitigate overfitting. Nevertheless, semantically aligned routing consistently outperforms permuted routing across all three baselines, with both nearest-neighbor matching and affinity-weighted prompt synthesis yielding higher H scores. These results show that LDP benefits not merely from additional prompt capacity, but from the correct correspondence between classes and latent domain bases.

Table 4: Semantic routing strategies. Ori., Perm., and NN. denote the original baseline, LDP with class-permuted affinity vectors, and nearest-neighbor matching, respectively.

Method	Setting	Base	New	H
CoOp w/LDP	Ori.	81.19	70.25	75.32
	Perm.	81.56	69.46	75.02
	NN.	81.40	71.65	76.22
	$\lambda = 0.1$	81.15	71.92	76.26
	$\lambda = 1$	81.36	72.33	76.58
	$\lambda = 10$	81.28	72.48	76.63
MaPLe w/LDP	Ori.	83.34	73.92	78.35
	Perm.	83.10	73.19	77.83
	NN.	83.51	73.81	78.36
	$\lambda = 0.1$	83.52	73.88	78.41
	$\lambda = 1$	83.33	74.01	78.39
	$\lambda = 10$	83.51	74.03	78.49
PromptSRC w/LDP	Ori.	83.31	75.58	79.26
	Perm.	83.41	75.77	79.40
	NN.	83.43	75.94	79.51
	$\lambda = 0.1$	83.71	76.44	79.91
	$\lambda = 1$	83.64	76.63	79.98
	$\lambda = 10$	83.44	76.77	79.97

6 Conclusion

We identify the monolithic assumption as a structural source of the Base–New Trade-off and propose LDP, which decomposes the class semantic space into latent domains and synthesizes class-adaptive prompts through affinity weighting. Our analysis formalizes the shift from a base-supervised global compromise to local prompt mixtures; under approximate barycentric reconstruction, the geometry-dependent LDP term reduces to a second-order local moment. Experiments across base-to-new, cross-dataset, and cross-domain settings consistently demonstrate the effectiveness of LDP.

AI USE STATEMENT

In this work, we used generative AI tools for correcting grammar, improving readability, and refining the clarity of sentences, LaTeX formatting, and manuscript organization. We have not used generative AI tools for developing theoretical models or conceptual frameworks, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for AI tools with English-language editing. We have reviewed all AI-assisted work. Specifically, we checked translations and revised text for technical accuracy. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of LDP, the datasets, data splits, training settings, evaluation metrics, and dataset-specific values of K in the paper and supplementary material. The source code, configuration files, and experiment scripts will be released under the MIT license upon publication.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, pp. 446–461. Springer International Publishing, 2014.

Abhiroop Chatterjee, Susmita Ghosh, Ashish Ghosh, and Emmett Ientilucci. Caspa: Graph-structured concept anchors for modality-agnostic adaptation in vision-language models. In *CVPR*, pp. 31566–31576, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. DyTox: Transformers for continual learning with dynamic token eXpansion. In *CVPR*, pp. 9275–9285, 2022.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pp. 178–178, 2004.

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. *IJCV*, 132 (2):581–595, 2024.

Yuncheng Guo and Xiaodong Gu. MMRL++: Parameter-Efficient and Interaction-Aware Representation Learning for Vision-Language Models. *IJCV*, 134(1):11, 2025.

Fusheng Hao, Fengxiang He, Fuxiang Wu, Tichao Wang, Chengqun Song, and Jun Cheng. Task-aware clustering for prompting vision-language models. In *CVPR*, pp. 14745–14755, 2025.

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.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8320–8329, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pp. 15257–15266, 2021b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Cong Hua, Qianqian Xu, Zhiyong Yang, Zitai Wang, Shilong Bao, and Qingming Huang. OpenworldAUC: Towards unified evaluation and optimization for open-world prompt tuning. In *ICML*, volume 267, pp. 24975–25020. PMLR, 2025.

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 *ICML*, volume 139, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pp. 709–727. Springer Nature Switzerland, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, pp. 19113–19122, 2023a.

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 *ICCV*, pp. 15144–15154, 2023b.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCVW*, pp. 554–561, 2013.

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. DPC: Dual-prompt collaboration for tuning vision-language models. In *CVPR*, pp. 25623–25632, 2025a.

Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *ICCV*, pp. 3618–3627, 2025b.

Zhixing Li, Arsham Gholamzadeh Khoee, and Yinan Yu. Latent domain prompt learning for vision-language models. In *ICASSP*, pp. 12787–12791, 2026.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *ICLR*, 2019.

J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, 1967.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013.

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*, pp. 722–729, 2008.

Jinyoung Park, Juyeon Ko, and Hyunwoo J. Kim. Prompt learning via meta-regularization. In *CVPR*, pp. 26930–26940, 2024.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505, 2012.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, and Yuan Xie. Prompt gradient projection for continual learning. In *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, volume 97, pp. 5389–5400. PMLR, 2019.

Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012.

Senmao Tian, Xiang Wei, and Shunli Zhang. Neutral-reference prompting for vision-language models. In *ICML*, 2026.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. ArGue: Attribute-guided prompt tuning for vision-language models. In *CVPR*, pp. 28578–28587, 2024.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In *NeurIPS*, volume 35, pp. 5682–5695. Curran Associates, Inc., 2022a.

Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. *AAAI*, 38(6):5749–5757, 2024.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pp. 631–648. Springer Nature Switzerland, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pp. 139–149, 2022c.

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 *CVPR*, pp. 14723–14732, 2025.

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*, pp. 3485–3492, 2010.

Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive Visual Prompt Learning with Contrastive Feature Re-formation. *IJCV*, 133(2):511–526, 2025.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. MMA: Multi-modal adapter for vision-language models. In *CVPR*, pp. 23826–23837, 2024.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma++: Effective multi-modal adaptation for vision-language models. *IEEE TPAMI*, pp. 1–15, 2026.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pp. 6757–6767, 2023.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. DePT: Decoupled prompt tuning. In *CVPR*, pp. 12924–12933, 2024.

Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. Large language models are good prompt learners for low-shot image classification. In *CVPR*, pp. 28453–28462, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pp. 16795–16804, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *IJCV*, 130(9):2337–2348, 2022b.

Yang Zhou, Yongjian Wu, Jiya Saiyin, Bingzheng Wei, and Yan Xu. Sdpt: Synchronous dual prompt tuning for visual-language pre-trained models. *IEEE TPAMI*, pp. 1–15, 2026.

A Proofs and Theoretical Clarifications

A.1 Proof of Characterization of Monolithic Prompt Approximation Error

Theorem 1 (restated). Under Assumptions 1–3, for a textual representation $\mathbf{t} \in \mathcal{U}$, let $\mathbf{\Delta}(\mathbf{t}) = \mathbf{t} - \mathbf{c}_b$. Then

$$\mathbf{P} - \Phi^\star(\mathbf{t}) = -J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t}) + \boldsymbol{\xi}_m(\mathbf{t}), \tag{A.1}$$

where

$$\|\boldsymbol{\xi}_m(\mathbf{t})\|_F \le \varepsilon_m + CV_b + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2. \tag{A.2}$$

Consequently, with $\mathcal{E}_m(\mathbf{t}) = \|\mathbf{P} - \Phi^\star(\mathbf{t})\|_F$, we have

$$\left|\mathcal{E}_m(\mathbf{t}) - \|J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t})\|_F\right| \le \varepsilon_m + CV_b + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2. \tag{A.3}$$

In particular,

$$\mathcal{E}_m(\mathbf{t}) \le \varepsilon_m + CV_b + \|J_{\Phi^\star}(\mathbf{c}_b)\|\|\mathbf{\Delta}(\mathbf{t})\|_2 + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2. \tag{A.4}$$

Proof. Because $\mathcal{U}$ is convex and contains all base-class representations, $\mathbf{c}_b \in \mathcal{U}$. Expanding each base-class oracle prompt around $\mathbf{c}_b$ gives

$$\Phi^\star(\mathbf{t}_c) = \Phi^\star(\mathbf{c}_b) + J_{\Phi^\star}(\mathbf{c}_b)(\mathbf{t}_c - \mathbf{c}_b) + \mathbf{R}(\mathbf{t}_c, \mathbf{c}_b). \tag{A.5}$$

Averaging Eq. (A.5) over $c \in \mathcal{C}_b$ yields

$$\mathbf{P}_b^\circ = \Phi^\star(\mathbf{c}_b) + J_{\Phi^\star}(\mathbf{c}_b)\left(\frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}(\mathbf{t}_c - \mathbf{c}_b)\right) + \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\mathbf{R}(\mathbf{t}_c, \mathbf{c}_b). \tag{A.6}$$

By the definition of $\mathbf{c}_b$,

$$\frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}(\mathbf{t}_c - \mathbf{c}_b) = \mathbf{0}. \tag{A.7}$$

Hence, the first-order term in Eq. (A.6) cancels, giving

$$\mathbf{P}_b^\circ - \Phi^\star(\mathbf{c}_b) = \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\mathbf{R}(\mathbf{t}_c, \mathbf{c}_b). \tag{A.8}$$

Using the triangle inequality and the quadratic remainder bound, we obtain

$$\|\mathbf{P}_b^\circ - \Phi^\star(\mathbf{c}_b)\|_F \le \frac{1}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\|\mathbf{R}(\mathbf{t}_c, \mathbf{c}_b)\|_F \le \frac{C}{|\mathcal{C}_b|}\sum_{c\in\mathcal{C}_b}\|\mathbf{t}_c - \mathbf{c}_b\|_2^2 = CV_b. \tag{A.9}$$

Next, applying the Local Smoothness assumption to $(\mathbf{t}, \mathbf{c}_b)$ gives

$$\Phi^\star(\mathbf{t}) = \Phi^\star(\mathbf{c}_b) + J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t}) + \mathbf{R}(\mathbf{t}, \mathbf{c}_b). \tag{A.10}$$

Adding and subtracting $\mathbf{P}_b^\circ$ and $\Phi^\star(\mathbf{c}_b)$ yields

$$\mathbf{P} - \Phi^\star(\mathbf{t}) = -J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t}) + (\mathbf{P} - \mathbf{P}_b^\circ) + (\mathbf{P}_b^\circ - \Phi^\star(\mathbf{c}_b)) - \mathbf{R}(\mathbf{t}, \mathbf{c}_b). \tag{A.11}$$

Define

$$\boldsymbol{\xi}_m(\mathbf{t}) \coloneqq (\mathbf{P} - \mathbf{P}_b^\circ) + (\mathbf{P}_b^\circ - \Phi^\star(\mathbf{c}_b)) - \mathbf{R}(\mathbf{t}, \mathbf{c}_b). \tag{A.12}$$

Combining Eqs. (A.11) and (A.12) yields Eq. (A.1).

By the Monolithic Fitting Consistency assumption, Eq. (A.9), and the Local Smoothness assumption,

$$\begin{aligned}\|\boldsymbol{\xi}_m(\mathbf{t})\|_F &\le \|\mathbf{P} - \mathbf{P}_b^\circ\|_F + \|\mathbf{P}_b^\circ - \Phi^\star(\mathbf{c}_b)\|_F + \|\mathbf{R}(\mathbf{t}, \mathbf{c}_b)\|_F \\ &\le \varepsilon_m + CV_b + C\|\mathbf{t} - \mathbf{c}_b\|_2^2 \\ &= \varepsilon_m + CV_b + C\|\mathbf{\Delta}(\mathbf{t})\|_2^2.\end{aligned} \tag{A.13}$$

This proves Eq. (A.2).

Finally, the reverse triangle inequality applied to Eq. (A.1) gives

$$\left|\mathcal{E}_m(\mathbf{t}) - \|J_{\Phi^\star}(\mathbf{c}_b)\mathbf{\Delta}(\mathbf{t})\|_F\right| \le \|\boldsymbol{\xi}_m(\mathbf{t})\|_F, \tag{A.14}$$

which, together with Eq. (A.2), proves Eq. (A.3).

Similarly, the triangle inequality and the definition of the operator norm give

$$\begin{aligned}\mathcal{E}_m(\mathbf{t}) &\le \|J_{\Phi^\star}(\mathbf{c}_b)\boldsymbol{\Delta}(\mathbf{t})\|_F + \|\boldsymbol{\xi}_m(\mathbf{t})\|_F \\ &\le \|J_{\Phi^\star}(\mathbf{c}_b)\|\|\boldsymbol{\Delta}(\mathbf{t})\|_2 + \varepsilon_m + CV_b + C\|\boldsymbol{\Delta}(\mathbf{t})\|_2^2,\end{aligned} \tag{A.15}$$

which proves Eq. (A.4). □

Note: For a linear map $\mathbf{L} : \mathbb{R}^D \to \mathcal{H}$, $\|\mathbf{L}\|$ denotes the operator norm induced by the Euclidean norm on $\mathbb{R}^D$ and the Frobenius norm on $\mathcal{H}$:

$$\|\mathbf{L}\| \coloneqq \sup_{\mathbf{z}\neq\mathbf{0}} \frac{\|\mathbf{L}\mathbf{z}\|_F}{\|\mathbf{z}\|_2}. \tag{A.16}$$

A.2 Proof of Full-Vocabulary Coverage Rebalancing

Proposition 1 (restated). Let $\mathcal{C} = \mathcal{C}_b \dot\cup \mathcal{C}_n$, with both subsets nonempty. For the same number of anchors and the same distance metric, let $\mathbf{A}_b^\star \in \arg\min_{\mathbf{A}} \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A})$ and $\mathbf{A}^\star \in \arg\min_{\mathbf{A}} \mathcal{Q}_{\mathcal{C}}(\mathbf{A})$ be global minimizers of the base-vocabulary and full-vocabulary quantization risks, respectively. Then

$$\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}_b^\star) \le \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}^\star), \qquad \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}^\star) \le \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}_b^\star). \tag{A.17}$$

Proof. Let

$$\alpha = \frac{|\mathcal{C}_b|}{|\mathcal{C}|}, \qquad \beta = \frac{|\mathcal{C}_n|}{|\mathcal{C}|}. \tag{A.18}$$

For any anchor set $\mathbf{A}$,

$$\mathcal{Q}_{\mathcal{C}}(\mathbf{A}) = \alpha\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}) + \beta\mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}). \tag{A.19}$$

Because $\mathbf{A}_b^\star$ minimizes the base-vocabulary risk,

$$\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}_b^\star) \le \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}^\star). \tag{A.20}$$

Because $\mathbf{A}^\star$ minimizes the full-vocabulary risk,

$$\alpha\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}^\star) + \beta\mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}^\star) \le \alpha\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}_b^\star) + \beta\mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}_b^\star). \tag{A.21}$$

Rearranging gives

$$\beta\big(\mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}^\star) - \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}_b^\star)\big) \le -\alpha\big(\mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}^\star) - \mathcal{Q}_{\mathcal{C}_b}(\mathbf{A}_b^\star)\big) \le 0. \tag{A.22}$$

Since $\beta > 0$,

$$\mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}^\star) \le \mathcal{Q}_{\mathcal{C}_n}(\mathbf{A}_b^\star). \tag{A.23}$$

Together with the base-vocabulary optimality inequality, this proves Eq. (A.17). □

A.3 Proof of Affinity Weighted Gradient Sharing

Proposition 2 (restated). During training, the semantic affinities $\pi_{c,k}$ are fixed with respect to the prompt parameters. For a mini-batch $\mathcal{X}$ sampled from the base-class training set, the gradient of the loss with respect to the k-th latent domain base satisfies

$$\frac{\partial\mathcal{L}_{\mathcal{X}}}{\partial\mathbf{b}_k} = \sum_{(\mathbf{x},y)\in\mathcal{X}} \sum_{c\in\mathcal{C}_b} \pi_{c,k} \frac{\partial\ell(\mathbf{x},y)}{\partial\mathbf{p}_d(\mathbf{t}_c)}. \tag{A.24}$$

Note: $\mathcal{L}_{\mathcal{X}}$ denotes the summed mini-batch loss:

$$\mathcal{L}_{\mathcal{X}} \coloneqq \sum_{(\mathbf{x},y)\in\mathcal{X}} \ell(\mathbf{x},y). \tag{A.25}$$

If an averaged mini-batch loss is used in implementation, both sides of the gradient identity are scaled by $1/|\mathcal{X}|$, leaving the gradient-sharing mechanism unchanged.

Proof. For each base class c, the corresponding latent domain prompt is synthesized as

$$\mathbf{p}_d(\mathbf{t}_c) = \sum_{j=1}^{K} \pi_{c,j} \mathbf{b}_j. \tag{A.26}$$

Because the affinities $\pi_{c,j}$ depend only on the fixed textual representations and semantic anchors, they are constant with respect to the learnable prompt bases. Therefore,

$$\frac{\partial \mathbf{p}_d(\mathbf{t}_c)}{\partial \mathbf{b}_k} = \pi_{c,k} \mathbf{I}, \tag{A.27}$$

where $\mathbf{I}$ denotes the identity operator on the latent prompt space. For a training example $(\mathbf{x}, y)$, the classification loss $\ell(\mathbf{x}, y)$ depends on the class-specific prompts for all $c \in \mathcal{C}_b$. Applying the chain rule gives

$$\frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{b}_k} = \sum_{c \in \mathcal{C}_b} \frac{\partial \mathbf{p}_d(\mathbf{t}_c)}{\partial \mathbf{b}_k} \frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{p}_d(\mathbf{t}_c)} = \sum_{c \in \mathcal{C}_b} \pi_{c,k} \frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{p}_d(\mathbf{t}_c)}. \tag{A.28}$$

Summing the per-example gradients over the mini-batch $\mathcal{X}$ yields

$$\frac{\partial \mathcal{L}_{\mathcal{X}}}{\partial \mathbf{b}_k} = \sum_{(\mathbf{x},y) \in \mathcal{X}} \frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{b}_k} = \sum_{(\mathbf{x},y) \in \mathcal{X}} \sum_{c \in \mathcal{C}_b} \pi_{c,k} \frac{\partial \ell(\mathbf{x}, y)}{\partial \mathbf{p}_d(\mathbf{t}_c)}, \tag{A.29}$$

which proves Eq. (A.24). □

A.4 Proof of Local-Mixture Expansion for LDP

Theorem 2 (restated). Under Assumptions 1 and 3, for any $\mathbf{t} \in \mathcal{U}$ whose positively weighted anchors also lie in $\mathcal{U}$,

$$\mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t}) = \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{e}_k + J_{\Phi^\star}(\mathbf{t}) \mathbf{r}_1(\mathbf{t}) + \boldsymbol{\rho}_d(\mathbf{t}), \tag{A.30}$$

where

$$\|\boldsymbol{\rho}_d(\mathbf{t})\|_F \le C m_2(\mathbf{t}). \tag{A.31}$$

Consequently,

$$\mathcal{E}_d(\mathbf{t}) \coloneqq \|\mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t})\|_F \le \delta(\mathbf{t}) + \|J_{\Phi^\star}(\mathbf{t})\| \|\mathbf{r}_1(\mathbf{t})\|_2 + C m_2(\mathbf{t}). \tag{A.32}$$

Proof. Because $\mathbf{t} \in \mathcal{U}$ and every anchor with $\pi_k(\mathbf{t}) > 0$ also belongs to $\mathcal{U}$, Assumption 3 can be applied to each pair $(\mathbf{a}_k^\star, \mathbf{t})$. Therefore,

$$\Phi^\star(\mathbf{a}_k^\star) = \Phi^\star(\mathbf{t}) + J_{\Phi^\star}(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) + \mathbf{R}(\mathbf{a}_k^\star, \mathbf{t}), \tag{A.33}$$

where

$$\|\mathbf{R}(\mathbf{a}_k^\star, \mathbf{t})\|_F \le C \|\mathbf{a}_k^\star - \mathbf{t}\|_2^2. \tag{A.34}$$

By the definition of $\mathbf{e}_k$,

$$\mathbf{Q}_k = \Phi^\star(\mathbf{a}_k^\star) + \mathbf{e}_k. \tag{A.35}$$

Substituting Eq. (A.35) into the definition of $\mathbf{P}_d(\mathbf{t})$ gives

$$\begin{aligned} \mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t}) &= \sum_{k=1}^{K} \pi_k(\mathbf{t}) \left[\Phi^\star(\mathbf{a}_k^\star) + \mathbf{e}_k\right] - \Phi^\star(\mathbf{t}) \\ &= \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{e}_k + \sum_{k=1}^{K} \pi_k(\mathbf{t}) \left[\Phi^\star(\mathbf{a}_k^\star) - \Phi^\star(\mathbf{t})\right]. \end{aligned} \tag{A.36}$$

The second equality uses $\sum_{k=1}^{K} \pi_k(\mathbf{t}) = 1$.

Substituting Eq. (A.33) into Eq. (A.36) yields

$$\mathbf{P}_d(\mathbf{t}) - \Phi^\star(\mathbf{t}) = \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{e}_k + \sum_{k=1}^{K} \pi_k(\mathbf{t}) J_{\Phi^\star}(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) + \sum_{k=1}^{K} \pi_k(\mathbf{t}) \mathbf{R}(\mathbf{a}_k^\star, \mathbf{t}). \tag{A.37}$$

Because the Jacobian is linear and does not depend on k,

$$\sum_{k=1}^{K} \pi_k(\mathbf{t}) J_{\Phi^\star}(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) = J_{\Phi^\star}(\mathbf{t}) \sum_{k=1}^{K} \pi_k(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) = J_{\Phi^\star}(\mathbf{t})\mathbf{r}_1(\mathbf{t}), \tag{A.38}$$

where the normalization of the affinities gives

$$\sum_{k=1}^{K} \pi_k(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) = \sum_{k=1}^{K} \pi_k(\mathbf{t})\mathbf{a}_k^\star - \left(\sum_{k=1}^{K} \pi_k(\mathbf{t})\right)\mathbf{t} = \sum_{k=1}^{K} \pi_k(\mathbf{t})\mathbf{a}_k^\star - \mathbf{t} = \mathbf{r}_1(\mathbf{t}). \tag{A.39}$$

Define the affinity-weighted Taylor remainder as

$$\boldsymbol{\rho}_d(\mathbf{t}) \coloneqq \sum_{k=1}^{K} \pi_k(\mathbf{t})\mathbf{R}(\mathbf{a}_k^\star, \mathbf{t}). \tag{A.40}$$

Combining Eqs. (A.37), (A.38), and (A.40) yields Eq. (A.30).

Because the affinities are nonnegative and normalized, the triangle inequality and Eq. (A.34) give

$$\|\boldsymbol{\rho}_d(\mathbf{t})\|_F \le \sum_{k=1}^{K} \pi_k(\mathbf{t})\|\mathbf{R}(\mathbf{a}_k^\star, \mathbf{t})\|_F \le C \sum_{k=1}^{K} \pi_k(\mathbf{t})\|\mathbf{a}_k^\star - \mathbf{t}\|_2^2 = C m_2(\mathbf{t}), \tag{A.41}$$

which proves Eq. (A.31).

Finally, applying the triangle inequality to Eq. (A.30) gives

$$\begin{aligned} \mathcal{E}_d(\mathbf{t}) &\le \left\| \sum_{k=1}^{K} \pi_k(\mathbf{t})\mathbf{e}_k \right\|_F + \|J_{\Phi^\star}(\mathbf{t})\mathbf{r}_1(\mathbf{t})\|_F + \|\boldsymbol{\rho}_d(\mathbf{t})\|_F \\ &\le \delta(\mathbf{t}) + \|J_{\Phi^\star}(\mathbf{t})\|\|\mathbf{r}_1(\mathbf{t})\|_2 + C m_2(\mathbf{t}), \end{aligned} \tag{A.42}$$

where the second inequality follows from the definition of the operator norm and Eq. (A.31). This proves Eq. (A.32). □

A.5 Proof of Second-Order Local Geometry Bound

Corollary 1 (restated). Under the conditions of Theorem 2 and Assumption 4,

$$\mathcal{E}_d(\mathbf{t}) \le \delta(\mathbf{t}) + \big(\eta\|J_{\Phi^\star}(\mathbf{t})\| + C\big) m_2(\mathbf{t}). \tag{A.43}$$

Proof. By Theorem 2, the LDP approximation error satisfies

$$\mathcal{E}_d(\mathbf{t}) \le \delta(\mathbf{t}) + \|J_{\Phi^\star}(\mathbf{t})\|\|\mathbf{r}_1(\mathbf{t})\|_2 + C m_2(\mathbf{t}). \tag{A.44}$$

Applying the approximate barycentric reconstruction condition in Assumption 4 gives

$$\mathcal{E}_d(\mathbf{t}) \le \delta(\mathbf{t}) + \eta\|J_{\Phi^\star}(\mathbf{t})\| m_2(\mathbf{t}) + C m_2(\mathbf{t}) = \delta(\mathbf{t}) + \left(\eta\|J_{\Phi^\star}(\mathbf{t})\| + C\right) m_2(\mathbf{t}), \tag{A.45}$$

which proves Eq. (A.43). □

Remark A.1 (Role of Approximate Barycentric Reconstruction)**.** The approximate barycentric reconstruction condition is not implied by the distance-softmax affinities alone. Without this condition, the convexity of the squared Euclidean norm gives

$$\|\mathbf{r}_1(\mathbf{t})\|_2^2 = \left\| \sum_{k=1}^{K} \pi_k(\mathbf{t})(\mathbf{a}_k^\star - \mathbf{t}) \right\|_2^2 \le \sum_{k=1}^{K} \pi_k(\mathbf{t})\|\mathbf{a}_k^\star - \mathbf{t}\|_2^2 = m_2(\mathbf{t}). \tag{A.46}$$

Therefore,

$$\|\mathbf{r}_1(\mathbf{t})\|_2 \le \sqrt{m_2(\mathbf{t})}. \tag{A.47}$$

Combining this inequality with Theorem 2 gives the unconditional bound

$$\mathcal{E}_d(\mathbf{t}) \le \delta(\mathbf{t}) + \|J_{\Phi^\star}(\mathbf{t})\|\sqrt{m_2(\mathbf{t})} + C m_2(\mathbf{t}). \tag{A.48}$$

Thus, Theorem 2 remains valid without the approximate barycentric reconstruction condition, but the first-order reconstruction residual cannot generally be absorbed into the local second moment. Under Assumption 4, this residual becomes second-order in the local anchor radius, leading to Corollary 1. Equivalently, defining the affinity-weighted local radius as $h_\pi(\mathbf{t}) \coloneqq \sqrt{m_2(\mathbf{t})}$, the geometry-dependent contribution in Corollary 1 is $O(h_\pi(\mathbf{t})^2)$.

A.6 Proof of Proxy Approximation-Estimation Trade-off

The following statement makes explicit the quantities and the covering-number and effective-sample-size conditions summarized in Proposition 3; its conclusion is unchanged.

Proposition A.1 (Detailed Form of the Proxy Approximation-Estimation Trade-off)**.** Let $\mathbf{T}$ be a random textual representation supported on a nonempty compact set $\mathcal{M} \subset \mathbb{R}^D$. Let $\mathcal{N}(\mathcal{M}, r)$ denote the minimum number of closed Euclidean balls of radius r, with centers in $\mathbb{R}^D$, required to cover $\mathcal{M}$. Define the optimal K-anchor semantic quantization risk as

$$q_K \coloneqq \inf_{(\mathbf{a}_1,\dots,\mathbf{a}_K)\in(\mathbb{R}^D)^K} \mathbb{E}_{\mathbf{T}} \left[\min_{1\le k\le K} \|\mathbf{T}-\mathbf{a}_k\|_2^2 \right]. \tag{A.49}$$

Assume that there exist constants $A > 0$, $d > 0$, and $r_0 > 0$, independent of K and N, such that

$$\mathcal{N}(\mathcal{M}, r) \le A r^{-d}, \qquad 0 < r \le r_0. \tag{A.50}$$

Let N denote the total number of base-class training samples. For the k-th prompt prototype, let $\widehat{\mathbf{Q}}_k$ denote its learned estimate and $\mathbf{Q}_k^\circ$ its population reference. For the i-th training sample with class label y_i, define its fixed affinity weight for the k-th prototype as

$$w_{i,k} \coloneqq \pi_k(\mathbf{t}_{y_i}), \tag{A.51}$$

and define the corresponding soft effective sample size as

$$N_{\mathrm{eff},k} \coloneqq \frac{\left(\sum_{i=1}^N w_{i,k}\right)^2}{\sum_{i=1}^N w_{i,k}^2}. \tag{A.52}$$

Because the affinities depend only on the fixed textual representations and semantic anchors, the weights $w_{i,k}$ are fixed with respect to the prompt-estimation randomness.

Define the expected estimation error of the k-th prototype and its average affinity mass as

$$\epsilon_k \coloneqq \mathbb{E}_{\mathrm{tr}} \left[\left\| \widehat{\mathbf{Q}}_k - \mathbf{Q}_k^\circ \right\|_F \right], \qquad \rho_k \coloneqq \mathbb{E}_{\mathbf{T}} \left[\pi_k(\mathbf{T}) \right], \tag{A.53}$$

where $\mathbb{E}_{\mathrm{tr}}$ denotes expectation over the training and estimation randomness, conditional on the fixed affinities, and $\mathbf{T}$ is independent of this randomness. The affinity-averaged prototype estimation error is

$$e_K \coloneqq \sum_{k=1}^K \rho_k \epsilon_k. \tag{A.54}$$

Since the affinities are nonnegative and normalized, $\rho_k \ge 0$ and $\sum_{k=1}^K \rho_k = 1$.

Assume that there exist constants $G > 0$ and $\nu > 0$, independent of K and N, such that

$$\epsilon_k \le \frac{G}{\sqrt{N_{\mathrm{eff},k}}}, \qquad N_{\mathrm{eff},k} \ge \nu \frac{N}{K}, \qquad k = 1,\dots,K. \tag{A.55}$$

Then there exist constants $C_1, C_2 > 0$, independent of K and N, such that

$$q_K \le C_1 K^{-\frac{2}{d}}, \qquad e_K \le C_2 \sqrt{\frac{K}{N}}. \tag{A.56}$$

Consequently, for any fixed $\gamma > 0$, independent of K and N, the model-selection proxy

$$\mathcal{J}_p(K) \coloneqq q_K + \gamma e_K \tag{A.57}$$

satisfies

$$\mathcal{J}_p(K) \le C_1 K^{-\frac{2}{d}} + \gamma C_2 \sqrt{\frac{K}{N}}. \tag{A.58}$$

Proof. We first bound the semantic quantization risk q_K. Let

$$K_0 \coloneqq \max\left\{1, \lceil A r_0^{-d} \rceil\right\}. \tag{A.59}$$

For $K \geq K_0$, define

$$r_K \coloneqq \left(\frac{A}{K}\right)^{\frac{1}{d}}. \tag{A.60}$$

Then $r_K \leq r_0$ and, by Eq. (A.50),

$$\mathcal{N}(\mathcal{M}, r_K) \leq A r_K^{-d} = K. \tag{A.61}$$

Hence, there exist $m \leq K$ centers $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^D$ whose closed radius-r_K balls cover $\mathcal{M}$. Repeating arbitrary centers when $m < K$ gives a feasible K-tuple of anchors satisfying

$$\min_{1 \leq k \leq K} \|\mathbf{t} - \mathbf{a}_k\|_2 \leq r_K \qquad \text{for every } \mathbf{t} \in \mathcal{M}. \tag{A.62}$$

Therefore, by the definition of q_K,

$$q_K \leq \mathbb{E}_{\mathbf{T}}\left[\min_{1 \leq k \leq K} \|\mathbf{T} - \mathbf{a}_k\|_2^2\right] \leq r_K^2 = A^{\frac{2}{d}} K^{-\frac{2}{d}}, \qquad K \geq K_0. \tag{A.63}$$

Because $\mathcal{M}$ is nonempty and compact, its diameter

$$D_{\mathcal{M}} \coloneqq \sup_{\mathbf{u}, \mathbf{v} \in \mathcal{M}} \|\mathbf{u} - \mathbf{v}\|_2 \tag{A.64}$$

is finite. Choosing any $\mathbf{a}_0 \in \mathcal{M}$ and repeating it K times gives

$$q_K \leq \mathbb{E}_{\mathbf{T}} \|\mathbf{T} - \mathbf{a}_0\|_2^2 \leq D_{\mathcal{M}}^2 \tag{A.65}$$

for every $K \geq 1$. Thus, defining

$$C_1 \coloneqq \max\left\{A^{\frac{2}{d}}, D_{\mathcal{M}}^2 K_0^{\frac{2}{d}}\right\}, \tag{A.66}$$

we obtain

$$q_K \leq C_1 K^{-\frac{2}{d}} \tag{A.67}$$

for all $K \geq 1$.

We next bound the affinity-averaged prototype estimation error. By Eq. (A.55),

$$\epsilon_k \leq \frac{G}{\sqrt{\nu}} \sqrt{\frac{K}{N}} \qquad \text{for every } k. \tag{A.68}$$

Using $\rho_k \geq 0$ and $\sum_{k=1}^K \rho_k = 1$, we obtain

$$e_K = \sum_{k=1}^K \rho_k \epsilon_k \leq \frac{G}{\sqrt{\nu}} \sqrt{\frac{K}{N}} \sum_{k=1}^K \rho_k = \frac{G}{\sqrt{\nu}} \sqrt{\frac{K}{N}}. \tag{A.69}$$

Hence, the second inequality in Eq. (A.56) holds with

$$C_2 \coloneqq \frac{G}{\sqrt{\nu}}. \tag{A.70}$$

Finally, substituting the two component bounds into $\mathcal{J}_p(K) = q_K + \gamma e_K$ yields

$$\mathcal{J}_p(K) \leq C_1 K^{-\frac{2}{d}} + \gamma C_2 \sqrt{\frac{K}{N}}, \tag{A.71}$$

which proves Eq. (A.58). □

Remark A.2. The quantity e_K also upper-bounds the expected estimation error of the affinity-weighted prompt mixture. Indeed, because the affinities are fixed with respect to the training randomness and $\mathbf{T}$ is independent of that randomness, the triangle inequality gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{T}}\mathbb{E}_{\mathrm{tr}}\left[\left\|\sum_{k=1}^{K}\pi_k(\mathbf{T})\left(\widehat{\mathbf{Q}}_k-\mathbf{Q}_k^{\circ}\right)\right\|_F\right]\\
&\le \mathbb{E}_{\mathbf{T}}\sum_{k=1}^{K}\pi_k(\mathbf{T})\mathbb{E}_{\mathrm{tr}}\left[\left\|\widehat{\mathbf{Q}}_k-\mathbf{Q}_k^{\circ}\right\|_F\right] \qquad \text{(A.72)}\\
&=\sum_{k=1}^{K}\rho_k\epsilon_k = e_K.
\end{aligned}$$

The proxy $\mathcal{J}_p(K) = q_K + \gamma e_K$ combines semantic resolution and prototype estimation. The quantity q_K is defined through nearest-anchor quantization, whereas $m_2(\mathbf{t})$ is the soft-affinity local moment used in the local-mixture analysis. Under Eq. (A.55), the prototype estimation term follows the stated rate in terms of the weight-based effective sample size $N_{\mathrm{eff},k}$. Thus, $\mathcal{J}_p(K)$ provides a theoretical rationale for balancing semantic resolution against estimation difficulty. In practice, K is selected through validation, which determines the realized empirical performance.

B Additional Experimental Details and Results

B.1 Experiment Settings

Table 5: Summary of the datasets used for base-to-new, cross-dataset, and cross-domain generalization. The table reports dataset statistics; the upper block lists the 11 datasets used for base-to-new and cross-dataset generalization, while the lower block lists the ImageNet variants used for cross-domain generalization.

Dataset	Classes	Train	Val	Test
Base-to-New / Cross-Dataset Generalization				
ImageNet	1,000	1,281,167	N/A	50,000
Caltech101	100	4,128	1,649	2,465
EuroSAT	10	13,500	5,400	8,100
SUN397	397	15,880	3,970	19,850
DTD	47	2,820	1,128	1,692
UCF101	101	7,639	1,898	3,783
FGVC Aircraft	100	3,334	3,333	
OxfordPets	37	2,944	736	3,669
StanfordCars	196	6,509	1,635	8,041
Flowers102	102	4,093	1,633	2,463
Food101	101	50,500	20,200	30,300
Cross-Domain Generalization				
ImageNet-V2	1,000	N/A	N/A	10,000
ImageNet-Sketch	1,000	N/A	N/A	50,889
ImageNet-A	200	N/A	N/A	7,500
ImageNet-R	200	N/A	N/A	30,000

Table 6: Summary of the datasets used for base-to-new, cross-dataset, and cross-domain generalization. The table reports dataset tasks and textual templates; the upper block lists the 11 datasets used for base-to-new and cross-dataset generalization, while the lower block lists the ImageNet variants used for cross-domain generalization.

Dataset	Task	Textual Template
Base-to-New / Cross-Dataset Generalization		
ImageNet	General object classification	a photo of a [CLASS].
Caltech101	General object classification	a photo of a [CLASS].
EuroSAT	Satellite image classification	a centered satellite photo of [CLASS].
SUN397	Scene image classification	a photo of a [CLASS].
DTD	Texture pattern classification	[CLASS] texture.
UCF101	Human action classification	a photo of a person doing [CLASS].
FGVC Aircraft	Aircraft type classification	a photo of a [CLASS], a type of aircraft.
OxfordPets	Pet breed classification	a photo of a [CLASS], a type of pet.
StanfordCars	Vehicle model classification	a photo of a [CLASS].
Flowers102	Flower species classification	a photo of a [CLASS], a type of flower.
Food101	Food category classification	a photo of a [CLASS], a type of food.
Cross-Domain Generalization		
ImageNet-V2	Resampling distribution shift	a photo of a [CLASS].
ImageNet-Sketch	Sketch style shift	a sketch of a [CLASS].
ImageNet-A	Natural adversarial shift	a photo of a [CLASS].
ImageNet-R	Rendition style shift	a rendition of a [CLASS].

Datasets. Following CoCoOp (Zhou et al., 2022a), we evaluate our method under base-to-new generalization, cross-dataset generalization, and cross-domain generalization settings. Tables 5 and 6 summarize the dataset statistics, tasks, and textual templates. For base-to-new and cross-dataset evaluation, we use a comprehensive suite of 11 datasets, including ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVCAir-

craft (Maji et al., 2013), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012), DTD (Cimpoi et al., 2014), and EuroSAT (Helber et al., 2019). These datasets cover diverse recognition scenarios, such as generic objects, fine-grained categories, scenes, textures, food, actions, and remote-sensing imagery. For cross-domain generalization, we use ImageNet (Deng et al., 2009) as the source domain for training and evaluate on 4 ImageNet variants with distinct visual shifts: ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021a), and ImageNet-R (Hendrycks et al., 2021b). These target domains introduce distribution changes caused by resampling, sketch-style abstraction, naturally hard examples, and artistic renditions, respectively, allowing us to assess the robustness of learned prompts under unseen domain shifts.

Table 7: Dataset-specific numbers of latent domains (K) selected by the proposed latent domain number selection mechanism.

Dataset	IN	C101	Pets	Cars	Flowers	Food	FGVCA	SUN397	DTD	EuroSAT	UCF101
K	6	5	4	5	3	7	4	6	3	7	5

Implementation Details. For LDP, the number of latent domains is selected from $K \in \{2, 3, 4, 5, 6, 7\}$. The final dataset-specific values of K are reported in Table 7. We determine K for each dataset using the latent domain number selection mechanism. Unless otherwise stated, the temperature parameter is fixed to $\lambda = 0.1$. We additionally evaluate the sensitivity to λ in the ablation study, including nearest neighbor matching and $\lambda \in \{0.1, 1, 10\}$. The main experiments are conducted on NVIDIA RTX 3090 and RTX 4090 GPUs. The source code, configuration files, and scripts for reproducing the experiments will be made publicly available upon publication under the MIT license.

Metrics. Following prior prompt learning works (Zhou et al., 2022a), we report the classification accuracy on base classes Acc_b, the classification accuracy on new classes Acc_n, and their harmonic mean (H). The H score is defined as

$$\mathrm{H} = \frac{2 \cdot \mathrm{Acc}_b \cdot \mathrm{Acc}_n}{\mathrm{Acc}_b + \mathrm{Acc}_n}. \tag{B.1}$$

Compared with the arithmetic mean, the harmonic mean assigns a lower score when either base-class accuracy or new-class accuracy is poor. Therefore, it is well suited for evaluating the Base–New Trade-off: a method can achieve a high H score only when it maintains strong performance on base classes while also generalizing well to new classes.

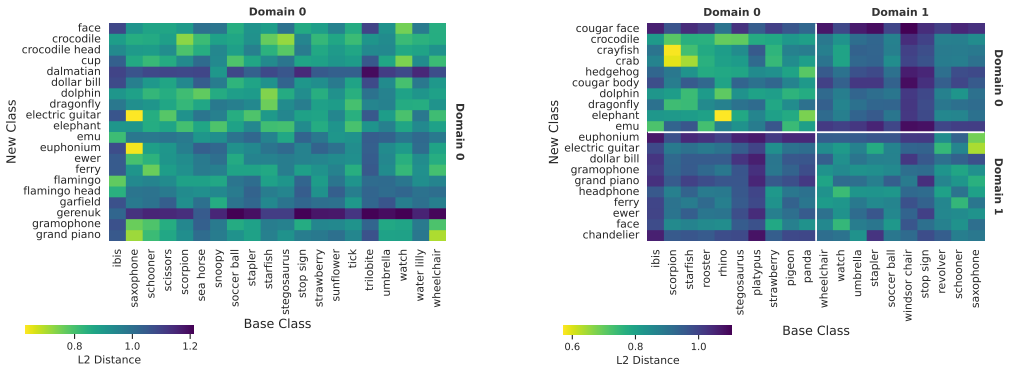
B.2 Experimental Results

Table 8: Qualitative interpretation of latent domains on Caltech101. We report representative class labels assigned to each latent domain obtained from frozen class-name representations when $K = 5$. The domain indices are arbitrary identifiers, but the assignments show semantically coherent groupings that broadly separate natural categories from man-made artifacts.

Domain ID	Representative Classes	Inferred Semantics
0	*beaver, brontosaurus, cougar body, stegosaurus, wild cat*	Wild and prehistoric animals
1	*gerenuk, okapi*	African ungulates
2	*ant, butterfly, crab, crayfish, dragonfly, trilobite, water lilly*	Invertebrates and plants
3	*accordion, chair, chandelier, wheelchair, windsor chair*	Furniture, interior objects, and instruments
4	*airplane, anchor, barrel, bass, binocular, wrench, yin yang*	General artifacts and symbols

Semantic Interpretability. To examine whether LDP captures meaningful semantic structures beyond arbitrary partitions, we conduct a qualitative case study on Caltech101 (Fei-Fei et al., 2004).

Table 8 summarizes representative classes assigned to each latent domain obtained from frozen class-name representations when $K = 5$. Although the domain indices are only cluster identifiers, the assignments exhibit coherent semantic tendencies. Natural categories are mainly grouped into organism-related domains, including wild and prehistoric animals, African ungulates, and invertebrates or plants. In contrast, man-made categories are mostly separated into furniture, interior objects, instruments and more general artifacts or symbols. This qualitative pattern suggests that LDP can recover local semantic regularities from class-name representations, providing qualitative support for the use of domain-specific prompts for transfer rather than relying on a single global prompt.



(a) CoOp (b) LDP

Figure 5: Qualitative comparison of prompt-induced textual feature distances between base and new classes. The displayed subset illustrates differences in local geometry between CoOp and LDP. Darker colors indicate larger ℓ_2 distance within each panel.

Qualitative Analysis. Figure 5 compares the pairwise ℓ_2 distance between prompt-induced textual representations of base and new classes. A smaller distance indicates stronger semantic proximity. In CoOp Zhou et al. (2022b), the learned global prompt produces entangled relations: semantically unrelated pairs, such as *crocodile* and *saxophone*, or *elephant* and *watch*, can become undesirably close. Meanwhile, related categories such as *garfield* and *ibis* are not clearly emphasized. In contrast, LDP forms a clearer block structure. Animal-related new classes such as *crocodile* and *elephant* are closer to related base classes such as *rhino* and *panda*, while instrument or object classes such as *euphonium* and *grand piano* are separated from animal categories and aligned with *saxophone*. These examples suggest that LDP mitigates the semantic entanglement caused by a monolithic prompt and better preserves local base-new semantic neighborhoods.

Base-to-New Generalization. Table 9 provides the detailed base-to-new generalization results, where each entry is reported as the mean and standard deviation over three seeds. The paired comparison uses the same few-shot protocol, data splits, and training configurations as the corresponding baselines. As shown in the average results, LDP consistently improves the harmonic mean (H) across CoOp (Zhou et al., 2022b), MaPLe (Khattak et al., 2023a), PromptSRC (Khattak et al., 2023b), and MMRL++ (Guo & Gu, 2025).

Table 9: Comparison of base-to-new generalization on 11 datasets. We report the accuracy on base classes, new classes, and their harmonic mean (H), along with standard deviations over three seeds. LDP improves the average H score for each baseline.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	81.19 ±0.38	70.25 ±0.92	75.32 ±0.54	76.23 ±0.09	69.74 ±0.28	72.84 ±0.17	97.93 ±0.14	93.56 ±0.64	95.69 ±0.27	95.22 ±0.30	97.54 ±0.21	96.36 ±0.25
w/LDP	82.13 ±0.25	73.77 ±0.74	**77.73** ±0.47	76.86 ±0.07	70.07 ±0.16	**73.31** ±0.06	97.98 ±0.21	95.38 ±0.28	**96.66** ±0.05	95.78 ±0.09	97.52 ±0.35	**96.64** ±0.20
MaPLe	83.34 ±0.03	73.92 ±0.31	78.35 ±0.18	77.02 ±0.10	69.39 ±0.17	73.00 ±0.13	98.30 ±0.06	94.76 ±0.27	**96.50** ±0.14	95.09 ±0.26	97.39 ±0.46	96.23 ±0.33
w/LDP	83.56 ±0.13	74.72 ±1.00	**78.89** ±0.50	77.21 ±0.16	69.29 ±0.25	**73.04** ±0.11	98.09 ±0.12	94.61 ±0.36	96.32 ±0.15	95.66 ±0.41	97.39 ±0.14	**96.52** ±0.24
PromptSRC	83.31 ±0.13	75.58 ±0.58	79.26 ±0.32	77.36 ±0.07	70.66 ±0.19	73.86 ±0.07	97.83 ±0.13	94.43 ±0.16	96.10 ±0.10	95.34 ±0.18	97.22 ±0.33	96.27 ±0.07
w/LDP	83.67 ±0.10	76.65 ±0.51	**80.01** ±0.23	77.50 ±0.07	70.79 ±0.28	**73.99** ±0.13	98.06 ±0.19	94.43 ±0.18	**96.21** ±0.02	95.59 ±0.04	97.18 ±0.34	**96.38** ±0.17
MMRL++	85.44 ±0.09	77.78 ±0.42	81.43 ±0.26	77.73 ±0.17	71.37 ±0.12	74.41 ±0.08	98.90 ±0.08	94.40 ±0.22	96.60 ±0.11	95.43 ±0.49	96.97 ±0.33	96.19 ±0.18
w/LDP	85.65 ±0.25	78.23 ±0.21	**81.77** ±0.21	77.90 ±0.08	71.67 ±0.09	**74.65** ±0.02	99.17 ±0.12	94.80 ±0.08	**96.93** ±0.04	95.97 ±0.26	97.20 ±0.36	**96.58** ±0.19

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	72.83 ±0.62	73.04 ±0.78	72.93 ±0.69	95.79 ±0.62	72.65 ±1.21	82.63 ±1.00	90.46 ±0.06	91.88 ±0.24	**91.17** ±0.11	35.85 ±1.47	29.95 ±6.80	32.64 ±4.90
w/LDP	74.17 ±0.19	73.06 ±0.73	**73.61** ±0.31	97.12 ±0.24	74.96 ±0.29	**84.62** ±0.27	90.52 ±0.15	91.70 ±0.18	91.11 ±0.11	36.99 ±0.43	33.67 ±1.11	**35.25** ±0.75
MaPLe	76.52 ±0.26	71.78 ±1.95	74.08 ±1.09	97.34 ±0.16	71.82 ±1.01	82.65 ±0.65	90.21 ±0.08	91.68 ±0.15	**90.94** ±0.04	39.38 ±0.15	34.29 ±1.03	36.66 ±0.59
w/LDP	76.69 ±0.35	72.89 ±0.52	**74.74** ±0.16	96.96 ±0.35	73.19 ±0.55	**83.42** ±0.30	90.36 ±0.08	91.32 ±0.23	90.84 ±0.09	40.52 ±0.64	35.43 ±0.61	**37.80** ±0.08
PromptSRC	75.82 ±0.60	75.15 ±0.42	75.49 ±0.46	97.79 ±0.12	76.00 ±0.59	85.53 ±0.35	90.68 ±0.08	91.74 ±0.14	**91.21** ±0.11	39.18 ±0.86	35.87 ±1.35	37.45 ±0.97
w/LDP	76.69 ±0.21	75.13 ±0.81	**75.90** ±0.39	97.53 ±0.28	76.99 ±0.33	**86.05** ±0.15	90.62 ±0.05	91.79 ±0.05	91.20 ±0.02	39.88 ±0.15	37.63 ±0.22	**38.72** ±0.16
MMRL++	81.23 ±0.54	75.23 ±0.09	78.12 ±0.25	98.50 ±0.16	77.47 ±0.83	**86.73** ±0.47	90.50 ±0.00	91.70 ±0.08	91.10 ±0.04	46.47 ±0.70	38.50 ±0.73	**42.11** ±0.43
w/LDP	81.73 ±0.25	75.27 ±0.29	**78.37** ±0.15	98.00 ±0.36	77.70 ±0.16	86.68 ±0.08	90.50 ±0.08	91.87 ±0.05	**91.18** ±0.05	47.53 ±0.50	36.83 ±0.19	41.50 ±0.31

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	80.51 ±0.40	75.45 ±0.51	77.90 ±0.33	78.90 ±0.54	46.46 ±0.95	58.48 ±0.88	86.84 ±1.97	50.33 ±3.27	63.73 ±2.49	82.52 ±0.70	72.13 ±2.11	76.98 ±0.89
w/LDP	81.42 ±0.39	75.77 ±0.09	**78.49** ±0.16	79.09 ±0.66	54.47 ±1.84	**64.51** ±1.44	90.01 ±1.98	69.64 ±1.95	**78.53** ±1.26	83.47 ±0.21	75.25 ±1.80	**79.15** ±1.03
MaPLe	81.75 ±0.31	77.10 ±0.76	**79.36** ±0.46	82.37 ±1.10	58.33 ±0.94	**68.30** ±0.30	94.18 ±1.57	69.76 ±2.10	80.15 ±1.73	84.63 ±0.42	76.78 ±0.22	80.51 ±0.16
w/LDP	81.78 ±0.20	76.90 ±0.49	79.27 ±0.18	81.56 ±0.29	57.61 ±1.07	67.52 ±0.83	94.78 ±1.18	75.56 ±5.41	**84.09** ±3.08	85.52 ±0.25	77.77 ±1.71	**81.46** ±0.96
PromptSRC	82.35 ±0.12	78.85 ±0.28	80.56 ±0.19	82.91 ±0.80	60.95 ±2.96	70.25 ±2.17	91.29 ±0.25	71.56 ±5.85	80.23 ±3.70	85.82 ±0.13	79.00 ±0.14	82.27 ±0.11
w/LDP	82.58 ±0.11	78.98 ±0.16	**80.74** ±0.10	82.71 ±0.66	61.43 ±1.06	**70.50** ±0.49	92.60 ±0.49	80.13 ±1.55	**85.91** ±0.68	86.64 ±0.20	78.71 ±1.04	**82.48** ±0.50
MMRL++	82.93 ±0.33	79.57 ±0.05	81.22 ±0.18	85.07 ±0.29	65.83 ±0.40	74.22 ±0.17	95.73 ±0.73	84.17 ±2.22	89.58 ±1.03	87.37 ±0.54	80.43 ±0.49	83.76 ±0.14
w/LDP	83.23 ±0.09	79.70 ±0.08	**81.43** ±0.02	84.87 ±0.66	66.67 ±0.78	**74.67** ±0.73	95.77 ±0.76	87.33 ±0.74	**91.36** ±0.63	87.53 ±0.31	81.50 ±0.29	**84.41** ±0.03