UNDERSTANDING PROMPT LEARNING AS LOW-RANK MANIFOLD ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models like CLIP achieve strong zero-shot performance, yet prompt learning methods such as CoOp that adapt these models remain poorly understood: why can optimizing a handful of continuous vectors in a high-dimensional embedding space work even with only a few training examples? We answer this question through low-rank cross-modal alignment. Under a linear shared-latent model with frozen linear encoders and conditionally zero-mean text noise outside the shared subspace, we show that the population contrastive objective admits an optimal prompt operator with the rank smaller than the shared latent dimension, for any positive temperature. For nonlinear representations, we establish a latent factorization of the centered cross-covariance operator, with a rank bound under low-dimensional conditional-mean structure and a spectral-tail bound when that structure is approximate. HSIC provides a one-way dependence guarantee for centered trace alignment, rather than an equivalent objective. We also derive a generalization bound for explicitly rank-constrained prompt classes. Synthetic diagnostics distinguish exact trace-objective recovery from the nonzero spectral tails of normalized InfoNCE, while real-CLIP measurements provide operator-level evidence of spectral concentration. Inspired by the shared-coordinate structure, LR-MaPLe couples visual and textual prompts through shared factorization codes. Across 10 benchmarks and three seeds, it achieves accuracy comparable to MaPLe with approximately $19\times$ fewer trainable parameters. The factorization width is an architectural choice, not an estimate of the latent dimension.

1 INTRODUCTION

Vision-language models such as CLIP (Radford et al., 2021) learn a joint image-text embedding space and transfer to downstream tasks via prompts. Context Optimization (CoOp) (Zhou et al., 2022b) replaces hand-crafted templates of prompt such as "a photo of a [class]" with learnable continuous vectors, where prompts can be learned from only a few training examples. Subsequent methods improve generalization via conditional prompts (Zhou et al., 2022a), gradient and embedding regularization (Zhu et al., 2023; Yao et al., 2023), self-regularization and feature decoupling (Khattak et al., 2023b; Zhang et al., 2024), or prompting both modalities in MaPLe (Khattak et al., 2023a). However, a fundamental question remains: *why does prompt learning work so well?*

CoOp optimizes prompts in a high-dimensional embedding space, typically with dimensionality equal to 512 or higher, yet can perform well with fewer than 16 labeled examples per class. Classical learning theory would suggest that such a severely overparameterized regime should lead to catastrophic overfitting. The base-to-new performance gap reveals a genuine generalization challenge. Information-theoretic analyses (Wu et al., 2023) and simplified head/prompt-tuning models (Wei et al., 2021) do not characterize the cross-modal alignment geometry underlying vision-language prompt learning. We aim to investigate what controls the effective dimensionality of prompt optimization without overfitting and how the learned prompt relates geometrically to the pre-trained representations.

In this work, we study a sufficient mechanism for effective prompt learning through *low-rank cross-modal alignment*. We model paired observations using a linear shared-latent model, with a shared variable $z \in \mathbb{R}^k$ and independent modality-specific noise. For a population contrastive objective with frozen linear encoders, we show that removing conditionally centered text nuisance cannot increase

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

the loss. Thus an optimal linear prompt operator exists with $\mathrm{rank}(P^\star) \le k$, independently of the ambient dimension and at any positive temperature (Proposition 3.1). This is an existence statement for a controlled population model, not a claim that every optimizer or an unconstrained empirical training trajectory has rank k.

For nonlinear representations, we analyze the *centered* cross-covariance operator. Conditional independence yields an exact factorization through the shared latent variable. When the centered conditional means lie in subspaces of dimension at most k, the operator has rank at most k; approximate structure instead controls its spectral tail (Theorem 3.4). This conditional-mean assumption is substantive: a low-dimensional latent input alone does not imply a low-rank nonlinear covariance operator. The Hilbert–Schmidt Independence Criterion (HSIC) supplies a one-way dependence certificate for centered trace alignment, not an equivalent optimization objective or a kernel-CCA identification.

The experiments distinguish the objects and objectives involved. A closed-form linear trace diagnostic tests exact rank recovery, whereas normalized InfoNCE exhibits nonzero spectral tails and is not claimed to recover a rank-k prompt. On real CLIP features, OxfordPets cross-covariance measurements provide quantitative evidence of spectral concentration and increased CKA, without estimating the population latent dimension. Inspired by the shared-coordinate structure, we instantiate LR-MaPLe, which factorizes modality-specific prompts through shared codes. This layerwise transformer design is motivated by, rather than proved optimal by, the theory. In an 11-dataset, three-seed comparison, LR-MaPLe achieves Macro HM 77.69 ± 0.44 versus 77.27 ± 0.54 for MaPLe, with 184K rather than 3.55M trainable parameters. The evidence supports comparable accuracy with substantially fewer parameters, not a statistically significant accuracy gain.

Summary of contributions. (i) We establish the existence of a rank-at-most-k population optimizer for the linear prompt surrogate under conditional centering (Proposition 3.1), and characterize centered nonlinear cross-covariance rank and spectral tails under explicit conditional-mean assumptions (Theorem 3.4). (ii) For an explicitly rank-constrained class, we obtain a Rademacher-based generalization rate of order $O(\sqrt{kd_y/N})$, versus $O(d_y/\sqrt{N})$ for the analogous unconstrained-class bound (Proposition 3.8); this does not by itself characterize unconstrained empirical optimization. (iii) We present LR-MaPLe as a parameter-efficient, theory-inspired instantiation of shared-coordinate prompting. Its factorization width r is distinct from both the latent dimension k and the numerical rank of the token matrices. We make no priority claim for the factorization, which overlaps algebraically with MMLoP (Ghiasvand et al., 2026). (iv) We report synthetic objective-specific diagnostics, quantitative real-CLIP covariance measurements, width and active-rank sensitivity studies, and an 11-dataset evaluation with uncertainty estimates. These results support MaPLe-level accuracy with approximately $19\times$ fewer trainable parameters, rather than state-of-the-art or Pareto-frontier performance.

2 Related Work

Prompt Learning for Vision-Language Models. Building on CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), text-side methods replace hand-crafted templates with continuous tokens (CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a)) and regularize against zero-shot drift (ProGrad (Zhu et al., 2023), KgCoOp (Yao et al., 2023)). VPT (Jia et al., 2022) extends prompting to the vision branch, while MaPLe (Khattak et al., 2023a) couples modalities at multiple layers. PromptSRC (Khattak et al., 2023b) combines independent visual and textual prompts with self-regulating consistency, Gaussian prompt aggregation, and textual-diversity regularization. PromptKD (Li et al., 2024) uses a larger adapted teacher and unlabeled target-domain images; it therefore has a different supervision and resource budget from labeled 16-shot adaptation. These are important full-system comparators, although neither isolates the shared-factor parameterization studied here.

Recent complementary mechanisms include LLM-generated interpretable prompts in IPO (Du et al., 2024), control of feature-gradient propagation in Skip Tuning (Wu et al., 2025), modality-agnostic representation tokens in MMRL (Guo & Gu, 2025), dual-prompt collaboration in DPC Li et al. (2025), bidirectional hierarchical knowledge transfer in HiCroPL (Zheng et al., 2025), and sample-specific variational prompts in VaMP (Cheng & Han, 2025). These methods use complementary adaptation mechanisms. Our focused novelty claim is a conditional characterization of low-rank

shared-latent cross-modal alignment and a parameter-efficient, theory-inspired instantiation—not priority for factorized prompting or state-of-the-art accuracy.

Low-Rank Prompting and Parameter-Efficient Adaptation. MMLoP (Ghiasvand et al., 2026) is the closest overlapping mechanism: its shared up-projection and LR-MaPLe's shared latent factor belong to the same algebraic family. We do not claim that factorized multimodal prompting is new. Our contribution is the shared-latent theoretical characterization and a minimal implementation used to study this mechanism. Full MMLoP additionally combines self-regulating consistency loss (SCL), uniform drift correction (UDC), a zero-shot template ensemble, and a different training configuration. We distinguish that system from our CE-only shared-factor control. CLIP-LoRA (Zanella & Ben Ayed, 2024) adapts model weights through low-rank updates rather than inserted prompt tokens. Adapter methods, including MMA (Yang et al., 2024), CLIP-Adapter (Gao et al., 2024), Tip-Adapter (Zhang et al., 2022), and TaskRes (Yu et al., 2023), provide further parameter-efficient alternatives. The MaPLe compression result is not a claim of minimum parameter count among these methods.

Theoretical Understanding of Prompt and Parameter-Efficient Tuning. Wei et al. (2021) analyzed head and prompt tuning under latent-variable language models, Petrov et al. (2023) characterized expressivity limits of soft prompts, and InfoPrompt (Wu et al., 2023) adopted an information-theoretic view. Kernel views of fine-tuning (Malladi et al., 2023), attention-based prompt dynamics (Oymak et al., 2023), and linear multimodal contrastive models (Nakada et al., 2023) illustrate the use of controlled abstractions. Our linear operator likewise isolates a sufficient geometric mechanism, rather than claiming an exact linearization of transformer token prompting. Contrastive pre-training has also been connected to spectral structure (HaoChen et al., 2021) and multi-view information (Tsai et al., 2020; Tian et al., 2020). Analyses of low-rank adaptation (Zeng & Lee, 2023) and multi-modal multi-index learning (Gui et al., 2025) are closely related. In contrast to learning encoders from scratch, we keep the encoders frozen. Our linear result shows that a low-rank population optimum exists without imposing a low-rank parameterization; it does not establish rank collapse along an unconstrained optimization trajectory.

3 Prompt Learning as Cross-Modal Alignment

We analyze prompt learning under a controlled generative model in which both modalities arise from a shared low-dimensional latent. The main results show that, in the low-temperature regime of the contrastive loss, the optimal prompt aligns the two modalities along the latent subspace: in the linear case (§3.2) this corresponds to a population-level (non-whitened) canonical correlation problem and yields a rank bound determined by the latent dimension; in the nonlinear case (§3.3) the same rank bound persists for the cross-covariance operator in RKHS, and §3.4 converts the rank bound into a Rademacher-based sample-complexity guarantee.

3.1 Setup: Linear Shared-Latent Model

Data model. We adopt a *linear shared-latent model* for paired image–text data: a latent semantic factor $z \in \mathbb{R}^k$ generates the two modalities via

$$x = A_x z + \xi_x, \qquad y = A_y z + \xi_y, \tag{1}$$

with loading matrices $A_x \in \mathbb{R}^{d_x \times k}$, $A_y \in \mathbb{R}^{d_y \times k}$, and zero-mean noise ξ_x, ξ_y independent of z and of each other. We assume finite second moments, $\mathbb{E}[z] = 0$, and $\Sigma_z := \mathbb{E}[zz^\top] \succ 0$. Equation (1) is the two-view linear factor model underlying probabilistic CCA (Bach & Jordan, 2005) and is a linear instance of the multi-modal multi-index framework studied in (Gui et al., 2025). Here, k describes the shared latent dimension, rather than the intrinsic dimension of the noisy joint distribution. The linearity is restrictive but already exposes the essential alignment geometry; §3.3 extends the analysis to nonlinear encoders and conditional mean representations.

Frozen encoders. The pre-trained image and text encoders are modeled as fixed linear maps

$$f(x) = W_x x \in \mathbb{R}^d, \qquad g(y) = W_y y \in \mathbb{R}^d, \tag{2}$$

with $W_x \in \mathbb{R}^{d \times d_x}$ and $W_y \in \mathbb{R}^{d \times d_y}$ held fixed during prompt learning. We analyze unnormalized dot-product similarities and impose an explicit Frobenius bound $\|P\|_F \le B$ on the prompt operator.

This constraint is not equivalent to feature normalization; cosine-normalized similarities require a separate nonlinear analysis.

Prompt parameterization. We abstract the soft prompt as a learnable linear operator $P \in \mathbb{R}^{d_y \times d_y}$ acting on the text representation:

$$g(y; P) := W_y(Py). \tag{3}$$

This is a linear surrogate for CoOp's continuous-token prompt that preserves a learnable transformation in the text branch with the encoder kept frozen, while admitting a tractable analysis. It is motivated by linearized views of adaptation (Malladi et al., 2023), but is not an exact parameterization of continuous-token prompts. The nonlinear extension is treated in §3.3.

Contrastive objective. We study a population InfoNCE surrogate for CoOp. For $C \geq 2$ independent image–text pairs $\{(x_i, y_i)\}_{i=1}^{C}$ drawn from (1), define

$$\mathcal{L}_C(P) = \mathbb{E}\left[-\frac{1}{C}\sum_{i=1}^{C} \log \frac{\exp\big(s_{i,i}(P)/\tau\big)}{\sum_{j=1}^{C}\exp\big(s_{i,j}(P)/\tau\big)}\right], \quad s_{i,j}(P) := f(x_i)^\top g(y_j; P), \tag{4}$$

where $\tau > 0$ and the expectation is over the sampled pairs. The optimization is $P^\star \in \arg\min_{\|P\|_F \leq B} \mathcal{L}_C(P)$. Here, C denotes the number of sampled candidates. This population objective differs from empirical CoOp training with fixed class text prototypes.

3.2 Linear Case: Cross-Modal Covariance Maximization

The following result is the central observation of this section: under a conditional centering assumption, projecting the prompt operator onto the shared text subspace cannot increase the population contrastive loss. Consequently, an optimal prompt operator exists whose rank is bounded by the shared latent dimension.

Proposition 3.1 (Low-Rank Optimality for Linear Encoders)**.** *Adopt the linear shared-latent model* (1) *with frozen linear encoders and prompt parameterization* (3)*. Let* Π *be the orthogonal projector onto* $\mathrm{range}(A_y)$*, and assume*

$$\mathbb{E}[y \mid x, \Pi y] = \Pi y. \tag{5}$$

Then, for every $\tau > 0$*, every* $C \geq 2$*, and every prompt operator* P*,*

$$\mathcal{L}_C(P\Pi) \leq \mathcal{L}_C(P), \qquad \|P\Pi\|_F \leq \|P\|_F. \tag{6}$$

In particular, there exists a minimizer $P^\star$ *over* $\|P\|_F \leq B$ *satisfying*

$$P^\star = P^\star\Pi, \qquad \mathrm{rank}(P^\star) \leq \mathrm{rank}(\Pi) = \mathrm{rank}(A_y) \leq k, \tag{7}$$

independently of the ambient dimension d_y*.*

Remark 3.2 (Geometric interpretation: shared-subspace alignment). The centered cross-covariance of the encoded modalities is

$$K_P := \mathrm{Cov}\big(f(x), g(y; P)\big) = W_x A_x \Sigma_z A_y^\top P^\top W_y^\top, \tag{8}$$

and hence $\mathrm{rank}(K_P) \leq k$. This identifies a low-dimensional cross-modal signal, but does not make the contrastive objective equivalent to trace maximization: the full loss also depends on the distribution of the logits. Proposition 3.1 instead uses convexity to show that removing conditionally centered text nuisance cannot increase the loss. No low-temperature approximation or whitening is required.

The rank bound (7) establishes the existence of a low-rank population optimizer; it does not assert that every optimizer has rank k, or that empirical training automatically produces a low-rank solution. An explicitly rank-constrained prompt class has $O(kd_y)$ rather than $O(d_y^2)$ degrees of freedom, as considered in §3.4. Transferring this capacity reduction to unconstrained empirical training requires additional analysis.

Remark 3.3 (When does conditional centering hold?). Condition (5) holds, for example, when $\xi_y \sim \mathcal{N}(0, \sigma_y^2 I)$ is independent of (z, ξ_x). Indeed,

$$y = \Pi y + \eta, \qquad \eta = (I - \Pi)\xi_y,$$

where η has zero mean and is independent of $(x, \Pi y)$. The latent variable z need not be Gaussian. More generally, only the conditional mean restriction $\mathbb{E}[\eta \mid x, \Pi y] = 0$ is needed. Unconditional noise centering and independence from the latent variable alone do not guarantee this restriction.

3.3 NONLINEAR EXTENSION: CROSS-COVARIANCE RANK BOUND IN RKHS

We now extend the cross-covariance analysis to nonlinear encoders, replacing W_x, W_y with measurable feature maps $\phi_x : \mathbb{R}^{d_x} \to \mathcal{H}$ and $\phi_y : \mathbb{R}^{d_y} \to \mathcal{H}$ into a shared separable real reproducing kernel Hilbert space $\mathcal{H}$ (an abstraction of CLIP's shared $\mathbb{R}^d$ embedding space). Assume

$$\mathbb{E}\|\phi_x(x)\|_{\mathcal{H}}^2 < \infty, \qquad \mathbb{E}\|\phi_y(Py)\|_{\mathcal{H}}^2 < \infty$$

for every feasible P. Let $\mu_x := \mathbb{E}[\phi_x(x)]$ and $\mu_y(P) := \mathbb{E}[\phi_y(Py)]$, and define the centered cross-covariance operator

$$C_{xy}(P) := \mathbb{E}\big[(\phi_x(x) - \mu_x) \otimes (\phi_y(Py) - \mu_y(P))\big]. \tag{9}$$

The next theorem extends the linear cross-covariance characterization to nonlinear features: under a low-dimensional conditional-mean assumption, the rank of $C_{xy}(P)$ is bounded by k for *every feasible* P, by virtue of the latent factorization.

Theorem 3.4 (Kernel Rank Bound)**.** *Under the data model* (1)*, with mutually independent z, ξ_x, ξ_y and nonlinear encoders $f(x) = \phi_x(x)$, $g(y; P) = \phi_y(Py)$, the operator $C_{xy}(P)$ is trace-class, and the alignment functional satisfies the identity*

$$\mathcal{A}(P) := \mathbb{E}\big[\langle \phi_x(x),\, \phi_y(Py)\rangle_{\mathcal{H}}\big] = \langle \mu_x, \mu_y(P)\rangle_{\mathcal{H}} + \operatorname{tr}\big(C_{xy}(P)\big). \tag{10}$$

Moreover, the conditional independence $x \perp y \mid z$ implied by the data model yields the latent factorization

$$\begin{aligned} C_{xy}(P) &= \mathbb{E}_z\big[(m_x(z) - \mu_x) \otimes (m_y(z; P) - \mu_y(P))\big], \\ m_x(z) &:= \mathbb{E}[\phi_x(x) \mid z], \qquad m_y(z; P) := \mathbb{E}[\phi_y(Py) \mid z]. \end{aligned} \tag{11}$$

If we further assume

(CM) Low-dimensional conditional means: *there exists a linear subspace $V_x \subseteq \mathcal{H}$ of dimension at most k such that $m_x(z) - \mu_x \in V_x$ for P_z-almost every z, and for each feasible P there exists a linear subspace $V_y(P) \subseteq \mathcal{H}$ of dimension at most k such that $m_y(z; P) - \mu_y(P) \in V_y(P)$ for P_z-almost every z,*

then $C_{xy}(P)$ has rank at most k for every feasible P.

Equation (10) is an algebraic decomposition, not an equivalence between contrastive-loss minimization and alignment maximization. The rank bound concerns $C_{xy}(P)$, rather than the prompt operator P itself.

Remark 3.5 (On (CM): when does the conditional-mean assumption hold?). Assumption (CM) is automatic in the linear case of Proposition 3.1: with $\phi_x = W_x$ and $\phi_y = W_y$, the conditional means are $m_x(z) = W_x A_x z + W_x \mathbb{E}[\xi_x \mid z] = W_x A_x z$ and $m_y(z; P) = W_y P A_y z$, lying in the subspaces $V_x = \operatorname{range}(W_x A_x)$ and $V_y(P) = \operatorname{range}(W_y P A_y)$, respectively, each of dimension at most k. For nonlinear encoders, (CM) is genuinely restrictive: it requires the centered conditional means to lie in low-dimensional linear subspaces, rather than merely to be functions of a k-dimensional latent variable. Approximate affine behavior of the encoders over the relevant data region may motivate an approximate version of (CM), but small observation noise alone does not guarantee it. Without (CM), the operator rank of $C_{xy}(P)$ can exceed k, despite the conditional independence of the modalities given z.

Remark 3.6 (Connection to HSIC). For the kernels induced by the feature maps, the Hilbert–Schmidt Independence Criterion (Gretton et al., 2005) is $\mathrm{HSIC}(X, Y(P)) = \|C_{xy}(P)\|_{\mathrm{HS}}^2$. By the trace-norm bound, Cauchy–Schwarz, and Theorem 3.4,

$$\big|\operatorname{tr}\big(C_{xy}(P)\big)\big| \;\le\; \sqrt{\operatorname{rank}(C_{xy}(P))}\, \|C_{xy}(P)\|_{\mathrm{HS}} \;\le\; \sqrt{k}\, \sqrt{\mathrm{HSIC}(X, Y(P))}.$$

Thus nonzero centered trace alignment certifies positive HSIC and hence dependence between the modalities. This is a one-way relation: positive HSIC does not imply nonzero trace alignment. It does not make either quantity an equivalent contrastive objective or justify identifying prompt learning with kernel CCA.

Remark 3.7 (Empirical implication). When (CM) holds, Theorem 3.4 implies that the singular values of the population cross-covariance vanish beyond index k. An empirical cross-covariance can nevertheless have nonzero tail singular values due to sampling error; a visible gap at index k additionally requires a non-negligible kth population singular value relative to the estimation error. When (CM) holds only approximately, a quantitative soft-cutoff claim requires control of the conditional-mean approximation error. This motivates examining cross-covariance spectra in the synthetic and real-CLIP settings of §4.3 and §F.1. Spectra of the prompt operator alone do not directly test this theorem, and empirical spectral concentration does not by itself establish (CM).

3.4 Sample Complexity from the Rank Bound

The rank bound of Proposition 3.1 converts directly into a generalization guarantee that scales with the latent dimension rather than the ambient one. Let $\mathcal{H}_{B,k} := \{P \in \mathbb{R}^{d_y \times d_y} : \mathrm{rank}(P) \le k,\ \|P\|_F \le B\}$ be the effective hypothesis class identified by Proposition 3.1, and let $\widehat{\mathcal{L}}_N$ denote the empirical loss on N pairs.

Proposition 3.8 (Generalization bound from low rank)**.** *Assume the contrastive loss is L_ℓ-Lipschitz in P on the Frobenius ball of radius B (which holds whenever the encoders and inputs are norm-bounded). Then with probability at least $1-\delta$ over the draw of N i.i.d. pairs, every $P \in \mathcal{H}_{B,k}$ satisfies*

$$\mathcal{L}_{\mathrm{CoOp}}(P) \;-\; \widehat{\mathcal{L}}_N(P) \;\le\; c_1\, L_\ell\, B \sqrt{\frac{k d_y}{N}} \;+\; c_2 \sqrt{\frac{\log(1/\delta)}{N}}, \tag{12}$$

for absolute constants $c_1, c_2 > 0$.

Remark 3.9 (Comparison with the unconstrained class). For the unconstrained class $\mathcal{H}_B = \{P : \|P\|_F \le B\}$, the analogous bound replaces $k d_y$ with d_y^2, giving an excess risk of order $L_\ell B \sqrt{d_y^2/N}$. The improvement factor is $\sqrt{k d_y / d_y^2} = \sqrt{k/d_y}$, which for CLIP ($d_y = 512$) and a typical intrinsic dimension on the order of $k = 10$–30 predicts roughly a 4–7$\times$ reduction in required samples—broadly consistent with the empirical observation that CoOp performs strongly with 1–16 shots per class. Proposition 3.8 thus converts the geometric rank bound into a quantitative answer to the few-shot puzzle posed in the introduction.

4 Experiments

Our experiments connect the proposed mechanism to three questions: can shared-factor prompting preserve few-shot transfer with fewer parameters; do synthetic and real representations exhibit the predicted alignment structure; and how do width and active rank affect the design? We retain the original per-dataset comparison and add the completed benchmark, uncertainty, covariance, and sensitivity studies.

4.1 Shared-Coordinate Prompting: LR-MaPLe

We implement the shared-coordinate principle with

$$P_{\mathrm{text}}^{(\ell)} = Z_\ell (U_{\mathrm{text}}^{(\ell)})^\top, \qquad P_{\mathrm{vis}}^{(\ell)} = Z_\ell (U_{\mathrm{vis}}^{(\ell)})^\top, \tag{13}$$

where $Z_\ell \in \mathbb{R}^{n_{\mathrm{ctx}} \times r}$ is shared and the bases are modality-specific. These factors use $r(n_{\mathrm{ctx}} + d_{\mathrm{text}} + d_{\mathrm{vis}})$ parameters per layer. Sharing coordinates is more specific than independent bottlenecks; its use at every transformer layer remains an architectural choice. The standard configuration uses $n_{\mathrm{ctx}} = 2$, $r = 16$, and nine prompted layers, so

$$\mathrm{rank}(P_{\mathrm{mod}}^{(\ell)}) \le \min(n_{\mathrm{ctx}}, r) = 2. \tag{14}$$

Thus r is a factorization width, not prompt rank or an estimate of k. For $r \ge 2$, it changes the parameterization without changing the rank ceiling. Unfactorized two-token MaPLe prompts also have rank at most two: the savings concern the shared parameterization, not a reduction from high-rank to low-rank two-row matrices. LR-MaPLe trains with cross-entropy for five epochs, keeping CLIP frozen. Algorithm 1 summarizes training; initialization notes appear in Appendix D.3.

Algorithm 1 LR-MaPLe: Low-Rank Multi-modal Prompt Learning

Require: Pre-trained CLIP (frozen), class names $\mathcal{C}$, latent rank r (default $r = 16$), prompt depth L, training data $\mathcal{D}$
Ensure: Learned low-rank prompt parameters
Initialize shared latent codes:
  Shallow: $Z_s \in \mathbb{R}^{n_{\text{ctx}} \times r}$ via SVD of word embeddings
  Deep: $Z_i \in \mathbb{R}^{n_{\text{ctx}} \times r}$, $i = 1, \ldots, L-1$, random init
Initialize modality-specific bases:
  Text bases: $U_{\text{text}}^{(i)} \in \mathbb{R}^{d_{\text{text}} \times r}$, $i = 0, \ldots, L-1$
  Vision bases: $U_{\text{vis}}^{(i)} \in \mathbb{R}^{d_{\text{vis}} \times r}$, $i = 0, \ldots, L-1$
for each training iteration **do**
  Generate prompts from shared latent codes:
    $P_{\text{text}}^{(i)} = Z_i \, U_{\text{text}}^{(i)\top} \in \mathbb{R}^{n_{\text{ctx}} \times d_{\text{text}}}$ ▷ rank $\leq r$
    $P_{\text{vis}}^{(i)} = Z_i \, U_{\text{vis}}^{(i)\top} \in \mathbb{R}^{n_{\text{ctx}} \times d_{\text{vis}}}$ ▷ rank $\leq r$
  Forward pass: inject $\{P_{\text{text}}^{(i)}\}$, $\{P_{\text{vis}}^{(i)}\}$ into CLIP transformer layers
  Compute loss: $\mathcal{L} = \mathcal{L}_{\text{CE}}(\mathbf{f}_{\text{img}}, \mathbf{f}_{\text{text}})$
  Update $\{Z_i, U_{\text{text}}^{(i)}, U_{\text{vis}}^{(i)}\}$ via SGD (CLIP frozen)
end for

Table 1: **Original base-to-new comparison on 10 datasets.** ViT-B/16, 16-shot; the original CoCoOp, MaPLe, and LR-MaPLe results are retained. LR-MaPLe uses 184K versus MaPLe's 3.55M trainable parameters. The expanded three-seed results are reported separately in Table 2.

	CoCoOp			**MaPLe**			**LR-MaPLe (Ours)**		
Dataset	Base	New	HM	Base	New	HM	Base	New	HM
ImageNet	75.98	70.43	73.10	70.73	74.00	72.33	76.73	70.73	**73.61**
Caltech101	97.96	93.81	95.84	97.74	94.36	96.02	97.87	95.17	**96.50**
OxfordPets	95.20	97.69	96.43	95.43	97.76	**96.58**	95.50	97.67	96.57
StanfordCars	70.49	73.59	72.01	72.94	74.00	**73.47**	72.83	74.10	73.46
OxfordFlowers	94.87	71.75	81.71	95.92	72.46	82.56	96.37	73.77	**83.57**
Food101	90.70	91.29	90.99	90.71	92.05	**91.38**	90.50	91.87	91.18
FGVCAircraft	33.41	23.71	27.74	36.72	35.61	36.15	36.50	36.67	**36.58**
DTD	77.01	56.00	64.85	80.36	59.18	**68.16**	76.50	60.73	67.71
EuroSAT	87.49	60.04	71.21	94.07	73.23	**82.35**	91.57	74.50	82.16
UCF101	82.33	73.45	77.64	83.00	78.66	80.77	85.30	77.07	**80.97**
Average	80.54	71.18	75.15	81.76	75.13	77.98	81.97	75.23	**78.23**

4.2 Base-to-New Generalization and Parameter Efficiency

Protocol and original comparison. We use CLIP ViT-B/16 and 16 labeled images per base class. Base/New denote accuracy on seen/unseen classes, and HM their harmonic mean. Table 1 retains the original 10-dataset comparison with CoCoOp and MaPLe, including every original per-dataset score. The expanded evaluation in Table 2 uses all 11 standard datasets, including SUN397, and seeds 1/2/3. These are separate result sets: the original rows are not combined with the updated aggregates. For the expanded study, dataset HM is computed after averaging Base/New accuracies over seeds; Macro HM averages these dataset HMs. Error bars are sample SDs of the three seed-level macro aggregates (Appendix D).

Expanded comparison. In addition to MaPLe, we evaluate independent vision–language prompting (IVLP) and a shared-factor MMLoP CE-only control. The latter retains rank-one prompts and shared up-projection but disables SCL, UDC, and layer-0 text-anchor initialization. It is not the full MMLoP system. LR-MaPLe obtains 77.69 ± 0.44 Macro HM versus MaPLe's 77.27 ± 0.54, with 184K rather

Table 2: **Expanded 11-dataset, three-seed evaluation** (ViT-B/16, 16-shot). Reported central values ± sample SDs; aggregation details in Appendix D. MMLoP is the shared-factor CE-only control.

Method	Macro Base	Macro New	Macro HM	SUN397 HM
LR-MaPLe	81.84 ± 0.37	74.52 ± 0.64	77.69 ± 0.44	79.37 ± 0.30
MaPLe	81.91 ± 0.21	73.78 ± 0.78	77.27 ± 0.54	79.44 ± 0.26
IVLP	81.76 ± 0.18	73.32 ± 1.49	76.97 ± 1.00	79.73 ± 0.51
MMLoP (CE-only)	83.82 ± 0.30	71.72 ± 1.07	76.94 ± 0.80	77.99 ± 0.63

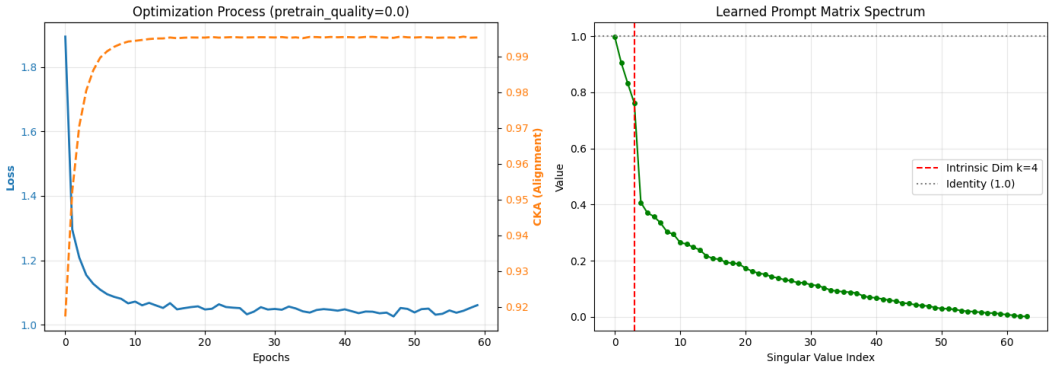

Figure 1: **Original normalized-InfoNCE diagnostic.** Loss decreases and CKA increases (left); P has a nonzero spectral tail (right), not an exact cutoff at $k = 4$ or the spectrum of $\Delta P = P - P_0$. Exact trace recovery is separate (Appendix C); this normalized experiment does not verify Proposition 3.1.

than 3.55M parameters, a 94.8% reduction. On SUN397, its HM is within 0.07 points of MaPLe; the complete Base/New/HM results appear in Appendix D.

Uncertainty and comparison scope. Paired dataset-level HM differences are $+0.43$ with 95% Student-t CI $[-0.35, 1.21]$ versus MaPLe and $+0.72$ with CI $[-0.28, 1.72]$ versus IVLP. Both include zero: the evidence supports similar observed accuracy, not a significant gain or a formal equivalence test. The aggregate gap over the MMLoP CE-only control is $+0.75$, without a significance claim. Full MMLoP reports 79.70 HM with 11.5K parameters, but adds consistency losses, drift correction, a template ensemble, and a different training budget. Our result tests MaPLe compression; it does not establish a Pareto frontier or universal superiority. Appendix D separates these controls from richer adaptation systems and teacher-plus-unlabeled-data methods.

4.3 Synthetic Diagnostics: Alignment and Objective-Specific Rank

Exact trace-objective recovery. For an exact-rank matrix M with $k \in \{2, 4, 8, 16\}$, the Frobenius-constrained trace problem has $P^\star_{\mathrm{tr}} = BM/\|M\|_F$. Across five seeds, rank caps at least k recover this optimizer to the reported precision: relative Frobenius error 0.0000, matrix cosine 1.0000, and trace ratio 1.0000; smaller caps give nonzero error. This tests the exact trace geometry, not the all-temperature population loss of Proposition 3.1. The recovered object is P; $\Delta P = P - P_0$ can have a different spectrum and rank. Appendix C defines the diagnostic and its metrics.

Normalized InfoNCE. The original visualization uses $d_x = d_y = d = 64$, $k = 4$, 1024 pairs, Gaussian latent variables and loading matrices, noise variance 0.1, orthogonal encoders, and $P_0 = I$. Training uses normalized symmetric InfoNCE for 60 epochs with Adam, learning rate 5×10^{-3}, and $\tau = 0.07$. Figure 1 shows decreasing loss, CKA Kornblith et al. (2019) increasing from approximately 0.92 to 0.99, and a nonzero prompt spectral tail. This supports an alignment interpretation, not exact rank recovery. The additional five-seed temperature study (Table 3) gives full-model effective-rank statistics near 50–51, not $k = 4$; lower full-model loss is consistent with the nested feasible classes. The reported statistic is not identified with covariance energy rank.

Table 3: **Normalized symmetric InfoNCE, $k = 4$, five seeds.** Uncertainties are as reported. The full-prompt effective-rank statistic is not a covariance rank or an estimate of k.

τ	Rank-4 loss	Full-model loss	Reported effective rank
0.03	6.2879 ± 0.4682	2.6304 ± 0.0571	50.40 ± 0.43
0.07	5.2812 ± 0.4046	3.3102 ± 0.0106	51.19 ± 0.30
0.20	5.5552 ± 0.0173	4.9462 ± 0.0049	50.03 ± 0.31

Table 4: **Factorization-width sweep, $n_{\mathrm{ctx}} = 2$.** Three-seed mean HM (%; uncertainties unreported). Rank denotes an upper bound, not a measured value.

r	Rank bound	ImageNet	DTD	EuroSAT	FGVCAircraft
1	1	73.60	66.60	77.11	28.25
2	2	73.43	65.65	77.57	36.54
4	2	73.59	65.27	77.30	35.61
8	2	73.61	64.07	80.86	34.26
16	2	73.45	66.71	77.92	34.80
32	2	73.85	67.25	81.50	34.86

4.4 Factorization Width versus Active Rank

Table 4 varies r under the standard $n_{\mathrm{ctx}} = 2$ configuration. Only $r = 1$ tightens the rank ceiling; for $r \geq 2$, this is a parameterization/optimization study, not an active-rank comparison. No width is uniformly best across ImageNet, DTD, EuroSAT, and FGVCAircraft.

To make the rank cap active, Table 5 uses $n_{\mathrm{ctx}} = 16$ and a matched unconstrained MaPLe control. The highest LR-MaPLe mean occurs at $r = 1$ on DTD and $r = 2$ on EuroSAT. MaPLe is weaker on DTD but stronger on EuroSAT, and several settings have large seed variation. Small active ranks can therefore suffice, but neither a universal optimum nor recovery of k is demonstrated. Together, the two sweeps support the shared-coordinate design while keeping numerical width selection an empirical architectural choice.

Table 5: **Active-rank diagnostic with $n_{\mathrm{ctx}} = 16$.** Three-seed HM results (mean $\pm$ sample SD, percent). The MaPLe control uses the same context length without the shared low-rank factorization.

Setting	DTD	EuroSAT
LR-MaPLe, $r = 1$	65.46 ± 2.50	74.32 ± 2.30
LR-MaPLe, $r = 2$	63.08 ± 1.39	76.48 ± 1.35
LR-MaPLe, $r = 4$	65.36 ± 0.68	69.68 ± 13.94
LR-MaPLe, $r = 8$	60.05 ± 4.11	72.14 ± 8.29
LR-MaPLe, $r = 16$	63.88 ± 0.55	72.32 ± 19.22
Matched MaPLe	60.76 ± 1.72	81.54 ± 4.40

5 Conclusion

Shared low-dimensional structure provides a sufficient mechanism for prompt adaptation: conditional centering yields a low-rank linear population optimum, and conditional-mean geometry controls nonlinear centered covariance rank and spectral tails. The associated constrained capacity bound motivates parameter-efficient learning. LR-MaPLe tests the shared-coordinate principle in a practical transformer architecture. The original benchmark comparison and added synthetic, covariance, sensitivity, and three-seed studies together support MaPLe-level accuracy with approximately $19\times$ fewer trainable parameters.

AI Use Statement

We used generative AI tools to assist with language editing, manuscript organization, formatting, and notation; to identify and summarize related literature; and to draft and revise text under the

authors' direction. These tools also assisted in exploring theoretical formulations and methodological alternatives, deriving inequalities, and drafting portions of proofs. In addition, we used AI-generated feedback to refine the presentation and interpretation of experimental results. The authors made the final decisions on the theoretical framework, methods, and conclusions; checked the cited sources and mathematical arguments; and reviewed and revised the AI-assisted material. The authors take full responsibility for the content, accuracy, and claims of this paper.

REFERENCES

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005. URL https://api.semanticscholar.org/CorpusID:17380611.

Silin Cheng and Kai Han. VaMP: Variational multi-modal prompt learning for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yingjun Du, Wenfang Sun, and Cees G. M. Snoek. Ipo: interpretable prompt optimization for vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024. ISBN 9798331314385.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Sajjad Ghiasvand, Haniyeh Ehsani Oskouie, Mahnoosh Alizadeh, and Ramtin Pedarsani. Mmlop: Multi-modal low-rank prompting for efficient vision-language adaptation. In *European conference on computer vision*. Springer, 2026.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. *arXiv preprint arXiv:2505.12473*, 2025.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25015–25025. IEEE, 2025.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15190–15200, 2023b.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25623–25632. IEEE, 2025.

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

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 4348–4380. PMLR, 25–27 Apr 2023.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pp. 26724–26768. PMLR, 2023.

Aleksandar Petrov, Philip HS Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. *arXiv preprint arXiv:2310.19698*, 2023.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34:16158–16170, 2021.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36:61060–61084, 2023.

Shihan Wu, Ji Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14723–14732. IEEE, 2025.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23826–23837, June 2024.

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

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10899–10909, 2023.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1593–1603, June 2024.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. Hierarchical cross-modal prompt learning for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1891–1901. IEEE, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

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

A Limitations and future work.

The linear prompt is a controlled abstraction of token prompting; its population theorem does not characterize normalized InfoNCE or unconstrained empirical optimization. The nonlinear result requires conditional structure and does not prove layerwise prompt optimality. Real-CLIP spectral evidence covers one dataset with a class-count ceiling; it neither estimates k nor selects r. The benchmark intervals are not equivalence tests, and broader matched comparisons remain necessary. Analyzing nonlinear prompt dynamics and deriving data-dependent width selection are natural next steps.

B Technical appendices

Proof of Proposition 3.1. Let Π be the orthogonal projector onto $\mathrm{range}(A_y)$, and define the score matrix

$$S(P) := \big(s_{i,j}(P)\big)_{i,j=1}^{C}, \qquad s_{i,j}(P) = f(x_i)^\top W_y P y_j.$$

For a score matrix $S \in \mathbb{R}^{C\times C}$, write

$$\Phi_\tau(S) := \frac{1}{C}\sum_{i=1}^{C}\left[\log\left(\sum_{j=1}^{C}\exp(S_{i,j}/\tau)\right) - \frac{S_{i,i}}{\tau}\right].$$

Then $\mathcal{L}_C(P) = \mathbb{E}[\Phi_\tau(S(P))]$. For every $\tau > 0$, the function Φ_τ is convex in S, since each summand is a log-sum-exp function minus a linear term.

Consider the sigma-algebra

$$\mathcal{F} := \sigma\big((x_i, \Pi y_i)_{i=1}^{C}\big).$$

Independence of the sampled pairs and the conditional centering assumption (5) imply, for each j,

$$\mathbb{E}[y_j \mid \mathcal{F}] = \mathbb{E}[y_j \mid x_j, \Pi y_j] = \Pi y_j.$$

Since $f(x_i)$ is $\mathcal{F}$-measurable, it follows that

$$\mathbb{E}[s_{i,j}(P) \mid \mathcal{F}] = f(x_i)^\top W_y P \Pi y_j = s_{i,j}(P\Pi),$$

or, equivalently,

$$\mathbb{E}[S(P) \mid \mathcal{F}] = S(P\Pi).$$

Conditional Jensen's inequality therefore yields

$$\Phi_\tau(S(P\Pi)) = \Phi_\tau\big(\mathbb{E}[S(P) \mid \mathcal{F}]\big) \le \mathbb{E}[\Phi_\tau(S(P)) \mid \mathcal{F}].$$

Taking expectations establishes

$$\mathcal{L}_C(P\Pi) \le \mathcal{L}_C(P). \tag{15}$$

Moreover, orthogonality of Π gives

$$\|P\|_F^2 = \|P\Pi\|_F^2 + \|P(I-\Pi)\|_F^2.$$

Thus $\|P\Pi\|_F \le \|P\|_F$, so projection preserves feasibility. It also gives

$$\operatorname{rank}(P\Pi) \le \operatorname{rank}(\Pi) = \operatorname{rank}(A_y) \le k.$$

To verify existence of a minimizer, observe that, uniformly over $\|P\|_F \le B$,

$$0 \le \Phi_\tau(S(P)) \le \log C + \frac{2B\|W_y\|_{\mathrm{op}}}{\tau}\left(\max_{1\le i\le C}\|f(x_i)\|\right)\left(\max_{1\le j\le C}\|y_j\|\right).$$

The right-hand side is integrable by the finite second-moment assumptions and the Cauchy–Schwarz inequality. Since $\Phi_\tau(S(P))$ is continuous in P for each sampled batch, dominated convergence implies that $\mathcal{L}_C$ is continuous on the Frobenius ball. This ball is compact, so there exists

$$\widehat{P} \in \arg\min_{\|P\|_F\le B} \mathcal{L}_C(P).$$

Set $P^\star := \widehat{P}\Pi$. By feasibility and (15),

$$\mathcal{L}_C(P^\star) \le \mathcal{L}_C(\widehat{P}) = \min_{\|P\|_F\le B}\mathcal{L}_C(P),$$

so $P^\star$ is also a minimizer. Finally, $\Pi^2 = \Pi$ implies

$$P^\star = P^\star\Pi, \qquad \operatorname{rank}(P^\star) \le k,$$

which proves the claim. □

Proof of Theorem 3.4. Fix any feasible prompt operator P, and write

$$u := \phi_x(x) - \mu_x, \qquad v := \phi_y(Py) - \mu_y(P).$$

We use the convention $(a \otimes b)h = \langle b, h\rangle_{\mathcal{H}} a$ for $a, b, h \in \mathcal{H}$.

Step 1: Trace-class property and alignment identity. The assumed finite second moments imply

$$\mathbb{E}\|u\|_{\mathcal{H}}^2 < \infty, \qquad \mathbb{E}\|v\|_{\mathcal{H}}^2 < \infty.$$

For a rank-one operator, $\|u \otimes v\|_1 = \|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}$, where $\|\cdot\|_1$ denotes the trace norm. Hence, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\|u \otimes v\|_1 = \mathbb{E}\big[\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}\big] \le \big(\mathbb{E}\|u\|_{\mathcal{H}}^2\big)^{1/2}\big(\mathbb{E}\|v\|_{\mathcal{H}}^2\big)^{1/2} < \infty.$$

Thus $C_{xy}(P) = \mathbb{E}[u \otimes v]$ is well-defined as a trace-class operator, and is therefore also Hilbert–Schmidt.

Since the trace is continuous with respect to the trace norm and $\mathrm{tr}(u \otimes v) = \langle u, v\rangle_{\mathcal{H}}$,

$$\begin{aligned}\mathrm{tr}\big(C_{xy}(P)\big) &= \mathbb{E}\big[\mathrm{tr}(u \otimes v)\big] \\ &= \mathbb{E}\big[\langle u, v\rangle_{\mathcal{H}}\big] \\ &= \mathbb{E}\big[\langle \phi_x(x), \phi_y(Py)\rangle_{\mathcal{H}}\big] - \langle \mu_x, \mu_y(P)\rangle_{\mathcal{H}}.\end{aligned}$$

Rearranging gives the alignment identity (10).

Step 2: Latent factorization. Under the data model (1), the independence of z, ξ_x, ξ_y implies $x \perp y \mid z$. Consequently, for fixed P, the features $\phi_x(x)$ and $\phi_y(Py)$ are conditionally independent given z.

Define

$$a(z) := m_x(z) - \mu_x = \mathbb{E}[u \mid z], \qquad b(z) := m_y(z; P) - \mu_y(P) = \mathbb{E}[v \mid z].$$

Conditional Jensen's inequality gives

$$\mathbb{E}_z\|a(z)\|_{\mathcal{H}}^2 \le \mathbb{E}\|u\|_{\mathcal{H}}^2, \qquad \mathbb{E}_z\|b(z)\|_{\mathcal{H}}^2 \le \mathbb{E}\|v\|_{\mathcal{H}}^2.$$

In particular, $\mathbb{E}_z[a(z) \otimes b(z)]$ is well-defined as a trace-class operator.

For arbitrary $h_1, h_2 \in \mathcal{H}$, conditional independence gives

$$\begin{aligned}&\mathbb{E}\big[\langle u, h_1\rangle_{\mathcal{H}}\langle v, h_2\rangle_{\mathcal{H}} \mid z\big] \\ &\quad= \mathbb{E}\big[\langle u, h_1\rangle_{\mathcal{H}} \mid z\big]\mathbb{E}\big[\langle v, h_2\rangle_{\mathcal{H}} \mid z\big] \\ &\quad= \langle a(z), h_1\rangle_{\mathcal{H}}\langle b(z), h_2\rangle_{\mathcal{H}}.\end{aligned}$$

Taking expectations and using the tensor-product convention,

$$\begin{aligned}\langle C_{xy}(P)h_2, h_1\rangle_{\mathcal{H}} &= \mathbb{E}_z\big[\langle a(z), h_1\rangle_{\mathcal{H}}\langle b(z), h_2\rangle_{\mathcal{H}}\big] \\ &= \big\langle \big(\mathbb{E}_z[a(z) \otimes b(z)]\big)h_2, h_1\big\rangle_{\mathcal{H}}.\end{aligned}$$

Since this holds for all $h_1, h_2 \in \mathcal{H}$,

$$C_{xy}(P) = \mathbb{E}_z[a(z) \otimes b(z)] = \mathbb{E}_z\big[(m_x(z) - \mu_x) \otimes (m_y(z; P) - \mu_y(P))\big],$$

which proves (11).

Step 3: Rank bound under (CM). By (CM), $a(z) \in V_x$ and $b(z) \in V_y(P)$ for P_z-almost every z, where $\dim(V_x) \le k$ and $\dim(V_y(P)) \le k$. For any $h \in \mathcal{H}$, the latent factorization gives

$$C_{xy}(P)h = \mathbb{E}_z\big[\langle b(z), h\rangle_{\mathcal{H}}\, a(z)\big].$$

The integrand is integrable by the Cauchy–Schwarz inequality and belongs to V_x almost surely. Since V_x is a closed finite-dimensional subspace, its expectation also belongs to V_x. Therefore,

$$\mathrm{range}\big(C_{xy}(P)\big) \subseteq V_x.$$

Similarly,

$$C_{xy}(P)^*h = \mathbb{E}_z\big[\langle a(z), h\rangle_{\mathcal{H}}\, b(z)\big] \in V_y(P),$$

so

$$\mathrm{range}\big(C_{xy}(P)^*\big) \subseteq V_y(P).$$

It follows that

$$\mathrm{rank}\big(C_{xy}(P)\big) \le \min\{\dim(V_x), \dim(V_y(P))\} \le k.$$

Since P was arbitrary, the rank bound holds for every feasible prompt operator. □

Proof of Proposition 3.8. To make the covering-number argument transparent, we work with general matrix dimensions $P \in \mathbb{R}^{m\times n}$ and specialize to the paper's setting $m = n = d_y$ at the end.

Setup. Let $\mathcal{H}_{B,k} := \{P \in \mathbb{R}^{m\times n} : \operatorname{rank}(P) \le k,\ \|P\|_F \le B\}$ and let $\ell(P;(x,y))$ denote the per-sample contrastive loss in (4). We assume:

(B) (Boundedness) $0 \le \ell(P;(x,y)) \le M$ for all $P \in \mathcal{H}_{B,k}$ and all (x,y) in the support of the data distribution.

(L) (Lipschitz) $|\ell(P;(x,y)) - \ell(P';(x,y))| \le L_\ell \|P - P'\|_F$ uniformly in (x,y).

Both hold under the standard CLIP setup with ℓ_2-normalized features and bounded inputs; we verify this at the end of the proof.

Step 1: Symmetrization and McDiarmid. By a standard symmetrization argument followed by McDiarmid's bounded-differences inequality (Shalev-Shwartz & Ben-David, 2014, Thm. 26.5), with probability at least $1-\delta$ over the draw of $\{(x_i,y_i)\}_{i=1}^N$,

$$\sup_{P\in\mathcal{H}_{B,k}} \big[\, \mathcal{L}_{\mathrm{CoOp}}(P) - \widehat{\mathcal{L}}_N(P) \,\big] \ \le\ 2\,\mathcal{R}_N(\ell\circ\mathcal{H}_{B,k}) + M\sqrt{\tfrac{2\log(1/\delta)}{N}}, \tag{16}$$

where $\mathcal{R}_N(\ell\circ\mathcal{H}_{B,k}) := \mathbb{E}\Big[\sup_{P\in\mathcal{H}_{B,k}} \frac{1}{N}\sum_{i=1}^N \sigma_i\,\ell(P;(x_i,y_i))\Big]$ is the Rademacher complexity of the loss class and $\sigma_i \in \{\pm 1\}$ are i.i.d. Rademacher signs.

Step 2: Reduction to covering of $\mathcal{H}_{B,k}$ via Dudley. Under (L), the empirical L_2 metric on the loss class is dominated by the Frobenius metric on $\mathcal{H}_{B,k}$:

$$\Big(\tfrac{1}{N}\sum_{i=1}^N |\ell(P;(x_i,y_i)) - \ell(P';(x_i,y_i))|^2\Big)^{1/2} \ \le\ L_\ell\,\|P-P'\|_F.$$

Hence the empirical L_2 covering numbers obey $\mathcal{N}(\epsilon, \ell\circ\mathcal{H}_{B,k}, \widehat{L}_2) \le \mathcal{N}(\epsilon/L_\ell, \mathcal{H}_{B,k}, \|\cdot\|_F)$. Dudley's entropy integral (Wainwright, 2019, Sec. 5.3) then gives

$$\mathcal{R}_N(\ell\circ\mathcal{H}_{B,k}) \ \le\ \frac{12}{\sqrt{N}}\int_0^{2L_\ell B}\sqrt{\log\mathcal{N}(\epsilon,\ell\circ\mathcal{H}_{B,k},\widehat{L}_2)}\,d\epsilon \ \le\ \frac{12L_\ell}{\sqrt{N}}\int_0^{2B}\sqrt{\log\mathcal{N}(u,\mathcal{H}_{B,k},\|\cdot\|_F)}\,du, \tag{17}$$

where the second inequality uses the substitution $u = \epsilon/L_\ell$ and the diameter bound $\sup_{P,P'\in\mathcal{H}_{B,k}}\|P-P'\|_F \le 2B$.

Step 3: Covering number for low-rank Frobenius balls.

Lemma B.1. *For $\mathcal{H}_{B,k} = \{P \in \mathbb{R}^{m\times n} : \operatorname{rank}(P)\le k,\ \|P\|_F \le B\}$ and any $\epsilon \in (0,B]$,*

$$\log\mathcal{N}(\epsilon,\mathcal{H}_{B,k},\|\cdot\|_F) \ \le\ k(m+n+1)\,\log\Big(\tfrac{9B}{\epsilon}\Big). \tag{18}$$

Proof. Every $P \in \mathcal{H}_{B,k}$ admits a thin SVD $P = USV^\top$ with $U \in \mathcal{O}_{m,k} := \{U\in\mathbb{R}^{m\times k} : U^\top U = I_k\}$, $V\in\mathcal{O}_{n,k}$, and $S = \operatorname{diag}(s_1,\dots,s_k)$ with $s_i \ge 0$ and $\|S\|_F = \|P\|_F \le B$. Standard volumetric arguments (Vershynin, 2018, Cor. 4.2.13) give ϵ-coverings $\mathcal{C}_U \subset \mathcal{O}_{m,k}$ (in operator norm) and $\mathcal{C}_V \subset \mathcal{O}_{n,k}$ with $|\mathcal{C}_U| \le (3/\epsilon)^{mk}$ and $|\mathcal{C}_V| \le (3/\epsilon)^{nk}$, and $\mathcal{C}_S \subset \{S\in\mathbb{R}^k_{\ge 0} : \|S\|_F \le B\}$ with $|\mathcal{C}_S| \le (3B/\epsilon)^k$. Form $\mathcal{C} := \{\widehat{U}\widehat{S}\widehat{V}^\top : \widehat{U}\in\mathcal{C}_U,\ \widehat{S}\in\mathcal{C}_S,\ \widehat{V}\in\mathcal{C}_V\}$.

For any $P = USV^\top$, choose $\widehat{U}\in\mathcal{C}_U$, $\widehat{V}\in\mathcal{C}_V$, $\widehat{S}\in\mathcal{C}_S$ with $\|U-\widehat{U}\|_2 \le \epsilon/(3B)$, $\|V-\widehat{V}\|_2 \le \epsilon/(3B)$, $\|S-\widehat{S}\|_F \le \epsilon/3$. Triangle inequality gives

$$\begin{aligned}\|USV^\top - \widehat{U}\widehat{S}\widehat{V}^\top\|_F &\le \|U(S-\widehat{S})V^\top\|_F + \|U\widehat{S}(V-\widehat{V})^\top\|_F + \|(U-\widehat{U})\widehat{S}\widehat{V}^\top\|_F \\ &\le \|S-\widehat{S}\|_F + B\,\|V-\widehat{V}\|_2 + B\,\|U-\widehat{U}\|_2 \ \le\ \epsilon,\end{aligned}$$

using $\|U\|_2 = \|V\|_2 = 1$ and $\|\widehat{S}\|_F \le B$. Hence $\mathcal{C}$ is an ϵ-net of size at most $(3B/\epsilon)^{mk}\cdot(3B/\epsilon)^{nk}\cdot(3B/\epsilon)^k = (3B/\epsilon)^{k(m+n+1)}$, and (18) follows after absorbing the factor 3 into 9 for slack. □

Step 4: Evaluating the entropy integral. Substituting Lemma B.1 into (17),

$$\mathcal{R}_N(\ell \circ \mathcal{H}_{B,k}) \;\le\; \frac{12L_\ell}{\sqrt{N}}\sqrt{k(m+n+1)}\int_0^{2B}\sqrt{\log(9B/u)}\,du.$$

Substitute $t = \log(9B/u)$ so that $u = 9B\,e^{-t}$, $du = -9B\,e^{-t}\,dt$, with $u = 2B \Rightarrow t = \log(9/2)$ and $u \to 0^+ \Rightarrow t \to \infty$:

$$\int_0^{2B}\sqrt{\log(9B/u)}\,du = 9B\int_{\log(9/2)}^{\infty}\sqrt{t}\,e^{-t}\,dt \;\le\; 9B\int_0^{\infty}\sqrt{t}\,e^{-t}\,dt = 9B\cdot\tfrac{\sqrt{\pi}}{2} \;\le\; 8B,$$

using $\Gamma(3/2) = \sqrt{\pi}/2$. Therefore

$$\mathcal{R}_N(\ell \circ \mathcal{H}_{B,k}) \;\le\; \frac{96\,L_\ell\,B}{\sqrt{N}}\sqrt{k(m+n+1)} \;\le\; c_1' L_\ell B\sqrt{\tfrac{k(m+n)}{N}} \tag{19}$$

for an absolute constant c_1' (e.g., $c_1' = 192$ suffices since $m+n+1 \le 2(m+n)$ when $m+n \ge 1$).

Step 5: Combining. Plugging (19) into (16), with probability at least $1-\delta$, every $P \in \mathcal{H}_{B,k}$ satisfies

$$\mathcal{L}_{\mathrm{CoOp}}(P) - \widehat{\mathcal{L}}_N(P) \;\le\; \underbrace{2c_1'}_{=:c_1} L_\ell\,B\sqrt{\tfrac{k(m+n)}{N}} + \underbrace{M\sqrt{2}}_{=:c_2}\sqrt{\tfrac{\log(1/\delta)}{N}}.$$

Specializing to $m = n = d_y$ recovers $\mathcal{L}_{\mathrm{CoOp}}(P) - \widehat{\mathcal{L}}_N(P) \le c_1 L_\ell B\sqrt{2kd_y/N} + c_2\sqrt{\log(1/\delta)/N}$, which is the bound stated in Proposition 3.8 after absorbing the factor $\sqrt{2}$ into c_1. □

Verification of (B) and (L). Under ℓ_2-normalized features, $|s_{i,c}(P)| \le 1$ for all i, c, P, so the contrastive loss satisfies $0 \le \ell(P;(x,y)) \le \log C + 2/\tau =: M$, giving (B). For (L), the gradient of $s_{i,c}(P) = f(x_i)^\top W_y P y_c$ in P is $\nabla_P s_{i,c} = (W_y^\top f(x_i))\,y_c^\top$, with Frobenius norm $\le \|W_y\|_{\mathrm{op}}\,\|f(x_i)\|\,\|y_c\| \le \|W_y\|_{\mathrm{op}}\,r_y$, where $r_y := \sup\|y\|$. Composing with the $1/\tau$-Lipschitz softmax gradient yields $\|\nabla_P \ell\|_F \le 2\|W_y\|_{\mathrm{op}}\,r_y/\tau =: L_\ell$, giving (L). □

C Additional Synthetic Diagnostics

C.1 Exact recovery for a separate trace objective

To isolate the rank statement for the trace objective, consider a nonzero matrix M of exact rank k and the problem

$$\max_{\|P\|_F \le B}\operatorname{tr}(P^\top M), \qquad P_{\mathrm{tr}}^\star = \frac{BM}{\|M\|_F}, \quad B > 0. \tag{20}$$

The optimizer follows directly from Frobenius Cauchy–Schwarz and has rank k. We use $k \in \{2,4,8,16\}$, five seeds, and rank caps below, equal to, and above k. For a learned P, the diagnostics are

$$\begin{aligned} E_F(P) &= \frac{\|P - P_{\mathrm{tr}}^\star\|_F}{\|P_{\mathrm{tr}}^\star\|_F}, \\ \cos_F(P, P_{\mathrm{tr}}^\star) &= \frac{\langle P, P_{\mathrm{tr}}^\star\rangle_F}{\|P\|_F\|P_{\mathrm{tr}}^\star\|_F}, \qquad R_{\mathrm{tr}}(P) = \frac{\langle P, M\rangle_F}{B\|M\|_F}. \end{aligned} \tag{21}$$

Across the tested runs, rank caps at least k recover $P_{\mathrm{tr}}^\star$ to the precision in Table 6; caps below k have strictly positive approximation error. With a nonzero initialization P_0, the update $\Delta P = P - P_0$ need not have the rank or spectrum of the recovered P. An identity initialization therefore cannot be subtracted and then used interchangeably with the optimizer in a rank comparison.

This is an exact diagnostic for (20), not a claim that either normalized InfoNCE or the full population loss (4) reduces to this trace objective. Proposition 3.1 instead proves low-rank existence through conditional centering and convexity at arbitrary temperature; it does not supply $P_{\mathrm{tr}}^\star$ as its closed-form optimizer.

Table 6: **Trace-objective recovery diagnostics.** Summary for five seeds and $k \in \{2, 4, 8, 16\}$ once the rank cap is at least k. Values are reported to four decimal places; a printed zero is a numerical-precision statement, not an exact arithmetic claim.

Relative Frobenius error	Matrix cosine	Trace ratio
0.0000	1.0000	1.0000

C.2 Normalized InfoNCE does not give exact rank recovery

Table 3 reports the additional five-seed normalized symmetric InfoNCE diagnostic at latent dimension $k = 4$. The full model obtains lower loss than the rank-4 model, consistent with the nesting of their feasible classes. Its reported effective-rank statistic is approximately 50–51, not four. Thus these runs must not be presented as exact rank-k recovery. The statistic in this table is distinct from the explicitly defined covariance 95%-energy rank used in §F.1; no numerical identification is made.

D Evaluation Protocol and Additional Benchmark Results

D.1 Datasets, aggregation, and statistical uncertainty

The completed evaluation uses ImageNet, Caltech101, OxfordPets, StanfordCars, OxfordFlowers, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, and UCF101. Training uses the base classes and 16 labeled examples per class; evaluation reports base-class and new-class accuracy with CLIP ViT-B/16. Seeds are 1, 2, and 3. LR-MaPLe uses nine prompted layers, two context tokens per modality, factorization width 16, and five epochs of cross-entropy-only training.

Let b_{ds} and n_{ds} denote base and new accuracy on dataset d and seed s, in percent. For $D = 11$ datasets and $S = 3$ seeds, set

$$\bar{b}_d = \frac{1}{S}\sum_{s=1}^{S} b_{ds}, \qquad \bar{n}_d = \frac{1}{S}\sum_{s=1}^{S} n_{ds}, \qquad H_d = \frac{2\bar{b}_d\bar{n}_d}{\bar{b}_d + \bar{n}_d}. \tag{22}$$

The central values in Table 2 are

$$\mathrm{MacroBase} = \frac{1}{D}\sum_{d} \bar{b}_d, \qquad \mathrm{MacroNew} = \frac{1}{D}\sum_{d} \bar{n}_d, \qquad \mathrm{MacroHM} = \frac{1}{D}\sum_{d} H_d. \tag{23}$$

In particular, Macro HM is not the harmonic mean of Macro Base and Macro New. For seed-specific aggregates, define

$$B_s = \frac{1}{D}\sum_{d} b_{ds}, \quad N_s = \frac{1}{D}\sum_{d} n_{ds}, \quad H_s = \frac{1}{D}\sum_{d} \frac{2b_{ds}n_{ds}}{b_{ds} + n_{ds}}. \tag{24}$$

The uncertainty for each metric is the sample standard deviation of its three seed-specific aggregates:

$$\mathrm{SD}(Q) = \left[\frac{1}{S-1}\sum_{s=1}^{S}(Q_s - \bar{Q})^2\right]^{1/2}, \qquad \bar{Q} = \frac{1}{S}\sum_{s} Q_s, \quad Q \in \{B, N, H\}. \tag{25}$$

Because harmonic averaging is nonlinear, the central Macro HM in (23) need not equal the mean of the H_s in (24). We keep the stated central-value convention separate from the seed-level dispersion calculation. These standard deviations are not standard errors or confidence intervals.

Paired dataset-level comparisons. For methods A and B, let $\Delta_d = H_d^A - H_d^B$, $\bar{\Delta} = D^{-1}\sum_d \Delta_d$, and let s_Δ be the sample standard deviation of the D paired differences. The reported 95% Student-t interval is

$$\bar{\Delta} \pm t_{0.975,D-1}\frac{s_\Delta}{\sqrt{D}}. \tag{26}$$

The pairing is across the 11 benchmark datasets, after seed aggregation; it is not an interval based on only three seed-wise differences. Table 7 gives the reported values. Both intervals include zero. This supports the description of similar observed accuracy, without demonstrating superiority, statistical equivalence, or non-inferiority under a pre-specified margin. The paired comparison and rounded macro values are retained at their reported precision.

Table 7: **Paired Macro HM differences across 11 datasets.** Differences and interval endpoints are in percentage points.

Comparison	Mean difference	95% Student-t interval
LR-MaPLe − MaPLe	$+0.43$	$[-0.35,\ 1.21]$
LR-MaPLe − IVLP	$+0.72$	$[-0.28,\ 1.72]$

D.2 Completed SUN397 evaluation

SUN397 was initially absent because the local dataset copy did not match the standard Zhou split paths. Resolving those paths enabled the completed three-seed evaluation in Table 8. LR-MaPLe's HM is within 0.07 percentage points of MaPLe on this dataset, while its trainable parameter count remains approximately $19\times$ smaller. SUN397 is included in all 11-dataset aggregates in the main text; it is not averaged into the preliminary ten-dataset table.

Table 8: **SUN397 base-to-new generalization.** ViT-B/16, 16-shot, seeds 1/2/3. Entries are reported central values with sample standard deviations over the corresponding seed-level results. MMLoP here denotes the shared-factor CE-only control.

Method	Base	New	HM
LR-MaPLe	80.53 ± 0.15	78.23 ± 0.46	79.37 ± 0.30
MaPLe	80.90 ± 0.00	78.03 ± 0.49	79.44 ± 0.26
IVLP	80.97 ± 0.40	78.53 ± 0.65	79.73 ± 0.51
MMLoP (shared, CE-only)	81.67 ± 0.12	74.64 ± 1.24	77.99 ± 0.63

D.3 LR-MaPLe implementation

Algorithm 1 gives the shared-factor implementation in §4.1. The CLIP weights remain frozen; the learned parameters are the shared codes and modality-specific bases. The standard configuration uses two context tokens at nine layers, width $r = 16$, and five epochs of cross-entropy training.

D.4 What the baseline comparisons do and do not isolate

MaPLe and IVLP. The primary design question is whether a shared-factor parameterization can retain MaPLe-level base-to-new accuracy with fewer trainable parameters. The 11-dataset evaluation includes IVLP to contextualize independent visual and textual prompting. The result is an accuracy–parameter comparison at the stated operating points, not proof that the factorization is the only possible efficient design.

MMLoP: shared-factor control versus complete system. LR-MaPLe and MMLoP Ghiasvand et al. (2026) share the same broad algebraic mechanism, with different native configurations. LR-MaPLe uses two prompt tokens per modality, width $r = 16$, nine layers, and five epochs with cross-entropy only. The full MMLoP configuration uses four tokens, rank one, nine layers, and 30 epochs; it also includes feature- and logit-level SCL, UDC during training and inference, and a 60-template zero-shot ensemble. Its reported 79.70 HM and 11.5K parameters describe that complete system, not the control in Table 2. Its own low-rank-prompting-only ablation reports 77.06 HM, also distinct from our CE-only control.

Our control retains low-rank prompting and shared up-projection but removes SCL, UDC, and the layer-0 text-anchor initialization. It yields 76.94 ± 0.80 Macro HM versus 77.69 ± 0.44 for LR-MaPLe. The control has higher Base but lower New accuracy. This comparison helps separate bare shared-factor prompting from the complete MMLoP system, but does not isolate every token, initialization, or training configuration difference. We make no significance claim for the 0.75-point aggregate gap and no priority claim for the factorization.

Broader adaptation systems and supervision budgets. PromptSRC Khattak et al. (2023b), CLIP-LoRA Zanella & Ben Ayed (2024), MMA Yang et al. (2024), MMRL Guo & Gu (2025), and

VaMP Cheng & Han (2025) provide relevant stronger or complementary adaptation systems. Their absence from the controlled table limits any broader empirical ranking. IPO Du et al. (2024), Skip Tuning Wu et al. (2025), DPC Li et al. (2025), and HiCroPL Zheng et al. (2025) are also discussed in Related Work. PromptKD Li et al. (2024), with a larger adapted teacher and unlabeled target-domain images, belongs to a separately identified teacher-plus-unlabeled-data category rather than the same resource budget as pure labeled 16-shot adaptation. The current comparisons support the focused MaPLe compression result, not superiority over these full systems.

E Factorization-Width and Active-Rank Sensitivity

E.1 Standard two-token configuration

Table 4 reports three-seed mean HM for $n_{\mathrm{ctx}} = 2$. The factorization width varies over $r \in \{1, 2, 4, 8, 16, 32\}$. Only $r = 1$ restricts the prompt rank below the two-row ceiling; all $r \geq 2$ permit rank at most two. Differences among the latter settings therefore reflect parameterization and optimization rather than different maximum output ranks. No width is uniformly best. In particular, the default $r = 16$ is a fixed architectural choice, not a value inferred from k.

E.2 Active-rank diagnostic with sixteen tokens

Increasing the context length to $n_{\mathrm{ctx}} = 16$ allows the rank constraint to vary over $r \in \{1, 2, 4, 8, 16\}$. The comparison in Table 5 includes a matched unconstrained MaPLe control at this context length. Among the LR-MaPLe settings, the largest mean HM occurs at $r = 1$ on DTD and $r = 2$ on EuroSAT. MaPLe is weaker on DTD but stronger on EuroSAT. Several EuroSAT settings have substantial seed variation. These results show that small active ranks can suffice in the tested settings, not that one rank is universally optimal, that the best rank is statistically resolved, or that a numerical estimate of the latent dimension has been recovered.

F Qualitative Grad-CAM Illustrations

The original qualitative study uses frozen CLIP ViT-B/16 with a CoOp-style soft prompt trained on ImageNet under the standard 16-shot protocol Zhou et al. (2022b). It compares that prompt with the manual template "a photo of a [class]." Grad-CAM Selvaraju et al. (2017) is computed with respect to the image–text cosine similarity. In the two examples in Figure 2, the learned prompts have reported similarity gains of $+0.075$ (crab) and $+0.128$ (cellular phone), with broader highlighted object regions. These examples illustrate changes in feature attribution; two selected images cannot establish that spurious correlations are absent or verify the nonlinear rank theorem. The quantitative covariance and CKA measurements in §F.1, rather than Grad-CAM, provide the principal real-CLIP diagnostic.

F.1 Quantitative Covariance Diagnostics on Real CLIP

To test the operator-level prediction of Theorem 3.4, we measure centered image–class-text covariance on balanced 512-image OxfordPets base/new manifests with CLIP ViT-B/16:

$$\widehat{C}_{xy}(P) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - \bar{f})(g_P(y_i) - \bar{g}_P)^\top. \tag{27}$$

Here f suppresses the image branch's prompt dependence. Define $\rho_{95}(C) = \min\{q : \sum_{j \leq q} \sigma_j(C)^2 \geq 0.95 \sum_j \sigma_j(C)^2\}$. Table 9 shows similar low covariance energy ranks and higher observed CKA for LR-MaPLe on both splits. Because the text feature is class-determined, $\mathrm{rank}(\widehat{C}_{xy}) \leq C_{\mathrm{cls}} - 1$; the concentration is relative to this ceiling as well as ambient dimension 512.

Covariance updates $\widehat{C}_{xy}(P) - \widehat{C}_{xy}(P_0)$ have energy ranks 10/9 on base and 10/10 on new classes for LR-MaPLe/MaPLe; a difference need not preserve either operator's rank. Separately, all 18 LR-MaPLe token matrices have numerical rank at most two. These observations support concentrated cross-modal structure but neither verify (CM) nor estimate k. Width, token rank, covariance energy

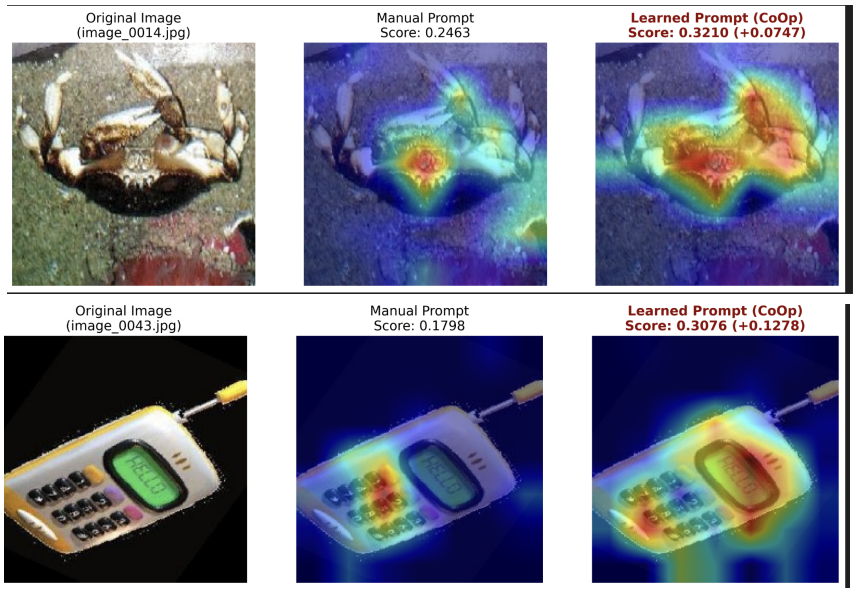

Figure 2: **Qualitative Grad-CAM examples on ImageNet.** Each row compares the original image, the manual-template attribution, and the learned CoOp-style prompt attribution with its image–text similarity. The reported similarity gains are $+0.075$ for the crab and $+0.128$ for the phone. These are illustrative examples, not tests of exact rank, recovery of k, or exclusion of spurious correlations.

Table 9: **Centered covariance diagnostics on OxfordPets.** Energy ranks describe $\widehat{C}_{xy}$, not prompt rank or an estimate of k. Each split uses a balanced 512-image manifest.

	95%-energy rank		CKA	
Split	LR-MaPLe	MaPLe	LR-MaPLe	MaPLe
Base	7	8	0.620	0.580
New	8	8	0.622	0.605

rank, and latent dimension are different quantities. Grad-CAM remains a qualitative illustration in Appendix F.