Decoupled Mode Connectivity for Base-to-Novel Generalization in Vision-Language Models

Imad Eddine Marouf[1] **Khalid Oublal**[2] **Enzo Tartaglione**[1,3] **Stéphane Lathuilière**[1,4]
[1]Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
[2]École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France
[3]CNR-ISASI, Lecce, Italy [4]Inria, Univ. Grenoble Alpes, CNRS, LJK, France

Abstract

Prompt learning adapts vision-language models such as CLIP by optimizing a short learnable prompt. The cross-entropy objective pulls this prompt toward base-class specialization, whereas generalization to unseen classes benefits from staying close to CLIP's zero-shot feature space. Because both objectives act on the same prompt, their gradients conflict at convergence, and in our experiments no single-prompt loss design clearly improves on the cosine baseline. We propose **Decoupled Mode Connectivity (DMC)**, which gives each objective its own prompt. Starting from a base-trained prompt, DMC trains a specialization prompt and a generalization prompt, and merges their text features with a fixed coefficient at test time, at the inference cost of a single prompt. A linear mode connectivity (LMC) corridor in text-feature space keeps the classification loss low for every classifier interpolated between the two endpoints, and a *Visual Anchor* KL term preserves CLIP's zero-shot class-similarity distribution of each image during specialization. We further argue that such regularizers should satisfy *class-permutation invariance* (CPI), and show via Fano's inequality, in an idealized model, that violating CPI lower-bounds the drop in novel accuracy by a term proportional to the mutual information between the prompt and the base-class labels. Across 11 datasets, DMC improves the average harmonic mean of base and novel accuracy over MERGETUNE for both CoOp and KgCoOp, and its interpolation curves pass beyond the single-prompt operating point, shifting the base–novel frontier rather than trading along it. Code is available at `https://github.com/IemProg/DMC`.

1 Introduction

Vision-language models such as CLIP [Radford et al., 2021] acquire broad zero-shot recognition through large-scale pretraining on image-text pairs, and prompt learning [Zhou et al., 2022b,a, Yao et al., 2023, Khattak et al., 2023a,b, Yang et al., 2024] adapts these models to downstream tasks with very few additional parameters. This adaptation comes with a tension between two objectives: supervised training on base classes pushes the learned prompt toward base-class discriminative directions, while generalization to unseen (novel) classes benefits from keeping the prompt close to CLIP's pretrained feature space. Fine-tuning on base classes thus erodes part of CLIP's zero-shot ability to recognize the novel classes, which are absent from the fine-tuning data.

Existing methods address this tension with regularizers that anchor the prompt to the zero-shot model. Cosine penalties pull the learned text features toward CLIP's frozen text features [Yao et al., 2023, Wang et al., 2026]; self-distillation constraints preserve CLIP's predictions during fine-tuning [Khattak et al., 2023b, Roy and Etemad, 2024]; and linear mode connectivity (LMC) paths enforce smooth transitions between fine-tuned and zero-shot solutions [Wang et al., 2026]. These designs differ in form, but share a common structure: a single prompt optimized under a sum of cross-entropy and a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

regularization term. We argue that this shared structure, not the choice of regularizer, is the bottleneck. At a converged single prompt, the gradients of the two terms point in opposing directions on all eight dataset–baseline pairs we measure (four datasets, CoOp and KgCoOp; Figure 4), so the prompt sits at a compromise where reweighting the terms only moves along a trade-off. Consistently, none of seven single-prompt variants we test, including Fisher-weighted regularization and self-distillation, improves on the cosine baseline by clearly more than retuning its weight (Appendix E).

We propose **Decoupled Mode Connectivity (DMC)**, which changes the parameterization instead of the loss. DMC builds on MERGETUNE [Wang et al., 2026], a continued-training stage that refines a fine-tuned prompt with cross-entropy, a cosine regularizer toward CLIP's zero-shot text features, and an LMC path back to the fine-tuned model. DMC replaces this single refined prompt with two: a specialization prompt $\mathbf{c}_{\mathrm{spec}}$ trained with cross-entropy and a generalization prompt $\mathbf{c}_{\mathrm{gen}}$ trained with the regularizer. Each can then move toward its own optimum, and an LMC corridor in CLIP's text-feature space makes every mixture of the two a good base-class classifier. At test time, DMC merges the two text features once with a fixed coefficient, so inference costs the same as with a single prompt. Unlike post-hoc interpolation such as WiSE-FT [Wortsman et al., 2022b] or training-free merging [Yadav et al., 2023, Yu et al., 2024], the two prompts are trained jointly so that the path between them has low loss; training-free merging lowers HM in our experiments (Section 4.2). Figure 1 illustrates this geometry on Flowers102.

Freed from the cosine regularizer, the specialization prompt still needs a way to retain CLIP's knowledge. We argue that any regularizer used for this should be *class-permutation invariant* (CPI): it should not depend on which classes happen to be in the base set, because it must transfer to classes it never sees. The cosine regularizer is CPI but would pull $\mathbf{c}_{\mathrm{spec}}$ back to the zero-shot features; in an idealized model a CPI violation lower-bounds the drop in novel accuracy. Following this principle, DMC regularizes $\mathbf{c}_{\mathrm{spec}}$ with a label-free *Visual Anchor* (VA), a KL term that preserves CLIP's zero-shot class-similarity distribution of each image.

Our contributions are:

(1) We introduce **Decoupled Mode Connectivity (DMC)**, which assigns the specialization and generalization objectives to two prompts coupled by an LMC corridor in CLIP's text-feature space, at the inference cost of a single prompt (Section 3).

(2) We argue that regularizers should satisfy class-permutation invariance (CPI), show in an idealized model that a CPI violation lower-bounds the drop in novel accuracy (Proposition 1), and use this principle to design the Visual Anchor.

(3) Across 11 datasets, DMC improves the average harmonic mean (HM) of base and novel accuracy over MERGETUNE by $+0.78$ (CoOp) and $+0.33$ (KgCoOp), with gains on 9 and 8 of the 11 datasets, and its interpolation curves pass beyond the single-prompt operating point (Section 4, Figure 2).

The gains require the two prompts to stay distinct. On the adapter-based MMA [Yang et al., 2024], their text features nearly coincide ($\cos \approx 0.99$) and the regularizer and VA no longer help, which delimits the scope of DMC (Appendix H).

2 Related Work

Prompt learning for vision-language models. Prompt learning adapts frozen vision-language models by optimizing continuous prompt vectors instead of discrete templates. CoOp [Zhou et al., 2022b] introduced learnable text prompts for CLIP, achieving strong few-shot performance on base classes at the cost of novel-class generalization. CoCoOp [Zhou et al., 2022a] addressed this by conditioning prompts on image features, improving transferability but adding inference overhead. MaPLe [Khattak et al., 2023a] extended prompting to both vision and language branches, improving cross-modal alignment. PromptSRC [Khattak et al., 2023b] introduced self-regularization constraints that preserve pretrained knowledge during adaptation. KgCoOp [Yao et al., 2023] added a cosine penalty between learned and hand-crafted prompt features. More recently, CoPrompt [Roy and Etemad, 2024] enforced consistency between trainable and pretrained predictions. All these methods optimize a single set of prompts under a combined loss, and face the trade-off between base specialization and novel generalization that we target.

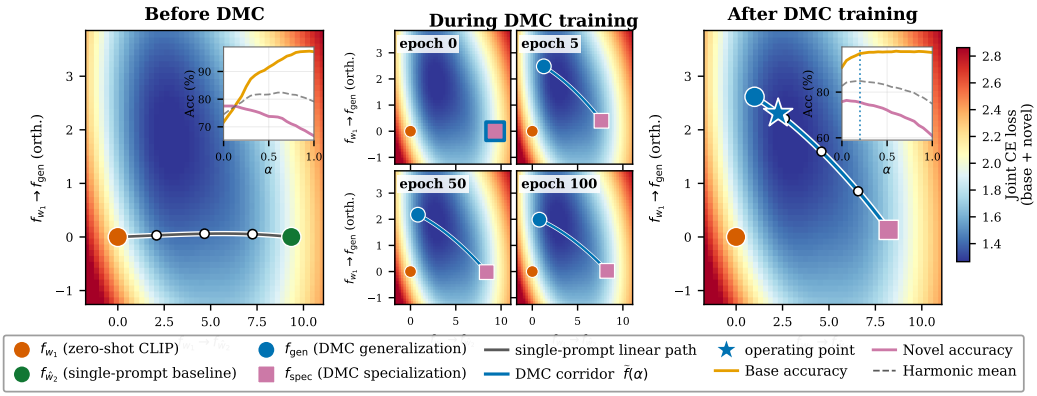

Figure 1: **Decoupled Mode Connectivity on Flowers102** (joint CE loss landscape). The 2D plane in CLIP text-feature space is anchored at f_{w_1} (zero-shot CLIP); the x-axis spans toward $f_{\hat{w}_2}$ (fine-tuned CoOp checkpoint) and the y-axis is the Gram–Schmidt orthogonal component of f_{gen} (text features of DMC's converged generalization prompt) relative to f_{w_1}. Each grid point is re-normalized per class before the joint base+novel CE loss is evaluated; construction details in Appendix G. **Before DMC** (left): the single-prompt baseline traces the **linear path**; the inset shows the base–novel trade-off, with **base** and **novel** accuracies crossing over as α varies. **During DMC** (middle): the two prompts start collapsed at $f_{\hat{w}_2}$ (epoch 0) and separate during training. **After DMC** (right): the converged **DMC corridor** lies inside the low-loss valley, with the operating point at $\alpha = 0.20$ marked by the **star**; the inset confirms base and novel accuracies are both high and stable along the corridor.

Regularization for base-to-novel generalization. The base-to-novel setting requires the fine-tuned model to generalize to classes unseen during training. Prior approaches regularize the learned representation toward the pretrained model: PromptSRC [Khattak et al., 2023b] constrains prompt, feature, and logit spaces simultaneously. MERGETUNE [Wang et al., 2026] combines an isotropic cosine score with a linear mode connectivity path to merge zero-shot and fine-tuned solutions. Fisher-based approaches [Kirkpatrick et al., 2017] weight regularization by parameter importance but introduce class-specific bias, which we show violates class-permutation invariance; a Fisher-weighted cosine brings no gain beyond retuning the cosine weight (Appendix I). A common limitation is that all single-prompt methods enforce competing objectives on the same parameters. Multi-task gradient analysis methods such as PCGrad [Yu et al., 2020] manage such conflicts by modifying gradient updates, but still optimize a single shared prompt. DMC instead assigns the two objectives to separate prompts.

Model merging and mode connectivity. Garipov et al. [2018] and Draxler et al. [2018] established that independently trained solutions can be connected by low-loss paths in weight space. Practical methods built on this insight: WiSE-FT [Wortsman et al., 2022b] interpolates between zero-shot and fine-tuned weights; model soups [Wortsman et al., 2022a] average multiple fine-tuned checkpoints; task arithmetic [Ilharco et al., 2023] composes task-specific weight vectors through addition and subtraction; TIES-Merging [Yadav et al., 2023] and DARE [Yu et al., 2024] resolve parameter interference during merging. These methods operate in weight space and merge models after training. By contrast, DMC merges in text-feature space and trains the LMC corridor as an objective, so both endpoints are optimized jointly for corridor quality.

3 Method

3.1 Background and Problem Statement

CLIP prompt learning. CLIP [Radford et al., 2021] encodes images and text into a shared ℓ_2-normalized embedding space via frozen visual and text encoders V and T. An image x is classified by computing cosine similarities between its visual feature $v = \ell_2(V(x))$ and the text features $\{f_k = \ell_2(T(p_k))\}_{k=1}^{C}$, where p_k is a text prompt for class k. Prompt learning [Zhou et al., 2022b] replaces hand-crafted templates with a learnable prompt $\mathbf{c} \in \mathbb{R}^{N \times D}$, shared across all classes, and

optimizes it on base-class data while keeping both encoders frozen. Throughout, we write $\ell_2(\cdot)$ for row-wise ℓ_2-normalization, $f_{w_1} = \ell_2(T(p_{\text{hand}})) \in \mathbb{R}^{C\times D}$ for the zero-shot text features obtained from hand-crafted templates, and γ for CLIP's learned logit scale. A prompt $\mathbf{c}$ induces the text features $f(\mathbf{c}) = \ell_2(T(\mathbf{c}))$, whose row f_k is the feature of class k; f_{gen} and f_{spec} denote the features of the two DMC prompts.

Single-prompt training and MERGETUNE. Let $\hat{w}_2$ denote the prompt of a model (CoOp or KgCoOp) fine-tuned on base classes, with induced features $f_{\hat{w}_2} = \ell_2(T(\hat{w}_2))$. The generalization regularizer $\mathcal{R}$ anchors the learned features to the zero-shot features f_{w_1} by measuring their mean cosine misalignment:

$$\mathcal{R}(f,\, f_{w_1}) = 1 - \frac{1}{C}\sum_{k=1}^{C} \cos\big(f_k,\; f_{w_1,k}\big). \tag{1}$$

$\mathcal{R}$ is defined on any normalized text-feature matrix $f \in \mathbb{R}^{C\times D}$; single-prompt training uses $f = f(\mathbf{c})$. MERGETUNE [Wang et al., 2026] continues training from $\hat{w}_2$. It learns a prompt $\mathbf{c}$ that combines $\mathcal{R}$ with cross-entropy and a linear mode connectivity (LMC) term:

$$\mathcal{L}_{\text{MERGETUNE}}(\mathbf{c}) = \mathcal{L}_{\text{CE}}(f(\mathbf{c})) \;+\; \lambda\,\mathcal{R}(f(\mathbf{c}),\, f_{w_1}) \;+\; \beta\,\mathbb{E}_\alpha\,\mathcal{L}_{\text{CE}}\big(\ell_2\big(f_{\hat{w}_2} + \alpha\,(f(\mathbf{c}) - f_{\hat{w}_2})\big)\big), \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy on base-class training data, $\mathcal{R}$ is the cosine regularizer from Eq. (1), and the third term enforces a low-loss LMC path from $f_{\hat{w}_2}$ to $f(\mathbf{c})$. The weight λ balances specialization against generalization and β weights the LMC path; DMC reuses these terms, with the corridor redefined between its two prompts, and writes the regularizer weight as λ_{gen}.

One prompt, two objectives. Specializing for base classes moves the text features away from their zero-shot positions, and $\mathcal{R}$ pulls them back. With one shared prompt $\mathbf{c}$, both gradients act on the same parameters, and they conflict: at the converged MERGETUNE prompt, the cosine ρ between $\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}$ and $\nabla_{\mathbf{c}}\mathcal{R}$ is negative on all eight dataset–baseline pairs we measure (Figure 4). Every stationary point of $\mathcal{L}_{\text{CE}} + \lambda\,\mathcal{R}$ satisfies $\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}} = -\lambda\,\nabla_{\mathbf{c}}\mathcal{R}$, which gives the following lemma (proof in Appendix M).

Lemma 1 (Pareto stationarity of single-prompt solutions)**.** *Let $\mathcal{L}_{\text{CE}}$ and $\mathcal{R}(\cdot,\, f_{w_1})$ be continuously differentiable in $\mathbf{c}$ (we write $\mathcal{L}_{\text{CE}}(\mathbf{c})$ and $\mathcal{R}(\mathbf{c}, f_{w_1})$ for $\mathcal{L}_{\text{CE}}(f(\mathbf{c}))$ and $\mathcal{R}(f(\mathbf{c}), f_{w_1})$), and let $\mathbf{c}^*$ be a stationary point of $\mathcal{L}(\mathbf{c}) = \mathcal{L}_{\text{CE}}(\mathbf{c}) + \lambda\,\mathcal{R}(\mathbf{c},\, f_{w_1})$ for some $\lambda > 0$. Then $\mathbf{c}^*$ is Pareto-stationary for the bi-objective problem $(\mathcal{L}_{\text{CE}},\, \mathcal{R})$: no direction in $\mathbb{R}^{N\times D}$ simultaneously decreases both objectives to first order. If additionally $\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}(\mathbf{c}^*) \neq \mathbf{0}$ and $\nabla_{\mathbf{c}}\mathcal{R}(\mathbf{c}^*) \neq \mathbf{0}$, then neither objective is individually minimized at $\mathbf{c}^*$.*

Changing λ, or replacing $\mathcal{R}$ by another regularizer, therefore only selects another compromise inside the same prompt. The lemma does not show that two prompts do better; we show this empirically: none of seven single-prompt variants improves on the baseline by clearly more than retuning λ (Appendix E), while DMC does on CoOp and KgCoOp (Section 4).

A scalar model shows what decoupling changes (Appendix C). With one parameter and two quadratic objectives minimized at $a \neq b$, every weighting gives a point between a and b that minimizes neither objective. With one parameter per objective, each reaches its own minimum, and interpolating between the two recovers every single-parameter compromise, so the trade-off becomes a choice of an interpolation coefficient at test time instead of a choice of λ at training time. In the high-dimensional prompt space, the interpolation path between two prompts need not follow the single-prompt path, so it can also reach points that no single prompt reaches; Figure 2 measures this.

3.2 Decoupled Mode Connectivity

DMC replaces the continued training of one prompt in MERGETUNE with two prompts, one per objective, both initialized from $\hat{w}_2$:

- **Specialization prompt $\mathbf{c}_{\text{spec}}$**: trained by cross-entropy on base-class data, without the cosine pull toward f_{w_1}, and regularized only by the label-free Visual Anchor (below).
- **Generalization prompt $\mathbf{c}_{\text{gen}}$**: trained by $\mathcal{R}$ to stay close to f_{w_1}. It is learned rather than fixed to f_{w_1}, because the learned endpoint keeps discriminative structure that the zero-shot features lack (Section 4.3).

Neither prompt receives both endpoint losses, so each can move toward its own optimum. For the adapter-based MMA [Yang et al., 2024], the two endpoints are two copies of its adapters; only their text features are interpolated, and images are encoded once with the specialization adapters.

Coupling the prompts: the LMC corridor. Two independently trained prompts give two unrelated classifiers. DMC couples them with an LMC corridor that requires low classification loss at every re-normalized convex combination of their text features. Because classification scores are $v^\top f_k$ and T is nonlinear, a linear path between prompts is not a linear path between classifiers, so the corridor operates in text-feature space:

$$\widetilde{f}(\alpha) = \ell_2\Big(f_{\text{gen}} + \alpha\,\big(f_{\text{spec}} - f_{\text{gen}}\big)\Big), \qquad \alpha \in [0,1]. \tag{3}$$

The full DMC objective is

$$\mathcal{L}_{\text{DMC}} = \underbrace{\mathcal{L}_{\text{CE}}(f_{\text{spec}}) + \lambda_{\text{VA}}\,\mathcal{L}_{\text{VA}}}_{\mathbf{c}_{\text{spec}}\text{ only}} + \underbrace{\lambda_{\text{gen}}\,\mathcal{R}(f_{\text{gen}},\, f_{w_1})}_{\mathbf{c}_{\text{gen}}\text{ only}} + \underbrace{\beta\,\mathbb{E}_{\alpha\sim\mathcal{U}[0,1]}\,\mathcal{L}_{\text{CE}}\big(\widetilde{f}(\alpha)\big)}_{\text{LMC corridor (both)}}, \tag{4}$$

where $\mathcal{L}_{\text{VA}}$ is defined in Eq. (5) below. The corridor term ($\mathcal{L}_{\text{LMC}}$ in Algorithm 1) is the only coupling: it passes gradients, including cross-entropy gradients, to both prompts, while the endpoint losses act on one prompt each. Sampling α over the whole interval trains every point of the corridor to classify base data well, so the classifier deployed at test time was itself optimized during training. At inference, α is fixed to $\alpha_{\text{fix}} = 0.20$, selected on two development datasets (Section 4.1); a small value keeps the classifier close to f_{gen}, which favors novel classes. The merged features $\widetilde{f}(\alpha_{\text{fix}})$ are computed once, so inference costs the same as with a single prompt. Figure 1 shows this geometry on Flowers102: f_{gen} and f_{spec} separate during training, and the converged corridor lies inside a low-loss region of the cross-entropy landscape that the path from zero-shot to the fine-tuned CoOp checkpoint does not reach.

Regularizing $\mathbf{c}_{\text{spec}}$: class-permutation invariance. Without $\mathcal{R}$, cross-entropy alone tends to widen the angles between classes to maximize the base-class margin, which discards the smaller similarities that f_{w_1} encodes between related classes. Which regularizer should replace $\mathcal{R}$ on $\mathbf{c}_{\text{spec}}$? It must transfer to classes it never sees, so it should not depend on which classes are in the base set. We formalize this as class-permutation invariance (CPI).

Definition 1 (Class-permutation invariance)**.** *A regularizer $\mathcal{R}(f,\, f_{w_1})$ is* class-permutation invariant *(CPI) if*

$$\mathcal{R}(\Pi f,\, \Pi f_{w_1}) = \mathcal{R}(f,\, f_{w_1})$$

for every $C \times C$ permutation matrix Π, where $f \in \mathbb{R}^{C\times D}$ is the text-feature matrix with rows indexed by class.

Proposition 1 (CPI violation lower-bounds the novel-accuracy drop; informal)**.** *Let $\mathbf{c}$ be the learned prompt of a training procedure that uses a regularizer $\mathcal{R}$, and $\mathbf{c}^\star$ the learned prompt of the same procedure with $\mathcal{R}$ replaced by a CPI regularizer. Under four idealized assumptions, if $\mathcal{R}$ violates CPI, the novel-class accuracy of $\mathbf{c}$ is lower than that of $\mathbf{c}^\star$ by at least $I(\mathbf{c}; Y_{\text{base}})/\log|\mathcal{Y}_{\text{novel}}| > 0$, where $I(\mathbf{c}; Y_{\text{base}})$ is the mutual information between the prompt and a uniformly drawn base label.*

Appendix M.2 gives the formal statement and proof (where the prompt is written θ), and Appendix I.5 discusses when the assumptions fail. Empirically, the non-CPI regularizers we measure violate the symmetry by orders of magnitude more than the CPI ones, the one we trained (Fisher-weighted cosine) gives no gain beyond retuning λ, and CPI alone does not guarantee transfer (Appendix I). The cosine regularizer weights all classes equally:

Corollary 1 (Cosine regularizer satisfies CPI)**.** *The cosine regularizer $\mathcal{R}$ in Eq.* (1) *satisfies Definition 1. It uses no image labels, so $\mathcal{R}$ itself adds no base-label information to the prompt it regularizes.*

Visual Anchor. We regularize $\mathbf{c}_{\text{spec}}$ with the *Visual Anchor* (VA), a CPI regularizer that preserves CLIP's zero-shot class-similarity distribution of each image without using its label. For each training image x, let $v_a = \ell_2(V(\text{aug}(x)))$ and $v_b = \ell_2(V(\text{aug}'(x)))$ denote frozen visual features of two

Algorithm 1 Decoupled Mode Connectivity (DMC)

Require: Base-class data $\mathcal{D}_{\text{base}}$, frozen CLIP encoders V, T, zero-shot features f_{w_1}, fine-tuned checkpoint $\hat{w}_2$, hyperparameters $\lambda_{\text{gen}}, \lambda_{\text{VA}}, \beta, \tau$, inference coefficient α_{fix} (we use λ_{gen}=8, λ_{VA}=0.5, β=4, τ=2, α_{fix}=0.20)

Ensure: Test-time classifier

INITIALIZATION

$\mathbf{c}_{\text{spec}} \leftarrow \hat{w}_2$, $\mathbf{c}_{\text{gen}} \leftarrow \hat{w}_2$ ▷ Both initialized from the fine-tuned checkpoint

Freeze $V, T, \hat{w}_2$; mark only $\mathbf{c}_{\text{spec}}, \mathbf{c}_{\text{gen}}$ as learnable

TRAINING

for each mini-batch $(x, y) \sim \mathcal{D}_{\text{base}}$ **do**

  $f_{\text{spec}} \leftarrow \ell_2(T(\mathbf{c}_{\text{spec}}))$, $f_{\text{gen}} \leftarrow \ell_2(T(\mathbf{c}_{\text{gen}}))$, $v_a \leftarrow \ell_2(V(\text{aug}(x)))$, $v_b \leftarrow \ell_2(V(\text{aug}'(x)))$

  Compute $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{VA}}$ from f_{spec} (VA also uses f_{w_1}, v_a, v_b) ▷ Eqs. (4), (5)

  Compute $\mathcal{R}$ from f_{gen} ▷ Eq. (1)

  Sample $\alpha \sim \mathcal{U}[0, 1]$; compute $\mathcal{L}_{\text{LMC}}$ from $\widetilde{f}(\alpha)$ via Eq. (3) ▷ Eq. (4), couples both

  $\mathcal{L}_{\text{DMC}} \leftarrow \mathcal{L}_{\text{CE}} + \lambda_{\text{VA}}\,\mathcal{L}_{\text{VA}} + \lambda_{\text{gen}}\,\mathcal{R} + \beta\,\mathcal{L}_{\text{LMC}}$ ▷ Eq. (4)

  Update $\mathbf{c}_{\text{spec}}, \mathbf{c}_{\text{gen}}$ via $\nabla\mathcal{L}_{\text{DMC}}$

end for

INFERENCE (test image x)

$f_{\text{spec}} \leftarrow \ell_2(T(\mathbf{c}_{\text{spec}}))$, $f_{\text{gen}} \leftarrow \ell_2(T(\mathbf{c}_{\text{gen}}))$ ▷ Recompute from final trained prompts

$\widetilde{f} \leftarrow \ell_2\big(f_{\text{gen}} + \alpha_{\text{fix}}(f_{\text{spec}} - f_{\text{gen}})\big)$ ▷ Merge endpoints once; $\alpha_{\text{fix}} = 0.20$

return $\hat{y} = \arg\max_k\ \gamma\, \widetilde{f}_k^{\top}\, \ell_2(V(x))$

independently augmented views. VA penalizes the KL divergence between the class-similarity distributions induced by f_{w_1} (zero-shot) and f_{spec}:

$$\mathcal{L}_{\text{VA}} = \mathbb{E}_x\ \text{KL}\Big(\underbrace{\sigma\big(\gamma\, f_{w_1}\, v_a\,/\,\tau\big)}_{\text{teacher (zero-shot, view } a)} \ \Big\|\ \underbrace{\sigma\big(\gamma\, f_{\text{spec}}\, v_b\,/\,\tau\big)}_{\text{student } (\mathbf{c}_{\text{spec}}\text{, view } b)} \Big), \tag{5}$$

where $f_{w_1} v_a \in \mathbb{R}^C$ denotes the vector of inner products $[f_{w_1,k}^{\top} v_a]_{k=1}^{C}$ (likewise for $f_{\text{spec}} v_b$), $\sigma(\cdot)$ is the softmax function, γ puts both logit vectors on CLIP's zero-shot scale, and τ is a distillation temperature that softens both distributions. Permuting the classes permutes both distributions in the same way and leaves the KL unchanged, so VA satisfies CPI. Proposition 2 (Appendix M.4) relates this two-view loss to the single-view case $v_a = v_b$. Algorithm 1 summarizes the full procedure.

4 Experiments

4.1 Setup

Datasets and protocol. We follow the base-to-novel generalization protocol of Zhou et al. [2022b]. Each dataset's classes are split into base (seen) and novel (unseen) halves. Models are trained on 16-shot base-class data and evaluated on both base and novel test sets. We report Base accuracy, Novel accuracy, and their Harmonic Mean (HM), $\text{HM} = 2\,\text{Base} \cdot \text{Novel}/(\text{Base} + \text{Novel})$, which is high only when both accuracies are high. MERGETUNE and DMC rows are means over 3 seeds (DMC at $\alpha = 0.20$ and its best base-validation checkpoint; MERGETUNE evaluates its final prompt) (std in Table 4); prior-work rows are the published numbers. We evaluate on 11 benchmarks: ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, and UCF101.

Baselines. We compare against prompt learning methods: CoOp [Zhou et al., 2022b], CoCoOp [Zhou et al., 2022a], KgCoOp [Yao et al., 2023], MaPLe [Khattak et al., 2023a], PromptSRC [Khattak et al., 2023b], MMA [Yang et al., 2024]; methods using external knowledge: CoPrompt [Roy and Etemad, 2024], ATPrompt [Li et al., 2025]; and model merging baselines: TIES [Ya-

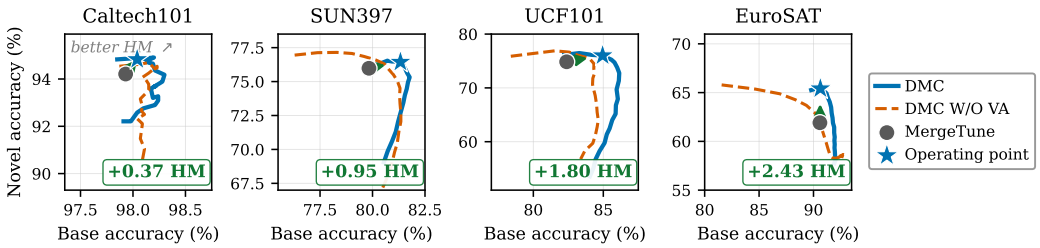

Figure 2: **Empirical interpolation curves** on Caltech101, SUN397, UCF101, and EuroSAT (CoOp base, 3-seed mean). Each curve traces the family of classifiers from f_{gen} (α=0, near zero-shot) to f_{spec} (α=1, fully specialized). Solid: **DMC** (full method); dashed: **DMC w/o VA**. **Gray dots** mark the **MERGETUNE baseline**; **blue stars** mark the operating point; the **green arrow** shows the shift from baseline to operating point. The boxed ΔHM value reports the gap (operating point vs. baseline), read off the seed-averaged curves; it can differ from Table 1 by up to 0.23 HM.

dav et al., 2023], DARE [Yu et al., 2024]. MERGETUNE [Wang et al., 2026] is a continued fine-tuning stage applied post hoc to existing parameter-efficient fine-tuning (PEFT) methods. DMC replaces this stage: it starts from the same fine-tuned checkpoint and trains two prompts instead of one.

Implementation details. All methods use a CLIP ViT-B/16 backbone. DMC trains two prompts initialized from a fine-tuned CoOp, KgCoOp, or MMA checkpoint, with 4×512 parameters per prompt on the prompt-tuning baselines ($\sim$2K parameters per prompt). Training hyperparameters: $\lambda_{\text{gen}} = 8.0$, $\beta = 4.0$, $\lambda_{\text{VA}} = 0.5$, $\tau = 2.0$; learning rate and epochs follow MERGETUNE's defaults; batch size in Appendix J. The interpolation coefficient α is sampled uniformly from $[0, 1]$ at every training step. At inference we use a single fixed $\alpha = 0.20$, selected on Flowers102 and Caltech101, which are also included in all reported averages. On the eight held-out datasets (ImageNet excluded), $\alpha = 0.20$ is on average 0.18 HM below the per-dataset oracle (Appendices J and K). The choice is not specific to this pair: over all 55 two-dataset development pairs, 47 select an α in $[0.15, 0.25]$, and 54 stay within 0.50 HM of the oracle on their held-out datasets (Table 18). Only the interpolated text feature $\widetilde{f}(\alpha)$ is used, so DMC adds no per-image inference cost over a single-prompt method. Training adds only the second prompt (4,096 trainable parameters in total): one run on Flowers102 (100 epochs, batch size 128, as in the ablations; the main runs use 32) takes 15 min 34 s and 5.9 GB of peak memory on a single A100 (Appendix J). The β sweep is in Appendix L.1.

4.2 Base-to-novel generalization

We apply DMC to three fine-tuning baselines (CoOp, KgCoOp, MMA) and compare it with the MERGETUNE continued-training stage and with two training-free merging methods (TIES, DARE). Table 1 reports HM on 11 datasets; Base and Novel accuracies are in Appendix A. In Table 3 (best-of-3-seed rows, Appendix A), DMC raises Base and Novel together over MERGETUNE on 9 of 11 datasets for CoOp, so the CoOp HM gains mostly come from a Pareto shift. For KgCoOp this holds on only 4 of 11 datasets, and the gain comes mainly from Base. MMA also raises both on 8 of 11, but its two DMC endpoints collapse (see below). For MMA, both MERGETUNE and DMC remain below the published MMA; for KgCoOp, DMC roughly matches the published KgCoOp (+0.05 HM when both are averaged over per-dataset HMs), while MERGETUNE stays below it. Figure 2 shows DMC's interpolation curves on four datasets.

Training-free merging. On average, TIES and DARE reduce HM relative to the fine-tuned baseline: CoOp loses 5.34 and 1.07 HM, KgCoOp 4.45 and 1.84, and MMA 10.46 and 8.06. Training-free merging therefore cannot replace the trained corridor.

Comparison with MERGETUNE. In the same pipeline, DMC improves HM over MERGETUNE on 9 of 11 datasets for CoOp and 8 of 11 for KgCoOp. The largest gains are on EuroSAT (+2.30), FGVCAircraft (+1.90) and StanfordCars (+1.83) for CoOp. Every CoOp and KgCoOp regression is within seed variance. Averaged over the 11 datasets, the gain is +0.78 HM (CoOp), +0.33 (KgCoOp) and +0.58 (MMA).

Table 1: Base-to-novel HM on 11 datasets. MERGETUNE and DMC: mean over 3 seeds ($\alpha = 0.20$). Bold: better of MERGETUNE and DMC within each base method. Avg: mean of the 11 per-dataset HMs for our runs; for published, TIES and DARE rows, HM of the average Base and Novel, as in Table 3. †: uses LLM or teacher-model knowledge.

Method	Avg	HM per Dataset										
		IN	Cal	Pets	Cars	Flo	Food	Air	SUN	DTD	Euro	UCF
CLIP (ICML 21)	71.70	70.22	95.40	94.12	68.65	74.83	90.66	31.09	72.23	56.37	60.03	73.85
CoOp (IJCV 22)	71.66	71.92	93.73	94.47	68.13	74.06	85.19	28.75	72.51	54.24	68.69	67.46
KgCoOp (CVPR 23)	77.01	72.78	96.03	96.18	73.36	83.65	91.10	34.83	78.36	64.35	73.48	79.66
MaPLe (CVPR 23)	78.55	73.47	96.02	96.58	73.47	82.56	91.38	36.50	79.75	68.16	82.35	80.77
PromptSRC (ICCV 23)	79.97	74.01	96.02	96.30	76.58	85.95	91.10	40.15	80.52	71.75	82.32	82.74
MMA (CVPR 24)	79.87	74.02	96.15	96.72	75.70	85.48	90.71	38.33	80.38	73.38	83.87	82.20
CoPrompt† (ICLR 24)	80.48	74.33	96.55	96.87	75.66	85.71	91.40	39.76	81.30	72.79	85.84	83.07
CoOp + ATPrompt† (ICCV 25)	74.65	73.33	95.74	95.67	71.58	79.77	88.09	32.52	74.24	58.22	71.96	73.45
CoOp + TIES (NeurIPS 23)	66.32	65.60	92.50	92.56	60.56	69.27	88.02	29.54	62.70	48.40	58.40	58.30
CoOp + DARE (ICML 24)	70.59	68.93	93.67	94.84	64.57	74.92	87.50	31.03	66.47	57.04	67.39	66.28
CoOp + MERGETUNE (ICLR 26)	76.45	72.89	96.27	**96.13**	73.38	83.92	91.05	33.33	77.85	**64.26**	73.45	78.42
CoOp + DMC (Ours)	**77.23**	**72.90**	**96.41**	96.04	**75.21**	**84.47**	**91.13**	**35.23**	**78.75**	63.53	**75.75**	**80.07**
KgCoOp + TIES (NeurIPS 23)	72.56	70.93	95.83	95.23	69.73	73.39	90.91	29.86	73.46	56.93	67.02	74.18
KgCoOp + DARE (ICML 24)	75.17	71.78	95.97	94.73	70.33	79.33	91.02	31.85	76.12	62.01	72.19	78.71
KgCoOp + MERGETUNE (ICLR 26)	76.43	72.82	95.94	**96.28**	74.90	83.39	91.08	**34.10**	78.03	62.66	**73.44**	78.08
KgCoOp + DMC (Ours)	**76.76**	**73.09**	**96.11**	96.23	**75.38**	**83.67**	**91.19**	33.17	**78.85**	**64.34**	73.38	**78.90**
MMA + TIES (NeurIPS 23)	69.41	69.66	93.77	92.17	65.08	72.07	89.31	28.65	67.56	51.63	66.80	65.60
MMA + DARE (ICML 24)	71.81	70.27	93.70	92.96	68.01	73.92	88.72	29.25	72.02	57.74	70.17	70.78
MMA + MERGETUNE (ICLR 26)	72.58	**73.15**	**95.75**	93.22	70.16	75.22	90.18	32.72	76.70	58.07	59.57	73.59
MMA + DMC (Ours)	**73.16**	72.90	95.74	**93.38**	**70.51**	**76.37**	**90.33**	**33.10**	**77.12**	**59.62**	**60.57**	**75.11**

Dependence on the base method. The average gain is smallest on KgCoOp, whose built-in cosine regularization already anchors the prompt to zero-shot features, and largest on CoOp, which has no such anchor. This is a method-level pattern only: per dataset, the magnitude of the gradient conflict does not predict the gain (Pearson $r = -0.05$, Appendix D). StanfordCars, where zero-shot CLIP is weak, improves for all three baselines.

When does DMC help? DMC helps only if the two endpoints stay apart. If $f_{\text{gen}} \approx f_{\text{spec}}$, every $\widetilde{f}(\alpha)$ is the same classifier. In the runs of Table 1, $\cos(f_{\text{gen}}, f_{\text{spec}})$ averages 0.628 (CoOp) and 0.704 (KgCoOp) over 10 datasets (ImageNet not measured). On MMA, an adapter-based method, it averages 0.996 and exceeds 0.99 on 9 of 10 datasets (Table 10). The two endpoints collapse, so the corridor has almost no effect on MMA. MMA+DMC still gains 0.58 HM over MMA+MERGETUNE, but these gains do not track the remaining separation. Both stay below published MMA on every dataset, and the component ablation on MMA removes or reverses every gain of $\mathcal{R}$ and VA seen on CoOp (Appendix H).

Interpolation curves. DMC's interpolation curve passes northeast of the single-prompt operating point on every panel of Figure 2. The single-prompt MERGETUNE point does not reach these (Base, Novel) points, consistent with Lemma 1: rescaling λ only moves a shared prompt along its Pareto-stationary set.

4.3 Ablation study

DMC component analysis. Table 2 toggles one component per row on four datasets. The anchor $\mathcal{R}$ is what makes decoupling work: without it, both prompts collapse to the same base-specialized solution (Appendix L.2) and DMC falls below MERGETUNE on every dataset, so a second prompt alone does not explain the gain. With $\mathcal{R}$, DMC beats single-prompt MERGETUNE on three of four datasets ($+1.47$ on Flowers102); the EuroSAT gap (-0.40) is within its seed spread (± 4). VA is a small refinement, positive on two of four datasets. HM plateaus for $\lambda_{\text{gen}} \geq 4$ (Appendix L.2).

Capacity and view controls. DMC uses two 2K-parameter prompts; a single CoOp prompt with $N_{\text{ctx}} = 16$ (8K parameters) still trails CoOp+DMC on all four datasets, by 1.07 to 5.07 HM (Table 7).

Table 2: Component ablation (HM, 3 seeds, $\beta = 4$ for all rows; batch size 128, whereas Table 1 uses 32, so values differ from it). Δ: change from the row above; bold: best per column. Rows without $\mathcal{R}$ omit VA.

					Flowers102		Caltech101		StanfordCars		EuroSAT		Average	
Method	**CE**	$\mathcal{R}$	**LMC**	**VA**	HM	Δ	HM	Δ	HM	Δ	HM	Δ	HM	Δ
CoOp	✓				80.51±1.54	—	94.42±0.60	—	71.98±0.70	—	71.18±5.75	—	79.52	—
MERGETUNE w/o $\mathcal{R}$	✓		✓		78.19±2.40	−2.32	94.43±1.89	+0.01	71.04±1.21	−0.94	73.39±4.93	+2.21	79.26	−0.26
MERGETUNE	✓	✓	✓		82.86±0.17	+4.67	96.03±0.21	+1.60	73.60±0.02	+2.56	**73.45**±3.48	+0.06	81.48	+2.22
DMC w/o $\mathcal{R}$	✓		✓		78.66±1.43	−4.20	94.60±0.41	−1.43	70.23±1.89	−3.37	68.79±3.72	−4.66	78.07	−3.42
DMC w/o VA	✓	✓	✓		83.91±0.46	+5.25	**96.25**±0.04	+1.65	73.97±0.44	+3.74	73.24±0.96	+4.45	81.84	+3.77
DMC (Ours)	✓	✓	✓	✓	**84.33**±0.84	+0.42	96.15±0.38	−0.10	**74.28**±0.52	+0.31	73.05±4.02	−0.19	**81.95**	+0.11

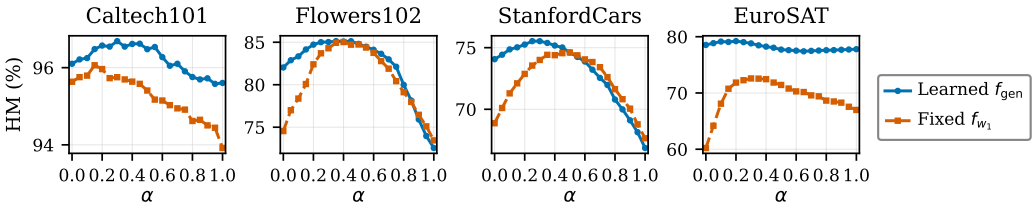

Figure 3: **Ablation: learned f_{gen} vs. fixed f_{w_1}.** Both variants retrain $\mathbf{c}_{\text{spec}}$ with the same loss (CE + VA + LMC), differing only in the generalization endpoint. At small α, including $\alpha = 0.20$, the **learned** f_{gen} (solid) gives a higher HM than the **fixed** f_{w_1} (dashed).

For VA, neither the second augmented view nor the augmentation policy matters much: all variants lie within about 1 HM of each other, mostly within one seed standard deviation (Tables 8 and 9).

Is the learnable generalization prompt necessary? As discussed in Section 3.2, f_{gen} retains task-discriminative structure from the CoOp initialization while recovering zero-shot geometry, making it a stronger interpolation endpoint than the fixed zero-shot features f_{w_1}. We verify this empirically by removing $\mathbf{c}_{\text{gen}}$ and replacing the generalization endpoint with the fixed f_{w_1}, while retraining $\mathbf{c}_{\text{spec}}$ with the same loss (CE + VA + LMC). The corridor is now anchored to f_{w_1}: $\widetilde{f}(\alpha) = \ell_2\big(f_{w_1} + \alpha(f_{\text{spec}} - f_{w_1})\big)$. Note that this configuration makes the LMC corridor structurally identical to MERGETUNE's: a path from a fixed reference to a single learned prompt. Figure 3 compares the HM curves. At $\alpha = 0.20$, the learned f_{gen} gives a higher HM on all four datasets. The gap is $+2.5$ HM on Flowers102, $+1.8$ on StanfordCars, $+7.4$ on EuroSAT, and smaller on Caltech101. It is largest on EuroSAT, where zero-shot CLIP is weakest among the four datasets (Table 1).

Gradient conflict at convergence. At the converged MERGETUNE checkpoint $\mathbf{c}^*$, we accumulate $\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}$ and $\nabla_{\mathbf{c}}\mathcal{R}$ over the full training set and report $\rho = \cos(\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}, \nabla_{\mathbf{c}}\mathcal{R})$, following Yu et al. [2020]; $\rho < 0$ means that the two steepest-descent directions are locally opposed. All eight values in Figure 4 are negative (per-dataset values in Appendix D), with the strongest conflict on Flowers102 ($\rho \approx -0.95$). KgCoOp's conflict is weaker than CoOp's on three of four datasets, consistent with its built-in anchor, and the magnitude of ρ does not predict DMC's gain. DMC removes the direct conflict: $\mathcal{R}$ acts only on $\mathbf{c}_{\text{gen}}$, the endpoint cross-entropy only on $\mathbf{c}_{\text{spec}}$, and the prompts interact only through the corridor.

Decoupling dynamics. Figure 5 shows the training dynamics of DMC on four datasets (one run, seed 1). Both prompts start at $\hat{w}_2$. f_{gen} returns quickly to the zero-shot geometry ($\cos(f_{\text{gen}}, f_{w_1})$ reaches 0.95–0.98 within 10 epochs), so $\mathcal{R}$ dominates early training, while f_{spec} moves away from f_{w_1}, with oscillations from the LMC coupling. The final separation is largest on Flowers102 and smallest on the near-ceiling OxfordPets.

Single-prompt loss modifications. Seven single-prompt loss modifications, evaluated on Caltech101 and Flowers102, produce no meaningful improvement over the cosine baseline: gains are at most $+0.40$ HM, similar to retuning λ, and never clear on both datasets. For the same Flowers102

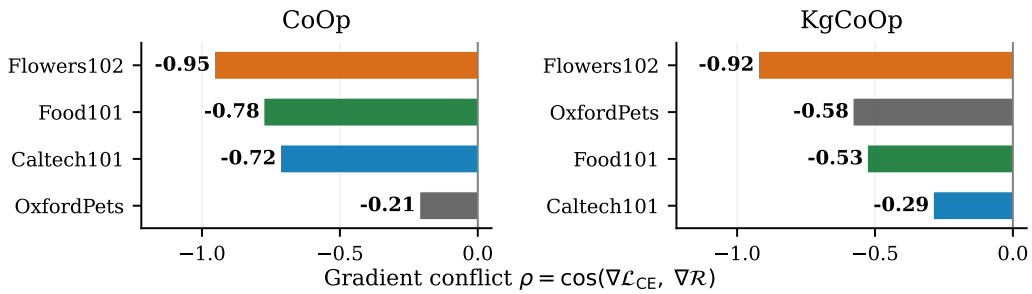

Figure 4: **Gradient conflict at convergence.** $\rho = \cos(\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}},\ \nabla_{\mathbf{c}}\mathcal{R})$ at the converged single-prompt checkpoint, averaged over the full training set. $\rho < 0$ on every dataset-baseline pair; Flowers102 shows the strongest conflict ($\rho \approx -0.95$ for CoOp).

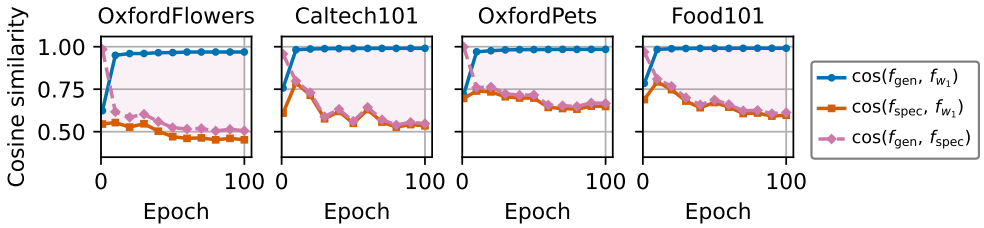

Figure 5: **DMC decoupling dynamics during training**. Both prompts are initialized from $\hat{w}_2$. **Blue**: $\cos(f_{\text{gen}}, f_{w_1})$ rises sharply within 10 epochs and saturates near 0.97–0.99, confirming rapid convergence toward zero-shot geometry. **Orange**: $\cos(f_{\text{spec}}, f_{w_1})$ decreases overall under CE supervision, with oscillations from the LMC coupling. Final values range from $\sim$0.45 (Flowers102, strongest separation) to $\sim$0.65 (OxfordPets, weakest). **Purple dashed**: $\cos(f_{\text{gen}},\ f_{\text{spec}})$ reflects the feature separation, decreasing from $\sim$1.0 to 0.50–0.67. The shaded region visualizes the decoupling gap. Single seed, with VA; Table 10 reports the 3-seed endpoint cosines (without VA; with-VA averages in its caption).

setup (seed 1, $\beta = 1$, best α, matching the best-over-sweep protocol), DMC reaches $+0.86$ HM over the cosine baseline (Table 22), more than twice the best single-prompt gain on this dataset ($+0.36$). Full results are in Appendix E.

5 Conclusion

Single-prompt methods for base-to-novel generalization settle at a Pareto-stationary compromise between the two objectives: the two objectives send opposing gradients to one shared prompt, and none of the single-prompt loss modifications we test escapes this constraint. DMC resolves this by assigning each objective to a dedicated prompt connected via an LMC corridor in text-feature space. A Visual Anchor KL term keeps CLIP's zero-shot class-similarity distribution of each image, and in an idealized information-theoretic model, violating class-permutation invariance lower-bounds the drop in novel accuracy. On CoOp and KgCoOp, DMC improves HM over MERGETUNE on most datasets. On MMA, the two endpoints converge to nearly the same features and DMC's regularizers no longer help.

Limitations. DMC adds one prompt ($\sim$2K parameters) over the single-prompt baseline. The coefficient α is chosen offline and fixed at 0.20 across datasets, since the HM-optimal α cannot be identified from base classes alone. On the eight datasets not used to select it (ImageNet excluded), the fixed value is on average 0.18 HM below the per-dataset oracle (Appendix K). DMC also needs the two prompts to stay distinct. On near-ceiling datasets they converge to similar features and the HM gain is small, and on MMA they collapse. Proposition 1 holds only under idealized assumptions (Appendix I.5).

Acknowledgments and Disclosure of Funding

This work benefited from Hi! PARIS and State funding managed by the French National Research Agency (ANR) under the France 2030 program, reference ANR-23-IACL-0005.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1309–1318, 2018.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kaiming He, Haoqing Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, Nov 2020. doi: 10.1109/JSAIT.2020.3040992.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023a.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective / multi-task learning framework induced by pareto stationarity. In *International Conference on Machine Learning*, pages 15895–15907, 2022.

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

Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *International Conference on Learning Representations*, 2024.

Wenqing Wang, Da Li, Xiatian Zhu, and Josef Kittler. MergeTune: Continued fine-tuning of vision-language models. In *International Conference on Learning Representations*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998, 2022a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022b.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. MMA: Multi-modal adapter for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024.

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

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 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 Overview

The appendix follows the order of the main argument: extended results first, then the evidence for the single-prompt ceiling, the ablations of DMC, its scope, the regularizer theory, the design choices, and finally the proofs. Unless stated otherwise, numbers for the rows we trained are means over seeds 1–3 (DMC at the fixed $\alpha = 0.20$, MERGETUNE at its final prompt), and prior-work numbers are as published.

Extended results

A. **Full Base-to-Novel Results** (Appendix A). Base, Novel and HM for every method on all 11 datasets, per-cell seed variability, and an analysis of the weak cases.

B. **Extended Interpolation Curves** (Appendix B). The α-sweep curves on the datasets not shown in Figure 2, including ImageNet.

Why a single prompt hits a ceiling

C. **Quadratic Model of the Single-Prompt Ceiling** (Appendix C). A scalar model of the ceiling and Lemma 1.

D. **Gradient Conflict per Dataset** (Appendix D). Measured gradient cosines at convergence for CoOp and KgCoOp.

E. **Single-Prompt Loss Modifications** (Appendix E). Seven alternative losses on Caltech101 and Flowers102.

Ablations of DMC

F. **Supporting Tables for the Ablations** (Appendix F). The larger single-prompt control and the two Visual Anchor studies.

G. **HM Landscape** (Appendix G). Harmonic-mean landscapes on Flowers102 and Caltech101, complementing Figure 1.

Scope

H. **When Does DMC Compose?** (Appendix H). The endpoint-separation condition, the MMA analysis, and a falsification test.

Regularizer theory

I. **Class-Permutation Invariance** (Appendix I). Numerical CPI test, informal justification, the ZSDD case, and the applicability of the bound.

Design choices and reproducibility

J. **Training and Inference Protocol** (Appendix J).

K. **Interpolation Coefficient Selection** (Appendix K). Eight base-only selection criteria, the oracle α per dataset, and robustness of the fixed $\alpha = 0.20$.

L. **Hyperparameter Sensitivity** (Appendix L). Sweeps over β and λ_{gen}.

Proofs

M. **Proofs** (Appendix M). Lemma 1, Proposition 1, Corollary 1, Proposition 2, and a summary of all assumptions.

A Full Base-to-Novel Results

This section asks whether DMC's HM gains come from Base, Novel or both, how stable they are across seeds, and where they fail. Table 3 reports Base, Novel, and HM for every method across all 11 datasets. The main paper (Table 1) shows HM only; this table provides the full breakdown for readers who wish to verify the Base and Novel trends independently. For the MERGETUNE and DMC rows it reports the *best of three seeds*, so that each Base/Novel pair comes from a single run, as

in the prior-work rows. These numbers are therefore higher than the 3-seed means of Table 1 (e.g., CoOp+DMC average HM 79.04 here versus 77.23 there) and should not be compared with them; all headline numbers in the paper are the seed means.

Table 4 gives the seed-resolved view: HM mean $\pm$ sample standard deviation for every row we trained. Of the 33 DMC cells, 23 have $\sigma_{\mathrm{HM}} \leq 0.5$ and 27 have $\sigma_{\mathrm{HM}} \leq 1.0$. The seed-fragile cells are concentrated on EuroSAT, DTD, FGVCAircraft and UCF101, where the corresponding MERGETUNE runs are equally noisy (e.g., $\sigma_{\mathrm{HM}} = 3.81$ and 4.03 on EuroSAT), so DMC inherits this variance rather than introducing it. MMA+DMC is markedly seed-stable (maximum $\sigma_{\mathrm{HM}} = 0.45$), consistent with the endpoint collapse of Appendix H: a degenerate corridor leaves little to vary.

A.1 Failure and weak-case analysis

DMC does not gain uniformly across datasets or baselines. We list the cases where the gain over MERGETUNE shrinks, disappears, or reverses (Table 4), and explain each.

Small regressions. Seven of the 33 cells are below MERGETUNE: CoOp on OxfordPets (-0.09) and DTD (-0.73); KgCoOp on OxfordPets (-0.05), FGVCAircraft (-0.93) and EuroSAT (-0.06); MMA on ImageNet (-0.25) and Caltech101 (-0.01). All but the MMA ImageNet cell lie within one combined seed standard deviation of the two cells. The two largest prompt-tuning regressions (CoOp DTD, KgCoOp FGVCAircraft) occur on datasets whose MERGETUNE runs are themselves seed-fragile ($\sigma_{\mathrm{HM}} = 1.93$ and 1.35).

Near-ceiling regime: Caltech101 and OxfordPets. Both datasets have zero-shot HM already close to 95. The interpolation frontier is flat there: base and novel accuracy move together across α, so the gap between the single-prompt operating point and the DMC corridor is small and DMC's margin shrinks to at most 0.17 HM in either direction. We attribute this to the ceiling rather than to a failure of the method.

MMA endpoint collapse. On MMA the two learned endpoints converge to nearly identical text features ($\cos(f_{\mathrm{gen}}, f_{\mathrm{spec}}) > 0.99$ on 9 of 10 datasets), and the interpolation curve carries little of DMC's effect. It is the clearest failure case in our study and is analyzed in full in Appendix H.

Seed-fragile cells. Six DMC cells have $\sigma_{\mathrm{HM}} > 1$: EuroSAT (CoOp 3.65, KgCoOp 3.16), DTD (CoOp 1.37, KgCoOp 1.13), FGVCAircraft (CoOp 1.37) and UCF101 (CoOp 1.47). Four have positive mean gains over MERGETUNE (CoOp on EuroSAT, FGVCAircraft and UCF101; KgCoOp on DTD) and two are slightly negative (CoOp on DTD, KgCoOp on EuroSAT), but in all but one case the seed spread overlaps zero (CoOp on FGVCAircraft, $+1.90$, is just outside one combined standard deviation), so we do not claim a reliable effect on these cells individually. EuroSAT is the extreme case: with five classes per split, a single class changing its predictions moves HM by several points.

Table 3: Base, Novel and HM on 11 datasets. For the MERGETUNE and DMC rows we report the *best of 3 seeds*, so these numbers are higher than, and not directly comparable with, the 3-seed means of Table 1, which are the headline results; prior-work rows are published numbers. Parentheses: average-HM change relative to the base method. †: Using large language model or teacher model's knowledge.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML 21)	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp (IJCV 22)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
KgCoOp (CVPR 23)	80.73	73.61	77.01	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe (CVPR 23)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC (ICCV 23)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
MMA (CVPR 24)	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
CoPrompt† (ICLR 24)	84.00	77.23	80.48	77.67	71.27	74.33	98.27	94.90	96.55	95.67	98.10	96.87
CoOp + ATPrompt† (ICCV 25)	82.68	68.04	74.65 (+2.99)	76.27	70.60	73.33	97.95	93.63	95.74	94.77	96.59	95.67
CoOp + TIES (NeurIPS 23)	66.95	65.71	66.32 (-5.34)	70.09	61.65	65.60	93.63	91.34	92.50	90.12	95.13	92.56
CoOp + DARE (ICML 24)	75.91	65.97	70.59 (-1.07)	74.25	64.32	68.93	96.36	91.12	93.67	93.41	96.33	94.84
CoOp + MERGETUNE (ICLR 26)	80.82	73.97	77.24 (+5.58)	75.96	69.91	72.81	97.91	94.65	96.25	95.09	97.75	96.40
CoOp + DMC (Ours)	83.67	74.89	79.04 (+7.38)	76.13	69.93	72.90	98.32	95.41	96.84	95.37	98.32	96.82
KgCoOp + TIES (NeurIPS 23)	73.03	72.09	72.56 (-4.45)	74.06	68.06	70.93	97.27	94.43	95.83	93.99	96.51	95.23
KgCoOp + DARE (ICML 24)	78.15	72.41	75.17 (-1.84)	75.10	68.73	71.78	97.53	94.47	95.97	92.22	97.37	94.73
KgCoOp + MERGETUNE (ICLR 26)	81.85	74.46	77.98 (+0.97)	76.49	69.35	72.75	97.87	94.91	96.37	95.32	97.76	96.53
KgCoOp + DMC (Ours)	83.58	74.19	78.61 (+1.60)	76.32	70.12	73.09	98.19	94.98	96.56	95.43	97.82	96.61
MMA + TIES (NeurIPS 23)	70.39	68.46	69.41 (-10.46)	73.16	66.48	69.66	95.36	92.23	93.77	91.24	93.11	92.17
MMA + DARE (ICML 24)	74.25	69.52	71.81 (-8.06)	74.17	66.76	70.27	94.95	92.49	93.70	90.52	95.53	92.96
MMA + MERGETUNE (ICLR 26)	70.59	75.07	72.76 (-7.11)	73.68	69.02	71.27	97.48	94.43	95.93	90.01	96.92	93.34
MMA + DMC (Ours)	72.14	75.52	73.79 (-6.08)	76.41	71.06	73.64	97.48	94.43	95.93	90.06	97.20	93.49

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML 21)	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp (IJCV 22)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
KgCoOp (CVPR 23)	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.10	36.21	33.55	34.83
MaPLe (CVPR 23)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
PromptSRC (ICCV 23)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
MMA (CVPR 24)	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
CoPrompt† (ICLR 24)	76.97	74.40	75.66	97.27	76.60	85.71	90.73	92.07	91.40	40.20	39.33	39.76
CoOp + ATPrompt† (ICCV 25)	77.43	66.55	71.58	97.44	67.52	79.77	88.74	87.44	88.09	40.38	27.22	32.52
CoOp + TIES (NeurIPS 23)	57.94	63.42	60.56	73.60	65.41	69.27	88.42	87.64	88.02	30.29	28.83	29.54
CoOp + DARE (ICML 24)	67.82	61.62	64.57	90.88	63.73	74.92	87.78	87.22	87.50	34.41	28.25	31.03
CoOp + MERGETUNE (ICLR 26)	71.94	75.12	73.49	95.32	74.30	83.51	90.56	91.75	91.15	36.17	34.35	35.24
CoOp + DMC (Ours)	79.19	74.70	76.88	97.91	76.38	85.82	90.84	92.01	91.42	40.94	35.03	37.76
KgCoOp + TIES (NeurIPS 23)	65.79	74.17	69.73	74.26	72.53	73.39	90.25	91.58	90.91	30.81	28.97	29.86
KgCoOp + DARE (ICML 24)	67.84	73.02	70.33	87.02	72.88	79.33	90.44	91.60	91.02	33.49	30.37	31.85
KgCoOp + MERGETUNE (ICLR 26)	73.21	75.14	74.17	96.49	74.68	84.20	90.51	91.83	91.17	37.09	35.11	36.07
KgCoOp + DMC (Ours)	78.66	74.75	76.66	98.01	74.61	84.72	90.86	91.89	91.37	39.44	32.21	35.46
MMA + TIES (NeurIPS 23)	60.72	70.12	65.08	73.46	70.74	72.07	88.79	89.83	89.31	29.27	28.06	28.65
MMA + DARE (ICML 24)	64.82	71.52	68.01	77.15	70.95	73.92	89.85	87.61	88.72	30.35	28.23	29.25
MMA + MERGETUNE (ICLR 26)	65.74	75.61	70.33	73.03	78.23	75.55	89.67	90.77	90.22	31.57	34.13	32.80
MMA + DMC (Ours)	66.62	75.46	70.76	76.54	77.94	77.23	89.91	90.95	90.43	32.11	35.03	33.51

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML 21)	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp (IJCV 22)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
KgCoOp (CVPR 23)	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.66
MaPLe (CVPR 23)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC (ICCV 23)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
MMA (CVPR 24)	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
CoPrompt† (ICLR 24)	82.63	80.03	81.30	83.13	64.73	72.79	94.60	78.57	85.84	86.90	79.57	83.07
CoOp + ATPrompt† (ICCV 25)	80.84	68.64	74.24	80.83	45.49	58.22	90.34	59.79	71.96	84.49	64.96	73.45
CoOp + TIES (NeurIPS 23)	62.41	62.99	62.70	57.33	41.87	48.40	50.17	69.85	58.40	62.45	54.68	58.30
CoOp + DARE (ICML 24)	70.33	63.01	66.47	72.49	47.02	57.04	69.71	65.22	67.39	77.61	57.84	66.28
CoOp + MERGETUNE (ICLR 26)	80.37	76.19	78.22	78.51	56.84	65.93	85.15	64.51	73.40	82.04	78.33	80.14
CoOp + DMC (Ours)	82.14	77.25	79.62	82.29	55.07	65.98	90.81	70.31	79.26	86.45	79.34	82.74
KgCoOp + TIES (NeurIPS 23)	73.05	73.88	73.46	60.77	53.54	56.93	68.66	65.45	67.02	74.46	73.91	74.18
KgCoOp + DARE (ICML 24)	77.47	74.81	76.12	72.46	54.19	62.01	84.48	63.03	72.19	81.63	75.99	78.71
KgCoOp + MERGETUNE (ICLR 26)	80.53	77.28	78.87	79.90	57.33	66.75	89.56	67.89	77.22	83.35	77.76	80.45
KgCoOp + DMC (Ours)	82.06	76.47	79.17	81.25	56.52	66.67	93.05	71.03	80.56	86.14	75.66	80.56
MMA + TIES (NeurIPS 23)	69.46	65.77	67.56	55.38	48.36	51.63	65.69	67.95	66.80	71.74	60.42	65.60
MMA + DARE (ICML 24)	71.43	72.61	72.02	67.65	50.37	57.74	77.42	64.16	70.17	78.46	64.47	70.78
MMA + MERGETUNE (ICLR 26)	75.20	78.64	76.88	57.06	60.75	58.85	52.10	70.59	59.95	70.99	76.69	73.73
MMA + DMC (Ours)	76.95	78.87	77.90	59.03	61.23	60.11	53.69	71.21	61.22	74.77	77.29	76.01

Table 4: Seed variability for every row we trained ourselves: HM mean ± sample standard deviation over 3 seeds (DMC at $\alpha = 0.20$; MERGETUNE at its final prompt). Per-cell companion to Table 1, which reports the same runs as means. Cells with $\sigma_{\mathrm{HM}} > 1$ are **bold** and flagged as seed-fragile.

	CoOp		KgCoOp		MMA	
Dataset	+MERGETUNE	+DMC	+MERGETUNE	+DMC	+MERGETUNE	+DMC
ImageNet	72.89 ± 0.18	72.90 ± 0.18	72.82 ± 0.10	73.09 ± 0.26	73.15 ± 0.10	72.90 ± 0.01
Caltech101	96.27 ± 0.26	96.41 ± 0.15	95.94 ± 0.03	96.11 ± 0.08	95.75 ± 0.16	95.74 ± 0.16
OxfordPets	96.13 ± 0.66	96.04 ± 0.29	96.28 ± 0.10	96.23 ± 0.35	93.22 ± 0.11	93.38 ± 0.01
StanfordCars	73.38 ± 0.37	75.21 ± 0.10	74.90 ± 0.29	75.38 ± 0.24	70.16 ± 0.24	70.51 ± 0.04
Flowers102	83.92 ± 0.57	84.47 ± 0.58	83.39 ± 0.27	83.67 ± 0.33	75.22 ± 0.40	76.37 ± 0.23
Food101	91.05 ± 0.11	91.13 ± 0.15	91.08 ± 0.11	91.19 ± 0.10	90.18 ± 0.06	90.33 ± 0.09
FGVCAircraft	33.33 ± 0.42	**35.23 ± 1.37**	**34.10 ± 1.35**	33.17 ± 0.90	32.72 ± 0.10	33.10 ± 0.04
SUN397	77.85 ± 0.35	78.75 ± 0.54	78.03 ± 0.53	78.85 ± 0.03	76.70 ± 0.17	77.12 ± 0.12
DTD	**64.26 ± 1.93**	**63.53 ± 1.37**	**62.66 ± 2.83**	**64.34 ± 1.13**	58.07 ± 0.68	59.62 ± 0.13
EuroSAT	**73.45 ± 3.81**	**75.75 ± 3.65**	**73.44 ± 4.03**	**73.38 ± 3.16**	59.57 ± 0.48	60.57 ± 0.45
UCF101	**78.42 ± 1.13**	**80.07 ± 1.47**	78.08 ± 0.90	78.90 ± 0.94	73.59 ± 0.21	75.11 ± 0.45
Average	76.45	77.23	76.43	76.76	72.58	73.16

B Extended Interpolation Curves

This section shows whether the operating-point shift of Figure 2 holds on the remaining datasets. Each figure shows the interpolation curves traced by **DMC** (solid blue) and **DMC W/O VA** (dashed vermillion) across $\alpha \in [0, 1]$, with the MERGETUNE baseline (gray dot), the shift to the operating point (green arrow) and the operating point (blue star), on the CoOp base. The boxed ΔHM value reports the gap between the operating point and the MERGETUNE baseline, read off the plotted seed-averaged curves; it differs from the corresponding entry of Table 4 by at most 0.18 HM (OxfordPets: $+0.09$ here, -0.09 in the table).

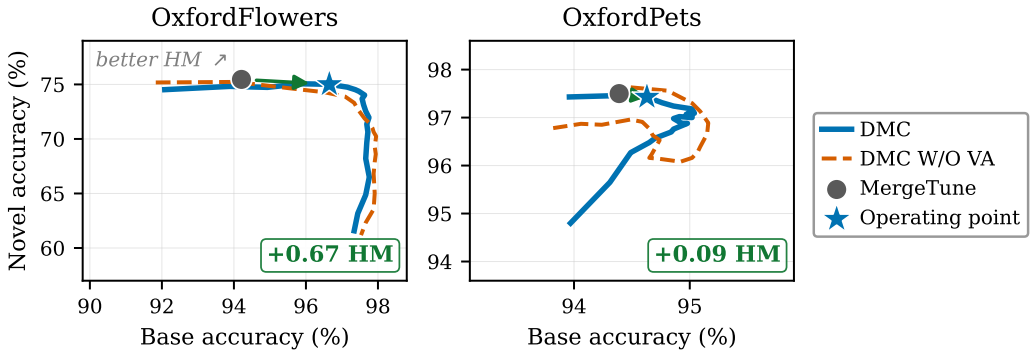

Figure 6: **Flowers102 and OxfordPets.** Flowers102 shows a shift toward higher base accuracy at nearly unchanged novel accuracy ($+0.67$ HM). OxfordPets has a near-ceiling baseline; DMC is within 0.1 HM of it ($+0.09$ on the plotted curves, -0.09 in Table 4).

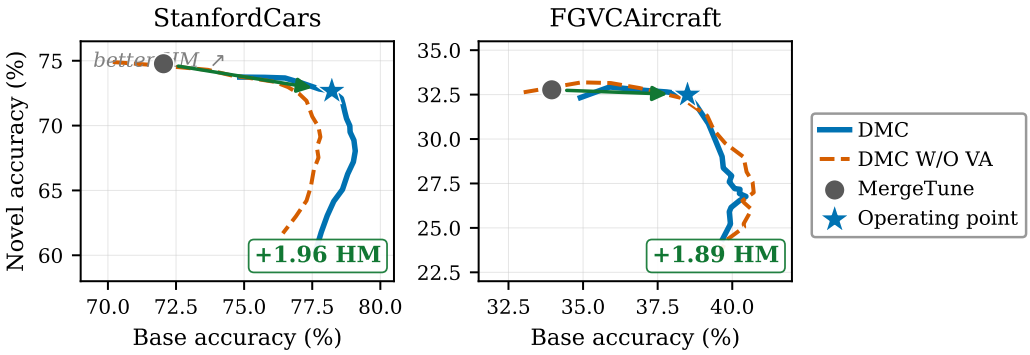

Figure 7: **StanfordCars and FGVCAircraft.** Both fine-grained datasets show clear operating-point shifts: +1.96 HM on StanfordCars and +1.89 on FGVCAircraft.

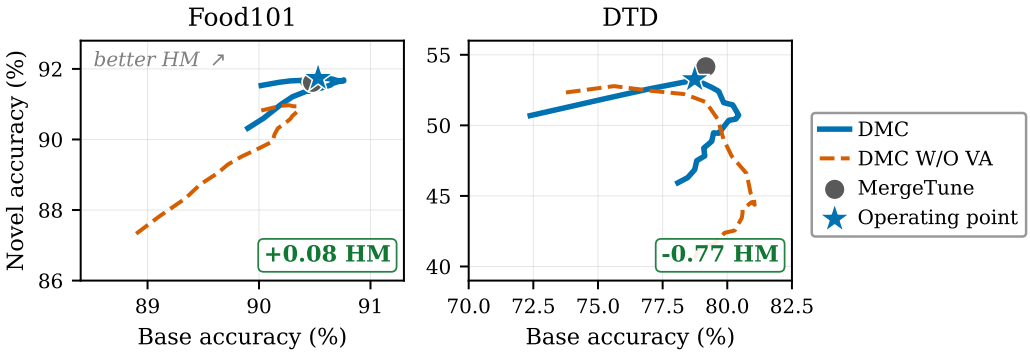

Figure 8: **Food101 and DTD.** Food101 has a flat frontier; DMC shifts the operating point slightly (+0.08 HM). DTD is the only dataset in these figures where DMC is slightly below the baseline (−0.77 HM).

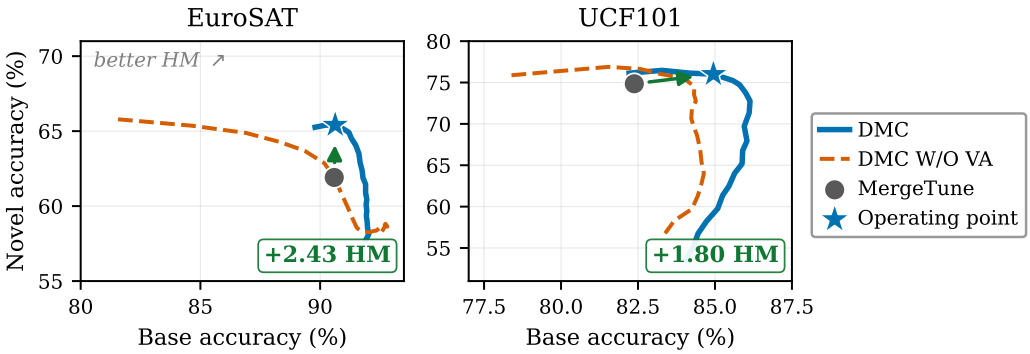

Figure 9: **EuroSAT and UCF101.** EuroSAT shows the largest improvement among these datasets (+2.43 HM). UCF101 achieves +1.80 HM. (Both datasets are also shown in the main paper, Figure 2.)

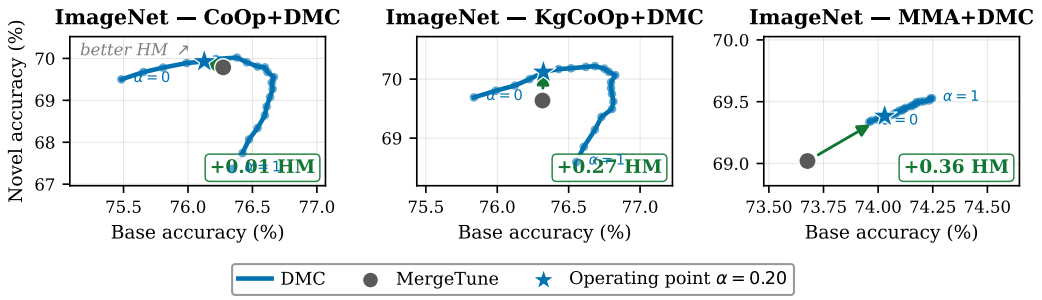

Figure 10: **ImageNet — CoOp+DMC, KgCoOp+DMC, MMA+DMC.** Interpolation curves on ImageNet (seed mean of the 21-point α sweep). The blue star marks the operating point at the fixed $\alpha = 0.20$ (the same value used for all held-out datasets, selected on Flowers102+Caltech101). At the operating point, DMC yields +0.01 HM on CoOp+DMC (72.89 $\rightarrow$ 72.90), +0.27 HM on KgCoOp+DMC (72.82 $\rightarrow$ 73.09), and -0.25 HM on MMA+DMC (73.15 $\rightarrow$ 72.90; the MMA panel shows an earlier single run, whose boxed value differs) over MergeTune. The operating-point shift on ImageNet is small relative to the specialized datasets (Figure 2); this is consistent with ImageNet's high-diversity 1000-class set, where the single-prompt envelope already covers most of the achievable operating points and decoupling has less headroom than on specialized tasks like EuroSAT or StanfordCars. Unlike the other interpolation-curve figures in this appendix, only the DMC (with VA) curve is shown: a DMC-without-VA ablation was not run on ImageNet.

C Quadratic Model of the Single-Prompt Ceiling

This section asks why a single prompt trained on a weighted sum of two objectives must settle for a compromise. To build intuition for the structural bottleneck in Section 3, consider a minimal scalar model with two competing objectives:

$$L_1(\theta) = (\theta - a)^2 \quad \text{(specialization, minimized at } a\text{)},$$
$$L_2(\theta) = (\theta - b)^2 \quad \text{(generalization, minimized at } b \neq a\text{)}.$$

Single-prompt scalarization. The weighted objective $(1-\lambda)L_1(\theta) + \lambda L_2(\theta)$ has the unique stationary point $\theta^*(\lambda) = (1-\lambda)a + \lambda b$ for $\lambda \in (0,1)$. By Lemma 1 below, $\theta^*(\lambda)$ is Pareto-stationary, and neither objective is individually minimized:

$$L_1\big(\theta^*(\lambda)\big) = \lambda^2(b-a)^2 > 0, \qquad L_2\big(\theta^*(\lambda)\big) = (1-\lambda)^2(b-a)^2 > 0.$$

As λ varies over $(0,1)$, the pair $(L_1(\theta^*(\lambda)),\ L_2(\theta^*(\lambda)))$ traces a quadratic arc that never reaches either axis of the (L_1, L_2) plane. The individual optima $(0,\ (b-a)^2)$ and $((b-a)^2,\ 0)$ are unreachable by any single-prompt scalarization.

Decoupled two-parameter model. With two independent parameters (θ_1, θ_2), decoupled optimization of $L_1(\theta_1) + L_2(\theta_2)$ gives $\theta_1^* = a$ and $\theta_2^* = b$. Each objective is then exactly minimized by its own parameter, so the joint objective reaches 0, whereas any single θ incurs $L_1(\theta) + L_2(\theta) \geq (b-a)^2/2$. The interpolated predictor $\tilde{\theta}(\alpha) = (1-\alpha)\theta_2^* + \alpha\theta_1^*$ (as in Eq. (3), $\alpha = 0$ is the generalization end) recovers every single-parameter compromise ($\tilde{\theta}(1-\lambda) = \theta^*(\lambda)$), so decoupling loses nothing and turns the trade-off into a post-hoc choice of α instead of a training-time choice of λ. In one dimension the interpolation path and the single-parameter path coincide, so this model explains why a single prompt must compromise, not how far a decoupled corridor can move beyond the single-prompt operating point; the latter is an empirical property of the high-dimensional prompt space, which Figure 2 and Appendix B measure.

The same picture holds for the prompt (Lemma 1, Section 3.1).

The proof is in Appendix M. The lemma is the high-dimensional analogue of the scalar picture: whenever both gradients are nonzero at convergence, the single prompt is a compromise that minimizes neither objective. Appendix D checks the nonzero-and-opposed gradient condition empirically.

D Gradient Conflict: Per-dataset Measurements

This section asks whether the two single-prompt gradients actually conflict at convergence, and whether the size of the conflict predicts DMC's gain. Figure 4 in the main paper reports the gradient conflict statistic $\rho = \cos(\nabla_{\mathbf{c}}\mathcal{L}_{\mathrm{CE}}, \nabla_{\mathbf{c}}\mathcal{R})$ at the converged MERGETUNE checkpoint. Table 5 provides the exact per-dataset values alongside the corresponding DMC HM gain.

Table 5: Gradient conflict $\rho = \cos(\nabla_{\mathbf{c}}\mathcal{L}_{\mathrm{CE}}, \nabla_{\mathbf{c}}\mathcal{R})$ at convergence, grouped by baseline. $\rho < 0$ on all eight dataset–baseline pairs. The DMC gain is the HM gain over MERGETUNE in the runs where ρ was measured; it can differ from the seed means of Table 1.

Dataset	ρ	DMC gain
CoOp baseline		
Flowers102	-0.955	$+0.23$
Caltech101	-0.716	$+0.36$
Food101	-0.776	$+0.08$
OxfordPets	-0.211	$+0.22$
KgCoOp baseline		
Flowers102	-0.923	$+0.33$
Caltech101	-0.289	$+0.32$
Food101	-0.528	$+0.11$
OxfordPets	-0.580	$+0.04$
Pearson r (ρ vs gain)		-0.05

The conflict is present everywhere, but its magnitude does not predict DMC's gain: Caltech101 gains $+0.36$ HM at $\rho = -0.72$, while Food101 gains only $+0.08$ at $\rho = -0.78$, and across the eight pairs the correlation is negligible (Pearson $r = -0.05$). A negative ρ indicates that the single prompt sits at a first-order trade-off (Lemma 1; at an exact stationary point of $\mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{R}$ alone, $\rho = -1$, and values above -1 reflect approximate convergence and the additional terms of the MERGETUNE objective); how much DMC gains depends on the shape of the interpolation curve beyond that operating point, which ρ does not measure. We therefore use ρ as evidence that the ceiling exists, not as a predictor of where decoupling helps most.

E Single-Prompt Loss Modifications

This section asks whether a better single-prompt loss, rather than a second prompt, could deliver DMC's gain. We test whether targeted single-prompt loss modifications can match the DMC frontier. Seven variants are evaluated on Caltech101 and Flowers102 as drop-in replacements or additions to the cosine score regularizer in MERGETUNE's objective:

1. Fisher-weighted cosine. Replace isotropic cosine distance with Fisher-diagonal-weighted cosine distance, with two normalizations (*max*, *per-class*) and $\lambda \in \{4, 8, 10, 16\}$. *2. Prompt-space EWC.* Add a Fisher-weighted ℓ_2 penalty on prompt parameters toward $\hat{w}_2$, swept over $\lambda_{\mathrm{pf}} \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$. *3. KL path.* Enforce KL-divergence preservation along the $f_{w_1} \to f(\mathbf{c})$ path, swept over $\lambda_{\mathrm{kl}} \in \{0.1, 0.3, 0.5, 1.0\}$. *4. ZSDD.* Distill the zero-shot prediction distribution at temperature τ, swept over $\lambda_{\mathrm{zsdd}} \in \{0.5, 1, 2, 4\}$ and $\tau \in \{2, 4\}$. *5. Weight scheduling.* Anneal the cosine weight λ over training, with linear or cosine schedules and final $\lambda_{\min} \in \{0, 2, 5\}$. *6. Visual Anchor (single-prompt).* Apply VA on the same prompt that receives CE, instead of the decoupled placement used in DMC. *7. TTAI.* Test-time adaptive interpolation with per-image α predicted from visual features.

Table 6 reports the best HM achieved by each modification on each dataset (best across the swept hyperparameters, seed 1, $\beta = 1.0$, the MERGETUNE trainer with a CoOp initialization, ViT-B/16). The cosine-score baseline at $\lambda = 8$ achieves 96.03 HM on Caltech101 and 83.80 HM on Flowers102; as single-seed numbers, they differ from the seed means of Table 1.

The largest single-dataset gain is $+0.40$ HM (ZSDD on Caltech101), and the largest gain on Flowers102 is $+0.36$ HM (single-prompt Visual Anchor). Both are within seed-variance scale and of the same order as what cosine-weight tuning alone achieves: sweeping $\lambda \in \{4, 8, 10, 16\}$ moves

Table 6: Single-prompt loss modifications on Caltech101 and Flowers102. For each modification we report the best HM across the hyperparameter sweep; Δ HM is the gap to the cosine-score baseline at $\lambda = 8$. No modification improves on the baseline by clearly more than retuning λ does, and modifications that gain on one dataset typically regress on the other.

	Caltech101				Flowers102			
Modification	Base	Novel	HM	Δ	Base	Novel	HM	Δ
Cosine score, $\lambda = 8$ (MERGETUNE)	97.93	94.21	96.03	—	94.21	75.46	83.80	—
Fisher-weighted cosine	97.42	95.09	96.24	$+0.21$	94.11	75.89	84.02	$+0.22$
Prompt-space EWC	97.87	94.32	96.06	$+0.03$	94.59	75.25	83.82	$+0.02$
KL path	97.87	94.10	95.95	-0.08	94.02	75.60	83.81	$+0.01$
ZSDD	97.48	95.41	96.43	$+0.40$	87.46	76.17	81.43	-2.37
Weight scheduling	98.13	92.79	95.39	-0.64	94.68	72.84	82.34	-1.46
Visual Anchor (single-prompt)	97.61	94.43	95.99	-0.04	95.25	75.39	84.16	$+0.36$
TTAI	97.93	94.43	96.15	$+0.12$	90.98	76.45	83.08	-0.72

the cosine baseline by $+0.22$ HM on Caltech101 and $+0.26$ HM on Flowers102; they exceed it by only 0.18 and 0.10 HM, respectively. Modifications that gain on one dataset typically regress on the other: ZSDD ($+0.40$ Caltech101, -2.37 Flowers102), single-prompt VA (-0.04 Caltech101, $+0.36$ Flowers102), TTAI ($+0.12$ Caltech101, -0.72 Flowers102). Fisher-weighted cosine is the only variant that gains noticeably on both datasets, by $+0.21$ and $+0.22$ HM, i.e., by about as much as retuning λ. Fisher-weighted cosine, which violates CPI, also fails to break the ceiling, consistent with Appendix I. The CPI test there measures a feature-space EWC variant, which is distinct from the prompt-space EWC evaluated here.

For comparison, on the same Flowers102 setup (seed 1, $\beta = 1.0$, best α, matching the best-over-sweep protocol above), DMC reaches HM 84.66 (Table 22), a $+0.86$ HM gain over the cosine baseline that is two to three times larger than the best single-prompt modification on this dataset. The pattern supports the structural-bottleneck claim: with the parameterization fixed to a single prompt, none of the seven regularizers we tested produces gains comparable to architectural decoupling.

F Supporting Tables for the Ablations

This section collects the tables behind two ablation questions of Section 4.3: does extra single-prompt capacity explain DMC's gain, and how much does the Visual Anchor contribute? (The four-dataset component ablation is Table 2 in the main text.)

Larger single prompt. Table 7 trains CoOp with $n_{\text{ctx}} \in \{8, 16\}$. At $n_{\text{ctx}} = 16$ the single prompt has twice DMC's parameter count, yet CoOp+DMC leads by 5.07, 1.07 and 5.00 HM on Flowers102, Caltech101 and StanfordCars; on EuroSAT the 3.46 HM lead is within the seed spread. Parameter count alone does not account for the gain.

Visual Anchor. Tables 8 and 9 vary the VA views and augmentation. The configurations span about 1.0 HM on Flowers102 and 0.32 HM on Caltech101, and most differences between them are within one seed standard deviation, consistent with the main-text finding that VA is a small refinement rather than the source of DMC's gain. The exact single-view case of Proposition 2 ($v_a = v_b$, preset *none*) is slightly below the two-view default on both datasets. These runs retrain the configuration of Table 1 (batch size 32): Caltech101 reproduces exactly (96.41), while Flowers102 gives 84.17 versus 84.47, within the seed spread. The *No VA* row comes from the component-ablation runs (batch size 128; Table 2).

Table 7: Extra single-prompt capacity does not match CoOp+DMC. Base-to-novel harmonic mean (HM) on four datasets. We train CoOp with enlarged context length $n_{\text{ctx}} \in \{8, 16\}$ using the same protocol as the standard $n_{\text{ctx}} = 4$ CoOp; $n_{\text{ctx}} = 16$ gives a single prompt with $16 \times 512 = 8192$ tunable parameters, strictly larger than the $2 \times 4 \times 512 = 4096$ parameters of CoOp+DMC. Values are mean $\pm$ sample standard deviation over 3 seeds; the $n_{\text{ctx}} = 4$ row is the single number reported in the main paper. CoOp+DMC still leads on every dataset; on EuroSAT the margin is within the seed spread.

Method	Flowers102	Caltech101	StanfordCars	EuroSAT
CoOp ($n_{\text{ctx}} = 4$, from Table 1)	74.06	93.73	68.13	68.69
CoOp ($n_{\text{ctx}} = 8$)	77.84 ± 0.36	92.19 ± 3.20	68.35 ± 1.23	69.95 ± 4.91
CoOp ($n_{\text{ctx}} = 16$)	79.40 ± 1.50	95.34 ± 0.64	70.21 ± 1.05	72.29 ± 9.52
CoOp + DMC (Ours)	$\mathbf{84.47 \pm 0.58}$	$\mathbf{96.41 \pm 0.15}$	$\mathbf{75.21 \pm 0.10}$	$\mathbf{75.75 \pm 3.65}$

Table 8: Decomposing the Visual Anchor (VA) benefit into KL anchoring vs. augmentation consistency. *No VA* disables the term entirely. *Single-view VA*: the student branch sees the raw image and the teacher branch one train-transform view; this is the closest variant to the single-view case of Proposition 2, which it matches exactly only without augmentation (preset *none* in Table 9). *Two-view VA* is the default used in all reported experiments: both branches see independent train-transform views. Δ is the HM change relative to the row above. All cells at $\alpha = 0.20$, $\beta = 4.0$, 3 seeds. Trainer: CoOp + DMC. *No VA* is from the component-ablation runs (batch size 128); the VA rows use batch size 32, so the first Δ also includes the batch-size change.

	Flowers102		Caltech101	
Configuration	HM	Δ	HM	Δ
No VA	83.91 ± 0.46	—	96.25 ± 0.04	—
Single-view VA	84.23 ± 0.16	$+0.32$	96.26 ± 0.25	$+0.01$
Two-view VA (default)	84.17 ± 0.84	-0.06	96.41 ± 0.15	$+0.15$

Table 9: Sensitivity of CoOp + DMC to the visual-anchor augmentation preset. Presets are: *none* (raw image, no augmentation), *weak* (mild crop and jitter, no grayscale), *default* (the paper setting: RandomResizedCrop 0.7–1.0, HorizontalFlip 0.5, ColorJitter $0.4/0.4/0.2/0.1$, RandomGrayscale 0.1), and *strong* (aggressive crop 0.2–1.0, strong jitter, RandomGrayscale 0.2). All cells are evaluated at $\alpha = 0.20$ and reported as mean $\pm$ sample standard deviation over 3 seeds. Trainer: CoOp + DMC.

	Flowers102			Caltech101		
Preset	Base	Novel	HM	Base	Novel	HM
none	96.65 ± 0.64	73.88 ± 0.74	83.74 ± 0.70	97.89 ± 0.32	94.36 ± 0.33	96.09 ± 0.30
weak	96.68 ± 0.17	75.44 ± 0.27	84.75 ± 0.14	97.96 ± 0.38	94.58 ± 0.71	96.24 ± 0.24
default	97.02 ± 0.79	74.32 ± 1.10	84.17 ± 0.84	98.13 ± 0.30	94.76 ± 0.55	96.41 ± 0.15
strong	97.12 ± 0.24	74.73 ± 0.84	84.46 ± 0.53	98.23 ± 0.20	94.58 ± 0.28	96.37 ± 0.13

G HM Landscape: Complementary View

This section asks whether the loss-landscape picture of Figure 1 carries over to the metric we report, the harmonic mean.

Plane construction. The 2D plane shared by Figure 1 (joint CE loss) and the HM landscapes below is anchored at the zero-shot CLIP features f_{w_1} in concatenated base∪novel text-feature space ($\mathbb{R}^{C\times 512}$, flattened). Two basis directions are derived from a trained DMC checkpoint: d_1 is the unit vector toward $f_{\hat{w}_2}$ (single-prompt CoOp baseline), and d_2 is the Gram–Schmidt orthogonalization of $f_{\text{gen}} - f_{w_1}$ against d_1. Each grid point (x, y) corresponds to a per-class feature matrix $f(x, y) = \ell_2\big(f_{w_1} + x\, d_1 + y\, d_2\big)$, ℓ_2-normalized per class, where x is in units of $\|f_{\hat{w}_2} - f_{w_1}\|$ and y is in units of $\|(f_{\text{gen}} - f_{w_1})_\perp\|$. At each grid point we recompute the classification logits with the frozen CLIP image encoder over the held-out base and novel splits and report (i) the joint base+novel cross-entropy (Figure 1) or (ii) the harmonic mean of base and novel accuracies (Figures 11, 12). By construction, f_{w_1} sits at the origin, $f_{\hat{w}_2}$ on the x-axis at $(1, 0)$, and f_{gen} in the upper half-plane; f_{spec}, the operating point at $\alpha = 0.20$, and the DMC corridor are projected onto the same plane and overlaid on top of the heatmap. The single-prompt linear path corresponds to varying $x \in [0, 1]$ at $y = 0$.

HM and CE views agree. The CE landscape (main-paper Figure 1) and the HM landscape share the same plane and overlays; only the heatmap value differs (CE loss vs. HM accuracy). Where the CE landscape shows a low-loss *valley*, the HM landscape shows a high-HM *peak*. Two structural observations transfer cleanly between the two views: (i) the high-HM region is offset from the single-prompt linear path, so moving the single-prompt solution toward (or away from) f_{w_1} does not reach it; (ii) the DMC corridor sits inside the high-HM region, as the LMC term encourages by keeping classification loss low along the interpolation. Figure 11 (Flowers102) and Figure 12 (Caltech101) display these effects on two datasets with very different dynamic ranges.

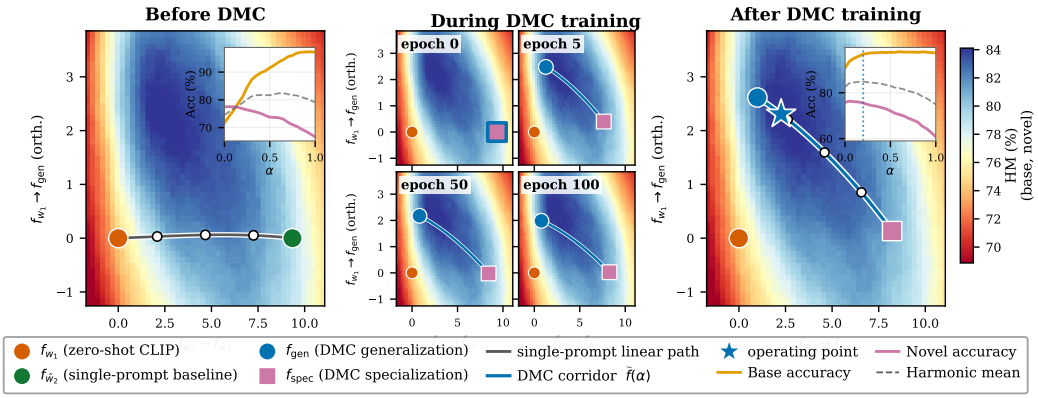

Figure 11: **HM landscape on Flowers102.** Same 2D plane and overlays as the main-paper Figure 1, with the heatmap recolored by harmonic mean of base and novel accuracy (high HM in blue). The DMC corridor lies on the high-HM peak; the operating point at $\alpha = 0.20$ sits near the corridor's HM maximum ≈ 84. The single-prompt linear path along the bottom of the plane traverses an HM ridge that peaks below ≈ 80 at every mixing weight, consistent with the single-prompt ceiling of Lemma 1 along this path.

Flowers102: large dynamic range. The HM on Flowers102 spans roughly 20 points over the plotted plane ($[64.9, 84.6]$), with the DMC operating point at ≈ 84, so the offset between the high-HM peak and the linear single-prompt path is visually dramatic. The DMC corridor traces the rim of this peak, and the fixed operating point at $\alpha = 0.20$ sits near its HM maximum.

Caltech101: near-ceiling regime, same geometry. Caltech101 covers an HM range of ~ 4 points because zero-shot accuracy is already high. The compressed dynamic range obscures the visual offset, yet the same geometry holds: the DMC corridor sits on a high-HM ridge that the single-prompt linear path never reaches. The corridor's offset thus scales with the dataset's HM range, which mirrors the

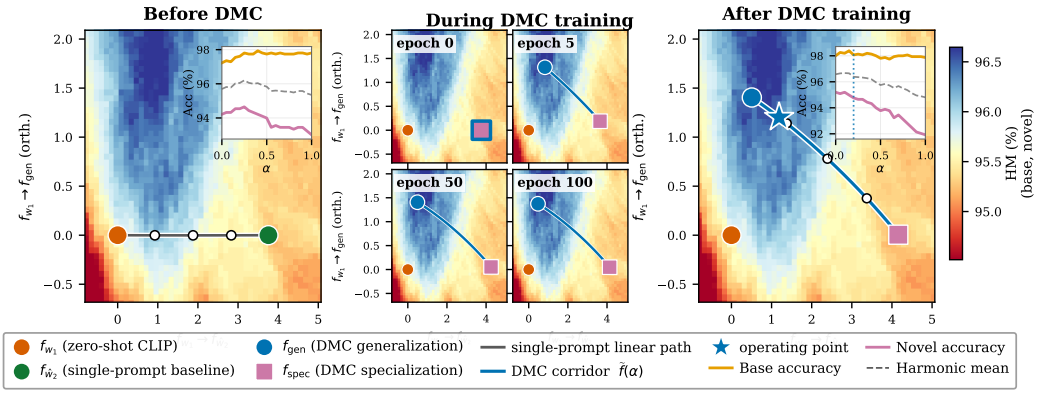

Figure 12: **HM landscape on Caltech101.** Caltech101 is a near-ceiling task (zero-shot CLIP attains HM $\approx$ 95), so the landscape varies over a much narrower range, $[92.8, 96.9]$, than Flowers102 ($[64.9, 84.6]$). Despite the compressed dynamic range, the same geometry is visible: the DMC corridor lies on a high-HM ridge offset from the linear single-prompt path, and the operating point at $\alpha = 0.20$ sits near the ridge maximum. Flat-but-shifted performance is the expected near-ceiling behavior of decoupling: when single-prompt HM is already close to the achievable upper bound, DMC's gain shrinks (CoOp+DMC gains $+0.14$ HM over MERGETUNE on Caltech101, Table 1).

dataset-level pattern in Table 1: large gains on fine-grained tasks, small gains on near-ceiling tasks. Both figures support the same conclusion as the loss landscape: DMC's benefit comes from reaching regions of this plane that the single-prompt linear path does not visit, rather than from a different choice along the same curve.

H When Does DMC Compose? The Endpoint-Separation Condition

This section asks when DMC can help a given base method. DMC interpolates between two learned endpoints; if they coincide, every $\tilde{f}(\alpha)$ is the same classifier and the corridor carries no information. We make this condition measurable, use it to explain the MMA result, test its predictions, and check it on a third parameterization.

H.1 The diagnostic

We measure $\cos(f_{\text{gen}}, f_{\text{spec}})$ at the evaluated (best-validation) checkpoint on 10 datasets (Table 10). The three base methods separate cleanly. CoOp averages 0.600 and KgCoOp 0.719, with no dataset above 0.99. MMA averages 0.993 and exceeds 0.99 on 9 of 10 datasets (all but SUN397, at 0.940); on Caltech101 its α-sweep is flat to within 0.02 HM (Table 17). The ordering CoOp $<$ KgCoOp $<$ MMA matches the ordering of DMC's average HM change relative to the published base method ($+6.40$, $+0.05$ and -6.56, averaging the per-dataset HMs for every row; Table 1), and the diagnostic is measurable at training time without any novel-class data.

H.2 Gains within the collapsed regime

MMA+DMC still improves on MMA+MERGETUNE (by 0.58 HM on average over 11 datasets). If these gains came from the corridor, the datasets that retain some separation should gain most. Table 11 does not bear this out: SUN397, the only dataset below 0.99, gains $+0.42$ HM, while the three largest gains (DTD $+1.55$, UCF101 $+1.52$, Flowers102 $+1.15$) occur at $\cos \geq 0.997$, and $\cos$ and gain are essentially uncorrelated (Pearson $r = 0.14$). Because interpolation cannot change the classifier when the endpoints coincide, we attribute these gains to the different training objective of the (shared) endpoint rather than to the corridor. The collapse accounts for the scope limit, not for these residual gains: on every dataset MMA+DMC stays below published MMA (Table 1).

Table 10: Endpoint-collapse diagnostic for DMC. We report $\cos(f_{\text{gen}}, f_{\text{spec}})$, the cosine similarity between the two DMC endpoints (class text features, averaged over classes) at the evaluated best-validation checkpoint, as 3-seed mean $\pm$ sample std. Runs: $\lambda_{\text{gen}} = 8$, $\beta = 4$, without VA; with VA (the runs of Table 1) the averages are 0.628, 0.704 and 0.996, with SUN397 again the only MMA dataset below 0.99 (0.964). Values close to 1.0 mean the two endpoints have collapsed to the same solution, so the linear connector degenerates and the α-sweep is flat. CoOp and KgCoOp endpoints remain distinct on every dataset (maximum 0.987). MMA endpoints exceed 0.99 on 9 of 10 datasets (all but SUN397, at 0.940); this collapse delimits DMC's scope: on MMA it cannot close the gap to published MMA (Section 4).

Dataset	CoOp + DMC	KgCoOp + DMC	MMA + DMC
Caltech101	0.608 ± 0.065	0.919 ± 0.034	$\mathbf{1.000 \pm 0.000}$
OxfordPets	0.785 ± 0.054	0.971 ± 0.022	$\mathbf{1.000 \pm 0.000}$
StanfordCars	0.673 ± 0.038	0.701 ± 0.009	$\mathbf{0.999 \pm 0.000}$
Flowers102	0.542 ± 0.030	0.606 ± 0.027	$\mathbf{0.997 \pm 0.002}$
Food101	0.741 ± 0.007	0.987 ± 0.014	$\mathbf{0.999 \pm 0.000}$
FGVCAircraft	0.514 ± 0.058	0.521 ± 0.005	$\mathbf{1.000 \pm 0.000}$
SUN397	0.720 ± 0.030	0.888 ± 0.040	$\mathbf{0.940 \pm 0.009}$
DTD	0.377 ± 0.142	0.348 ± 0.043	$\mathbf{1.000 \pm 0.000}$
EuroSAT	0.490 ± 0.147	0.684 ± 0.106	$\mathbf{1.000 \pm 0.000}$
UCF101	0.550 ± 0.070	0.561 ± 0.036	$\mathbf{1.000 \pm 0.000}$
Average	0.600	0.719	**0.993**

Table 11: Per-dataset gap decomposition for MMA + DMC: endpoint cosine $\cos(f_{\text{gen}}, f_{\text{spec}})$ (Table 10) and HM of MMA + MERGETUNE and MMA + DMC (mean over 3 seeds, Table 1). ImageNet has no entry in Table 10 and is omitted; the average is over the 10 datasets shown (the 11-dataset averages are 72.58 and 73.16). The gain does not track the residual separation (Pearson $r = 0.14$ between cos and gain): the largest gains occur on DTD, UCF101 and Flowers102, where $\cos \geq 0.997$.

Dataset	$\cos(f_{\text{gen}}, f_{\text{spec}})$	MMA + MERGETUNE HM	MMA + DMC HM	Gain
Caltech101	1.000	95.75	95.74	-0.01
OxfordPets	1.000	93.22	93.38	$+0.16$
StanfordCars	0.999	70.16	70.51	$+0.35$
Flowers102	0.997	75.22	76.37	$+1.15$
Food101	0.999	90.18	90.33	$+0.15$
FGVCAircraft	1.000	32.72	33.10	$+0.38$
SUN397	0.940	76.70	77.12	$+0.42$
DTD	1.000	58.07	59.62	$+1.55$
EuroSAT	1.000	59.57	60.57	$+1.00$
UCF101	1.000	73.59	75.11	$+1.52$
Average (10 datasets)	0.993	72.52	73.19	$+0.67$

H.3 Intervening on capacity

The condition predicts that giving the generalization endpoint more capacity should break the collapse. We re-ran MMA+DMC on Flowers102 and DTD with a larger generalization-adapter scale (10^{-2} and 10^{-1}, up from 10^{-3}) and with a fresh Kaiming initialization of that adapter (Table 12). Capacity does govern whether the endpoints *can* diverge: cos falls to 0.95 at scale 10^{-2} and to 0.78 at 10^{-1}. Moderate separation helps where it is achieved without overfitting (Flowers102 HM $+2.81$ at scale 10^{-2}; DTD unchanged within noise), but more separation is not better: at scale 10^{-1} the enlarged endpoint overfits the base classes (Flowers102 Base 93.16, Novel 61.09) and DTD loses 11.3% relative HM. Fresh initialization does not break the collapse at all, and no configuration we tested recovers MMA-alone. Endpoint separation is therefore necessary but not sufficient: the endpoints must be distinct *and* each must remain a good solution to its own objective.

Table 12: MMA adapter-capacity ablation for DMC. We test two interventions against the canonical MMA+DMC endpoint collapse ($\cos(f_{\text{gen}}, f_{\text{spec}}) \approx 1.00$): Config A enlarges the generalization-adapter scale from 10^{-3} to 10^{-2} or 10^{-1}; Config B re-initializes the generalization adapter with a fresh Kaiming draw instead of copying the base adapter. HM is base-to-novel harmonic mean at $\alpha = 0.20$; cos is $\cos(f_{\text{gen}}, f_{\text{spec}})$ at end of training; both reported as mean $\pm$ sample std over 3 seeds. $\cos \approx 1$ marks endpoint collapse. MMA-alone is the published number; MMA + MERGETUNE is the mean over 3 seeds from Table 1. A scale of 10^{-2} partly separates the endpoints ($\cos \approx 0.95$) and raises Flowers102 HM by 2.81; a scale of 10^{-1} separates them further ($\cos \approx 0.78$) but the enlarged endpoint overfits the base classes and HM drops. Fresh initialization does not break the collapse. No intervention closes the gap to MMA-alone.

	Flowers102		DTD	
Method	HM	cos	HM	cos
MMA-alone (published)	85.48	—	73.38	—
MMA + MERGETUNE	75.22	—	58.07	—
MMA + DMC (canonical)	76.37 ± 0.23	0.997 ± 0.002	59.62 ± 0.13	1.000 ± 0.000
Config A: enlarge generalization-adapter scale				
scale $= 10^{-2}$	79.18 ± 1.45	0.951 ± 0.004	59.92 ± 1.50	0.980 ± 0.002
scale $= 10^{-1}$	73.78 ± 1.75	0.780 ± 0.017	52.90 ± 1.31	0.795 ± 0.017
Config B: fresh Kaiming init for generalization adapter				
fresh init	73.71 ± 0.16	0.999 ± 0.000	57.49 ± 0.12	1.000 ± 0.000

H.4 A falsification test

A stronger prediction is that on a collapsed base the DMC components should be inert or harmful, reversing the pattern of Table 2. We repeated that component ablation with MMA as the base method, with the same seeds and code paths (Table 13; the two MERGETUNE rows use $\beta = 1$). The outcome matches the prediction: every Δ of $\mathcal{R}$ or VA that is positive on CoOp is zero or negative on MMA. Adding $\mathcal{R}$ lowers HM in seven of eight cases (from -0.04 to -3.25 HM; 0.00 for MERGETUNE on Caltech101), and VA changes HM by at most 0.06. The best MMA row on all four datasets is DMC without $\mathcal{R}$, and that row still sits 6.75 HM below published MMA on Flowers102. Because the deltas reverse sign rather than merely shrink, the result favors the endpoint-separation explanation over a generic "MMA is harder to improve" explanation.

H.5 Beyond prompts and adapters

To check that the condition concerns separation rather than prompts specifically, we implemented LoRA on the text encoder (Q, K, V projections, all 12 layers, ranks 8 and 16) as a third parameterization, on two datasets (Table 14). On Flowers102, $\cos = 0.45$ at both ranks and LoRA+DMC gains $+10.3$ to $+11.3$ HM over LoRA alone; on Caltech101, $\cos = 0.86$–0.91 and DMC adds only $+0.3$ to $+0.4$. Ordering all measurements by cos: LoRA Flowers102 0.45 ($+10$–11), CoOp 0.60 ($+6.40$), KgCoOp 0.72 ($+0.05$), LoRA Caltech101 0.86–0.91 ($+0.3$–0.4) and MMA 0.99 (-6.56), where the prompt and adapter numbers are 11-dataset average changes relative to the published base method. Clear separation goes with large gains and near-total collapse with the one large loss; in between (KgCoOp, LoRA Caltech101) the gains are small and not strictly ordered. We therefore use $\cos < 0.9$ at convergence as a coarse screening heuristic, not as a predictor of the gain. Three caveats apply: the LoRA evidence covers two datasets; our LoRA baseline is under-tuned, so the LoRA Δ values overstate what a well-tuned LoRA would leave on the table (the cos measurement is unaffected); and the condition is symmetric between the two encoders, so vision-side LoRA or adapters of comparable capacity should also satisfy it, which we have not verified.

Table 13: Component ablation with MMA as the base method (HM, mean ± sample std over 3 seeds; DMC rows at $\alpha = 0.20$). Same components as Table 2; the two MERGETUNE rows use $\beta = 1$, the DMC rows $\beta = 4$. Bottom row: published MMA (Table 1).

Method	Flowers102	Caltech101	StanfordCars	EuroSAT
MMA+MERGETUNE w/o $\mathcal{R}$	76.63 ± 0.64	95.75 ± 0.16	70.71 ± 0.16	61.32 ± 0.28
MMA+MERGETUNE	75.22 ± 0.40	95.75 ± 0.16	70.16 ± 0.24	59.57 ± 0.48
MMA+DMC w/o $\mathcal{R}$	$\mathbf{78.73 \pm 0.46}$	$\mathbf{95.77 \pm 0.15}$	$\mathbf{71.62 \pm 0.25}$	$\mathbf{63.78 \pm 0.34}$
MMA+DMC w/o VA	76.38 ± 0.13	95.73 ± 0.17	70.57 ± 0.13	60.53 ± 0.49
MMA+DMC (full)	76.37 ± 0.23	95.74 ± 0.16	70.51 ± 0.04	60.57 ± 0.45
MMA (published)	*85.48*	*96.15*	*75.70*	*83.87*

Table 14: LoRA on the text encoder (Q, K, V, all 12 layers) with and without DMC ($\lambda_{\text{gen}} = 8$, $\beta = 4$, no VA, 50 epochs). HM mean ± sample std over 3 seeds; DMC at $\alpha = 0.20$; $\cos = \cos(f_{\text{gen}}, f_{\text{spec}})$. The LoRA baseline is under-tuned (Flowers102 base accuracy 69.5).

Dataset	Rank	LoRA	LoRA+DMC	Δ	cos
Flowers102	8	73.03 ± 0.03	84.37 ± 0.33	$+11.34$	0.454 ± 0.010
Flowers102	16	73.18 ± 0.11	83.51 ± 0.41	$+10.33$	0.454 ± 0.015
Caltech101	8	95.69 ± 0.00	96.01 ± 0.48	$+0.32$	0.861 ± 0.146
Caltech101	16	95.68 ± 0.05	96.06 ± 0.30	$+0.38$	0.908 ± 0.092

I Class-Permutation Invariance: Extended Analysis

This section tests which candidate regularizers satisfy class-permutation invariance (CPI) and whether the size of a violation relates to how the regularizer behaves. Table 15 reports $\delta_{\max}$ computed over 10,000 random class permutations on Caltech101 ($C = 50$). CPI terms (cosine score, Visual Anchor, ZSDD, and the Drift-Penalty criterion) return $\delta_{\max} \leq 2.4 \times 10^{-7}$ (float32 round-off), consistent with exact floating-point invariance. Non-CPI regularizers return values orders of magnitude larger: Fisher-weighted cosine yields $\delta_{\max} = 0.014$ and EWC yields $\delta_{\max} = 1.07$. Fisher-weighted cosine does not break the single-prompt ceiling (Appendix E). We did not train feature-space EWC as a regularizer, so we make no claim about its accuracy; the δ values only show that both non-CPI regularizers violate the symmetry by orders of magnitude more than the CPI ones. ZSDD is CPI ($\delta_{\max} = 2.4 \times 10^{-7}$) but works only when the zero-shot teacher is accurate; CPI is a desirable symmetry property, not a sufficient one.

Table 15: Class-permutation invariance (CPI) over 10,000 random permutations on Caltech101 ($C = 50$). ΔHM: change of the best single-prompt configuration over the cosine baseline on Caltech101 / Flowers102 (Table 6); —, symmetry test only, not trained. ‡Drift-Penalty is an α-selection criterion (Appendix K), not a training regularizer.

Regularizer	**CPI**	$\boldsymbol{\delta_{\max}}$	$\boldsymbol{\delta_{\text{mean}}}$	**ΔHM (Cal / Flo)**
Cosine score [Wang et al., 2026]	✓	1.2×10^{-7}	4×10^{-8}	baseline
Visual Anchor (ours)	✓	6×10^{-8}	2×10^{-8}	-0.04 / $+0.36$
Drift-Penalty (ours)‡	✓	1.2×10^{-7}	4×10^{-8}	—
ZSDD	✓	2.4×10^{-7}	7×10^{-8}	$+0.40$ / -2.37
Fisher-weighted cosine	×	1.4×10^{-2}	3.3×10^{-3}	$+0.21$ / $+0.22$
EWC (feature space)	×	1.07	0.297	—

I.1 Informal justification

The formal statement is Proposition 1; here we give the intuition.

A regularizer $\mathcal{R}(f, f_{w_1})$ is estimated on base classes $\mathcal{Y}_{\text{base}}$ and must benefit generalization to novel classes $\mathcal{Y}_{\text{novel}}$, with $\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{novel}} = \emptyset$.

If $\mathcal{R}$ is not CPI, it assigns asymmetric importance to specific base-class directions. This asymmetry depends on base-class statistics (curvature, frequency, confidence) that have no defined value for novel-class directions. The regularizer therefore preserves a property specific to $\mathcal{Y}_{\text{base}}$ that may conflict with novel-class preservation.

If $\mathcal{R}$ is CPI, it preserves a property that holds for any permutation of the class set, and therefore any subset of the same size. Since $|\mathcal{Y}_{\text{novel}}| \approx |\mathcal{Y}_{\text{base}}|$ (equal, or differing by one class) in standard protocols, a CPI regularizer preserves a property equally valid for $\mathcal{Y}_{\text{novel}}$.

I.2 Interpreting the magnitude of δ

The $\delta_{\max}$ value measures how much the regularizer changes under class relabeling. A large $\delta_{\max}$ indicates strong class-specific encoding. When applied to novel classes, this class-specific information acts as noise: the regularizer pulls the prompt toward geometry optimal for specific base classes, away from the geometry needed for novel classes. This is a heuristic account; we have not tested whether larger δ predicts a larger accuracy loss.

EWC ($\delta_{\max} = 1.07$) computes its penalty from the full gradient of the classification loss, encoding the specific direction each base class contributes. Fisher-weighted cosine ($\delta_{\max} = 0.014$) uses only the Fisher diagonal as a per-class weight, less specific but still asymmetric. The cosine score ($\delta_{\max} = 1.2 \times 10^{-7}$) uses a uniform weight of $1/C$ for every class, achieving exact CPI.

I.3 The ZSDD special case

ZSDD is CPI ($\delta_{\max} = 2.4 \times 10^{-7}$) because its softmax treats all class features symmetrically. However, ZSDD fails on fine-grained datasets where zero-shot CLIP has low accuracy. On Flowers102 (zero-shot base $\approx 72\%$), the teacher distribution has high entropy, teaching the student to be similarly uncertain and capping base-class confidence. On Caltech101 (zero-shot base $\approx 97\%$), the teacher is sharp and ZSDD produces a small gain.

CPI is therefore not sufficient on its own. The cosine score avoids the teacher-quality issue by anchoring to geometric structure rather than output predictions. Visual Anchor is also a distillation-style term and shares this caveat in principle; in DMC it acts only on $\mathbf{c}_{\text{spec}}$, alongside CE, and its measured effect is small (Appendix F).

I.4 Reproducibility

The CPI test requires no training. Given a text encoder T and zero-shot features $f_{w_1} \in \mathbb{R}^{C \times D}$, compute:

$$\delta_{\max}(\mathcal{R}) = \max_{\Pi \in S} |\mathcal{R}(\Pi f, \Pi f_{w_1}) - \mathcal{R}(f, f_{w_1})|, \tag{6}$$

where S is a random sample of $10{,}000$ permutation matrices from the symmetric group of degree C, and $f = \ell_2(T(\mathbf{c}))$ is the current text-feature matrix. For $C = 50$ classes, this takes under one second.

I.5 Applicability and limitations of the CPI bound

Proposition 1 is a capacity-allocation argument that holds under four assumptions (Appendix M.2): distinct minimizers (0), equal prompt entropy (i), conditional label independence (ii) and baseline efficiency (iii). Each is a modeling choice that can fail in practice. We state what each failure does to the bound and what the evidence still supports.

(0) Distinct minimizers. A CPI-violating penalty changes the objective under relabeling, but two different objectives can share a minimizer; if they do, $I(\theta; Y_{\text{base}})$ can be zero and the bound is vacuous rather than wrong.

(i) Equal prompt entropy $H(\theta) = H(\theta^*) = \kappa$. This normalizes the comparison so that an accuracy gap reflects *how* capacity is spent rather than *how much*. It fails when the two algorithms use different regularization strengths. When entropies differ, the per-prompt inequality $I(\theta; Y_{\text{novel}}) \leq$

$H(\theta) - I(\theta; Y_{\text{base}})$ still holds, but Eq. (8) acquires the correction term $H(\theta) - H(\theta^*)$ on its right-hand side.

(ii) Conditional independence $Y_{\text{base}} \perp Y_{\text{novel}} \mid \theta$. Natural when base and novel labels share no fine-grained attributes. It fails on splits where both halves draw from the same higher-level taxa, since capacity that helps novel classes then also helps base classes and the additive decomposition underlying the capacity bound becomes approximate.

(iii) Baseline efficiency. We assume the CPI-satisfying baseline converts its capacity into novel-class information at the Fano ceiling. Real algorithms fall short of it; then Eq. (11) is no longer guaranteed and can overstate the gap, and the proposition indicates the direction of the effect rather than bounding its size.

Uniform label priors. Uniform priors are part of the setup rather than an assumption on A; they give the cleanest Fano expression. Long-tailed splits reduce the effective label entropy and loosen the bound by the entropy gap $\log |\mathcal{Y}| - H(Y)$.

What survives. Empirically, the non-CPI regularizers violate the symmetry by orders of magnitude more than the CPI ones (Appendix I), but the one we trained (Fisher-weighted cosine) is not clearly harmful in its best configuration (Table 6), so the evidence supports the direction of the proposition, not a measured gap. We therefore present Proposition 1 as an idealized bound that motivates the choice of CPI regularizers and predicts the direction of failure, and we do not rely on its constant.

J Training and Inference Protocol

This section specifies how DMC is trained and evaluated and what it costs. The LMC corridor is a degree-1 (linear) path in ℓ_2-normalized text-feature space; no polynomial or Bézier basis is used.

Training-time α sampling. DMC is evaluated at the checkpoint with the best base-class validation accuracy; MERGETUNE is evaluated at its final prompt. At each gradient step, the LMC path term is evaluated at a single α sampled uniformly from $[0, 1]$, with the endpoints $\alpha \in \{0, 1\}$ included in the sampling support. The expectation over α enforces low classification loss across the full interpolation range, rather than at any fixed mixing weight.

Inference cost. The two prompts are needed only to compute the text features f_{gen} and f_{spec}, which is performed once per dataset. At per-image inference, only the precomputed interpolated feature $\widetilde{f}(\alpha) = \ell_2\big(f_{\text{gen}} + \alpha\,(f_{\text{spec}} - f_{\text{gen}})\big)$ is used. Per-image cost is therefore identical to that of a single-prompt method.

Compute cost. We profiled one full DMC training cell (CoOp+DMC, Flowers102, seed 1, 51 base classes, 100 epochs, batch 128) with `/usr/bin/time -v` and a 2 s `nvidia-smi` sampler on a single NVIDIA A100-PCIE-40GB. Wall-clock is 15 min 34 s (9.34 s per epoch, 0.75 s per iteration), peak GPU memory is 5949 MiB (14.5% of the card), and the trainable parameter count is 4096 (two prompts of $N = 4$ tokens at $D = 512$). Extrapolating this cell to all 11 datasets, 3 baselines and 3 seeds gives roughly 26 A100-hours, an underestimate for larger datasets such as ImageNet. At test time the two prompts are collapsed into one interpolated text feature per class, so per-image inference matches single-prompt CoOp, KgCoOp and MMA exactly. DMC also inherits MERGETUNE's optimizer, schedule and epoch count unchanged, so the only added cost is the second prompt. The profiled cell uses batch size 128, as in the component ablation; the DMC runs of Table 1 use batch size 32, except CoOp and KgCoOp on ImageNet (128).

α selection on the development set. The fixed inference coefficient $\alpha = 0.20$ was selected on a two-dataset development set (Flowers102 and Caltech101) and evaluated on the remaining eight held-out datasets (ImageNet excluded). On the held-out set, $\alpha = 0.20$ achieves a mean HM gap to the per-dataset oracle of -0.18 (worst case -0.28). The method is robust to this choice: every $\alpha \in [0.15, 0.25]$ yields a mean oracle gap of at most 0.26 HM. A per-criterion comparison against eight alternative selection rules is in Appendix K.

K Interpolation Coefficient Selection

DMC produces a family of classifiers parameterized by $\alpha \in [0, 1]$. This section asks how to pick α without novel-class data, and how much the fixed $\alpha = 0.20$, chosen on Flowers102 and Caltech101, costs elsewhere. Here we analyze the α selection problem and evaluate eight automatic criteria.

Problem statement. We seek α^* maximizing $\mathrm{HM} = 2 \cdot \mathrm{Base} \cdot \mathrm{Novel}/(\mathrm{Base} + \mathrm{Novel})$, but novel-class data is unavailable at selection time. Any criterion evaluated on base classes tends to favor high α (more specialization), even when the HM-optimal α is low.

Criteria evaluated. We evaluate eight criteria on up to four datasets (Flowers102, Caltech101, StanfordCars, EuroSAT), using CoOp + DMC + VA checkpoints; the five criteria with large gaps on the two development datasets were evaluated only there.

- **Val-Acc.** Selects α with highest base-class validation accuracy. Fails because val accuracy increases monotonically with α, selecting $\alpha \approx 1.0$.
- **Proxy-HM.** Splits base classes into two random halves (10 splits, averaged), treating one as proxy novel classes. Selects α maximizing the proxy HM. Fails identically to Val-Acc: both halves were seen during training, so the proxy does not simulate the base-novel boundary.
- **Drift-Penalty.** Balances validation accuracy against cosine drift from f_{w_1}: $\alpha^* = \arg\max_\alpha[\overline{\mathrm{Acc}}_{\mathrm{val}}(\alpha) - \mu \cdot \overline{d}_{\cos}(\alpha)]$, with both terms min-max normalized to $[0, 1]$. Requires tuning μ per dataset.
- **Gram-Align.** Selects α maximizing Pearson correlation between the class-prototype Gram matrices at α and at f_{w_1}.
- **Effective Rank.** Computes the effective rank of the prototype matrix $\mathbf{F}(\alpha) \in \mathbb{R}^{C \times d}$ via $\mathrm{erank} = \exp(-\sum_i p_i \log p_i)$. Fails: effective rank *increases* with α because CE spreads prototypes apart, always selecting $\alpha = 1.0$.
- **Neighborhood Preservation.** For each class, measures the fraction of K-nearest neighbors (in f_{w_1} space) preserved at α. Selects highest α with NP above a threshold. Fails: NP starts below 1.0 at $\alpha = 0$, so threshold criteria select $\alpha = 0$.
- **KL Divergence.** $\alpha^* = \arg\max_\alpha[-\mathrm{CE}_{\mathrm{val}}(\alpha) - \nu \cdot \mathrm{KL}(q_\alpha \| q_{w_1})]$, with temperature τ for the softmax. Both terms in nats. KL grows much faster than CE, making the balance sensitive to ν and τ.
- **Pareto Knee.** Plots the Pareto front $(-\mathrm{CE}_{\mathrm{val}}(\alpha),\ \cos(\mathbf{f}(\alpha), f_{w_1}))$ and selects the point of maximum curvature. Hyperparameter-free, but overshoots on gentle fronts.

Results. Table 16 compares all criteria. Among adaptive criteria, Drift-Penalty (μ=4) performs best (-0.22 avg) but requires dataset-dependent μ. KL-based selection is principled but sensitive to ν/τ. No adaptive criterion beats the fixed $\alpha = 0.20$ (-0.19 avg). The Pareto knee overshoots on gentle fronts (StanfordCars -1.16, EuroSAT -2.07).

K.1 Robustness of the fixed α across datasets and dev-set choices

The analysis above evaluates α on four datasets. Two further experiments extend it to all 33 (baseline $\times$ dataset) cells and, for CoOp+DMC, to every alternative two-dataset development set.

Per-dataset oracle $\alpha^\star$. We swept α on a 21-point grid and computed the oracle of the seed-averaged curve per cell (Table 17). The median oracle is $\alpha^\star = 0.25$ on both prompt-tuning bases, one grid step from the fixed 0.20, and the mean HM sacrificed by using 0.20 instead of the per-dataset oracle is 0.08 (CoOp) and 0.17 (KgCoOp). MMA's median $\alpha^\star = 0.95$ is a symptom of the endpoint collapse of Appendix H — on a degenerate corridor the choice of $\alpha^\star$ is poorly determined (mean gap 0.26 HM, up to 0.66 on ImageNet) — not of a failure of the fixed choice.

Dev-set reshuffle. Our fixed α was chosen on Flowers102 and Caltech101, which raises the question of how much that particular pair matters. We repeated the selection for all $\binom{11}{2} = 55$ two-dataset development pairs and measured the held-out loss on the remaining 9 datasets (Table 18).

Table 16: α selection comparison (mean Δ to oracle HM, 3 seeds). Fixed $\alpha = 0.20$ achieves the best average gap. —: not evaluated (large gap already on the development datasets).

Criterion	**Flowers102**	**Caltech101**	**StanfordCars**	**EuroSAT**	**Avg**
Val-Acc	-3.50	-0.95	—	—	—
Proxy-HM	-5.29	-1.05	—	—	—
Gram-Align	-2.29	-0.24	—	—	—
EffRank-Max	-9.26	-1.54	—	—	—
NP$\geq$0.90	-2.29	-0.28	—	—	—
Drift-Pen (μ=4)	-0.25	-0.30	-0.18	-0.14	-0.22
KL (τ=2, ν=0.5)	-0.33	-0.09	-0.87	-0.39	-0.42
Pareto Knee	-0.27	-0.38	-1.16	-2.07	-0.97
Fixed α=0.10	-1.01	-0.19	-0.30	-0.14	-0.41
Fixed α=0.15	-0.44	-0.08	-0.20	-0.18	-0.22
Fixed α=0.20	$\mathbf{-0.17}$	-0.19	$\mathbf{-0.14}$	-0.28	$\mathbf{-0.19}$

Table 17: Per-dataset oracle-best $\alpha^\star$ for DMC across three baselines. HM is averaged over 3 seeds. Δ is $\mathrm{HM}(\alpha^\star) - \mathrm{HM}(0.20)$. Bottom rows: median $\alpha^\star$ across the 11 datasets and range $[\min, \max]$. For CoOp and KgCoOp the median oracle $\alpha^\star$ lies one grid step from the fixed $\alpha = 0.20$; for MMA the corridor is degenerate (Appendix H), so $\alpha^\star$ is poorly determined.

Dataset	CoOp+DMC				KgCoOp+DMC				MMA+DMC			
	$\alpha^\star$	HM*	HM$_{.20}$	Δ	$\alpha^\star$	HM*	HM$_{.20}$	Δ	$\alpha^\star$	HM*	HM$_{.20}$	Δ
Caltech101	0.15	96.52	96.41	+0.11	0.00	96.26	96.11	+0.15	0.00	95.76	95.74	+0.02
DTD	0.20	63.54	63.54	+0.00	0.25	64.51	64.37	+0.14	0.65	59.68	59.62	+0.06
EuroSAT	0.10	75.98	75.84	+0.14	0.00	74.12	73.39	+0.73	0.95	60.77	60.57	+0.20
FGVCAircraft	0.20	35.24	35.24	+0.00	0.05	33.33	33.22	+0.11	1.00	33.21	33.10	+0.10
Food101	0.45	91.22	91.13	+0.09	0.45	91.20	91.19	+0.01	0.15	90.33	90.33	+0.00
ImageNet	0.30	73.06	72.90	+0.17	0.45	73.32	73.09	+0.23	1.00	73.56	72.90	+0.66
Flowers102	0.25	84.49	84.47	+0.02	0.10	83.72	83.67	+0.05	0.95	77.01	76.37	+0.64
OxfordPets	0.25	96.09	96.04	+0.05	0.30	96.33	96.23	+0.09	1.00	93.44	93.38	+0.06
StanfordCars	0.25	75.34	75.21	+0.13	0.25	75.48	75.39	+0.10	0.90	70.70	70.51	+0.19
SUN397	0.25	78.81	78.75	+0.06	0.25	78.94	78.85	+0.08	0.80	77.70	77.12	+0.58
UCF101	0.15	80.23	80.08	+0.15	0.15	79.06	78.91	+0.15	0.95	75.48	75.11	+0.36
Median $\alpha^\star$	0.25	–	–	–	0.25	–	–	–	0.95	–	–	–
Range $\alpha^\star$	[0.10,0.45]	–	–	–	[0.00,0.45]	–	–	–	[0.00,1.00]	–	–	–
Mean Δ	–	–	–	+0.08	–	–	–	+0.17	–	–	–	+0.26

The selected $\alpha^\star$ lands in $[0.15, 0.25]$ for 47 of 55 pairs (85%), 47 pairs keep the held-out mean gap below 0.30 HM, and 54 of 55 stay below 0.50 HM. The median held-out gap is 0.15 HM. The paper's own pair sits inside this dominant mode with a gap of 0.09 HM (measured against the seed-averaged oracle and on 9 held-out datasets, so not directly comparable with the 0.18 of Table 20), so the choice of development set does not drive the results. The worst pair, Food101 with ImageNet, selects $\alpha^\star = 0.45$ and costs 1.19 HM.

Table 18: Dev-set reshuffle robustness for CoOp+DMC. For each of the $\binom{11}{2} = 55$ ways to pick a 2-dataset dev subset, we choose the fixed α that minimizes the mean gap-to-oracle on the dev pair, then measure the mean gap-to-oracle on the remaining 9 held-out datasets. The paper's dev pair is Flowers102+Caltech101. Gaps are to the oracle of the seed-averaged α-curve (Table 17).

Statistic over 55 dev pairs	Value
Pairs that pick $\alpha^\star \in [0.15, 0.25]$	47/55 (85%)
Pairs that pick $\alpha^\star \leq 0.10$ or ≥ 0.35	2/55 (4%)
Pairs with held-out mean gap ≤ 0.30 HM	47/55 (85%)
Pairs with held-out mean gap ≤ 0.50 HM	54/55 (98%)
Mean held-out gap across 55 pairs	0.17 HM
Median held-out gap across 55 pairs	0.15 HM
Best pair: EuroSAT+SUN397 ($\alpha^\star = 0.20$)	0.08 HM
Worst pair: Food101+ImageNet ($\alpha^\star = 0.45$)	1.19 HM

K.2 Drift-Penalty μ sweep on the development set

Table 19 sweeps the drift-penalty weight μ on the two development datasets (Flowers102, Caltech101; 3 seeds each, $\lambda_{\text{gen}} = 8.0$, $\beta = 4.0$), reporting both mean and worst-case oracle gap; these runs are separate from those of Table 16, so the numbers differ. Accuracy-based criteria fail: validation accuracy selects $\alpha = 1.0$ on Flowers102 (pure specialization), losing -7.19 HM versus oracle, because base-class validation accuracy increases monotonically with α. Proxy-HM (random base-class splits used as proxy novel classes) selects the same α as Val-Acc in every case, indicating that base-class splits cannot simulate the base-novel boundary. The drift-penalty criterion balances validation accuracy against cosine drift from f_{w_1}; with $\mu = 4$, the gap shrinks to -0.13 on Flowers102 and -0.02 on Caltech101. The fixed $\alpha = 0.20$ achieves comparable performance with no hyperparameters.

Table 19: Drift-Penalty μ sweep on the development set (Flowers102, Caltech101; 3 seeds, $\lambda_{\text{gen}} = 8.0$, $\beta = 4.0$). Mean and worst-case HM gap to oracle.

	Flowers102		**Caltech101**	
Criterion	Mean Δ	Worst Δ	Mean Δ	Worst Δ
Val-Acc	-7.19	-12.69	-0.29	-0.48
Proxy-HM	-7.19	-12.69	-0.29	-0.48
Gram-Align	-1.47	-1.91	-0.24	-0.45
ValAcc$\times$Gram	-0.78	-1.22	—	—
Drift-Pen (μ=1)	-0.78	-1.22	-0.16	-0.48
Drift-Pen (μ=2)	-0.16	-0.36	-0.16	-0.48
Drift-Pen (μ=4)	$\mathbf{-0.13}$	$\mathbf{-0.29}$	$\mathbf{-0.02}$	$\mathbf{-0.06}$

K.3 Cross-validated fixed α

We select α on two development datasets (Flowers102, Caltech101) and evaluate on 8 held-out datasets. On the dev set, $\alpha \in \{0.20, 0.25\}$ are tied (mean HM $= 90.44$); we choose $\alpha = 0.20$ as the more conservative candidate. Table 20 reports the oracle gap on held-out datasets.

Table 20: Generalization of fixed $\alpha = 0.20$ (selected on dev set) to 8 held-out datasets. Mean Δ to oracle HM.

α	Pets	Food	Cars	Aircraft	SUN	DTD	EuroSAT	UCF	**Avg**
0.10	-0.20	-0.30	-0.30	-0.66	-0.26	-1.33	-0.14	-0.33	-0.44
0.15	-0.19	-0.22	-0.20	-0.42	-0.18	-0.46	-0.18	-0.06	-0.24
0.20	$\mathbf{-0.17}$	$\mathbf{-0.13}$	$\mathbf{-0.14}$	$\mathbf{-0.22}$	$\mathbf{-0.07}$	$\mathbf{-0.23}$	$\mathbf{-0.28}$	$\mathbf{-0.21}$	$\mathbf{-0.18}$
0.25	-0.12	-0.09	-0.01	-0.44	-0.01	-0.43	-0.53	-0.41	-0.26
0.30	-0.15	-0.08	-0.15	-0.93	-0.02	-0.54	-0.77	-0.86	-0.44

Fixed $\alpha = 0.20$ achieves the best average gap (-0.18) on the eight held-out datasets. Any $\alpha \in [0.15, 0.25]$ yields an average gap of at most 0.26 HM. This stability is consistent with the consistently shaped interpolation curves of DMC, whose HM peak lies in a narrow band across datasets (Figure 14).

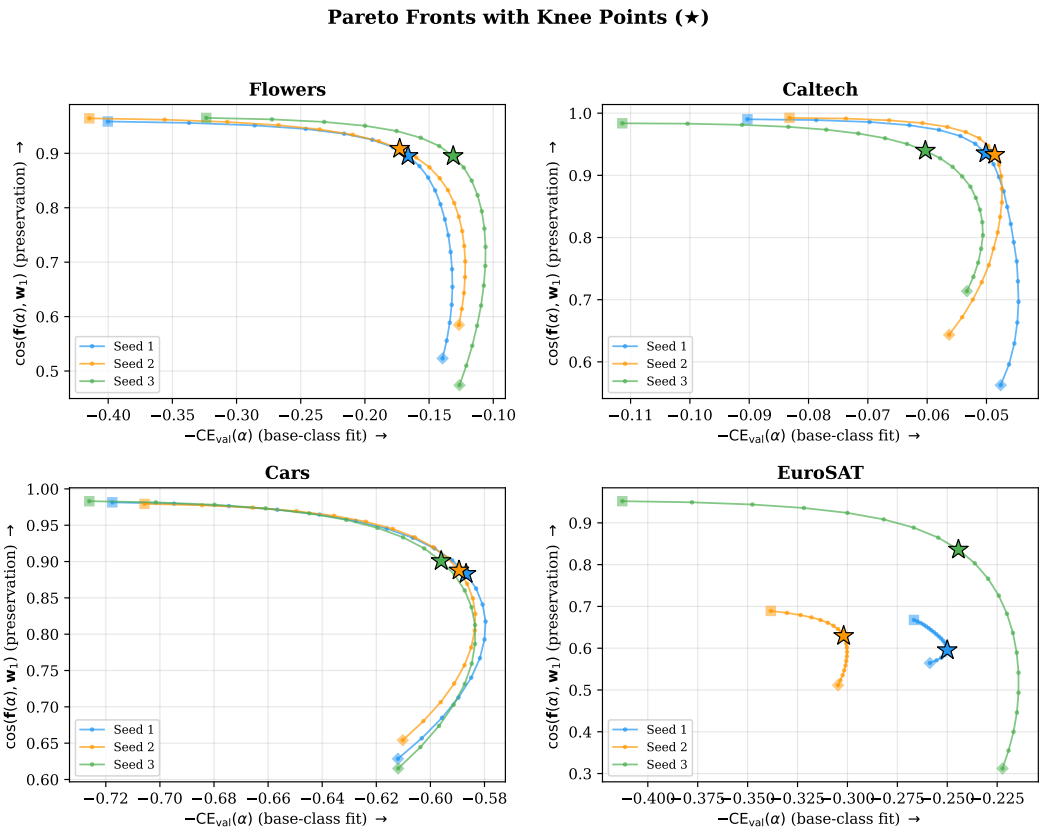

Figure 13: **Pareto fronts** for α selection on four datasets. Stars mark Pareto knee points. As α increases, features move right (better base fit) and down (more drift from f_{w_1}).

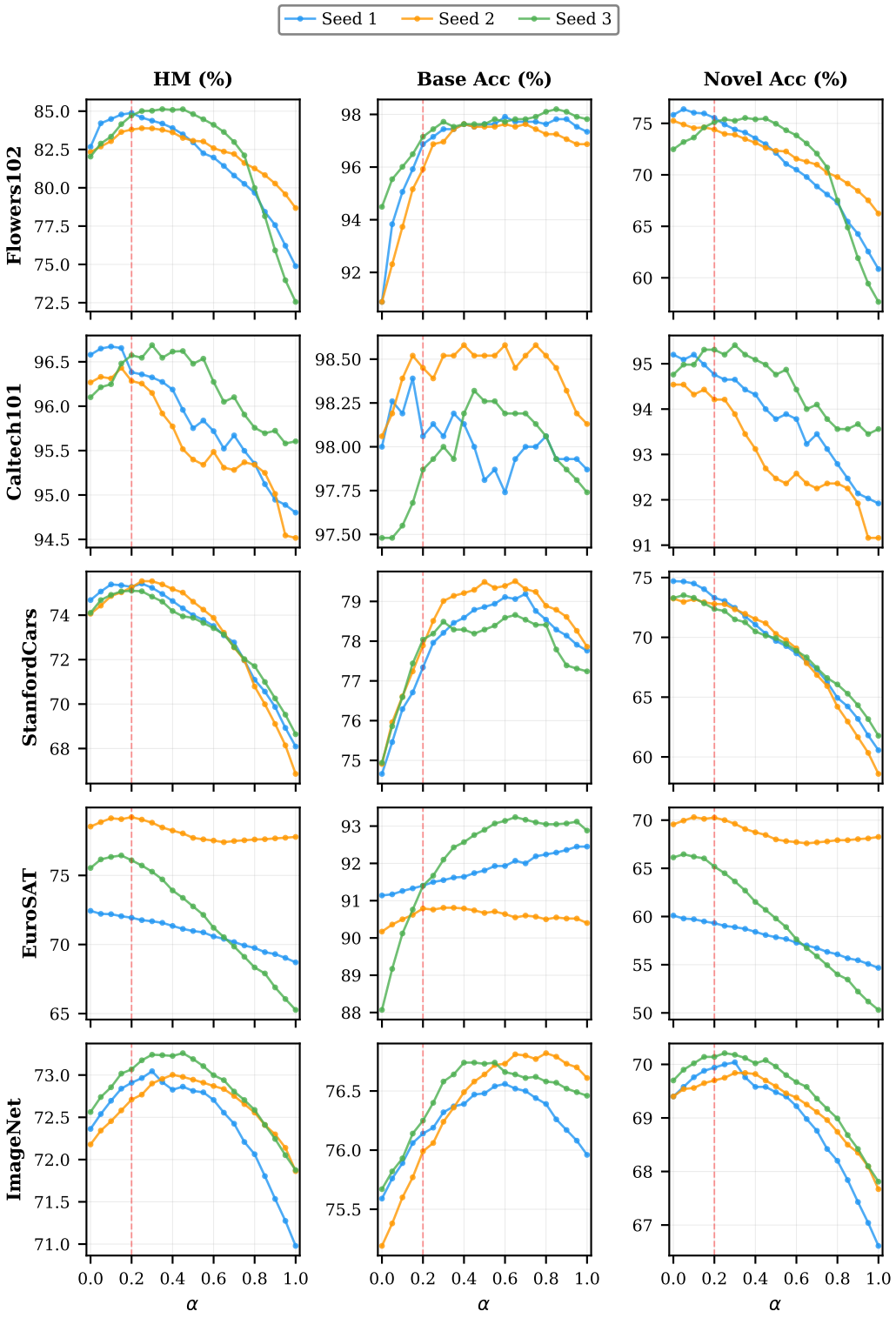

Figure 14: **HM, Base accuracy, and Novel accuracy vs** α on four datasets. Each row is one dataset; columns show HM, Base, and Novel accuracy. The dashed red line marks $\alpha = 0.20$. Base accuracy increases monotonically with α (the specialization prompt dominates), while Novel accuracy decreases (generalization degrades). The HM peak reflects the trade-off between these opposing trends and lies in $\alpha \in [0.10, 0.35]$ across all datasets.

L Hyperparameter Sensitivity

DMC's two main scalar weights are the LMC path weight β and the cosine regularizer weight λ_{gen}, the counterparts of MERGETUNE's β and λ. Both are selected on the two development datasets (Flowers102, Caltech101; 3 seeds each) and held fixed for all reported experiments. This section asks how sensitive DMC is to these two weights; the sweeps show broad plateaus around the selected values.

L.1 LMC Path Weight (β)

The training objective contains an LMC path term whose weight β controls how strongly the corridor between f_{spec} and f_{gen} is regularized. Table 21 evaluates β on Flowers102 and Caltech101 with $\lambda_{\text{gen}} = 8.0$ across three seeds, using DMC without VA ($\lambda_{\text{VA}} = 0$). At each run's oracle $\alpha^\star$, higher β reduces cross-seed variance: on Flowers102, $\beta = 4$ reduces the sample standard deviation from ± 1.50 to ± 0.37, stabilizing an outlier seed, at the cost of a lower mean (84.15 versus 84.44 at $\beta = 1$, within noise); its mean is also within noise of, and slightly below, the single-prompt baseline row (84.24 ± 0.25). On Caltech101, $\beta = 4$ improves both mean and stability. At the fixed inference coefficient $\alpha = 0.20$ used everywhere else, $\beta = 4$ has the smallest standard deviation on both datasets, the best Caltech101 mean (96.31 ± 0.10 versus 95.84 ± 0.32 at $\beta = 1$), and a Flowers102 mean within noise of the best (83.91 ± 0.46, versus 84.00 ± 0.91 at $\beta = 2$ and 82.78 ± 0.72 at $\beta = 1$). The case for $\beta = 4$ is therefore lower variance, not a higher mean. We therefore use $\beta = 4.0$ throughout.

Table 21: β (LMC path weight) sweep ($\lambda_{\text{gen}} = 8.0$, 3 seeds, DMC without VA). Per-seed HM at each run's oracle $\alpha^\star$; mean $\pm$ sample standard deviation. *Baseline*: single-prompt MERGETUNE objective ($\lambda = 10$, $\beta = 1$), same seeds. Higher β stabilizes performance; the selected $\beta = 4.0$ achieves the lowest cross-seed variance among the DMC rows on both datasets.

	Flowers102				**Caltech101**			
β	S1	S2	S3	Mean $\pm$ Std	S1	S2	S3	Mean $\pm$ Std
1.0	84.66	82.84	85.82	84.44 ± 1.50	96.15	96.37	95.76	96.09 ± 0.31
2.0	84.38	83.06	85.63	84.36 ± 1.29	96.20	96.39	95.76	96.12 ± 0.32
4.0	**84.26**	**83.73**	**84.45**	$\mathbf{84.15 \pm 0.37}$	**96.27**	**96.42**	**96.55**	$\mathbf{96.41 \pm 0.14}$
Baseline	*84.06*	*84.12*	*84.53*	84.24 ± 0.25	*96.00*	*96.24*	*95.92*	96.05 ± 0.17

L.2 Cosine Regularizer Weight (λ_{gen})

Table 22 reports HM as a function of λ_{gen}, the weight on the cosine regularizer $\mathcal{R}$ that anchors $\mathbf{c}_{\text{gen}}$ to f_{w_1} (Eq. (1)). On Flowers102 and Caltech101 with $\beta = 1.0$ (seed 1), performance plateaus for $\lambda_{\text{gen}} \geq 4$ (Flowers102 HM 84.45–84.66; Caltech101 95.72–96.15). Below this threshold, $\cos(f_{\text{gen}}, f_{w_1})$ drops from 0.98 to 0.92 and novel accuracy degrades. At $\lambda_{\text{gen}} = 0$ the cosine anchor is removed and decoupling collapses: both prompts converge to the same base-specialized solution. We select $\lambda_{\text{gen}} = 8.0$.

Table 22: Effect of λ_{gen} (cosine regularizer weight) in DMC training (seed 1, $\beta = 1.0$; HM at the best α^* of each run). At $\lambda_{\text{gen}} = 0$ the anchor is removed and decoupling collapses. Selected configuration is highlighted; the bottom row reports the single-prompt baseline.

	Flowers102				**Caltech101**			
λ_{gen}	α^*	Base	Novel	HM	α^*	Base	Novel	HM
0.0	0.15	97.72	65.32	78.30	1.00	98.13	92.90	95.44
1.0	0.10	97.25	70.99	82.07	0.40	97.42	93.45	95.39
2.0	0.25	97.15	73.12	83.44	0.40	97.68	93.56	95.58
4.0	0.30	96.77	74.96	84.48	0.25	97.87	93.67	95.72
8.0	**0.30**	**95.73**	**75.89**	**84.66**	**0.15**	**97.93**	**94.43**	**96.15**
10.0	0.30	94.97	76.03	84.45	0.20	97.81	94.43	96.09
Baseline	—	*94.21*	*75.46*	*83.80*	—	*97.93*	*94.21*	*96.03*

M Proofs

M.1 Proof of Lemma 1

Proof. At any stationary point $\mathbf{c}^*$ of $\mathcal{L}(\mathbf{c}) = \mathcal{L}_{\text{CE}}(\mathbf{c}) + \lambda\,\mathcal{R}(\mathbf{c},\, f_{w_1})$ with $\lambda > 0$, the first-order condition gives:

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}(\mathbf{c}^*) + \lambda\,\nabla_{\mathbf{c}}\mathcal{R}(\mathbf{c}^*) = \mathbf{0}.$$

Setting $\mu_1 = 1/(1+\lambda) > 0$ and $\mu_2 = \lambda/(1+\lambda) > 0$ (so $\mu_1 + \mu_2 = 1$), we obtain

$$\mu_1\,\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}(\mathbf{c}^*) + \mu_2\,\nabla_{\mathbf{c}}\mathcal{R}(\mathbf{c}^*) = \mathbf{0}.$$

This is precisely the Pareto-stationarity condition [Momma et al., 2022]: a point is Pareto-stationary for $(\mathcal{L}_{\text{CE}},\, \mathcal{R})$ if and only if there exist non-negative weights summing to one such that the weighted gradient vanishes. Since $\mu_1, \mu_2 > 0$, the condition holds and $\mathbf{c}^*$ is Pareto-stationary, meaning no direction in $\mathbb{R}^{N\times D}$ simultaneously decreases both objectives to first order.

If $\nabla_{\mathbf{c}}\mathcal{L}_{\text{CE}}(\mathbf{c}^*) \neq \mathbf{0}$, then $\mathbf{c}^*$ is not a critical point of $\mathcal{L}_{\text{CE}}$ alone; in particular, $\mathbf{c}^*$ does not minimize $\mathcal{L}_{\text{CE}}$. The analogous argument applies to $\mathcal{R}$ when its gradient is nonzero. □

M.2 Proof of Proposition 1

Proposition 1 (CPI violation lower-bounds the novel-accuracy drop). *Let $\theta = A(\mathcal{D}_{\text{base}})$ be the prompt produced by a training algorithm A that uses a regularizer $\mathcal{R}$, and let $\theta^* = A^*(\mathcal{D}_{\text{base}})$ be the prompt from an otherwise identical algorithm using a CPI-satisfying regularizer. Let $Y_{\text{base}} \sim \text{Unif}(\mathcal{Y}_{\text{base}})$ and $Y_{\text{novel}} \sim \text{Unif}(\mathcal{Y}_{\text{novel}})$. Under assumptions (i)–(iii) below (equal prompt entropy, conditional label independence, and baseline efficiency), the following capacity bound holds:*

$$I(\theta;\, Y_{\text{novel}}) \;\le\; I(\theta^*;\, Y_{\text{novel}}) - I(\theta;\, Y_{\text{base}}).$$

If $\mathcal{R}$ violates CPI, then, under assumption (0), $I(\theta;\, Y_{\text{base}}) > 0$ and

$$A_{\text{novel}}(\theta^*) - A_{\text{novel}}(\theta) \;\ge\; \frac{I(\theta;\, Y_{\text{base}})}{\log\,|\mathcal{Y}_{\text{novel}}|} > 0.$$

Setup. Let $\mathcal{D}_{\text{base}} \sim p_{\text{base}}^n$ denote the training set, and write $\theta = A(\mathcal{D}_{\text{base}})$ for the converged prompt produced by training algorithm A. The prompt $\theta \in \mathbb{R}^{N\times D}$ is a random variable whose randomness is inherited entirely from the draw of $\mathcal{D}_{\text{base}}$. Let $Y_{\text{base}} \sim \text{Unif}(\mathcal{Y}_{\text{base}})$ and $Y_{\text{novel}} \sim \text{Unif}(\mathcal{Y}_{\text{novel}})$ be the marginal label random variables, and let $M = |\mathcal{Y}_{\text{novel}}|$. Since $\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{novel}} = \emptyset$, marginal independence holds: $Y_{\text{base}} \perp Y_{\text{novel}}$. In this idealized model, the joint law of $(\theta, Y_{\text{base}}, Y_{\text{novel}})$ is induced by the random draw of the training data (and of which classes are labeled as base), so θ may carry information about the base labels. We treat θ as a discrete random variable (finite-precision parameters), so that $H(\theta)$ and all mutual informations below are Shannon quantities.

Part 1: CPI violation implies $I(\theta;\, Y_{\text{base}}) > 0$. We impose the following regularity condition on A:

(0) *Distinct minimizers.* For any $C \times C$ permutation matrix Π_0 such that $\mathcal{R}(\Pi_0 f, \Pi_0 f_{w_1}) \neq \mathcal{R}(f, f_{w_1})$, the algorithm A produces outputs that differ in distribution on $\mathcal{D}_{\text{base}}$ and on the relabeled dataset $\mathcal{D}_{\text{base}}^{(\Pi_0)}$: $A(\mathcal{D}_{\text{base}}) \neq_d A(\mathcal{D}_{\text{base}}^{(\Pi_0)})$. This holds whenever the loss landscape is not symmetric under Π_0. CPI violation guarantees this asymmetry: since $\mathcal{R}$ assigns different penalties to the two datasets, their optimization landscapes differ, and their minimizers differ generically.

The condition $I(\theta;\, Y_{\text{base}}) = 0$ would require $A(\mathcal{D}_{\text{base}}) =_d A(\mathcal{D}_{\text{base}}^{(\Pi)})$ for every permutation Π: the output distribution must be invariant to relabeling. Since $\mathcal{R}$ violates CPI (Definition 1), there exists Π_0 such that $\mathcal{R}(\Pi_0 f, \Pi_0 f_{w_1}) \neq \mathcal{R}(f, f_{w_1})$. By assumption (0), $A(\mathcal{D}_{\text{base}}) \neq_d A(\mathcal{D}_{\text{base}}^{(\Pi_0)})$, contradicting invariance. Therefore $I(\theta;\, Y_{\text{base}}) > 0$.

Part 2: The Fano bound. We invoke three assumptions:

(i) *Equal prompt entropy.* $H(\theta) = H(\theta^*) = \kappa$ for some $\kappa > 0$. Both prompts have the same entropy, so differences in novel accuracy are attributable to *what* the capacity encodes, not *how much*.

(ii) *Conditional label independence.* $Y_{\text{base}} \perp Y_{\text{novel}} \mid \theta$ and $Y_{\text{base}} \perp Y_{\text{novel}} \mid \theta^*$. Given the prompt, novel and base class labels are independent. We treat this as an assumption; it is natural when the two label sets are semantically disjoint, so no specific base-class label helps predict which novel class an image belongs to.

(iii) *Baseline efficiency.* $I(\theta^*;\, Y_{\text{novel}}) = H(\theta^*) = \kappa$ and $A^*_{\text{novel}} = (I(\theta^*;\, Y_{\text{novel}}) + I_X + 1)/\log M$, where $I_X = I(Y_{\text{novel}};\, X)$ is the image-side mutual information (identical for both prompts, since test images are drawn from the same distribution). This is a best-case assumption: we suppose the CPI-satisfying baseline fully converts its freed capacity into novel-class information so that the Fano upper bound is tight. If the baseline falls short of this ceiling, the lower bound on the accuracy gap is no longer guaranteed and can overstate the gap (Appendix I.5).

Capacity bound. Applying the chain rule and using marginal independence ($Y_{\text{base}} \perp Y_{\text{novel}}$) and assumption (ii):

$$\begin{aligned} I(\theta;\, Y_{\text{novel}} \mid Y_{\text{base}}) &= H(Y_{\text{novel}} \mid Y_{\text{base}}) - H(Y_{\text{novel}} \mid \theta,\, Y_{\text{base}}) \\ &= H(Y_{\text{novel}}) - H(Y_{\text{novel}} \mid \theta) = I(\theta;\, Y_{\text{novel}}). \end{aligned}$$

Therefore:

$$I(\theta;\, Y_{\text{base}},\, Y_{\text{novel}}) = I(\theta;\, Y_{\text{base}}) + I(\theta;\, Y_{\text{novel}}).$$

Since $I(U;\, V) \leq H(U)$ for any random variables U, V:

$$I(\theta;\, Y_{\text{base}}) + I(\theta;\, Y_{\text{novel}}) \leq H(\theta) = \kappa \quad \text{(by assumption (i))}.$$

This yields the capacity bound:

$$I(\theta;\, Y_{\text{novel}}) \;\leq\; \kappa - I(\theta;\, Y_{\text{base}}). \tag{7}$$

By assumption (iii), $I(\theta^*;\, Y_{\text{novel}}) = \kappa$; applied to θ^*, the same capacity bound then forces $I(\theta^*;\, Y_{\text{base}}) = 0$. This is an idealization: it ignores the base-label information that the cross-entropy term injects into both prompts, and only isolates the contribution of the regularizer. Substituting into (7):

$$I(\theta;\, Y_{\text{novel}}) \;\leq\; I(\theta^*;\, Y_{\text{novel}}) - I(\theta;\, Y_{\text{base}}). \tag{8}$$

This is the capacity bound behind Proposition 1.

Fano inequality. We assume that, given the prompt, the image carries the same information about its novel label as it does marginally, i.e. the prompt does not change what the image itself reveals:

$$I(Y_{\text{novel}};\, X \mid \theta) \;=\; I(Y_{\text{novel}};\, X) \;=:\; I_X. \tag{9}$$

Under (9), the chain rule gives:

$$I(Y_{\text{novel}};\, \theta,\, X) = I(Y_{\text{novel}};\, \theta) + I(Y_{\text{novel}};\, X \mid \theta)$$
$$= I(Y_{\text{novel}};\, \theta) + I(Y_{\text{novel}};\, X) = I(\theta;\, Y_{\text{novel}}) + I_X,$$

where the second equality uses (9). For any classifier $\hat{Y}_{\text{novel}} = g(\theta,\, X)$, Fano's inequality [Cover and Thomas, 2006] (Theorem 2.10.1) applied to the channel $Y_{\text{novel}} \to (\theta, X) \to \hat{Y}_{\text{novel}}$ gives (logarithms are base 2):

$$A_{\text{novel}} \;\le\; \frac{I(Y_{\text{novel}};\, \hat{Y}_{\text{novel}}) + 1}{\log M}.$$

Since $\hat{Y}_{\text{novel}}$ is a deterministic function of (θ, X), the data-processing inequality yields $I(Y_{\text{novel}};\, \hat{Y}_{\text{novel}}) \le I(Y_{\text{novel}};\, \theta,\, X)$, and therefore:

$$A_{\text{novel}} \;\le\; \frac{I(\theta;\, Y_{\text{novel}}) + I_X + 1}{\log M}. \tag{10}$$

Completing the bound. Substituting (8) into (10):

$$A_{\text{novel}} \;\le\; \frac{I(\theta^*;\, Y_{\text{novel}}) - I(\theta;\, Y_{\text{base}}) + I_X + 1}{\log M} \;=\; A^*_{\text{novel}} - \frac{I(\theta;\, Y_{\text{base}})}{\log M},$$

where the last equality uses assumption (iii) and I_X cancels because test images are drawn from the same distribution for both prompts. Rearranging:

$$\Delta A_{\text{novel}} = A^*_{\text{novel}} - A_{\text{novel}} \;\ge\; \frac{I(\theta;\, Y_{\text{base}})}{\log\, |\mathcal{Y}_{\text{novel}}|}. \tag{11}$$

If $\mathcal{R}$ violates CPI, Part 1 gives $I(\theta;\, Y_{\text{base}}) > 0$, so $\Delta A_{\text{novel}} > 0$. This is the bound stated informally in Proposition 1. □

M.3 Proof of Corollary 1

Corollary 1 (Cosine regularizer satisfies CPI). *The cosine regularizer $\mathcal{R}$ in Eq.* (1) *satisfies Definition 1: it uses no image labels, so $\mathcal{R}$ itself adds no base-label information to the prompt it regularizes.*

Proof. For any permutation matrix Π and feature matrices $f, f_{w_1} \in \mathbb{R}^{C\times D}$, permuting the rows of both simultaneously re-indexes the classes in the sum. Each cosine term satisfies $\cos\big((\Pi f)_k,\, (\Pi f_{w_1})_k\big) = \cos\big(f_{\Pi^{-1}(k)},\, f_{w_1,\Pi^{-1}(k)}\big)$, so summing over all k is equivalent to summing over the re-indexed classes. Since the sum in Eq. (1) runs uniformly over all C classes, $\mathcal{R}(\Pi f,\, \Pi f_{w_1}) = \mathcal{R}(f,\, f_{w_1})$ for every Π, confirming CPI (Definition 1). □

Scope. In DMC, $\mathbf{c}_{\text{gen}}$ also receives label-dependent cross-entropy gradients through the LMC corridor term of Eq. (4) (for every sampled $\alpha < 1$), so the trained $\mathbf{c}_{\text{gen}}$ can carry base-label information. The corollary therefore states that the cosine regularizer itself introduces no base-label dependence; it does not claim that $\mathbf{c}_{\text{gen}}$ is independent of the base labels under the full DMC objective.

Remark 1 (CPI violation and the Hellstrom generalization bound)**.** *By Theorem 2.2 of Hellström and Durisi [2020], for θ trained on $\mathcal{D}_{\text{base}}$ and a loss bounded in $[0, 1]$:*

$$\mathbb{E}\big[A_{\text{base,train}}(\theta) - A_{\text{base}}(\theta)\big] \;\le\; \sqrt{\frac{2\, I(\theta;\, \mathcal{D}_{\text{base}})}{n}},$$

where $A_{\text{base,train}}(\theta)$ is the training accuracy on base classes, $A_{\text{base}}(\theta)$ its population counterpart, and the expectation is over $\mathcal{D}_{\text{base}}$ and the training algorithm. When CPI is violated, $I(\theta;\, \mathcal{D}_{\text{base}}) \ge I(\theta;\, Y_{\text{base}}) > 0$, so the bound is strictly positive: a larger mutual information between the prompt and base training data permits a larger generalization gap. The bound concerns the base-class distribution; it does not by itself bound novel-class accuracy, and a positive upper bound permits but does not imply a gap. We include it as a complementary view, not as a second proof of Proposition 1.

M.4 Proof of Proposition 2

Proposition 2 (Visual Anchor as KL regularization)**.** *Define the Gibbs class-similarity distributions* $p_{K_0}(k \mid x) \propto \exp\big(\gamma\, f_{w_1,k}^\top V(x)/\tau\big)$ *and* $p_{K_\theta}(k \mid x) \propto \exp\big(\gamma\, f_{\mathrm{spec},k}^\top V(x)/\tau\big)$*, where* $f_{\mathrm{spec}} = \ell_2(T(\mathbf{c}_{\mathrm{spec}}))$*,* γ *is CLIP's logit scale, and* $V(x)$ *is shorthand for the* ℓ_2*-normalized visual feature* $\ell_2(V(x))$*. In the single-view case (*$v_a = v_b = V(x)$*), the VA loss (Eq.* (5)*) reduces exactly to:*

$$\mathcal{L}_{\mathrm{VA}}(\mathbf{c}_{\mathrm{spec}}) \;=\; \mathbb{E}_x\, \mathrm{KL}\big(p_{K_0}(\cdot \mid x) \,\|\, p_{K_\theta}(\cdot \mid x)\big). \tag{12}$$

Minimizing $\mathcal{L}_{\mathrm{VA}}$ *is thus equivalent to minimizing KL divergence from CLIP's pretrained class-similarity distribution. The two-view formulation in Eq.* (5) *additionally enforces augmentation consistency; for it, the proof establishes only that the loss converges to Eq.* (12) *as the augmentation variance vanishes.*

Exact equality in the single-view case. Define the Gibbs class-similarity distributions:

$$p_{K_0}(k \mid x) = \frac{\exp\big(\gamma\, f_{w_1,k}^\top V(x)/\tau\big)}{\sum_{k'} \exp\big(\gamma\, f_{w_1,k'}^\top V(x)/\tau\big)}, \qquad p_{K_\theta}(k \mid x) = \frac{\exp\big(\gamma\, f_{\mathrm{spec},k}^\top V(x)/\tau\big)}{\sum_{k'} \exp\big(\gamma\, f_{\mathrm{spec},k'}^\top V(x)/\tau\big)}.$$

In the single-view case ($v_a = v_b = V(x)$), the VA loss (Eq. (5)) reduces to:

$$\mathcal{L}_{\mathrm{VA}}(\mathbf{c}_{\mathrm{spec}}) = \mathbb{E}_x \sum_k p_{K_0}(k \mid x) \log \frac{p_{K_0}(k \mid x)}{p_{K_\theta}(k \mid x)} = \mathbb{E}_x\, \mathrm{KL}\big(p_{K_0}(\cdot \mid x) \,\|\, p_{K_\theta}(\cdot \mid x)\big).$$

This is an exact equality: since p_{K_0} and p_{K_θ} are both proper normalized distributions, the KL divergence is well-defined and no additive constant appears. This establishes Eq. (12). □

KL direction. The direction $\mathrm{KL}(p_{K_0} \| p_{K_\theta})$ (teacher ‖ student) penalizes f_{spec} for assigning low probability where the pretrained model f_{w_1} is confident, which regularizes f_{spec} toward the zero-shot geometry. The reverse direction $\mathrm{KL}(p_{K_\theta} \| p_{K_0})$ would instead penalize the student for assigning high probability to classes where the pretrained model is not confident; this mode-seeking behavior does not enforce coverage of all teacher-confident classes and is therefore a weaker regularizer toward the zero-shot distribution [Agarwal et al., 2024].

Two-view approximation. In practice (Eq. (5)), teacher and student use independently augmented views $v_a = V(\mathrm{aug}(x))$ and $v_b = V(\mathrm{aug}'(x))$, yielding the cross-view loss

$$\mathcal{L}_{\mathrm{VA}}^{(2)} = \mathbb{E}_{x,\, v_a,\, v_b}\, \mathrm{KL}\big(p_{K_0}(\cdot \mid x, v_a) \,\|\, p_{K_\theta}(\cdot \mid x, v_b)\big),$$

which differs from the single-view expression. As augmentation variance vanishes, i.e., $\mathbb{E}\|V(\mathrm{aug}(x)) - V(x)\|_2^2 \to 0$, both v_a and v_b converge in distribution to $V(x)$; by continuity of the KL divergence and dominated convergence (the logits are bounded because all features are ℓ_2-normalized), $\mathcal{L}_{\mathrm{VA}}^{(2)}$ converges to the single-view expression. The independent views serve a different purpose: by using distinct random crops $v_a \neq v_b$ of the same image, minimizing the cross-view KL requires f_{spec} to match the teacher's class-similarity structure across the augmentation distribution, not just for the specific views seen at each step. This is analogous to the cross-view consistency used in DINO [Caron et al., 2021] and MoCo [He et al., 2020]. For nonzero augmentation, $\mathcal{L}_{\mathrm{VA}}^{(2)}$ is still a KL divergence from CLIP's pretrained class-similarity distribution, but one evaluated on a different view than the student's; it equals Eq. (12) only in the zero-augmentation limit, and we make no exactness claim beyond that limit. □

M.5 Assumptions summary

For completeness, we list all assumptions used in the proofs above. Items 2–5 are the four assumptions of Proposition 1; Appendix I.5 discusses when each fails.

1. **Smoothness** (Lemma 1). $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{R}$ are continuously differentiable in $\mathbf{c}$. If the gradient of each is nonzero at $\mathbf{c}^*$, the conclusion that neither objective is individually minimized follows immediately.

2. **Distinct minimizers** (Proposition 1, assumption (0)). CPI violation changes the optimization landscape, producing distinct minimizers under different class relabelings. We expect this to hold generically, but it is not guaranteed: two different objectives can share a minimizer.
3. **Equal prompt entropy** (Proposition 1, assumption (i)). $H(\theta) = H(\theta^*) = \kappa$. This normalizes the comparison so that differences in novel accuracy are attributed to allocation, not amount, of capacity.
4. **Conditional label independence** (Proposition 1, assumption (ii)). $Y_{\text{base}} \perp Y_{\text{novel}} \mid \theta$ and given θ^*. This is an assumption, not a consequence of label-set disjointness alone. It is natural when the two class sets are semantically disjoint and no base-class label provides information about which novel class an image belongs to.
5. **Baseline efficiency** (Proposition 1, assumption (iii)). The CPI-satisfying baseline fully converts its freed capacity into novel-class information. If the baseline is sub-optimal, the bound (11) is no longer guaranteed and can overstate the accuracy gap.
6. **Single-view case** (Proposition 2). Eq. (12) holds exactly when $v_a = v_b$. The two-view formulation (5) is not covered exactly; it converges to the single-view expression as augmentation variance vanishes.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state three claims: (1) a single prompt settles at a Pareto-stationary compromise between the two objectives, and the single-prompt variants we test do not clearly improve on it; (2) DMC shifts the base–novel frontier for prompt-tuning baselines via a two-prompt architecture with a feature-space LMC corridor; (3) regularizers should satisfy CPI to transfer across the base-novel boundary, motivated by an idealized bound. Claim (1) is motivated by Lemma 1 and supported empirically (Figure 4, Appendix E); claim (3) is motivated by Proposition 1, an idealized bound, and claim (2) is supported by experiments on 11 datasets across three baselines (Section 4); the MMA scope limitation and the idealized nature of the CPI bound are stated where the claims are made.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph in Section 5 discusses: (1) an additional inference hyperparameter α with no base-only selection criterion that recovers the HM-optimal point without novel-class signal; (2) reduced Pareto shift on near-ceiling datasets where the two prompts converge to similar representations; (3) the theoretical assumptions underlying Proposition 1 (stated in Appendix M, with their failure modes in Appendix I.5); and (4) the requirement that the two endpoints stay separated, which fails on the adapter-based MMA (Appendix H).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All assumptions are stated in the formal statement (Appendix M.2) and listed in full in Appendix M.5. Proofs are given in Appendix M for Lemma 1 (Pareto stationarity), Proposition 1 (CPI–Fano lower bound), Corollary 1 (cosine satisfies CPI), Proposition 2 (Visual Anchor as a KL divergence), and Remark 1 (base-class generalization gap, a complementary view). Proposition 1 holds in an idealized model; Appendix I.5 discusses when its assumptions fail. Each result in the main text is cross-referenced to its appendix proof.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Algorithm 1 gives the complete training and inference procedure. Section 4 specifies the backbone (CLIP ViT-B/16), all hyperparameters ($\lambda_{\text{gen}} = 8.0$, $\beta = 4.0$, $\lambda_{\text{VA}} = 0.5$, $\tau = 2.0$, $\alpha = 0.20$), the 16-shot base-to-novel protocol, the α selection procedure (dev set: Flowers102 and Caltech101, 3 seeds), and the datasets. All other settings follow MergeTune [Wang et al., 2026] defaults, which are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/IemProg/DMC`. All 11 evaluation datasets are publicly available benchmarks following the standard CoOp protocol [Zhou et al., 2022b]. Full implementation details are provided in Section 4, Algorithm 1 and Appendix J.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4 specifies the base-to-novel split protocol (16-shot training, separate base and novel test sets), all DMC hyperparameters, and how β, λ_{gen} and α were selected

(dev-set sweeps on Flowers102 and Caltech101; λ_{VA} and τ are fixed, and we report no sweep for them), the CLIP ViT-B/16 backbone, and the baselines. Optimizer and remaining training settings (learning rate, epochs) follow MERGETUNE defaults; batch sizes are given in Appendix J. The α selection analysis is in Section 4.1 and Appendix K.

Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Every configuration behind the main results and ablation tables is run with 3 seeds; the single-prompt loss sweep, the λ_{gen} sweep and the training-dynamics figure use one seed and are labelled as such.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix J (Compute cost) reports the GPU type (one NVIDIA A100-PCIE-40GB), peak memory (5949 MiB), and wall-clock time of a full training cell (15 min 34 s), and estimates roughly 26 A100-hours for all 11 datasets, 3 baselines and 3 seeds. This covers the main-table DMC runs only; ablations, sweeps and preliminary experiments required additional compute of the same order per run. DMC trains approximately 2K parameters per prompt on top of a frozen CLIP ViT-B/16; training follows MERGETUNE's schedule.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper proposes a prompt learning method for image classification using publicly available datasets and models. No human subjects, sensitive data, or dual-use risks are involved.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This paper presents a foundational method for adapting vision-language models to downstream classification tasks. It does not target a specific application domain and has no direct path to high-risk societal harms such as surveillance, disinformation, or biased decision-making systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: No new datasets or high-risk pre-trained models are released. The method fine-tunes a small prompt context on top of publicly available CLIP checkpoints.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets (ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, UCF101) and all baseline methods are cited. The evaluation protocol follows Zhou et al. [2022b], and the CLIP backbone is used under its publicly released terms. CLIP is released under the MIT license; the datasets are standard academic benchmarks used under their original research terms of use, and we redistribute none of them.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The only new asset is the code, released at `https://github.com/IemProg/DMC`; the method, hyperparameters and protocol it implements are documented in Section 3, Section 4, Algorithm 1 and Appendix J. No new dataset or model checkpoint is released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper involves no human subjects and requires no IRB approval.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology. CLIP's text encoder is used as a frozen feature extractor, which is standard in the prompt learning literature.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.