# When Target Accuracy Hides Semantic Interface Drift in Vision–Language Adaptation

**Anonymous authors**
Paper under double-blind review

## Abstract

A target-adapted vision–language model can improve fine-grained accuracy while moving away from the pretrained image–text interface that supports zero-shot transfer. We define *semantic interface drift* through changes in direct prompt compatibility, prototype relations, local neighborhoods, and held-out zero-shot behavior. We study same-image frozen-feature anchoring in an identity-initialized visual adapter and test the role of the reference with a controlled intervention. The central experiment closely matches adaptation-validation accuracy while replacing the same-image reference with a within-minibatch shuffled feature. Across five-seed CLIP experiments, the same-image reference preserves the measured geometry and held-out transfer axes at comparable adaptation accuracy. The shuffled-reference control yields a 27.49-point mean transfer drop, compared with 3.78 points for the same-image reference. Three-seed replications recover the retention pattern with SigLIP2 and when EuroSAT replaces CIFAR-100 as the adaptation dataset. These results identify reference correspondence as a design axis that target accuracy alone does not measure.

## 1 Introduction

A target-adapted checkpoint can solve its labeled task while losing the transfer capability supplied by the pretrained CLIP interface. A vision–language model recognizes a new taxonomy by comparing image features with natural-language prompts (Radford et al., 2021). Adaptation improves discrimination on the target labels, but it can also change which prompts an image supports, which images remain neighbors, and how accurately the encoder recognizes unseen label sets without additional tuning. Target accuracy records specialization while hiding this transfer degradation.

Existing preservation methods constrain parameters, output features, or predictions during fine-tuning (Mukhoti et al., 2024). Similar adaptation accuracy can nevertheless conceal different changes to the image–text interface. We focus on the pretrained correspondence that supports later reuse. We call the loss of prompt compatibility, prototype relations, and local neighborhoods *semantic interface drift*. Conventional forgetting metrics track accuracy loss on previously learned tasks; this drift can grow while target accuracy improves. Checkpoints with comparable target utility can consequently retain very different transfer capability.

We vary the reference inside a feature-preservation objective. An identity-initialized visual adapter specializes to the target taxonomy while the text encoder remains fixed. The same-image objective anchors each adapted feature to its own frozen coordinate. Matched controls replace that coordinate with shuffled, class-related, or aggregate references while retaining the penalty form, optimization protocol, and target utility. Held-out transfer then reveals how reference assignment affects reuse. Figure 1 summarizes this intervention.



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.

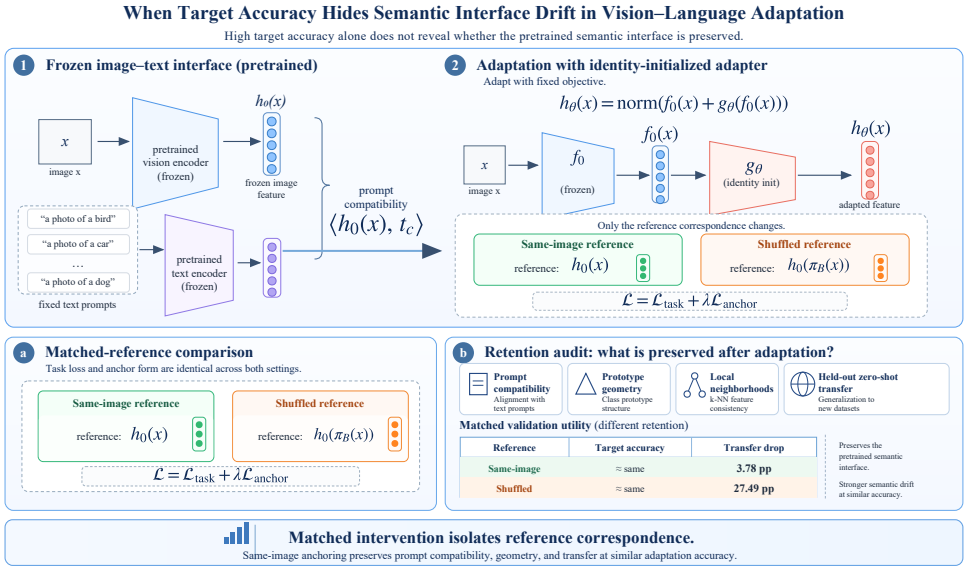



Figure 1: The central intervention. A frozen image–text interface supplies $h_0(x)$ and fixed text prompts. The adapter uses either the same-image reference $h_0(x)$ or a within-minibatch shuffled reference $h_0(\pi_B(x))$, while the task loss and anchor form remain fixed. At matched validation utility, same-image anchoring preserves prompt compatibility and geometry and limits the mean held-out transfer drop to 3.78 pp, compared with 27.49 pp for the shuffled reference.

Interface drift tracks transfer degradation across reference assignments: over 30 seed–reference pairs, loss of global prototype correlation has Spearman correlation .672 with mean held-out transfer drop. The matched comparison reveals a sharper separation. Shuffled-reference and same-image anchoring differ by only 0.14 target-accuracy points, yet their mean transfer drops differ by 23.70 points. Same-class references retain more transfer than aggregate or unrelated references, but less than the original image coordinate. EuroSAT adaptation reproduces this hierarchy at approximately 94.5–94.6% validation accuracy, where the four references yield transfer drops of 8.66, 23.98, 40.02, and 40.15 points.

Our contributions are:

- We formulate semantic interface drift as the loss of prompt compatibility and representation geometry during adaptation, and show that it tracks transfer degradation hidden by target accuracy.
- We isolate reference correspondence through a utility-matched intervention. Preserving each image's pretrained coordinate retains substantially more transfer than class-related, shuffled, or aggregate references under the same feature objective.
- We reproduce the reference ordering across CIFAR-100 and EuroSAT adaptation, CLIP ViT-B/32 and ViT-B/16, and SigLIP2; preservation baselines and anchor-strength controls distinguish correspondence from update magnitude.

## 2 Related Work

**Vision–language adaptation.** Prompt learning and parameter-efficient tuning adapt pretrained models with limited supervision (Zhou et al., 2022; Li et al., 2024). ProLIP tunes the visual embedding projector with zero-shot regularization (Fahes et al., 2026), illustrating how the intervention point affects adaptation. SigLIP 2 extends pretrained encoders with improved semantic understanding and dense features (Tschannen et al., 2025), making preservation relevant beyond the original CLIP family. These developments motivate evaluating both target specialization and the usability of the inherited interface. Our fixed-text visual adapter exposes this distinction by allowing image features to change while holding the text coordinates constant.

**Preserving pretrained behavior.** Feature distortion during fine-tuning can affect out-of-distribution performance (Kumar et al., 2022). Parameter anchoring constrains departure from initialization (Li et al., 2018; Kirkpatrick et al., 2017); WiSE-FT interpolates pretrained and adapted weights (Wortsman et al., 2022); LDIFS preserves frozen features (Mukhoti et al., 2024); and ARF uses semantic anchors for robust adaptation (Han et al., 2024). Lipsum-FT preserves image–text scores using random text guidance (Nam et al., 2024), while R-Adapter combines lightweight adapters with self-ensembling (Kim et al., 2024). Same-image L2 is a standard feature-preservation objective; the new object of study is the correspondence encoded by its reference. We hold the penalty family and target utility fixed while changing that correspondence, then measure which prompt relations, geometry, and held-out transfer remain usable. Native LDIFS and WiSE-FT checkpoints provide comparisons on six transfer datasets.

**Semantic geometry and correspondence.** Alignment and uniformity characterize contrastive representation geometry (Wang & Isola, 2020), while hyperbolic image–text representations encode hierarchical relations (Desai et al., 2023). More recently, MPS-Tuning preserves feature Gram matrices and shapes class separability during few-shot adaptation (Chen et al., 2025). This relational perspective complements the same-image coordinate constraint studied here: pairwise relations and image-specific references describe different properties of an adapted representation. With normalized features, squared L2 and cosine distance differ only by a constant factor. We fix the penalty family and vary reference correspondence, using prototype correlation, neighborhood overlap, and transfer to characterize the resulting interface.

# 3 METHOD

## 3.1 SETUP AND REFERENCE INTERVENTION

Let $f_0$ be a frozen image encoder and $g_\theta$ an identity-initialized visual adapter. For image $x$, the frozen and adapted normalized features are

$$h_0(x) = \mathrm{norm}(f_0(x)), \qquad h_\theta(x) = \mathrm{norm}\left(f_0(x) + g_\theta(f_0(x))\right). \tag{1}$$

The text encoder and prompt embeddings $\{t_c\}$ remain fixed. Target supervision uses the fine-class prompt loss

$$\mathcal{L}_{\mathrm{task}} = -\log \frac{\exp(s h_\theta(x)^\top t_y)}{\sum_c \exp(s h_\theta(x)^\top t_c)}, \tag{2}$$

where $s$ is a learned clipped logit scale.

For a minibatch $B$, same-image anchoring adds

$$\mathcal{L}_{\mathrm{anchor}} = \frac{1}{|B|} \sum_{x \in B} \|h_\theta(x) - h_0(x)\|_2^2, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda \mathcal{L}_{\mathrm{anchor}}. \tag{3}$$

For the question studied here, correspondence is encoded by the assignment $x \mapsto r(x)$ inside the same loss. The same-image choice $r(x) = h_0(x)$ preserves the pretrained coordinate assigned to that image while allowing the adapter to specialize. The hierarchy replaces it with the nearest frozen feature from the same class, a random same-class or same-superclass feature, an unrelated frozen feature, or a minibatch/global mean. These choices separate instance identity, semantic relevance, and aggregate direction without changing the penalty family. We compare these references at matched target utility or within a narrow common accuracy range.

Reference identity also changes the fixed point of the objective. At the pretrained representation, the same-image penalty has zero feature gradient because $h_\theta(x) = h_0(x)$. A mismatched reference instead applies a nonzero correction toward another image or an aggregate coordinate. Same-image anchoring can therefore preserve the original assignment while responding to the task loss; class-related and aggregate references encode different assignments even when their penalty magnitudes are similar.

All reference comparisons fix the image encoder, adapter architecture, text prompts, minibatch construction, and checkpoint rule. The original five-seed matched control uses $h_0(\pi_B(x))$, where $\pi_B$ is resampled within each minibatch. The expanded three-seed hierarchy uses $\lambda = 2$ for every reference. All frozen features are precomputed, and transfer labels remain outside training and checkpoint selection.

Table 1: Measurements used to separate target utility, semantic interface drift, and transfer degradation. Frozen CLIP defines the geometry and transfer reference.

| Role | Measurement | Direction |
|---|---|---|
| Target utility | Fine-prompt accuracy | ↑ |
| Target hierarchy | Fine prediction → parent class | ↑ |
| Interface drift | Direct prompt accuracy; $\rho$; NP@5 | ↑ |
| Transfer loss | Held-out zero-shot drop $\Delta_{\mathrm{tr}}$ | ↓ |

**Local score stability.** For fixed unit prompts, same-image displacement bounds the change in each prompt score, and a frozen top-1 prediction is unchanged when its margin exceeds twice that displacement. Appendix A gives the derivation and margin audit. Because this sufficient condition covers fewer than $0.4\%$ of CIFAR-100 examples, transfer retention is evaluated directly on held-out datasets.

### 3.2 Measuring interface drift and transfer degradation

We separate target utility, semantic interface drift, and transfer degradation. Target utility is fine-prompt accuracy on the adaptation dataset. Drift is observed through direct coarse-prompt accuracy, global prototype Spearman correlation, five-neighbor overlap (NP@5), and mean squared same-image displacement. Transfer degradation is the mean decrease from frozen zero-shot accuracy over a held-out dataset set $\mathcal{S}$,

$$\Delta_{\mathrm{tr}}(\theta) = \frac{1}{|\mathcal{S}|} \sum_{D \in \mathcal{S}} \left[ \mathrm{Acc}_D(f_0) - \mathrm{Acc}_D(h_\theta) \right]. \tag{4}$$

Mapped-coarse accuracy asks whether fine predictions recover the target's parent taxonomy, whereas direct prompt-coarse accuracy tests the original 20 coarse text prompts. On CIFAR-100, mapped-coarse first takes the 100-way fine argmax and applies the official map $\pi$; prompt-coarse takes the 20-way argmax against coarse text embeddings. Appendix B gives the exact prompt strings and mapping. Global prototype $\rho$ is Spearman correlation over upper-triangular pairwise cosine similarities of normalized class means. NP@5 is the mean overlap of their top-five prototype neighbors, excluding self. In the tables, *Drift* denotes $|V|^{-1} \sum_{x \in V} \|h_\theta(x) - h_0(x)\|_2^2$; semantic interface drift refers to the joint loss of prompt compatibility and global or local relations rather than this displacement alone.

Each axis tests a distinct link in the argument. Fine accuracy establishes that the adapted checkpoint solves the labeled task. Prompt and geometry metrics expose changes hidden by that accuracy. Held-out transfer tests whether the changed interface remains reusable on new labels and image distributions. Average displacement alone cannot identify which prompt decisions, class relations, or external datasets remain stable.

Table 1 summarizes the evaluation sequence: select by adaptation utility, then measure prompt compatibility, global and local geometry, held-out transfer, and reference correspondence.

Adaptation-dataset fine and coarse prompts use "a photo of a [name]" with underscores replaced by spaces. Held-out transfer uses the seven-template ensemble in Appendix B, with no adaptation-taxonomy mapping.

## 4 Experiments

### 4.1 Experimental Setup

We call the labeled fine-tuning task the *adaptation dataset* and reserve *held-out transfer dataset* for subsequent zero-shot evaluation. We adapt OpenAI CLIP ViT-B/32 on 40,000 CIFAR-100 images (Krizhevsky, 2009) and select checkpoints on a disjoint, stratified 10,000-image validation split from the official training partition. The official CIFAR-100 test split remains sealed for adaptation-task confirmation. Table 2 reports validation results used for diagnosis and operating-point selection. The diagnosis and matched-shuffle experiments use five seeds; the expanded reference hierarchy

and ViT-B/16 replication use three. CIFAR-10 (Krizhevsky, 2009), STL-10 (Coates et al., 2011), Oxford-IIIT Pets (Parkhi et al., 2012), Food-101 (Bossard et al., 2014), ImageNet-R, ImageNet-A, ImageNet-V2, and ImageNet-Sketch are held-out transfer datasets. We also evaluate official LDIFS and WiSE-FT checkpoints on the four ImageNet variants, DTD, and EuroSAT using dataset-specific prompts and raw-image re-encoding.

The non-CIFAR replication adapts CLIP ViT-B/32 on 21,600 EuroSAT training images (Helber et al., 2019), selects checkpoints on 2,700 validation images, and evaluates a disjoint 2,700-image test split over three seeds. Same-image and shuffled-reference anchoring use the previously fixed $\lambda = 2$, and checkpoints are selected by adaptation-validation cross-entropy. Transfer is evaluated after this selection on ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), and DTD (Cimpoi et al., 2014). Every transfer score uses fixed dataset-specific text prompts and independently generated frozen image/text caches. A three-seed SigLIP2-base-patch16-224 replication (Tschannen et al., 2025) includes a frozen CIFAR-10 sanity baseline.

The image branch is a 512-dimensional identity-initialized residual adapter with a 256-dimensional bottleneck. We use AdamW, batch size 256, a 30-epoch maximum, three warm-up epochs, gradient clipping at 5.0, and validation-only checkpoint selection. The text encoder, prompt embeddings, and coarse prompts remain fixed. Appendix B gives the exact prompt and mapping construction. The remaining grid, preprocessing, checkpoint paths, and seed-level artifacts are supplied in the appendix package.

After feature caching, same-image L2 adds $17.5\%$ to the recorded adapter training-loop time; complete timings and their scope are reported in Appendix C.

We define the controls as follows. *Semantic metric* is an anisotropic feature loss whose rank-32 protected basis is the leading PCA basis of the centered, normalized 100-fine-plus-20-coarse text matrix. The table uses protected weight 1, orthogonal weight $\eta = 0.5$, and $\lambda = 3$. *PCGrad–L2* uses the same-image L2 anchor with $\lambda = 1$. It separately computes the task and retention gradients, then projects the task gradient away from the retention gradient when their inner product is negative. These definitions make Table 5 reproducible without consulting the appendix implementation listing.

**Selection and statistical unit.** The seed is the unit of comparison. Table 2 reports adaptation-validation diagnostics; its checkpoint is selected on the disjoint validation split and the sealed adaptation test is not used for selection. For the five-seed matched-reference experiment, we evaluate the prespecified $4 \times 3$ shuffled-reference grid ($\lambda \in \{.3, .5, 1, 2\}$ and three learning-rate tags) and choose the candidate closest to the seed-matched same-image-$\lambda = 2$ validation accuracy. Transfer is evaluated after this selection. The expanded hierarchy fixes $\lambda = 2$ for all references and uses the same validation-only checkpoint rule. The fixed-control screen and logit-KL sweep use the listed operating points. Released LDIFS/WiSE-FT checkpoints were selected independently of the six transfer datasets. For the shuffled-reference analysis, the paired statistic is the per-seed mean held-out transfer drop, so the two-sided paired $t$-test has $n = 5$.

**Evaluation axes.** Fine accuracy measures adaptation utility. Direct coarse-prompt accuracy, global prototype $\rho$, NP@5, and transfer change measure complementary aspects of the retained interface. Operating points are selected by adaptation utility and then compared on these retention axes. The matched-reference result tests correspondence, while the discrete sweep and controls locate the effect across anchor strengths and intervention points.

### 4.2 Target Utility Hides Interface Drift and Transfer Loss

Table 2 makes the hidden state visible. Unregularized adaptation reaches 78.59% fine accuracy, while direct coarse-prompt accuracy is 57.38%, global prototype correlation is .523, and NP@5 is .708. Same-image L2 with $\lambda = 2$ reaches 79.25% fine accuracy and retains 63.27% direct prompt compatibility, .959 prototype correlation, and .851 NP@5. Similar target utility thus coexists with sharply different interface states.

Across 30 seed–reference pairs, mean transfer drop has Spearman correlation .672 with $1 - \rho$, .620 with same-image displacement, and .559 with $1 - \mathrm{NP@5}$. Prompt and geometry changes

Table 2: Five-seed ViT-B/32 validation diagnosis (mean ± standard deviation). Target accuracy alone does not reveal the degradation of direct prompt compatibility or frozen prototype relations.

| Method | Fine ↑ | Prompt-coarse ↑ | Global $\rho$ ↑ | NP@5 ↑ | Drift ↓ |
|---|---|---|---|---|---|
| Frozen CLIP | 63.97 | 58.04 | 1.000 | 1.000 | 0.000 |
| Unregularized | 78.59 ± 0.19 | 57.38 ± 0.42 | .523 ± .025 | .708 ± .007 | 1.501 ± .012 |
| Same-image L2, $\lambda = 2$ | 79.25 ± 0.07 | 63.27 ± 0.08 | .9591 ± .0004 | .851 ± .003 | .069 |
| Same-image L2, $\lambda = 3$ | 79.07 ± 0.05 | 63.25 ± 0.03 | .9746 ± .0002 | .874 ± .001 | .049 |

Table 3: Reference hierarchy after CIFAR-100 adaptation (mean ± standard deviation over three seeds). All rows use the same feature penalty with $\lambda = 2$ and achieve similar target accuracy. Transfer drop is averaged over ImageNet-R/A/V2/Sketch for ViT-B/32 and CIFAR-10/STL-10/Pets/Food101 for ViT-B/16. Lower drop is better.

| Reference assigned to $x$ | Fine ↑ | Global $\rho$ ↑ | NP@5 ↑ | B/32 drop ↓ | B/16 drop ↓ |
|---|---|---|---|---|---|
| Same image $h_0(x)$ | 79.63 ± .08 | .9574 ± .0004 | .858 ± .002 | **8.41 ± .06** | **5.45 ± .04** |
| Nearest same-class image | 79.50 ± .05 | .9588 ± .0003 | .852 ± .000 | 17.99 ± .16 | 14.04 ± .18 |
| Random same-class image | 79.34 ± .03 | .9187 ± .0013 | .799 ± .001 | 25.73 ± .14 | – |
| Random same-superclass image | 79.34 ± .06 | .8127 ± .0033 | .759 ± .001 | 28.05 ± .20 | – |
| Fixed random image | 79.45 ± .06 | .5862 ± .0018 | .735 ± .005 | 28.89 ± .16 | 28.09 ± .34 |
| Minibatch mean | 79.39 ± .06 | .5931 ± .0017 | .741 ± .003 | 28.99 ± .11 | 28.70 ± .30 |

therefore track the external behavior that target accuracy misses. The matched comparison next varies reference assignment at fixed target utility and penalty form.

### 4.3 Reference Correspondence under Matched Target Utility

The shuffled-minus-same-image paired difference is 23.70 points across five seeds (95% paired $t$ interval $[20.61, 26.80]$, $p = 2.9 \times 10^{-5}$, direction $5/5$). Both objectives use the same penalty and reach nearly identical validation fine accuracy. Reference assignment alone produces a large difference in held-out retention.

The expanded hierarchy identifies which part of the reference matters. Table 3 holds $\lambda = 2$ fixed and keeps CIFAR-100 fine accuracy within 0.29 points across all rows. The same-image reference has an 8.41-point mean drop over ImageNet-R/A/V2/Sketch. The drop rises to 17.99 points for the nearest same-class image and 25.73 points for a random same-class image. Class relevance retains some transfer, while the original image coordinate retains substantially more.

A minibatch-mean reference produces a 28.99-point drop, nearly identical to a fixed random image at 28.89 points; global-mean and centered-shuffle controls give the same pattern in the supplementary material. The shared direction does not recover the missing transfer. ViT-B/16 reproduces the ordering, with mean drops of 5.45, 14.04, 28.09, and 28.70 points for same-image, nearest same-class, fixed-random, and minibatch-mean references. Instance correspondence retains more transfer than class relevance or aggregate direction.

### 4.4 Comparison with Preservation Baselines

We group baselines by intervention point. Within the fixed-text image-adapter protocol, the controls compare parameter anchoring, an anisotropic semantic metric, PCGrad–L2, and frozen-teacher logit KL. LDIFS-like objectives constrain output features (Mukhoti et al., 2024), ClipFit-style PEFT constrains parameter scope (Li et al., 2024), ProLIP fine-tunes the vision encoder's projection matrix, whereas its RLA variant is an external regularized linear adapter for black-box access (Fahes et al., 2026); ARF adds external semantic anchors (Han et al., 2024), and WiSE-FT interpolates raw checkpoints (Wortsman et al., 2022). These methods preserve different objects; the reference experiment isolates correspondence within a common objective.

The raw-checkpoint comparison uses CLIP ViT-B/32 adapted on CIFAR-100 and six disjoint-label datasets held out during training and checkpoint selection. Table 4 reports each released checkpoint's adapted accuracy and change from frozen CLIP. LDIFS has a mean change of $-3.39$ pp and is positive only on EuroSAT, while WiSE-FT has a mean change of $-8.02$ pp across all six datasets.

Table 4: Official raw-checkpoint baseline results on six disjoint-label datasets. Both methods use CLIP ViT-B/32 adapted on CIFAR-100. Entries are adapted-minus-frozen top-1 changes in percentage points; transfer data are not used for training or checkpoint selection.

| Method | IN-R | IN-A | IN-V2 | IN-Sketch | DTD | EuroSAT | Mean | Cosine |
|---|---|---|---|---|---|---|---|---|
| LDIFS ($\alpha = 10$) | $-2.35$ | $-5.17$ | $-5.79$ | $-3.81$ | $-4.41$ | $+1.19$ | $-3.39$ | .687 |
| WiSE-FT ($\alpha = .75$) | $-5.68$ | $-10.49$ | $-10.03$ | $-7.18$ | $-8.67$ | $-6.07$ | $-8.02$ | .741 |

WiSE-FT has the higher mean feature cosine (.741 versus .687 for LDIFS) but the lower mean transfer change, showing that global proximity alone does not rank transfer retention consistently.

### 4.5 Anchor-strength Sweep and Controls

The anchor-strength sweep exposes a specialization–preservation trade-off. Across CIFAR-10, STL-10, and Pets, unregularized adaptation has a 19.89-point mean transfer drop. $\lambda = 2$ reduces it to 3.78 points and $\lambda = 3$ to 2.59 points. Figure 2 reports the discrete scan; prompt compatibility and prototype retention rise over most of the sampled range, transfer drop decreases, and fine accuracy is non-monotonic with its highest sampled value near $\lambda = 2$.

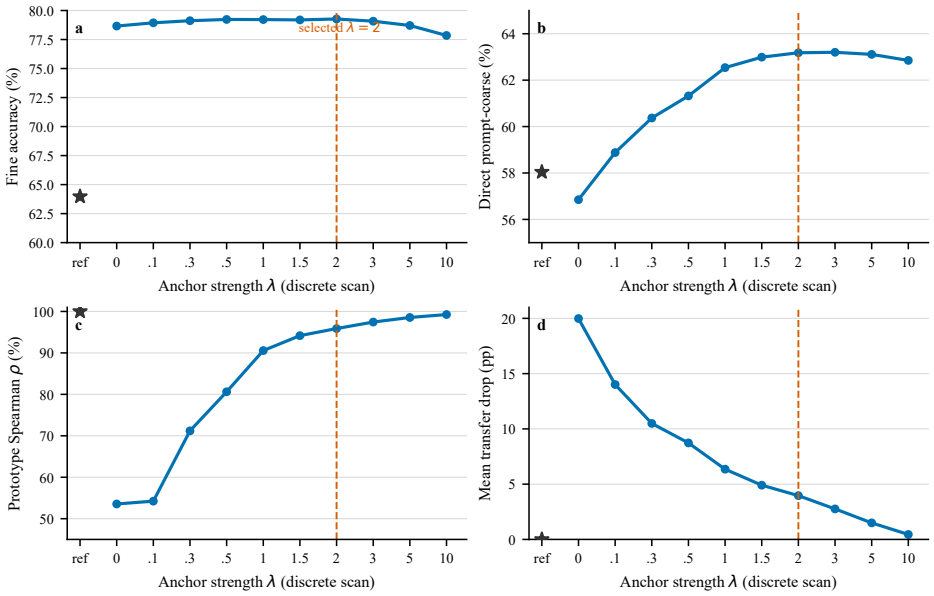



Figure 2: Discrete anchor-strength sweep under same-image anchoring. The points are the prespecified values $\lambda \in \{0, .1, .3, .5, 1, 1.5, 2, 3, 5, 10\}$ plus the frozen reference. Prompt compatibility and prototype geometry increase over most of the scan, held-out transfer drop decreases, and fine accuracy varies non-monotonically with its highest sampled value near $\lambda = 2$.

Table 5 compares operating points across the controls. The semantic metric retains less direct prompt compatibility, geometry, and held-out transfer than same-image L2 at the balanced point. PCGrad–L2 reaches stronger retention than $\lambda = 2$ while giving up 0.44 fine-accuracy points. Thus $\lambda = 2$ is the balanced same-image point, whereas $\lambda = 3$ and PCGrad–L2 trace a stronger-preservation trade.

A five-seed frozen-teacher logit-KL control preserves outputs on the 100 fine adaptation prompts. Across $T = .5$ and $\lambda \in \{.1, .3, 1, 3\}$, fine accuracy remains 79.16–79.19%, while adaptation-test geometry stays near the unregularized range ($\rho = .519$ to .511 and NP@5 from .720 to .719). The same-image-L2 checkpoint reaches $\rho = .958$ and NP@5 $= .857$ at 79.57% fine accuracy. This prompt-output control leaves geometry near its unregularized level; Table 6 gives the sweep.

Table 5: Control screen on the ViT-B/32 adaptation-validation split (mean ± standard deviation over five seeds). Same-image L2 gives the highest fine accuracy at the balanced point; PCGrad–L2 gives the strongest prompt and prototype retention at lower fine accuracy.

| Method | Fine accuracy ↑ | Prompt-coarse ↑ | Global $\rho$ ↑ | Mean transfer drop ↓ |
|---|---|---|---|---|
| Unregularized adapter | 78.59 ± 0.19 | 57.38 ± 0.42 | .523 ± .025 | 19.89 ± 0.30 |
| Same-image L2, $\lambda = 2$ | **79.25 ± 0.07** | 63.27 ± 0.08 | .9591 ± .0004 | 3.78 ± 0.15 |
| Semantic metric | 79.19 ± 0.08 | 62.75 ± 0.07 | .9503 ± .0004 | 4.15 ± 0.09 |
| PCGrad–L2 | 78.80 ± 0.13 | **63.31 ± 0.18** | **.974 ± .004** | 2.81 ± 0.25 |

Table 6: Five-seed frozen-teacher logit-KL control on the sealed ViT-B/32 adaptation test. Logit-KL leaves geometry near unregularized adaptation, whereas same-image L2 retains it.

| Method | Fine ↑ | Prompt-coarse ↑ | Global $\rho$ ↑ | NP@5 ↑ | Drift ↓ |
|---|---|---|---|---|---|
| Unregularized | 79.16 ± 0.23 | 56.57 ± 0.28 | .519 ± .024 | .720 ± .011 | 1.503 ± .012 |
| Logit KL, $\lambda = .1$ | 79.16 ± 0.23 | 56.57 ± 0.28 | .519 ± .024 | .720 ± .011 | 1.503 ± .012 |
| Logit KL, $\lambda = .3$ | 79.17 ± 0.24 | 56.55 ± 0.29 | .519 ± .024 | .720 ± .010 | 1.502 ± .012 |
| Logit KL, $\lambda = 1$ | 79.17 ± 0.22 | 56.53 ± 0.30 | .517 ± .024 | .720 ± .009 | 1.501 ± .012 |
| Logit KL, $\lambda = 3$ | 79.19 ± 0.21 | 56.37 ± 0.27 | .511 ± .024 | .719 ± .009 | 1.495 ± .012 |
| Same-image L2, $\lambda = 2$ | **79.57 ± 0.06** | **62.51 ± 0.06** | **.9576 ± .0003** | **.857 ± .002** | **.0692 ± .0002** |

### 4.6 Replication across Adaptation Datasets and Backbones

We first change the adaptation dataset. CLIP ViT-B/32 is adapted on EuroSAT with the same adapter and fixed $\lambda = 2$ objectives. Unregularized, same-image, and shuffled-reference adaptation reach 94.84%, 94.90%, and 94.79% adaptation-test accuracy, respectively. The near match enables a comparison of interface retention at similar adaptation utility.

Table 7: EuroSAT adaptation-dataset replication (three seeds). Same-image anchoring retains interface geometry and transfer at nearly identical adaptation accuracy.

**(a) EuroSAT adaptation task and geometry**

| Method | Test accuracy ↑ | Global $\rho$ ↑ | NP@5 ↑ | Drift ↓ |
|---|---|---|---|---|
| Unregularized | 94.84 ± .06 | .317 ± .019 | .660 | 1.790 ± .005 |
| Same-image L2, $\lambda = 2$ | **94.90 ± .06** | **.949 ± .002** | **.907 ± .012** | **.086 ± .001** |
| Shuffled L2, $\lambda = 2$ | 94.79 ± .06 | .295 ± .012 | .613 ± .012 | .2454 ± .0004 |

**(b) Transfer accuracy after EuroSAT adaptation**

| Method | IN-R | IN-A | IN-V2 | IN-Sketch | DTD | Mean $\Delta$ ↑ |
|---|---|---|---|---|---|---|
| Frozen CLIP | 65.48 | 29.24 | 55.01 | 40.96 | 43.19 | 0.00 |
| Unregularized | 23.93 ± .23 | 9.84 ± .21 | 16.32 ± .24 | 11.62 ± .12 | 23.55 ± .40 | −29.73 ± .21 |
| Same-image L2, $\lambda = 2$ | **53.37 ± .18** | **23.89 ± .18** | **44.46 ± .28** | **31.11 ± .08** | **37.78 ± .11** | **−8.66 ± .14** |
| Shuffled L2, $\lambda = 2$ | 9.58 ± .39 | 3.63 ± .08 | 6.40 ± .07 | 3.14 ± .09 | 10.96 ± .19 | −40.03 ± .08 |

**(c) Utility-matched reference hierarchy**

| Reference | Val. accuracy ↑ | Mean transfer drop ↓ |
|---|---|---|
| Same image | **94.63 ± .06** | **8.66 ± .14** |
| Nearest same-class image | 94.59 ± .06 | 23.98 ± .26 |
| Fixed random image | 94.49 ± .02 | 40.02 ± .13 |
| Minibatch mean | 94.51 ± .02 | 40.15 ± .11 |

Table 7 shows that same-image anchoring retains prototype correlation of .949 and limits the five-domain transfer drop to 8.66 points, whereas shuffled-reference anchoring reaches $\rho = .295$ and a 40.03-point drop. The expanded hierarchy keeps validation accuracy between 94.49% and 94.63%, yet transfer drops increase from 8.66 points for the same-image reference to 23.98 for the nearest same-class reference, 40.02 for a fixed random image, and 40.15 for the minibatch mean. Despite its smaller displacement than unregularized adaptation (.245 versus 1.790), the shuffled reference preserves less geometry and transfer. Same-image L2 has the smaller drop in all 15 seed–dataset comparisons.

ViT-B/16 independently reproduces the full reference ordering in Table 3. In the earlier unregularized comparison, same-image anchoring also changes prototype $\rho$ from $.486$ to $.926$ and reduces the three-dataset transfer drop from 21.77 to 3.72 points at similar sealed adaptation-test accuracy.

A parameter anchor reaches a 5.07-point drop on ViT-B/16 and lower direct prompt compatibility and prototype geometry than same-image anchoring. Under this adapter, the image-feature reference aligns more closely with the measured interface than the weight-space reference.

On Food101, same-image anchoring with $\lambda = 2$ reduces the transfer drop from 35.77 to 14.18 points.

The three-seed SigLIP2 replication uses a separately generated cache whose frozen CIFAR-10 accuracy is 93.85%. Unregularized adaptation reduces this accuracy to $82.44 \pm 0.39\%$, a drop of 11.41 points. Same-image anchoring retains $92.53 \pm 0.08\%$ and limits the drop to 1.32 points. On the CIFAR-100 adaptation task, unregularized and anchored adapters reach nearly identical fine accuracy (84.45% and 84.43%), while anchoring changes global $\rho$ from $.355$ to $.941$ and NP@5 from $.652$ to $.829$. The retention ordering is reproduced on this modern backbone under its native interface.

The replications vary complementary parts of the setting. EuroSAT changes the adaptation dataset from natural objects to satellite imagery, ViT-B/16 changes visual capacity, and SigLIP2 changes the pretraining objective and native image–text interface. Adaptation utility remains closely matched in each comparison, allowing interface retention to be separated from ordinary task gains.

The same pattern appears from several angles. On EuroSAT, shuffled-reference anchoring moves features less than unregularized adaptation but retains less prototype geometry and transfer. On CIFAR-100, nearest and random same-class references retain progressively less transfer than the same-image reference, while aggregate-direction controls behave like random assignment. Across both settings, transfer retention follows the image–coordinate association more closely than update magnitude, class membership, or a shared direction.

## 5 Discussion

Semantic interface drift links representation change to transfer loss. Target accuracy admits many solutions, but only some remain compatible with frozen text labels. Reference identity determines the fixed point: same-image anchoring keeps each pretrained coordinate, same-class references permit rearrangement, and random or aggregate references impose different assignments. Their ordering across CIFAR-100, EuroSAT, and two CLIP backbones ties transfer retention to the original image–coordinate correspondence.

## 6 Limitations

The experiments cover CIFAR-100 and EuroSAT adaptation, two CLIP backbones, SigLIP2, and held-out transfer. Full-model fine-tuning and larger backbones would test whether the same relationship holds when adaptation changes the encoder itself.

## 7 Conclusion

Target accuracy can hide transfer degradation in vision–language adaptation. Semantic interface drift exposes the accompanying loss of prompt compatibility and representation geometry. At matched target utility, same-image references retain more transfer than shuffled, class-related, and aggregate references across the tested datasets and model families. Preserving the pretrained image–coordinate assignment therefore helps maintain a reusable interface after adaptation.

## Reproducibility Statement

Section 4 describes evaluation and checkpoint selection. Appendices A–C provide the derivations, prompts, label mapping, and additional controls. The accompanying records contain splits, preprocessing, seeds, grids, checkpoints, metrics, and transfer outputs. An anonymized release of the PGMG implementation, experiment entry points, and figure-generation code is available at `https://anonymous.4open.science/r/pgmg-anonymous-source-06BD/`. Datasets, pretrained weights, and generated caches are obtained from their official sources and are not redistributed.

## AI Use Statement

Generative AI tools assisted with refining research questions and hypotheses; developing conceptual and mathematical arguments, including proof derivations; designing experiments and implementing/debugging code; interpreting results; searching and summarizing literature and checking references; drafting and editing manuscript text; and assisting with scientific figures. The authors take responsibility for the final content.

## References

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, 2014. doi: 10.1007/978-3-319-10599-4_29.

Dexia Chen, Qianjie Zhu, Weibing Li, Yue Yu, Tong Zhang, and Ruixuan Wang. Preserve and sculpt: Manifold-aligned fine-tuning of vision-language models for few-shot learning. *arXiv preprint arXiv:2508.12877*, 2025. URL https://arxiv.org/abs/2508.12877.

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.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 215–223, 2011.

Karan Desai, Maximilian Nickel, Tushar Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 7694–7713, 2023.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Perez, and Raoul de Charette. Clip's visual embedding projector is a few-shot cornucopia. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3254–3264, 2026. doi: 10.1109/WACV61042.2026.00318.

Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Anchor-based robust finetuning of vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26909–26918, 2024. doi: 10.1109/cvpr52733.2024.02542.

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. doi: 10.1109/JSTARS.2019.2918242.

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 *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.

Sungyeon Kim, Boseung Jeong, Donghyun Kim, and Suha Kwak. Efficient and versatile robust fine-tuning of zero-shot models. In *European Conference on Computer Vision (ECCV)*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14394–14410, 2024. doi: 10.18653/v1/2024.emnlp-main.797.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. *arXiv preprint arXiv:1802.01483*, 2018. URL https://arxiv.org/abs/1802.01483.

Jishnu Mukhoti, Yarin Gal, Philip H. S. Torr, and Puneet K. Dokania. Fine-tuning can cripple your foundation model; preserving features may be the solution. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=kfhoeZCeW7.

Giung Nam, Byeongho Heo, and Juho Lee. Lipsum-ft: Robust fine-tuning of zero-shot models using random text guidance. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 20654–20687, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5389–5400, 2019.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9929–9939, 2020.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.