ON THE ROLE OF MECHANISTIC INTERPRETABILITY FOR VISION-LANGUAGE PROMPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

While mechanistic interpretability tools such as sparse autoencoders (SAEs) have been proposed for extracting monosemantic, human-understandable features from vision-language models (VLMs) such as CLIP, existing work focuses on using these tools for post-hoc diagnostic analysis. We posit that SAE-based interpretability methods are more than passive probing tools, but can actively guide adaptation of VLMs to downstream tasks. We propose IPL (Interpretability-Guided Prompt Learning), a framework which leverages SAE decoders to extract interpretable concept directions, composes them into prompt tokens via a learnable attention selector, and injects the resulting tokens into both the vision and text encoder layers of CLIP for adaptation. A projection-preservation regularizer keeps prompt-token geometry aligned with the underlying concept directions. Across 15 benchmark datasets and experimental settings covering base-to-novel generalization, domain generalization, cross-dataset transfer, and few-shot learning, IPL consistently outperforms prior prompt-learning methods, with unified (multimodal) concept directions achieving the strongest results, compared to vision-only and text-only concept directions.

1 INTRODUCTION

Mechanistic interpretability (MI) research has made rapid progress in the past years and sparse autoencoders (SAEs) (Lim et al., 2024; Shen et al., 2025; Bhalla et al., 2024) have emerged as a prominent tool for MI. SAEs have been shown to be able to decompose intermediate representations of vision-language models (VLMs) like CLIP (Radford et al., 2021) into human-understandable concept directions: dictionaries of features that activate selectively on coherent semantic categories such as object parts, textures, or scene types, with explicit per-input activation magnitudes that quantify each concept's presence in any given image. These tools have given researchers a better picture of what CLIP has learned. So far, however, this growing toolbox has been used almost exclusively for post-hoc purposes like *interpreting* model behavior after the model has been trained. Our central research question is *can interpretability tools guide the adaptation of VLMs like CLIP to downstream settings?* We study this question in the context of *prompt learning*, where interpretability can provide actionable signals for model adaptation.

Prompt learning has emerged as the dominant parameter-efficient strategy for adapting CLIP to downstream tasks, with approaches like CoOp (Zhou et al., 2022b), MaPLe (Khattak et al., 2023), and MMRL (Guo & Gu, 2025) introducing learnable prompt tokens at various layers of CLIP. While effective, these prompts are initialized and optimized as free-form continuous vectors, unconstrained by any representation the model already uses. Recent attempts like CPL (Zhang et al., 2024), ArGue (Tian et al., 2024), IntCoOp (Ghosal et al., 2024), and IPO (Du et al., 2024), constrains prompt content to named concepts, but they obtain those concepts from *outside* CLIP by querying LLMs or VLMs for natural-language attributes and hard-coding them as prompt tokens. The interpretability is thus outsourced to *external* tools; it does not engage with the rich *internal* concept structure that mechanistic interpretability methods (Lim et al., 2024; Bhalla et al., 2024; Shen et al., 2025) have shown already exists inside CLIP itself. This gap motivates our approach.

We treat SAE-derived concept dictionaries not as analytical tools but as a semantically grounded concept directions from which prompt tokens can be composed. To this end, we propose **IPL** (Interpretability-Guided Prompt Learning), a prompt-learning framework to adapt to a dataset. IPL extracts a concept dictionary from an SAE for that dataset, composes prompt tokens as learnable

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.

attention-weighted combinations of those concepts, and projects the resulting tokens into both vision and text encoder layers. A projection-preservation regularizer keeps prompt-token geometry aligned with the underlying concept directions, so that the learned prompts remain anchored to the interpretable concept directions they were composed from.

Our contributions and key findings can be summarized as follows:

- We introduce IPL, a prompt-learning framework that shows that mechanistic interpretability tools can be used as a training guide for vision-language prompt learning.
- IPL composes prompt tokens via attention over interpretable concept directions and preserves their geometry through a projection preservation regularizer.
- IPL is agnostic to the choice of the SAE, supporting vision-only, text-only, and unified concept dictionaries. We find unified signals from VL-SAE (Shen et al., 2025) produce the strongest gains.
- Extensive experiments across four prompt-learning downstream task settings and 15 evaluation datasets show that IPL (unified) outperforms prior interpretable prompt-learning methods by up to $+2.49\%$ in harmonic mean on base-to-novel generalization while remaining competitive with strong black-box prompt-learning baselines.

2 Related Work

Mechanistic interpretability of VLMs. MI aims to decompose neural computations into human-understandable circuits and features for post-hoc interpretation of a neural network's knowledge. . Sparse autoencoders (SAEs) (Huben et al., 2024) are a promising tool for this purpose: by reconstructing intermediate activations with a sparse overcomplete dictionary, SAE decoder directions often capture monosemantic concepts. PatchSAE (Lim et al., 2024) applies SAEs to CLIP vision-transformer residual streams to reveal patch-level visual concepts. SpLiCE (Bhalla et al., 2024) decomposes CLIP image embeddings into sparse non-negative combinations of text-aligned concept directions. VL-SAE (Shen et al., 2025) learns a unified latent concept space for image and text inputs with modality-specific decoders. In contrast, IPL uses SAE-derived concept directions as a structured basis for prompt learning, bridging mechanistic interpretability with downstream adaptation.

Prompt learning for VLMs. CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and related contrastive vision-language models achieve zero-shot recognition using hand-crafted text prompts. Prompt learning adapts these models efficiently to downstream tasks while preserving pretrained capabilities. CoOp (Zhou et al., 2022b) learns continuous text prompts and CoCoOp (Zhou et al., 2022a) makes them image-conditional to reduce base-class overfitting. KgCoOp (Yao et al., 2023) regularizes prompts toward hand-crafted templates, while ProDA (Lu et al., 2022) models prompt distributions for diverse class semantics. Multimodal methods such as MaPLe (Khattak et al., 2023) learn coupled vision-text prompts, while PromptSRC (Chen et al., 2023), MMA (Yang et al., 2024), TCP (Yao et al., 2024), 2SFS (Farina et al., 2025), and SkipT (Wu et al., 2025) improve generalization through regularization or architectural changes. More recently, MMRL (Guo & Gu, 2025) injects multimodal representation tokens into intermediate layers, and VaMP (Cheng & Han, 2026) learns sample-conditioned text prompts via variational inference. Despite their effectiveness, these methods learn opaque continuous prompt vectors without human-interpretable semantics.

Interpretable prompt learning. A separate line of work improves prompt interpretability by using human-readable concepts as prompt content. CPL (Zhang et al., 2024) caches class-relevant concept descriptions as context, while ArGue and ArGue-N (Tian et al., 2024) use GPT-3 to generate visual attributes per class and concatenate selected attributes into prompts. IntCoOp (Ghosal et al., 2024) extracts attribute labels with BLIP-2 and aligns them with class embeddings during prompt tuning. IPO (Du et al., 2024) optimizes prompts in discrete token space so learned tokens remain natural-language phrases. Overall, these methods rely on externally generated concepts or attributes, so interpretability comes from the prompt input form rather than CLIP's internal computation. In contrast, IPL derives concept directions using mechanistic interpretability methods, explicitly connecting prompt learning to CLIP's internal representations.

Distribution Shift and Novelty. Adapting models to novelty has a long history, such as through domain adaptation (Ganin et al., 2016; Sun et al., 2017; Zhao et al., 2019), domain generalization (Muandet et al., 2013; Li et al., 2017), meta-learning (Finn et al., 2017), and few-shot learning (Ravi & Larochelle, 2017). These formulations assume access to model weights and modify parameters of

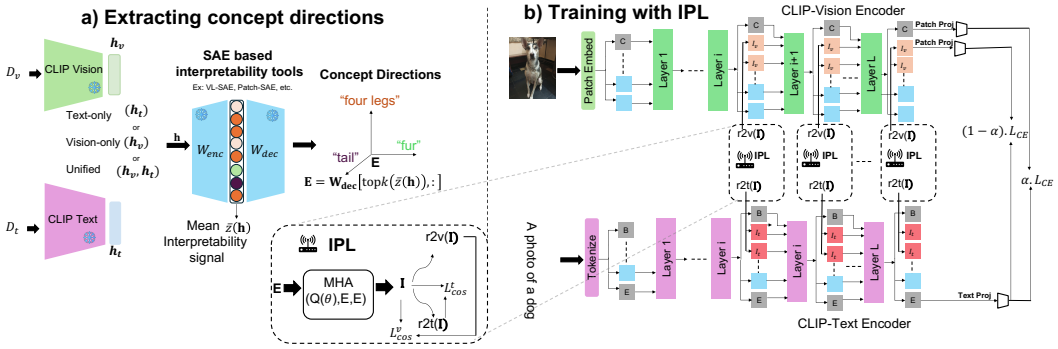


Figure 1: Overview of IPL: We extract concept directions from the data the model needs to adapt to, and use these to construct concept-guided prompt tokens, which are projected at different layers in both text and vision encoder of CLIP, while preserving concept geometry after projection.

task-specific models to bridge source-target gaps. In contrast, prompt-learning seeks to learn a small set of continuous prompt tokens without modifying any backbone parameters. Instead of learning a domain-invariant representation from data, the goal is to steer a pretrained representation that is already broadly capable toward task-, domain-, or class-specific behavior. Existing prompt-learning methods often optimize design choices for a single setting such as base-to-novel category generalization, domain generalization, cross-dataset transfer, and few-shot learning. IPL treats all four protocols as instances of a single underlying problem, adapting prompts under limited supervision while preserving CLIP's pretrained generalization, and addresses them with one unified mechanism.

3 METHOD

Interpretability-Guided Prompt Learning (IPL) leverages mechanistic interpretability to guide prompt learning for adapting VLMs to downstream tasks like generalizing from base to novel classes, domain generalization (input shifts at fixed label space), cross-dataset evaluation (zero-shot transfer to disjoint label spaces), and few-shot learning. We first introduce preliminaries elucidating zero-shot classification, prompt learning with CLIP, and SAE-based interpretability probing in Sec. 3.1. Then, we describe how IPL obtains interpretability signals in Sec. 3.2.1, how we use the interpretability signals to guide construction of prompt tokens in Secs. 3.2.2 and 3.2.3, and how we use these prompt tokens as learnable representations for training and inference in Sec. 3.2.4.

3.1 PRELIMINARIES

3.1.1 REVISITING CLIP ZERO-SHOT CLASSIFICATION.

CLIP consists of an image encoder f_v and text encoder f_t trained contrastively to align image and text embeddings in a shared latent space. During training, an input image is encoded into features $h_v = f_v(x)$ and a text prompt such as $p =$ *"a photo of a dog"* is tokenized and embedded to produce text features $h_t = f_t(p)$. Contrastive learning loss functions such as the InfoNCE loss (Oord et al., 2018) are used to align h_v and h_t closer to each other for matching image-text pairs and farther from each other for unrelated image-text pairs. At inference, CLIP's paired encoders are used for predicting categorical labels for an input image $\mathbf{x}$. Given a set of class labels $y \in \{1, \cdots, C\}$, text prompts are created using a template such as *"a photo of a* `[class]` *"*. The predicted category is decided by highest cosine similarity between the image features and one of the text prompt features:

$$\hat{y} = \arg\max_{y} \frac{\exp(\text{sim}(h_v, h_t)/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(h_v, h_t^i)/\tau)}; \quad \textit{using temperature coefficient } \tau. \tag{1}$$

3.1.2 UNDERSTANDING PROMPT LEARNING FOR EFFICIENT ADAPTATION.

Hand-engineered prompt templates like *"a photo of a* `[class]` *"* for CLIP are often sub-optimal when they are misaligned with the semantic context of complex and fine-grained downstream domains.

Fine-tuning CLIP on prompts catered to each domain can lead to catastrophic forgetting and is computationally expensive. The alternative, *prompt learning*, introduces continuous learnable tokens either in the input (Khattak et al., 2023; Zhou et al., 2022b;a; Du et al., 2024; Zhang et al., 2024) or intermediate layers (Yang et al., 2024). This approach allows adapting CLIP to downstream tasks without hand-engineering prompts and without updating CLIP's encoder. Existing approaches differ in the layers and encoders in which prompt tokens are inserted. **Language prompt learning** introduces N learnable prompt tokens $I_h^t \in \mathbb{R}^{N\times d}$ into H consecutive transformer layers of the text encoder, while, **vision prompt learning** introduces N learnable prompt tokens $I_h^v \in \mathbb{R}^{N\times d}$. In **multimodal prompt learning** prompt tokens are added to both text and vision branches. MaPLe (Khattak et al., 2023) passes language prompt tokens to a learnable coupling function to obtain vision prompts tokens. MMRL (Guo & Gu, 2025) proposes a shared representation space with tokens initialized by sampling from gaussian distribution and uses a linear projection layer to get vision and text branch prompt tokens. VaMP (Cheng & Han, 2026) injects sample specific text prompts obtained by variation inference conditioned on image features, and also simultaneously injects vision prompts.

3.1.3 Sparse-Autoencoder based mechanistic interpretability

Given an input image x, the vision encoder of CLIP produces features h. The latent space of VLMs is known to be densely entangled (Bhalla et al., 2024), with individual neurons encoding many unrelated concepts (*polysemanticity*). To disentangle this space, mechanistic interpretability studies (Huben et al., 2024) have shown that sparse autoencoders (SAEs) find interpretable *monosemantic* features in their latent space. The key idea is to train a wide autoencoder to reconstruct input features, regularized with a sparsity penalty. From features h, the SAE encoder with weights $W_{enc}\in\mathbb{R}^{d\times K}$ with $K\gg d$, produces sparse concept activations z. The SAE decoder with weights $W_{dec}\in\mathbb{R}^{K\times d}$ reconstructs h:

$$z = \sigma(hW_{enc} + b_{enc}) \in \mathbb{R}^K; \quad \text{and} \quad \hat{h} = zW_{dec} + b_{dec}. \tag{2}$$

Here, σ is a non-negative activation (ReLU or top-K) inducing sparsity, and $\mathbf{b_{enc}}, \mathbf{b_{dec}}$ are bias terms. Each z_i represents a distinct monosemantic concept. The SAE is trained to minimize $\|h-\hat{h}\|^2+\lambda\|z\|_1$ on a large corpus of representations (for *e.g.* VL-SAE (Shen et al., 2025) is trained with CC3M dataset). K is dictionary size; typically $K \gg d$ (for *e.g.* in PatchSAE (Lim et al., 2024) $K = 64 \times d$). We now define concept directions and concept strengths extracted from the SAE weights.

Definition 1 (Concept Directions and Strengths)**.** *Each unit-normalized row i of the decoder weight matrix, $v_i = \frac{W_{dec}[i,:]}{\|W_{dec}[i,:]\|} \in \mathbb{R}^d$, is the direction associated with concept i in the encoder's representation space. Given SAE activation z for input x, z_i is the strength of concept direction v_i in x.*

An image of a dog on grass might activate directions corresponding to *fur*, *four-legged*, *grass*, *etc.*, based on the training images on which the SAE activates strongly. Thus, each candidate concept is paired with an evidence set of training images supporting its semantic interpretation. The strength z_i is an input-dependent score for each concept – different images can have different strengths along the same direction. We term concept directions and strengths put together as "interpretability signals", since they are derived from mechanistic interpretability tools.

Modality of interpretability signals. The formulation applies for any SAE that can be trained over a model's internals, allowing concept extraction to be performed on the vision, text, or joint embeddings of a multimodal model, yielding modality-specific variants of our approach. *Vision interpretability signals* are obtained by running images through CLIP's vision encoder and reading concept activations from a vision-SAE such as PatchSAE (Lim et al., 2024) and VL-SAE-vision (Shen et al., 2025). *Text interpretability signals* are obtained by running per-class textual prompts through CLIP's text encoder and reading activations from a text SAE such as SAE-text (Shen et al., 2025). *Unified interpretability signals* constitute a unified concept set obtained by leveraging both image and text embeddings from CLIP such as VL-SAE-unified (Shen et al., 2025).

3.2 Interpretability Guided Prompt Learning

3.2.1 Extracting concepts from interpretability methods

Given a sparse autoencoder built on CLIP-style models, we construct an interpretable concept dictionary $E \in \mathbb{R}^{k\times d}, k < K$, as described below. Note that each row of E is a concept direction (Definition 1) corresponding to a human-meaningful concept in CLIP's residual-stream space.

Computing mean interpretability signal: Let $\mathcal{D}_{\text{base}}$ denote the base-class training set, comprising images and their associated class prompts. For each sample in $\mathcal{D}_{\text{base}}$, we compute the per-image interpretability signal z over the SAE's full dictionary of K concepts using Equation (2). We then compute the per-concept mean interpretability signal $\bar{z}$, averaged across all sample in $\mathcal{D}_{\text{base}}$. When we only use the images in $\mathcal{D}_{\text{base}}$ and the image encoder to produce $\bar{z}$, we refer to $\bar{z}$ as a *vision signal* and when we only use the prompts in $\mathcal{D}_{\text{base}}$ and the text encoder, we refer to $\bar{z}$ as a *text signal*. When we use both modalities, we refer to $\bar{z}$ as a unified interpretability signal.

Top-k selection. $\bar{z}_i$ denotes the strength of concept i. Based on these strengths, we retain only top-k concepts and store their indices in index-set $\mathcal{I}$. We obtain the resulting concept dictionary (modality variants: E^{vision}, E^{text}, and E^{unified} corresponding to the modality of the interpretability signal).

$$E = W_{\text{dec}}[\mathcal{I}, :] \in \mathbb{R}^{k \times d}. \tag{3}$$

3.2.2 Deriving Prompt Tokens from Interpretability Signals

We introduce N learnable representation queries $Q = [q_1; \ldots; q_N] \in \mathbb{R}^{N \times d}$, trained jointly with the prompt-learning objective. Unlike prior work (Guo & Gu, 2025; Cheng & Han, 2026) which uses Gaussian-initialized tokens, our queries select content from the frozen interpretable dictionary E via attention. For each injection layer $m \in \mathcal{M}$ transformer layers, we instantiate a multi-head attention selector $\text{MHA}^{(m)}$ that attends from queries Q to the concepts E to produce prompt tokens:

$$I^{(m)} = \text{MHA}^{(m)}(Q, E, E) \in \mathbb{R}^{N \times d}, \tag{4}$$

where Q provides the queries and E provides both keys and values. Each $I^{(m)}$ is thus an attention-weighted combination of concept directions. Thus, the prompt tokens by construction are expressible in the same interpretable basis as the model's own concept inventory.

3.2.3 Projecting prompt tokens into encoder layers

To project the prompt tokens ($\in \mathbb{R}^{N \times d}$) into each injection layer of both the vision and text encoders, we employ two per-layer projections: $\text{R2V}^{(m)}$, a linear map preserving residual width, and $\text{R2T}^{(m)}$, that projects to the text encoder width d_t.

$$I_v^{(m)} = \text{R2V}^{(m)}(I^{(m)}) \in \mathbb{R}^{N \times d_v} \quad \text{and} \quad I_t^{(m)} = \text{R2T}^{(m)}(I^{(m)}) \in \mathbb{R}^{N \times d_t}. \tag{5}$$

The projected tokens are concatenated to the existing token sequence at layer m in their respective encoders as described in Sec. 3.1.2

3.2.4 Learning with projected prompt tokens

Let h_v and $h_t^{(c)}$ denote [CLS] token projection and CLIP text projection features computed with prompt tokens injected at layers $\mathcal{M}$, and τ be CLIP's learned temperature coefficient. For an image-class pair $(\mathbf{x}, y)$, IPL produces two complementary classification probabilities:

$$\hat{y}_{\text{cls}}^{(c)}(x) = \frac{\exp(\text{sim}(h_v, h_t^{(c)})/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(h_v, h_t^{(i)})/\tau)}; \quad \text{and} \quad \hat{y}_{\text{rep}}^{(c)}(x) = \frac{\exp(\text{sim}(\bar{I}(x), h_t^{(c)})/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(\bar{I}(x), h_t^{(i)})/\tau)}. \tag{6}$$

The first probability, $\hat{y}_{\text{cls}}^{(c)}(x)$, reflects the standard image-text similarity using the `[CLS]` embedding for category c. The second probability, $\hat{y}_{\text{rep}}^{(c)}(x)$, derives image features from the prompt tokens themselves: we take the mean-pooled prompt-token output at the final injection layer, denoted $\bar{I}(x) = \text{mean}(I(x))$, and use it in place of the CLS-derived image feature.

Projection-preservation regularization. The naive linear projection in Sec. 3.2.3 can rotate concept structure arbitrarily, decoupling the prompt tokens' geometry from I interpretable structure. To preserve concept geometry after projection, we introduce a projection-preservation loss:

$$\mathcal{L}_{\text{proj}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \Big(\underbrace{1 - \cos\Big(\text{R2V}^{(m)}(E), E\Big)}_{\text{vision: cosine alignment}(\mathcal{L}_{\cos}^{v})} + \underbrace{\tfrac{1}{k^2} \big\| G\big(\text{R2T}^{(m)}(E)\big) - G(E) \big\|_F^2}_{\text{text: similarity-structure preservation}(\mathcal{L}_{\text{gram}}^{t})} \Big), \tag{7}$$

where $\cos(\cdot,\cdot)$ averages cosine similarity row-wise and and $G(\mathbf{X}) = (\mathbf{X})\,(\mathbf{X})^\top \in \mathbb{R}^{k\times k}$ is the row-normalized Gram matrix. $\mathcal{L}_{\text{proj}}$ encourages projection of $\mathbf{I}_v^{(m)}$ and $\mathbf{I}_v^{(m)}$ to themselves irrespective of dimensionality changes. The text term preserves pairwise similarity of $I_t^{(m)}$ under $\text{R2T}^{(m)}$ even though dimensionality changes via Gram matrix construction. Together, these constrain the projections to keep prompt-token geometry similar to concept directions.

Training Objective: Our training objective is the Lagrangian with α balancing the two cross-entropy objectives, and λ_{proj} is the projection-preservation loss weight. The trainable parameters are: queries Q in layer-specific selectors $\{\text{MHA}^{(m)}\}$, and projections $\{\text{R2V}^{(m)}, \text{R2T}^{(m)}\}$. CLIP's encoders, concepts E, and the interpretability methods' parameters remain frozen throughout training.

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{CE}}(\hat{y}_{\text{cls}}, y) + (1-\alpha)\;\mathcal{L}_{\text{CE}}(\hat{y}_{\text{rep}}, y) + \lambda_{\text{proj}}\,\mathcal{L}_{\text{proj}}. \tag{8}$$

Inference: For base-class evaluation, we use the fused probabilities $\alpha\,\hat{y}_{\text{cls}} + (1-\alpha)\,\hat{y}_{\text{rep}}$ to leverage both the standard and rep-token classification signals. For novel-class evaluation, we use only $\hat{y}_{\text{cls}}$, since rep tokens are trained on base classes and may bias predictions toward seen categories. This preserves the zero-shot transfer capability of the underlying CLIP alignment for unseen classes.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on the 15 standard image-classification datasets used by the prompt-learning literature: for Base to novel generalization, Domain Generalization Few-shot Learning, and cross-dataset evaluation we use 11 datasets ImageNet (Deng et al., 2009), Caltech101 (Li et al., 2022), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), OxfordFlowers (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012). For domain generalization we use ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a) as target domains.

Interpretability Methods. We instantiate IPL with three modality variants of VL-SAE (Shen et al., 2025): VL-SAE-vision (vision decoder), VL-SAE-text (text decoder), and VL-SAE-unified (concatenation of both decoders). Performance of other SAE based interpretability tools like PatchSAE(Lim et al., 2024) and SpLiCE(Bhalla et al., 2024) are discussed in appendix Sec. D.2.

Hyperparameter details. All IPL runs use $N = 5$ representation tokens, $\alpha = 0.7$, soft injection mode, and inject prompt tokens at layers $\mathcal{M} = \{6, 7, 8, 9, 10, 11, 12\}$ in both vision and text encoders. We train for 10 epochs with a learning rate of 0.0035 (cosine schedule, 1-epoch warmup), batch size 4, and SGD optimizer following the convention of MMRL (Guo & Gu, 2025). All our experiments use CLIP ViT-B/16 as the frozen vision-language backbone. Image and text encoder weights from OpenAI's released checkpoints are used throughout, with **no fine-tuning of CLIP at any point.** All experiments use 16-shot training and results reported over 3 seeds.

Baselines. We compare IPL against two categories of baselines. black-box prompt-learning: which inject randomly initialized prompt tokens into CLIP encoder layer and optimize over them like zero-shot CLIP, CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), MaPLe (Khattak et al., 2023), PromptSRC (Chen et al., 2023), MMA (Yang et al., 2024), TCP (Yao et al., 2024), 2SFS (Farina et al., 2025), SkipT (Wu et al., 2025),MMRL (Guo & Gu, 2025), and VaMP (Cheng & Han, 2026). interpretable prompt-learning: which inject the LLM queries attributes into CLIP encoder layers and optimize over them like CPL (Zhang et al., 2024), ArGue-N (Tian et al., 2024), IntCoOp (Ghosal et al., 2024), and IPO (Du et al., 2024). Numbers are reported from original papers.

4.2 Results and Discussion

Base to novel generalization. Each of the 11 datasets is split equally into base and novel classes. Base classes are used for training, and evaluation is performed on both base and novel classes. This setting allows us to examine how interpretability-signal-guided prompts affect VLM generalization to both seen and unseen classes. In Tab. 1, we report the performance of IPL using vision-only interpretability

Table 1: Base-to-novel generalization accuracies (B: Base, N: Novel, H: Harmonic Mean) compared to state-of-the-art interpretable and blackbox methods. Bold indicates best overall and underline indicates best interpretable method. **Note: the average reported in this table is over all 11 datasets, please refer to appendix Sec. A for detailed results with on other datasets.**

	Method	Average			ImageNet			StanfordCars			SUN397			UCF101		
		B	N	H	B	N	H	B	N	H	B	N	H	B	N	H
Interpretable	CPL (Zhang et al., 2024)	84.38	78.03	81.08	78.74	72.03	75.24	79.31	76.65	77.96	81.88	79.65	80.75	86.73	80.17	83.32
	IPO (Du et al., 2024)	79.92	80.51	80.21	77.83	72.45	75.04	73.42	75.71	74.55	81.25	80.92	81.08	85.32	80.92	83.06
	IntCoOp† (Ghosal et al., 2024)	83.20	78.72	80.75	75.99	72.67	74.29	77.04	76.32	76.67	81.63	79.33	80.46	86.76	79.42	82.92
	ArGue-N (Tian et al., 2024)	83.77	78.74	81.18	76.95	71.86	74.32	75.06	74.18	74.62	81.89	80.48	81.18	86.00	79.43	82.58
	IPL (Vision)	85.80	77.15	81.25	78.30	70.80	74.36	81.10	75.20	78.04	83.00	79.20	81.06	88.60	79.90	84.03
	IPL (Text)	85.75	76.59	80.91	78.40	70.70	74.35	81.40	75.50	78.34	83.10	79.40	81.21	88.80	80.00	84.17
	IPL (Unified)	86.26	**81.23**	**83.67**	**79.30**	**74.50**	**76.83**	81.90	**80.40**	81.14	83.00	**82.10**	**82.55**	**89.30**	**85.80**	**87.52**
Blackbox	CLIP (Radford et al., 2021)	69.34	74.22	71.70	72.43	68.14	70.22	63.37	74.89	68.65	69.36	75.35	72.23	70.53	77.50	73.85
	CoOp (Zhou et al., 2022b)	82.69	63.22	71.66	76.47	67.88	71.92	78.12	60.40	68.13	80.60	65.89	72.51	84.69	56.05	67.46
	CoCoOp (Zhou et al., 2022a)	80.47	71.69	75.83	75.98	70.43	73.10	70.49	73.59	72.01	79.74	76.86	78.27	82.33	73.45	77.64
	ProDA (Lu et al., 2022)	81.56	72.30	76.65	75.40	70.23	72.72	74.70	71.20	72.91	78.67	76.93	77.79	85.23	71.97	78.04
	KgCoOp (Yao et al., 2023)	80.73	73.60	77.00	75.83	69.96	72.78	71.76	75.04	73.36	80.29	76.53	78.36	82.89	76.67	79.65
	MaPLe (Khattak et al., 2023)	82.28	75.14	78.55	76.66	70.54	73.47	72.94	74.00	73.47	80.82	78.70	79.75	83.00	78.66	80.77
	PromptSRC (Chen et al., 2023)	84.26	76.10	79.97	77.60	70.73	74.01	78.27	74.97	76.58	82.67	78.47	80.52	87.10	78.80	82.74
	TCP (Yao et al., 2024)	84.13	75.36	79.51	77.27	69.87	73.38	80.80	74.13	77.32	82.63	78.20	80.35	87.13	80.77	83.83
	MMA (Yang et al., 2024)	83.20	76.80	79.87	77.31	71.00	74.02	78.50	73.10	75.70	82.27	78.57	80.38	86.23	80.03	82.20
	2SFS (Farina et al., 2025)	85.55	75.48	80.20	77.71	70.99	74.20	82.50	74.80	78.46	82.59	78.91	80.70	87.85	78.19	82.74
	SkipT (Wu et al., 2025)	85.04	77.53	81.11	77.73	70.40	73.89	82.93	72.50	77.37	82.40	79.03	80.68	87.30	82.47	84.81
	MMRL (Guo & Gu, 2025)	85.68	77.16	81.20	77.90	71.30	74.45	81.30	75.07	78.06	83.20	79.30	81.20	88.10	80.07	83.89
	VaMP (Cheng & Han, 2026)	**86.45**	78.67	82.37	78.98	73.45	76.11	**83.78**	80.14	**81.91**	**83.37**	78.95	81.09	88.52	78.99	83.48

Table 3: Cross-dataset transfer. All methods are trained on ImageNet (16-shot, ViT-B/16) and evaluated zero-shot on 10 downstream datasets. Top-1 accuracy (%) averaged over 3 seeds. The methods are grouped into *Interpretable* and *Blackbox*.

	Method	Source	Target										
		ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
Interpretable	CPL (Zhang et al., 2024)	73.53	95.52	91.64	66.17	73.35	**87.68**	27.36	68.24	48.96	51.25	70.52	68.07
	IPO (Du et al., 2024)	72.15	94.34	90.96	66.10	72.75	86.75	25.14	67.97	47.01	48.56	69.23	67.36
	IPL (Vision)	73.70±0.05	94.80±0.61	91.60±0.57	65.20±0.50	72.20±0.31	85.20±0.29	25.30±0.26	66.60±0.24	44.90±1.63	53.80±3.06	68.60±0.73	66.82±0.33
	IPL (Text)	73.60±0.05	94.90±0.42	91.30±0.37	65.00±0.45	72.80±0.83	84.90±0.12	26.00±0.59	66.60±0.31	45.00±0.37	54.50±4.21	68.40±0.46	66.94±0.45
	IPL (Unified)	**73.70**±0.08	**95.90**±0.39	91.30±0.33	**66.80**±0.79	**73.40**±0.08	87.60±0.17	**27.40**±0.90	**68.80**±0.43	45.90±0.43	**54.50**±4.24	**70.70**±0.66	**68.23**±0.60
Blackbox	CoOp (Zhou et al., 2022b)	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
	CoCoOp (Zhou et al., 2022a)	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
	MaPLe (Khattak et al., 2023)	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
	PromptSRC (Chen et al., 2023)	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
	TCP (Yao et al., 2024)	71.40	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73	66.29
	MMA (Yang et al., 2024)	71.00	93.80	90.30	66.13	72.07	86.12	25.33	68.17	46.57	49.24	68.32	66.61
	MMRL (Guo & Gu, 2025)	72.03	94.67	91.43	66.10	72.77	86.40	26.30	67.57	45.90	53.10	68.27	67.25
	VaMP (Cheng & Han, 2026)	72.83	94.96	**91.79**	66.10	73.18	86.97	26.76	68.04	46.82	53.82	68.93	67.74

signals (IPL(vision)), text-only interpretability signals (IPL (text)), and unified interpretability signals (IPL (unified)). We observe that IPL (vision) and IPL (text) achieve gains over previous interpretable prompt-learning methods on the base and novel classes for a few datasets, such as StanfordCars, UCF101, and Caltech101. On average, they also improve base-class performance. However, their performance comparatively drops on several other datasets, particularly on novel classes. This is likely because interpretability signals derived from a single modality may not be sufficient to produce well-aligned logits that improve generalization. In contrast, IPL with unified signals achieves the strongest average results on both base and novel classes, even outperforming the black-box state-of-the-art method VaMP on several datasets.

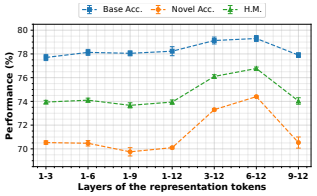
Table 2: Domain generalization. All methods are trained on ImageNet (16-shot, ViT-B/16) and evaluated on four ImageNet variants. Results are top-1 accuracy (%).

	Method	Source	Target				
		ImageNet	-V2	-Sketch	-A	-R	Avg.
Interpretable	IntCoOp† (Ghosal et al., 2024)	71.85	65.21	49.20	51.55	76.88	60.71
	ArGue-N† (Tian et al., 2024)	71.84	65.02	49.25	51.47	76.96	60.68
	IPL (Vision)	73.63 +1.78	64.93 -0.09	48.10 -1.15	51.90 +0.43	75.90 -1.06	60.20 -0.48
	IPL (Text)	73.57 +1.72	64.77 -0.25	48.07 -1.18	51.73 +0.26	75.67 -1.29	60.06 -0.62
	IPL (Unified)	**74.50** +2.65	**65.83** +0.81	**50.17** +0.92	**52.13** +0.66	77.87 +0.91	**61.50** +0.82
Blackbox	CLIP (Radford et al., 2021)	66.73	60.83	46.15	47.77	73.96	57.18
	CoOp (Zhou et al., 2022b)	71.51	64.20	47.99	49.71	75.21	59.28
	CoCoOp (Zhou et al., 2022a)	71.02	64.07	48.75	50.63	76.18	59.91
	MaPLe (Khattak et al., 2023)	70.72	64.07	49.15	50.90	76.98	60.26
	PromptSRC (Chen et al., 2023)	71.27	64.35	49.55	50.90	77.80	60.65
	MMRL (Guo & Gu, 2025)	72.03	64.47	49.17	51.20	77.53	60.59
	VaMP (Cheng & Han, 2026)	72.83	64.96	49.69	51.97	**78.01**	61.16

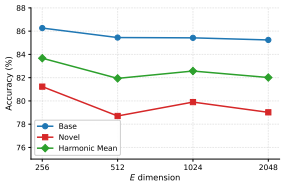
Domain generalization. We evaluate whether IPL enables VLMs to generalize to distribution shifts. We use ImageNet as

$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{CE-rep}}$	$\mathcal{L}^{v}_{\cos}$	$\mathcal{L}^{t}_{\cos}$	Base	Novel	HM
✓	✗	✗	✗	72.43 ± 0.90	68.70 ± 0.10	70.52
✓	✓	✗	✗	77.93 ± 0.29	69.40 ± 0.44	73.42
✓	✗	✓	✗	72.27 ± 0.71	68.77 ± 0.29	70.47
✓	✗	✗	✓	72.47 ± 0.67	70.30 ± 0.26	71.37
✓	✓	✓	✗	77.93 ± 0.15	69.27 ± 0.15	73.34
✓	✓	✗	✓	78.27 ± 0.25	70.60 ± 0.20	74.24
✓	✗	✓	✓	72.50 ± 0.66	70.30 ± 0.26	71.38
✓	✓	✓	✓	**79.30** ± 0.20	**74.50** ± 0.10	**76.83**

(a) Loss-components.



(b) Insertion layers.



(c) Concept dictionary size k.

Figure 3: **Ablation studies on ImageNet and downstream tasks.** (a) Loss-component ablation for IPL (unified). (b) Effect of representation-token insertion layers on base-to-novel generalization. (c) Effect of concept dictionary size k averaged across 11 datasets.

the source dataset for training and evaluate the learned model on target ImageNet variants, including ImageNet-V2, -Sketch, -A, and -R. IPL (unified) achieves strong results, outperforming prior state-of-the-art across all target domains and on the average target-domain accuracy.

Cross-dataset evaluation. We evaluate the transferability of IPL to different target datasets. Following the setting of Zhou et al. (2022a), we train the source model on ImageNet and evaluate its performance on 10 other datasets at inference time in the few-shot setting as shown in Tab. 3. IPL (unified) achieves the better overall performance across the target datasets.

Few-shot learning. We evaluate IPL's ability to adapt under different few-shot regimes using all the classes of ImageNet. Fig. 2 shows that IPL (unified) consistently outperforms previous baselines across different numbers of shots. Since only a limited number of prior methods report few-shot learning results, we compare IPL against the baselines for which such evaluations were reported. Results on other datasets in Sec. B.

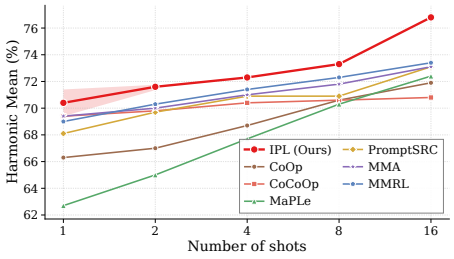


Figure 2: Few-shot performance on ImageNet under the base-to-novel protocol. IPL consistently improves harmonic mean over base and novel classes across shot counts. Shaded bands denote $\pm\sigma$ over 3 seeds.

4.3 Ablation Studies

Losses. Fig. 3a studies the effect of each loss component in IPL using the unified variant on ImageNet. The results show that the projection-preservation regularization losses ($\mathcal{L}^{v}_{\cos}$ and $\mathcal{L}^{t}_{\cos}$) improve generalization, with $\mathcal{L}_{\text{CE-rep}}$ further strengthening representation learning. The best performance is obtained when all losses are used together, indicating that task discrimination and preservation of interpretable concept directions are complementary for balancing base- and novel-class generalization.

Layers for prompt token insertion. Fig. 3b analyzes the effect of inserting IPL prompt tokens at different CLIP encoder layers on ImageNet. We observe that mid-to-late layer insertion performs best, with layers 6-12 achieving the strongest harmonic mean. Layers 3-12 show a similar trend, suggesting that interpretability-guided prompts are most effective when injected into semantic representations. In contrast, early-layer insertion leads to weaker generalization, likely because early layers encode lower-level visual features less suited for concept-guided adaptation.

Different sizes of k for E concepts. We analyze the impact of number of concept directions k used to contruct prompt tokens. Fig. 3c shows that $k = 256$ achieves better harmonic mean, indicating that a compact set of high-ranked interpretable concept directions is sufficient for IPL's prompt token construction. Performance comparatively degrades with larger k, suggesting that lower ranked concepts add noise rather than useful structure.

Replacing the Concept Dictionary. To determine whether the gains of IPL arise specifically from the SAE-derived concept directions, rather than simply constraining prompts to a fixed set of directions, we substitute E, with a matrix of same shape with either (i) row-normalized random Gaussian vectors, (ii) random orthogonal directions and (iii) the top-256 principal components by performing PCA of CLIP features on each dataset's base-class training set. As shown in Tab. 4, all three replacements

reduce base-to-novel (B2N) harmonic mean by about 2.9 points, with similar drops in domain generalization (DG) by 2.0 and cross-dataset transfer (CD) by 2.2 points respectively. Notably, PCA performs similar to the random controls, suggesting that the IPL's gain is not simply due to using data-adapted directions. Furthermore, the degradation is substantially larger on novel classes, *i.e.*, replacing E reduces base accuracy by only 0.8 points but novel accuracy by 4.8 points. This suggests that while arbitrary directions can suffice for adaptation to base classes, SAE-derived concept directions are important for generalization to unseen classes.

Table 4: SAE-derived concept dictionaries outperform Gaussian, orthogonal, and PCA concept dictionaries. Δ is IPL minus the control, so positive values favor IPL.

Task	Metric	IPL (Unified)	Gauss. Δ	Ortho. Δ	PCA Δ
B2N	Base	83.38	+0.80	+0.83	+0.78
B2N	Novel	80.70	+4.78	+4.76	+4.75
B2N	HM	82.01	+2.92	+2.92	+2.81
DG	OOD Acc.	61.50	+2.04	+1.98	+2.08
CD	Target Acc.	68.23	+2.19	+2.32	+1.99

Direct use of concept directions without prompt-learning. An alternative way to use SAE-derived concept directions is to inject them directly at inference, as in activation steering. We evaluate this direct-use alternative using the same dictionary E as IPL. We construct a steering vector as an activation-weighted sum of the top-N concept directions and add it to the CLS residual stream with strength γ, without any training. We sweep concept budget $N \in \{5, 20, 50, 100, 256\}$, steering strength $\gamma \in \{0.1, 0.5, 1, 2, 5\}$, injection layers $\{6, 9, 12\}$, and weighted or uniform pooling, and report the best of 150 configurations per dataset. As shown in Tab. 5, steering improves over zero-shot CLIP by 0.80 harmonic-mean points on average, but remains below IPL across all four datasets, with an average gap of 12.03 points. This comparison suggests that directly injecting SAE-derived directions provides limited adaptation, whereas using them to guide prompt learning yields substantially larger gains.

Table 5: Direct inference-time use of the same SAE-derived concept directions as IPL. Best steering results are reported over 150 configurations.

Dataset	Zero-shot	Steering	IPL	Δ
ImageNet	68.61	68.62	**76.83**	+8.21
StanfordCars	63.09	64.52	**81.14**	+16.62
SUN397	76.29	76.72	**82.55**	+5.83
UCF101	68.71	70.07	**87.52**	+17.45
Average	69.18	69.98	**82.01**	+12.03

4.4 Qualitative Visualization

Fig. 4 traces the interpretability chain that IPL exploits for prompt construction, illustrated on an example from Oxford Flowers dataset, using VL-SAE as the interpretability tool. Initially we rank concepts by mean SAE activation $\bar{z}_i$ over the dataset's base classes yields the unit-norm directions $v_i \in E^{\text{unified}}$ that populate IPL's concept dictionary, which are visually meaningful and dataset-relevant (*pointed floret tips*, *pinched bud*). Cosine contributions reveal that $I^{(m)}$ re-weights concepts beyond their raw activation ranking, indicating that the multi-head attention selector learns task-discriminative combinations rather than just copying $\bar{z}$. Comparing Grad-CAM saliency, vanilla CLIP attends to semantically irrelevant regions, while IPL concentrates on the flower head, suggesting that interpretability-guided prompts induce more semantically grounded attention.

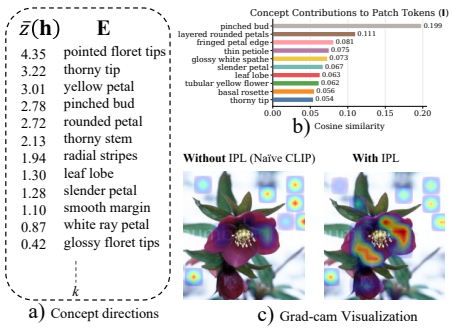


Figure 4: Illustration of (a) top-ranked concept directions in E with mean activations $\bar{z}$, (b) contributions of these concepts to a learned prompt token I, (c) Grad-CAM heatmaps showing that IPL focuses attention on the semantic features flower head.

5 Conclusions

We presented IPL, a prompt-learning framework that leverages SAE based mechanistic interpretability tools as active training guides. By extracting concept directions from frozen sparse autoencoders, composing them into prompt tokens via learnable attention selector, and preserving their geometry through a projection regularizer, IPL grounds prompt learning in CLIP's own interpretable structure rather than externally generated attributes. Across base-to-novel generalization, domain generalization, cross-dataset transfer, and few-shot learning, IPL with unified concept directions performs better in average accuracy compared to most of the previous state-of-the-art methods. As a limitation we acknowledge that IPL's effectiveness depends on the quality of the underlying interpretability tool.

AI USE STATEMENT

In this work, we used generative AI tools as Python programming assistants. We have not used generative AI tools for any other tasks. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code and artifacts are submitted as supplementary materials and will be publicly released (open-source) upon acceptance via GitHub or similar platforms.

REFERENCES

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *International conference on learning representations*, 2023.

Silin Cheng and Kai Han. Vamp: Variational multi-modal prompt learning for vision-language models. *Advances in Neural Information Processing Systems*, 38:89489–89515, 2026.

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*, 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 *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29989–29998, 2025.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Soumya Suvra Ghosal, Samyadeep Basu, Soheil Feizi, and Dinesh Manocha. Intcoop: Interpretability-aware vision-language prompt tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19584–19601, 2024.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25015–25025, 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, 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*, pp. 8340–8349, 2021a.

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*, pp. 15262–15271, 2021b.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

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*, pp. 4904–4916. PMLR, 2021.

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*, pp. 19113–19122, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. Ieee, 2017.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *arXiv preprint arXiv:2412.05276*, 2024.

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*, pp. 5206–5215, 2022.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

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*, pp. 8748–8763. PmLR, 2021.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

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

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision (CRCV), November 2012.

Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain adaptation in computer vision applications*, pp. 153–171. Springer, 2017.

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*, pp. 28578–28587, 2024.

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.

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*, 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. IEEE, 2010.

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*, pp. 23826–23837, 2024.

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

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*, pp. 23438–23448, 2024.

Yi Zhang, Ce Zhang, Ke Yu, Yushun Tang, and Zhihai He. Concept-guided prompt learning for generalization in vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7377–7386, 2024.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

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*, pp. 16816–16825, 2022a.

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, 2022b.

APPENDIX

A BASE TO NOVEL GENERALIZATION RESULTS

Table 6: Comparison with state-of-the-art methods on base-to-novel generalization across 11 datasets Methods are grouped into *Interpretable* and *Blackbox*. Our method (IPL) is reported using vision-only, text-only and unified interpretability signals using VL-SAE(Shen et al., 2025). Bold indicates best per column and underline indicates second-best per column. Delta rows show absolute accuracy-point change over ArGue-N, with positive gains in green and negative changes in red. †IntCoOp does not report DTD; its average in the source paper is over 10 datasets.

	Method	Average			Caltech101			OxfordPets			Flowers102		
		B	N	H	B	N	H	B	N	H	B	N	H
	CPL (Zhang et al., 2024)	84.38	78.03	81.08	98.35	95.13	96.71	95.86	98.21	97.02	98.07	80.43	88.38
	IPO (Du et al., 2024)	79.92	80.51	80.21	97.32	95.23	96.26	95.21	98.23	96.70	96.78	78.32	86.58
	IntCoOp† (Ghosal et al., 2024)	83.20	78.72	80.75	97.80	94.76	96.25	95.92	98.20	97.04	97.82	75.54	85.24
	ArGue-N (Tian et al., 2024)	83.77	78.74	81.18	98.63	94.70	96.63	96.23	98.59	97.40	98.62	77.96	87.08
Interpretable	**IPL (Vision)**	85.80±0.42	77.15±1.09	81.25	99.10±0.19	95.50±0.75	97.27	95.90±0.24	97.70±0.60	96.79	98.70±0.09	76.60±0.45	86.26
Interpretable	**Δ vs ArGue-N**	+2.03	-1.59	+0.07	+0.47	+0.80	+0.64	-0.33	-0.89	-0.61	+0.08	-1.36	-0.82
Interpretable	**IPL (Text)**	85.75±0.36	76.59±0.83	80.91	99.30±0.21	94.50±0.38	96.84	96.30±0.33	97.10±0.59	96.70	98.60±0.16	77.10±0.57	86.53
Interpretable	**Δ vs ArGue-N**	+1.98	-2.15	-0.27	+0.67	-0.20	+0.21	+0.07	-1.49	-0.70	-0.02	-0.86	-0.55
Interpretable	**IPL (Unified)**	86.26±0.40	**81.23±1.05**	**83.67**	**99.50±0.24**	**96.40±0.75**	**97.93**	96.60±0.62	**98.60±0.29**	**97.59**	98.80±0.37	81.80±0.38	89.50
Interpretable	**Δ vs ArGue-N**	+2.49	+2.49	+2.49	+0.87	+1.70	+1.30	+0.37	+0.01	+0.19	+0.18	+3.84	+2.42
Blackbox	CLIP (Radford et al., 2021)	69.34	74.22	71.70	96.84	94.00	95.40	91.17	97.26	94.12	72.08	77.80	74.83
Blackbox	CoOp (Zhou et al., 2022b)	82.69	63.22	71.66	98.00	89.81	93.73	93.67	95.29	94.47	97.60	59.67	74.06
Blackbox	CoCoOp (Zhou et al., 2022a)	80.47	71.69	75.83	97.96	93.81	95.84	95.20	97.69	96.43	94.87	71.75	81.71
Blackbox	ProDA (Lu et al., 2022)	81.56	72.30	76.65	98.27	93.23	95.68	95.43	97.83	96.62	97.70	68.68	80.66
Blackbox	KgCoOp (Yao et al., 2023)	80.73	73.60	77.00	97.72	94.39	96.03	94.65	97.76	96.18	95.00	74.73	83.65
Blackbox	MaPLe (Khattak et al., 2023)	82.28	75.14	78.55	97.74	94.36	96.02	95.43	97.76	96.58	95.92	72.46	82.56
Blackbox	PromptSRC (Chen et al., 2023)	84.26	76.10	79.97	98.10	94.03	96.02	95.33	97.30	96.30	98.07	76.50	85.95
Blackbox	TCP (Yao et al., 2024)	84.13	75.36	79.51	98.23	94.67	96.42	94.67	97.20	95.92	97.73	75.57	85.23
Blackbox	MMA (Yang et al., 2024)	83.20	76.80	79.87	98.40	94.00	96.15	95.40	98.07	96.72	97.77	75.93	85.48
Blackbox	2SFS (Farina et al., 2025)	85.55	75.48	80.20	98.71	94.43	96.52	95.32	97.82	96.55	98.29	76.17	85.83
Blackbox	SkipT (Wu et al., 2025)	85.04	77.53	81.11	98.50	95.33	96.89	95.70	97.87	96.77	98.57	75.80	85.70
Blackbox	MMRL (Guo & Gu, 2025)	85.68	77.16	81.20	98.97	94.50	96.68	95.90	97.60	96.74	**98.97**	77.27	86.78
Blackbox	VaMP (Cheng & Han, 2026)	**86.45**	78.67	82.37	98.95	95.96	97.43	**96.95**	95.24	96.08	98.96	**83.97**	**90.85**

	Method	Food101			FGVCAircraft			DTD			EuroSAT		
		B	N	H	B	N	H	B	N	H	B	N	H
	CPL (Zhang et al., 2024)	91.92	93.87	92.88	42.27	38.85	40.49	80.92	62.27	70.38	94.18	81.05	87.12
	IPO (Du et al., 2024)	90.92	93.08	91.99	41.21	41.42	41.31	82.14	66.81	73.69	94.25	80.11	86.61
	IntCoOp† (Ghosal et al., 2024)	91.45	91.99	91.72	38.55	35.90	37.17	–	–	–	95.26	78.01	85.77
	ArGue-N (Tian et al., 2024)	91.42	92.40	91.91	41.29	38.80	40.01	80.33	67.03	73.08	95.10	**90.68**	**92.84**
Interpretable	**IPL (Vision)**	90.70±0.12	91.50±0.05	91.10	46.30±0.66	38.20±1.02	41.86	86.10±0.61	64.50±1.55	73.75	96.00±0.53	79.60±6.25	87.03
Interpretable	**Δ vs ArGue-N**	-0.72	-0.90	-0.81	+5.01	-0.60	+1.85	+5.77	-2.53	+0.67	+0.90	-11.08	-5.81
Interpretable	**IPL (Text)**	90.90±0.22	91.50±0.00	91.20	45.90±1.25	37.90±0.62	41.52	85.00±0.39	64.70±0.95	73.47	95.60±0.42	74.10±4.78	83.49
Interpretable	**Δ vs ArGue-N**	-0.52	-0.90	-0.71	+4.61	-0.90	+1.51	+4.67	-2.33	+0.39	+0.50	-16.58	-9.35
Interpretable	**IPL (Unified)**	90.60±0.05	**97.70±0.05**	**94.02**	46.50±0.57	42.20±0.70	**44.25**	**86.30±0.29**	66.50±1.87	75.12	**97.10±0.48**	87.50±5.34	92.05
Interpretable	**Δ vs ArGue-N**	-0.82	+5.30	+2.11	+5.21	+3.40	+4.24	+5.97	-0.53	+2.04	+2.00	-3.18	-0.79
Blackbox	CLIP (Radford et al., 2021)	90.10	91.22	90.66	27.19	36.29	31.09	53.24	59.90	56.37	56.48	64.05	60.03
Blackbox	CoOp (Zhou et al., 2022b)	88.33	82.26	85.19	40.44	22.30	28.75	79.44	41.18	54.24	92.19	54.74	68.69
Blackbox	CoCoOp (Zhou et al., 2022a)	90.70	91.29	90.99	33.41	23.71	27.74	77.01	56.00	64.85	87.49	60.04	71.21
Blackbox	ProDA (Lu et al., 2022)	90.30	88.57	89.43	36.90	34.13	35.46	80.67	56.48	66.44	83.90	66.00	73.88
Blackbox	KgCoOp (Yao et al., 2023)	90.50	91.70	91.09	36.21	33.55	34.83	77.55	54.99	64.35	85.64	64.34	73.48
Blackbox	MaPLe (Khattak et al., 2023)	90.71	92.05	91.38	37.44	35.61	36.50	80.36	59.18	68.16	94.07	73.23	82.35
Blackbox	PromptSRC (Chen et al., 2023)	90.67	91.53	91.10	42.73	37.87	40.15	83.37	62.97	71.75	92.90	73.90	82.32
Blackbox	TCP (Yao et al., 2024)	90.57	91.37	90.97	41.97	34.43	37.83	82.77	58.07	68.25	91.63	74.73	82.32
Blackbox	MMA (Yang et al., 2024)	90.13	91.30	90.71	40.57	36.33	38.33	83.20	65.63	73.38	85.46	82.34	83.87
Blackbox	2SFS (Farina et al., 2025)	90.71	91.34	91.02	**47.48**	35.51	40.63	84.60	65.01	73.52	96.91	67.09	79.29
Blackbox	SkipT (Wu et al., 2025)	90.67	92.03	91.34	45.37	37.13	40.84	83.77	**67.23**	74.59	92.47	83.00	87.48
Blackbox	MMRL (Guo & Gu, 2025)	90.57	91.50	91.03	46.30	37.03	41.15	85.67	65.00	73.82	95.60	80.17	87.21
Blackbox	VaMP (Cheng & Han, 2026)	**92.77**	93.16	92.96	46.77	41.13	43.76	86.14	67.20	**75.50**	95.78	77.21	85.49

Table 6 reports IPL's base-to-novel results on the remaining datasets, complementing the subset shown in the main paper (Table 1). The trends mirror those in the main results: IPL with unified concept directions consistently achieves the strongest harmonic mean, while the vision-only and text-only variants remain competitive with prior interpretable prompt-learning methods.

B RESULTS ON FEW SHOT LEARNING OTHER DATASETS.

In Tab. 7, we study performance of few-shot learning with (1, 2, 4, 8, and 16) shots for different datasets. We observe that IPL performs better than previous blackbox based prompt learning methods. Note that interpretability based prompt learning methods do not provide results for few-shot learning.

Table 7: Few-shot classification accuracies across selected datasets under 1, 2, 4, 8, and 16 shots.

Method	Average					Caltech101					OxfordPets					Flowers102				
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
IPL (Vision)	69.51	72.54	77.06	81.15	84.62	93.93	94.63	95.57	96.37	96.63	88.10	91.33	92.70	93.20	93.77	76.50	83.23	88.90	94.57	96.93
IPL (Text)	69.47	72.82	76.15	80.75	84.95	93.97	94.43	95.77	96.13	96.63	88.30	90.97	93.03	92.53	93.80	76.40	83.53	89.97	94.50	96.97
IPL (Unified)	**74.03**	77.29	80.47	82.54	**87.07**	93.73	94.53	**96.80**	**96.53**	**97.73**	90.60	91.83	92.83	93.13	**94.80**	**86.37**	**91.33**	**95.77**	96.50	**98.90**
Linear probe CLIP	48.72	60.44	70.51	76.41	80.51	79.88	89.01	92.05	93.41	95.43	44.06	58.37	71.17	78.36	85.34	69.74	85.07	92.02	96.10	97.37
CoOp	67.34	71.37	74.85	77.87	80.99	92.60	93.07	94.40	94.37	95.57	90.37	89.80	92.57	91.27	91.87	77.53	87.33	92.17	94.97	97.07
CoCoOp	65.63	66.21	71.21	73.35	75.88	93.83	94.82	94.98	95.04	95.16	91.27	**92.64**	92.81	93.45	93.34	72.08	75.79	78.40	84.30	87.84
MaPLe	70.88	74.28	77.31	80.49	83.32	92.57	93.97	94.43	95.20	96.00	89.10	90.87	91.90	92.57	92.83	83.30	88.93	92.67	95.80	97.00
PromptSRC	73.36	76.42	79.72	82.04	84.40	93.67	94.53	95.27	95.67	96.07	**92.00**	92.50	**93.43**	**93.50**	93.67	85.93	91.17	93.87	96.27	97.60
MMA	69.55	72.48	77.51	80.91	84.38	92.90	94.00	94.33	95.37	96.33	91.23	91.97	92.23	92.77	93.23	83.60	90.30	93.00	95.97	97.97
MMRL	73.83	**77.37**	**80.71**	**82.95**	86.09	**94.17**	**94.83**	96.03	96.27	97.13	90.87	91.57	92.57	93.03	93.83	85.97	91.20	94.60	**96.60**	98.40

Method	Food101					FGVCAircraft					DTD					EuroSAT				
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
IPL (Vision)	85.13	85.47	86.23	86.97	**87.50**	29.10	34.40	37.47	44.17	52.70	52.43	56.93	64.93	69.58	73.87	61.40	61.77	73.63	83.20	90.93
IPL (Text)	85.07	85.47	86.27	**87.03**	87.43	28.80	34.43	37.40	44.27	54.33	50.07	56.90	64.63	69.87	74.23	63.67	64.03	65.97	80.93	91.27
IPL (Unified)	85.10	85.40	86.17	86.87	87.47	**29.17**	**35.30**	39.37	46.93	**59.60**	**56.50**	60.03	66.07	69.60	75.03	**76.73**	82.60	86.30	88.23	**95.97**
Linear probe CLIP	43.96	61.51	73.19	79.79	82.90	19.61	26.41	32.33	39.35	45.36	34.59	40.76	55.71	63.46	69.96	49.23	61.98	77.09	84.43	87.21
CoOp	84.33	84.40	84.47	82.67	84.20	21.37	26.20	30.83	39.00	43.40	50.23	53.60	58.70	64.77	69.87	54.93	65.17	70.80	78.07	84.93
CoCoOp	**85.65**	**86.22**	**86.88**	86.97	87.25	12.68	15.06	24.79	26.61	31.21	48.54	52.17	55.04	58.89	63.04	55.33	46.74	65.56	68.21	73.32
MaPLe	80.50	81.47	81.77	83.60	85.33	26.73	30.90	34.87	42.00	48.40	52.13	55.50	61.00	66.50	71.33	71.80	78.30	84.50	87.73	92.33
PromptSRC	84.87	85.70	86.17	86.90	**87.50**	27.67	31.70	37.47	43.27	50.83	56.23	59.97	65.53	69.87	72.73	73.13	79.37	86.30	**88.80**	92.43
MMA	83.03	82.50	82.13	83.00	84.57	28.73	31.90	37.57	44.83	52.70	52.27	56.90	63.93	67.97	73.47	55.07	59.80	79.40	86.47	92.37
MMRL	84.87	85.53	85.77	86.33	87.03	28.53	34.23	**40.47**	**48.07**	57.60	56.37	**61.37**	**67.87**	**71.60**	**75.30**	76.00	**82.87**	**87.67**	88.73	93.37

C Ablation Studies

C.1 Variants of the projection-preservation regularizer

We ablate the similarity operator used for projection preservation by independently varying the vision- and text-side regularizers between cosine alignment and similarity-structure preservation. As shown in Tab. 8, all four combinations perform within roughly 0.5 harmonic-mean points, indicating that IPL is not sensitive to the specific choice of preservation operator. Together with the loss ablation in Fig. 3a, these results show that constraining the projected prompt representations is more important than the particular form of the constraint.

Table 8: Projection-preservation is robust to the choice of similarity operator on the vision and text sides.

Vision term	Text term	Base	Novel	HM
Cosine	Gram-L2	78.20 ± 0.16	74.60 ± 0.08	76.36
Cosine	Cosine	79.30 ± 0.29	74.50 ± 0.06	76.83
Gram-L2	Cosine	78.27 ± 0.19	74.60 ± 0.08	76.39
Gram-L2	Gram-L2	78.27 ± 0.12	74.67 ± 0.05	76.43

C.2 Extended results for dictionary controls

Domain generalization and cross-dataset transfer. Tab. 9 expands the distribution-shift results reported in Tab. 4. Under domain generalization, replacing the SAE-derived dictionary consistently reduces accuracy across the four ImageNet variants, resulting in an approximately 2-point drop in average OOD accuracy. A similar pattern holds for cross-dataset transfer, where all three replacement dictionaries reduce average target accuracy by approximately 2 points. Notably, Gaussian, orthogonal, and PCA directions perform similarly in both settings, further indicating that the gains are specific to the SAE-derived concept directions rather than simply using a fixed or data-adapted basis.

Base-to-novel generalization. For completeness, Tab. 10 reports the per-dataset base-to-novel results underlying the averages presented in Tab. 4. Across all four datasets, replacing the SAE-derived dictionary has a substantially larger effect on novel-class accuracy than on base-class accuracy, consistent with the aggregate trend reported in the main text.

Table 9: SAE-derived concept dictionaries consistently outperform random and PCA dictionaries under domain generalization and cross-dataset transfer. Left: domain generalization, trained on ImageNet and evaluated on four ImageNet variants. Right: cross-dataset transfer, trained on ImageNet and evaluated zero-shot on 10 target datasets. Averaged across 3 seeds.

	Domain generalization					**Cross-dataset**	
E dictionary	-V2	-Sketch	-A	-R	OOD Avg	Source	10-target Avg
IPL (Unified)	**65.83**	**50.17**	**52.13**	**77.87**	**61.50**	73.70	**68.23**
Random Gaussian	64.90	48.17	48.80	75.97	59.46	73.57	66.04
Random Orthogonal	64.97	48.17	48.93	76.00	59.52	73.53	65.91
PCA (data-adapted)	64.77	48.30	48.77	75.83	59.42	**73.83**	66.24

Table 10: Per-dataset base-to-novel results for the dictionary-control ablation.

	IPL (Unified)			Random Gaussian			Random Orthogonal			PCA (data-adapted)		
Dataset	B	N	H	B	N	H	B	N	H	B	N	H
ImageNet	79.30	74.50	76.83	78.20	70.67	74.24	78.27	70.60	74.24	78.30	70.53	74.22
StanfordCars	81.90	80.40	81.14	80.97	74.63	77.67	81.07	74.80	77.81	81.13	74.83	77.85
SUN397	83.00	82.10	82.55	82.80	79.10	80.91	82.73	79.10	80.88	82.87	79.30	81.05
UCF101	89.30	85.80	87.52	88.33	79.27	83.55	88.10	79.27	83.45	88.03	79.10	83.32
Average	**83.38**	**80.70**	**82.01**	82.58	75.92	79.09	82.54	75.94	79.10	82.58	75.94	79.11

D Additional results

D.1 Concepts attended by the selector

We inspect the concept directions receiving the highest attention from the learned selector at each injection layer.

Tab. 11 reports the layer-wise top attended direction for three representative datasets. Specifically, we report the number of distinct top-1 concepts across the seven injection layers for against a randomly initialized selector as a control. The associated VL-SAE labels are qualitatively consistent with the dataset content.

For example, the top directions include vegetation- and road-related labels for EuroSAT, texture-related labels for DTD, and flower-related labels for Oxford Flowers. We provide these examples as a qualitative view of the concepts used by the learned selector. Note that similar to every other prompt learning methods, prompt tokens composed using our method IPL are class-independent concept contexts, injected for every input.

We further examine concept attribution at the class level. Although the selector's attention weights are class-independent, the influence of an attended concept on the output can vary across class predictions. We measure this using gradient sensitivity, defining the attribution of concept i to class c as its attention weight multiplied by the gradient of the class-c logit with respect to that weight. Tab. 12 reports representative examples from Oxford Flowers, where different classes receive their highest attribution from different concept directions. We will release per-class concept attributions for all datasets with the final code and artifacts.

D.2 Results on other Interpretability Methods

In Fig. 5, we study how different interpretability tools like PatchSAE(Lim et al., 2024) and splice (Bhalla et al., 2024) perform on base to novel generalization for different dataset. We observe that these tools perform on par with VL-SAE (vision).

Table 11: Top-attended concept direction across injection layers. We report the top-attended SAE direction and its VL-SAE label for three representative datasets. **Labels are provided by VL-SAE and are not used during training.**

Dataset	L6	L7	L8	L9	L10	L11	L12
EuroSAT	green mat (3708)	herb patch (3642)	road shoulder (702)	green mat (3378)	green mat (2931)	road shoulder (2644)	road shoulder (3562)
DTD	repeat dot (1463)	chessboard patch (2442)	gauze weave (3530)	tiled seam (1719)	alternating stripe (58)	tiled seam (1827)	overlapping strands (3683)
Flowers	pointed floret tips (1923)	pointed floret tips (1817)	pointed floret tips (4077)	pointed floret tips (1347)	pointed floret tips (1626)	yellow spadix (3978)	narrow linear leaves (2274)

Random-initialized selector (control): no concept is consistently top-ranked on any of the three datasets.

Table 12: Representative per-class concept attributions on Oxford Flowers. Attribution measures the gradient sensitivity of each class logit to its attended concept directions.

Class	Top-1 attributed concept
pink primrose	glossy white spathe
desert rose	thorny tip
passion flower	pointed floret tips

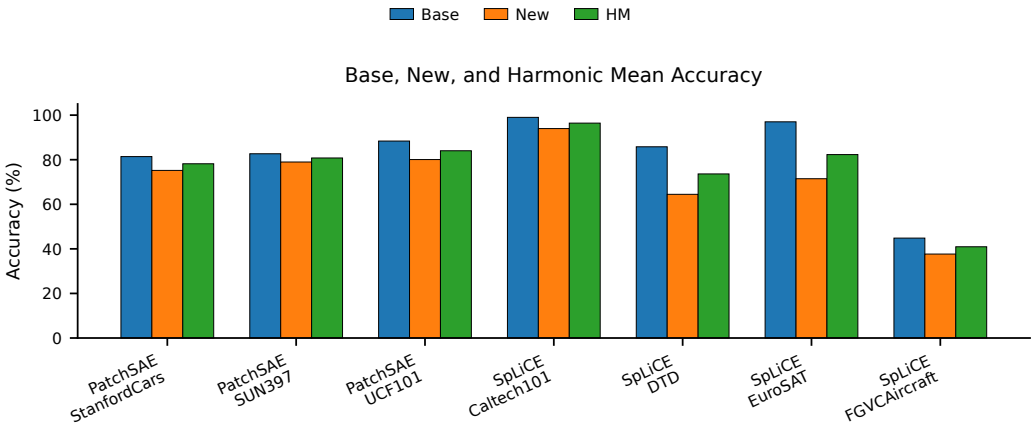


Figure 5: Performace of other interpretability tools like PatchSAE (Lim et al., 2024) and SpLiCE (Bhalla et al., 2024) on base to novel generalization for different datasets.

D.3 Representation-token count.

Fig. 6 shows that performance improves rapidly as the number of representation tokens increases from 1 to 5 and then saturates. We therefore use $N = 5$ throughout the main experiments, since larger values provide only marginal gains while increasing prompt length.

D.4 Impact of Base/Novel score fusion at inference

IPL produces two classification outputs in a single forward pass: $\hat{y}_{\text{cls}}$, computed from the standard CLIP image-text similarity with prompt tokens injected, and $\hat{y}_{\text{rep}}$, computed using the mean-pooled prompt-token representation. Since $\hat{y}_{\text{rep}}$ is trained with base-class supervision, our default inference protocol fuses the two outputs for base classes and uses only $\hat{y}_{\text{cls}}$ for novel classes.

Tab. 13 compares this protocol (R_{asym}) with two unified alternatives. Fusing both outputs for both splits (R_{fused}) leaves base accuracy unchanged but reduces novel accuracy by 1.25 points. In contrast, using $\hat{y}_{\text{cls}}$ for both splits (R_{cls}) preserves novel accuracy while reducing base accuracy by 0.96 points, yielding an HM of 83.21 compared with 83.67 for R_{asym}. Thus, the asymmetric protocol provides a

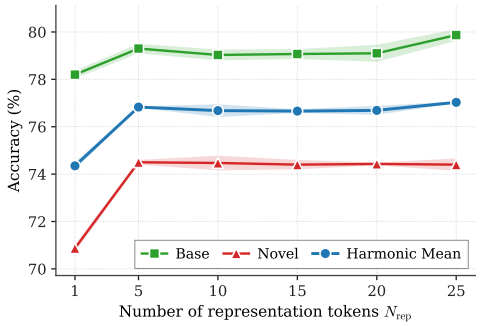


Figure 6: Effect of the number of representation tokens N_{rep} on ImageNet base-to-novel generalization. Performance saturates at $N_{\text{rep}} = 5$, which we use in all main experiments. Shaded bands denote $\pm$ std over 3 seeds.

Table 13: Using fused probabilities for base classes and $\mathbf{l}_{\text{cls}}$ for novel classes modestly improves base-to-novel harmonic mean.

Rule	Base	Novel	HM
R_{asym} (fused on base, $\hat{y}_{\text{cls}}$ on novel)	**86.26** $\pm$ 0.13	**81.23** $\pm$ 0.16	**83.67**
R_{fused} (fused on both, $\alpha = 0.7$)	**86.26** $\pm$ 0.13	79.98 $\pm$ 0.15	83.00
R_{cls} ($\hat{y}_{\text{cls}}$ on both)	85.30 $\pm$ 0.15	**81.23** $\pm$ 0.16	83.21

modest improvement in harmonic mean, while most of the performance is retained using the unified $\hat{y}_{\text{cls}}$-only rule.

E Analyzing the shift in concept activations before and after injection of IPL prompts

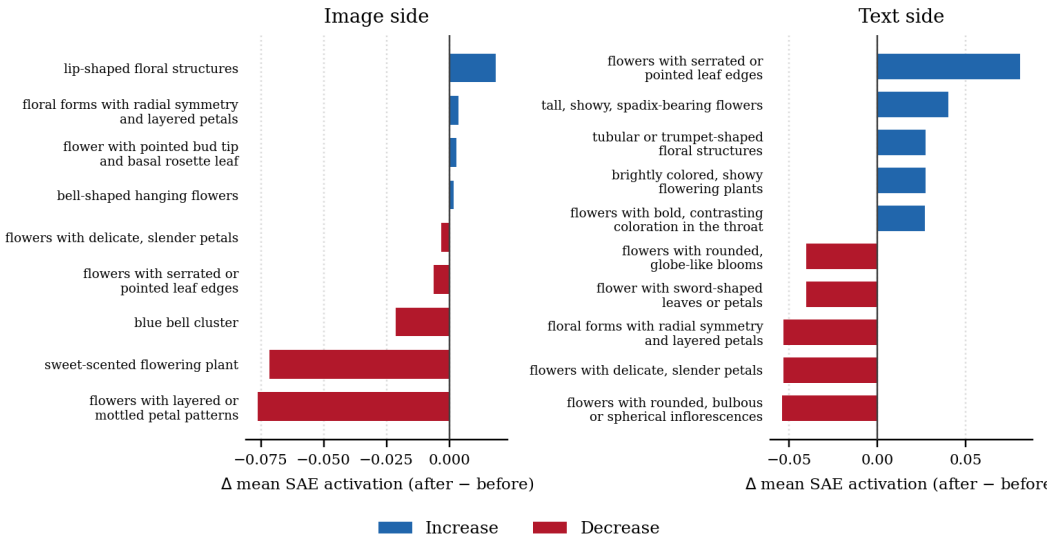


Figure 7: Analysis of how IPL's prompt tokens change VL-SAE concept activations on Oxford Flowers. We report the concepts with the largest shifts, on the image side (left) and text side (right). *Before* corresponds to vanilla CLIP with no prompt tokens, and *after* to the same frozen CLIP with IPL's trained prompt tokens injected. Here, both use same inputs.

In IPL, since $\mathbf{E}$ is extracted from the frozen SAE backbone and is not updated while training. Hence the concept basis itself is not entirely changed but how strongly each basis is represented changes. We measure this by reading VL-SAE activations at CLIP's final embedding layer, under two conditions with identical inputs: vanilla CLIP with no prompt tokens (*before*), and the same frozen

CLIP with IPL's trained prompt tokens injected (*after*). Figure 7 shows the largest shifts on Oxford Flowers. The changes are substantial in both directions, but every large mover is a floral morphology concept: on the image side *lip-shaped floral structures* rises ($0.055 \rightarrow 0.074$) while *flowers with layered or mottled petal patterns* falls ($0.266 \rightarrow 0.190$); on the text side *tall, showy, spadix-bearing flowers* rises ($0.027 \rightarrow 0.067$) while *flowers with rounded, bulbous or spherical inflorescences* falls ($0.203 \rightarrow 0.149$). Adaptation therefore re-weights which fine-grained attributes dominate within an already dataset-relevant region of the concept space, rather than displacing the representation toward a different kind of concept

F LIMITATIONS ANALYSIS AND FUTURE SCOPE

IPL introduces additional compute relative to existing prompt-learning baselines. As reported in Table 14, IPL requires approximately 23.1 GMAC per forward pass compared to 20.9 GMAC for MMRL (Guo & Gu, 2025) and 20.5 GMAC for ArGue (Tian et al., 2024), a $\sim$10–13% increase, attributable entirely to the per-layer multi-head attention selectors and R2V/R2T projection layers introduced for interpretability-guided prompt construction. We note that this overhead is a deliberate consequence of IPL's per-layer composition: the selector $\mathrm{MHA}^{(m)}$ and projections are instantiated independently at each rep-injection layer m, which provides per-layer flexibility in how concepts are composed at different depths but multiplicatively scales prompt-side FLOPs with $|\mathcal{M}_{\text{rep}}|$. Several straightforward extensions can mitigate this cost: (i) sharing MHA and projection parameters across injection layers, which would reduce prompt-side FLOPs by roughly $7\times$ in our default configuration at the cost of less per-layer specialization; (ii) injecting at fewer layers, which our ablation (Fig. 3b) shows incurs only a small accuracy drop; or (iii) lower-rank parameterizations of the projection layers. A systematic study of these efficiency-accuracy trade-offs is left to future work.

Table 14: Computational cost on CLIP ViT-B/16 (single forward pass). *Prompt* columns report FLOPs introduced by each method's prompt-token machinery. Δ is the percentage increase in total GMAC over ArGue.

Method	**Vision** (GMAC)	**Text** (GMAC)	**Prompt** (GMAC)	**Total** (GMAC)	**Δ** (%)
ArGue (Tian et al., 2024)	17.563	2.980	0.000	20.543	—
MMRL (Guo & Gu, 2025)	17.833	3.096	0.023	20.951	+1.99
IPL (Vision)	17.833	3.096	2.203	23.132	+12.60
IPL (Text)	17.833	3.096	2.203	23.132	+12.60
IPL (Unified)	17.833	3.096	2.203	23.132	+12.60