Spherical Interpolation for Backward-Compatible Multimodal Representations

Simone Ricci[1,2*] **Niccolò Biondi**[3] **Federico Pernici**[1,2]

[1]DINFO (Department of Information Engineering), University of Florence, Italy
[2]MICC (Media Integration and Communication Center)
[3]University of Trento, Italy

Website **Code**

Abstract

Contrastive vision-language models map visual and textual representations into a shared normalized embedding space, making cosine similarity the natural metric for cross-modal retrieval. A practical challenge arises during model upgrades: independently trained models generally produce incompatible representation spaces, so replacing a deployed model typically requires recomputing embeddings for the entire gallery, which is prohibitively expensive at scale. Orthogonal post-hoc alignment can partially mitigate this problem by mapping new-model queries into the old-model gallery space. However, because independently trained models can differ in fine-grained representation structure, the orthogonal alignment remains approximate, leaving a residual angular discrepancy between the old-model query and the aligned new-model query. We study whether interpolation along the spherical geodesic between these two normalized query representations can improve retrieval without re-indexing the gallery. We characterize when this path contains an interior query direction closer to an idealized retrieval-optimal direction than either endpoint, and connect this characterization to Recall@K through a local margin-based certification result. Experiments across multiple benchmarks and model families show that post-alignment spherical interpolation improves over orthogonal alignment alone, recovering backward-compatibility in most evaluated settings. Consistent with our geometric characterization, per-query oracle analysis shows that retrieval-favorable interior points occur frequently in practice.

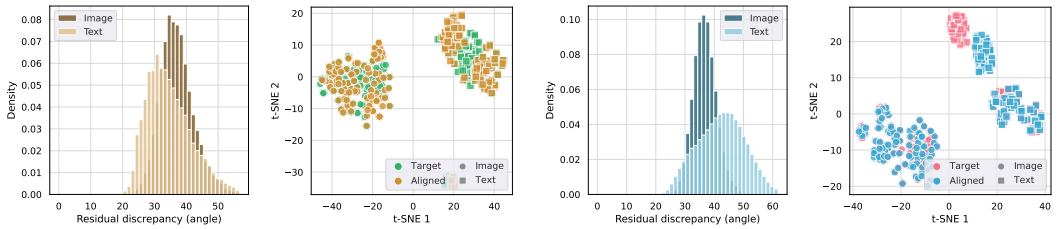
1 Introduction

Contrastive vision-language models (VLMs) have become a standard foundation for multimodal retrieval and zero-shot classification [1, 2, 3, 4, 5]. By mapping images and text into a shared normalized embedding space, these models leverage cosine similarity for cross-modal retrieval and prompt-based classification. As the ecosystem of pretrained models grows through public model releases and model hubs [6, 7, 8], a practical question arises: how can a deployed retrieval system benefit from a stronger model without recomputing all embeddings produced by the old one?

This question is particularly important in large-scale retrieval systems. In many production deployments, galleries containing millions or billions of items have already been encoded and indexed, making index reconstruction after every model update computationally prohibitive [9, 10, 11, 12, 13]. Moreover, re-indexing may be infeasible when the original raw data are no longer accessible due

*Corresponding author: simone.ricci@unifi.it.

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



(a) CC3M validation split, used to estimate the orthogonal map.

(b) Flickr30k dataset, used to evaluate the estimated orthogonal map in a zero-shot setting.

Figure 1: Residual discrepancy after orthogonal alignment from SigLIP2 to CLIP ViT-B/32. We show two complementary views: the distribution of residual angular mismatch between paired embeddings from the two models, and a 2D t-SNE projection of the corresponding feature spaces. The remaining angular mismatch after alignment suggests that the orthogonal map leaves a residual cross-model gap, motivating a post-alignment interpolation strategy.

to privacy constraints, storage limitations, or data retention policies [14, 15, 16]. In such settings, backward-compatible upgrades that preserve the existing index may be the only viable option. Even when re-indexing is feasible, progressive upgrades create a migration period requiring compatibility with the existing gallery.

This compatibility challenge is increasingly common as advances in architectures, optimization, and training data drive frequent model upgrades [17, 18, 19, 20, 21]. Yet such updates can compromise backward-compatibility [10] and alter the behavior of downstream applications in undesirable ways [10, 22, 11, 15, 23, 24, 25, 26]. Independently trained models rarely produce directly comparable representations [27], so directly replacing a query encoder can break compatibility with an existing gallery. Backward-compatible training [10] addresses this issue by imposing restrictive constraints on the new-model training that can degrade its performance [28, 29]. A recent alternative is post-hoc alignment, which decouples model improvement from compatibility constraints by training the new model independently and restoring compatibility afterward through a lightweight mapping between representation spaces [12, 13, 30]. This approach is motivated by the manifold hypothesis and related views on latent-space convergence, which suggest that functionally similar models may differ largely by simple transformations of a shared latent structure [31, 32, 33, 34, 30].

For contrastive VLMs, orthogonal Procrustes provides a natural alignment map: embeddings are normalized and compared by cosine similarity, so the resulting orthogonal transformation preserves inner products and therefore the image–text geometry underlying retrieval and zero-shot classification [35]. Moreover, agreement of the multimodal similarity kernel between independently trained contrastive VLMs identifies a single orthogonal map shared across image and text encoders, enabling cross-modal transfer of an alignment estimated from one modality [35]. Related unimodal literature further supports orthogonal transformations over unconstrained linear maps, particularly when the new model is more informative, as an isometry preserves rather than distorts its discriminative geometry [33, 36, 30]. However, independent training can induce representation differences that are not captured by a single global isometry. Orthogonal post-hoc alignment is therefore approximate: the aligned new-model query embedding will not, in general, coincide with the embedding that the old model would have produced for the same input, especially when models share coarse semantic structure but differ in fine-grained organization [37]. Consistent with this observation, Fig. 1 shows that a non-negligible residual angular mismatch remains after orthogonally aligning SigLIP2 to CLIP ViT-B/32, both on the alignment dataset and under zero-shot transfer.

Motivated by the residual angular discrepancy, we treat the old-model query and the aligned new-model query as endpoints in a shared normalized representation space. Rather than assuming that either endpoint is optimal for retrieval, we ask whether an intermediate direction along the spherical geodesic connecting them can be more effective against the old-model gallery. Because retrieval is based on cosine similarity between normalized embeddings, the corresponding spherical interpolation is the minor geodesic between the two query directions. We parameterize this path using spherical linear interpolation (SLERP) [38], which traverses the geodesic at constant angular speed and therefore gives the interpolation weight a consistent geometric interpretation across queries. We formalize this intuition by considering an idealized retrieval-optimal query direction, used only as an

analytical reference, and characterizing when the interpolation geodesic contains an interior point with a smaller angular gap to this direction than either endpoint. Geometrically, this occurs exactly when the normalized projection of the retrieval-optimal direction onto the interpolation plane lies in the relative interior of the arc; otherwise, the optimum along the arc is attained at an endpoint. We then connect this angular characterization to top-K retrieval through a margin perturbation bound and a local Recall@K certification result.

We propose an approach that requires no model retraining or gradient-based optimization: the deployed gallery remains unchanged, while the new-model query is mapped into the old-model space by orthogonal Procrustes alignment and then interpolated with the old-model query using a fixed weight selected on the alignment support set. Empirically, we evaluate backward-compatible cross-modal retrieval across multiple benchmarks and model families. Orthogonal alignment provides a useful shared coordinate system but often fails to achieve backward-compatibility on its own, whereas post-alignment spherical interpolation improves over orthogonal alignment and frequently restores compatibility. Controlled ablations show that endpoint combination accounts for most of the mean retrieval improvement, while support-set weight selection primarily improves compatibility robustness. The weight selected on the same CC3M alignment support set used to estimate the Procrustes map transfers across retrieval datasets, achieving performance close to the per-dataset optimum without target-test tuning. A per-query oracle analysis, reported only as an upper bound, shows that retrieval-favorable interior positions occur frequently in practice; flip-rate analysis further reveals that the retrieval gains arise from a favorable balance between positive and negative retrieval flips. Finally, the approach remains effective when the gallery is re-indexed and in zero-shot classification.

Our main contributions are:

- We geometrically characterize when spherical interpolation between old-model and aligned new-model queries contains an interior direction closer to an idealized retrieval-optimal direction than either endpoint, and connect this characterization to top-K retrieval.
- We introduce a query-side approach for contrastive VLM upgrades that combines orthogonal Procrustes alignment with spherical interpolation, requiring neither model retraining nor gradient-based optimization, using a fixed interpolation weight selected on the alignment support set, while leaving the deployed gallery unchanged.
- We validate the approach across CLIP, SigLIP1, and SigLIP2 on Flickr30k, COCO2014, and NoCaps, showing gains over orthogonal alignment and frequent backward-compatibility recovery. Ablations show that endpoint combination drives most of the improvement, while support-set weight selection transfers across datasets without target-test tuning.

2 Related Work

Backward-compatible representation learning. Backward-compatible training (BCT) [10] formalized how to upgrade a retrieval model without re-encoding a deployed gallery by requiring embeddings from the new model to remain directly comparable with those of the old one. Subsequent work extended this paradigm through class- and prototype-level alignment [11], neighborhood-level compatibility constraints [39], open-set and universal formulations [14], regression-aware hot-refresh updates [40], basis expansion [28], stationary representations [41, 15, 42], and orthogonal transformation layers [29]. In particular, stationary representations learned with d-Simplex fixed classifiers [43], the geometry that also emerges under neural collapse [44], provably satisfy the inequality constraints of the formal compatibility definition in unimodal retrieval [26]. Other works relax the uniform compatibility requirement through selective compatibility [45], hyperbolic embeddings constrained by entailment cones [46], or perturbed prototype targets that preserve discriminative structure while maintaining interoperability [47]. Recent work also shows that hyperspherical simplex representations derived for classifier outputs can be inherently backward-compatible and satisfy the formal definition of backward compatibility on average [48]. Related lifelong formulations have been studied in image-to-image retrieval and person re-identification, where compatibility avoids re-indexing across sequential updates [49, 50, 42, 15]. However, most of this literature addresses unimodal model upgrades. For vision-language models, XBT [51] extends backward-compatible learning to cross-modal retrieval by learning a projection module, pretraining it on text, injecting Gaussian noise, and fine-tuning the new VLM with LoRA using both text and image data. While effective,

this requires a dedicated training pipeline with large text and image–text corpora. In contrast, our approach requires no model retraining or gradient-based optimization: it estimates an orthogonal Procrustes map from a small alignment support set and performs query-side interpolation without modifying either model or re-indexing the gallery.

Model interpolation. Interpolation has been studied extensively in parameter space through mode connectivity, weight averaging, model soups, permutation-aware merging, and task arithmetic [52, 53, 54, 55, 56, 57, 58]. Recent works also use spherical interpolation, but for different objects and objectives. In zero-shot composed image retrieval, [59] uses SLERP to combine image and text embeddings into a composed query, together with text-anchored tuning. WARP [60], by contrast, applies SLERP in policy weight space to merge independently fine-tuned language-model policies, improving reward while controlling deviation from a reference policy. Unlike these works, we apply spherical interpolation on the query side for backward-compatible retrieval, interpolating between the old-model query and the Procrustes-aligned new-model query for the same input. We characterize when the resulting minor geodesic contains an interior direction closer to an idealized retrieval-optimal direction than either endpoint, and connect this geometry to Recall@K through margin-based guarantees.

We provide an extended discussion of post-hoc alignment, cross-model correspondence, and the geometry of contrastive vision-language representations in Appendix A.

3 Geometric Characterization of Spherical Interpolation for Compatible Retrieval

We consider a deployed cross-modal retrieval system whose gallery is fixed and encoded by an old contrastive VLM $\phi_{\text{old}} : \mathcal{X} \to \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} := \{z \in \mathbb{R}^d : \|z\| = 1\}$, and $\mathcal{X}$ denotes the input space, which may contain either images or text. Suppose that an independently trained new model $\phi_{\text{new}} : \mathcal{X} \to \mathbb{S}^{d-1}$ becomes available. Our goal is to use the new model on the query side while keeping the gallery indexed by the old model.

Let $\mathcal{A} = \{x_i\}_{i=1}^{N_a}$ be an alignment support set, disjoint from the gallery set $\mathcal{G} = \{g_j\}_{j=1}^{N_g}$ and the query set $\mathcal{Q} = \{x_k\}_{k=1}^{N_q}$. For each $x_i \in \mathcal{A}$, define $u_i := \phi_{\text{old}}(x_i)$ and $\bar{v}_i := \phi_{\text{new}}(x_i)$, and stack these embeddings row-wise into $U, \bar{V} \in \mathbb{R}^{N_a \times d}$. We estimate the new-to-old alignment map by solving the orthogonal Procrustes problem:

$$R^\star = \arg \min_{R^\top R = I} \|\bar{V}R - U\|_F^2. \tag{1}$$

If $\bar{V}^\top U = P\Sigma Q^\top$ is its singular value decomposition, the closed-form solution is $R^\star = PQ^\top$. Since $R^\star$ is orthogonal, the aligned representation preserves the inner-product geometry of the new model while expressing it in the old-model coordinate system. For clarity, we develop the main analysis in the equal-dimensional case $\phi_{\text{old}}, \phi_{\text{new}} : \mathcal{X} \to \mathbb{R}^d$; the extension to unequal-dimensional model pairs using rectangular Procrustes is described in Appendix B.

For any query $x \in \mathcal{Q}$, define

$$u(x) := \phi_{\text{old}}(x) \in \mathbb{S}^{d-1}, \qquad v(x) := \phi_{\text{new}}(x)R^\star \in \mathbb{S}^{d-1}, \tag{2}$$

as the old-model query and the aligned new-model query, respectively. Thus, $u(x)$ and $v(x)$ are two normalized representations of the same input, both directly comparable with the old-model gallery. Because the alignment is approximate for independently trained models, these query directions generally remain distinct, as illustrated in Fig. 1.

We now characterize when interpolation between $u(x)$ and $v(x)$ can yield a direction closer to a retrieval-favorable direction than either endpoint. For normalized embeddings $a, b \in \mathbb{S}^{d-1}$, define the spherical angle $\angle(a, b) := \arccos\big(\langle a, b\rangle\big) \in [0, \pi]$. For a fixed query $x \in \mathcal{Q}$, let $\theta := \angle(u(x), v(x)) \in (0, \pi)$.[2] Since the analysis is pointwise, we omit the dependence on x and write u, v, and θ. The problem then reduces to interpolation between two unit vectors along their minor geodesic.

[2]The condition $0 < \theta < \pi$ excludes the coincident case $u(x) = v(x)$ and the antipodal case $\cos(\theta) = -1$, for which the minor geodesic is not uniquely defined.

Definition 1 (SLERP). *For $u, v \in \mathbb{S}^{d-1}$ with $\theta = \angle(u, v) \in (0, \pi)$, the SLERP query at interpolation weight $\alpha \in [0, 1]$ is*

$$q_\alpha := \mathrm{slerp}(u, v; \alpha) := \frac{\sin((1-\alpha)\theta)}{\sin\theta}\, u + \frac{\sin(\alpha\theta)}{\sin\theta}\, v. \tag{3}$$

SLERP traces the unique minor geodesic from u to v at constant angular speed:

$$\angle(u, q_\alpha) = \alpha\theta, \qquad \angle(q_\alpha, v) = (1-\alpha)\theta, \qquad \angle(q_\alpha, q_\beta) = |\alpha - \beta|\,\theta. \tag{4}$$

Thus, α has a direct geometric interpretation as the normalized angular displacement from the old-model query toward the aligned new-model query. NLERP [61] traces the same minor geodesic up to a monotone reparameterization and therefore attains the same optimal directions under continuous weight selection. We use SLERP because its constant-angular-speed parameterization gives the interpolation weight a consistent geometric meaning across queries; Appendix C discusses both interpolations.

To formalize whether interpolation moves toward a retrieval-favorable direction, let $q^* \in \mathbb{S}^{d-1}$ denote an idealized retrieval-optimal query direction, used only as an analytical reference. For an interpolated query q_α, define the angular retrieval gap

$$G(\alpha) := \angle(q_\alpha, q^*), \qquad \alpha \in [0, 1]. \tag{5}$$

A strict interior angular improvement occurs if there exists $\alpha^* \in (0, 1)$ such that

$$G(\alpha^*) < \min\{G(0), G(1)\}. \tag{6}$$

Since every q_α lies on the minor geodesic joining u and v, the analysis reduces to the two-dimensional interpolation plane $\mathrm{span}(u, v)$. We use the orthonormal basis

$$e_1 := u, \qquad e_2 := \frac{v - \langle u, v\rangle u}{\|v - \langle u, v\rangle u\|} = \frac{v - \cos\theta\, u}{\sin\theta}, \tag{7}$$

so that

$$u = e_1, \qquad v = \cos\theta\, e_1 + \sin\theta\, e_2, \qquad q_\alpha = \cos(\alpha\theta) e_1 + \sin(\alpha\theta) e_2. \tag{8}$$

Hence, only the projection of q^* onto the interpolation plane affects the variation of $G(\alpha)$ along the geodesic.

Lemma 1 (Decomposition relative to the interpolation plane; proof in Appendix D). *Let $q^* \in \mathbb{S}^{d-1}$. There exist unique vectors $p \in \mathrm{span}(u, v)$ and $w_\perp \perp \mathrm{span}(u, v)$ such that $q^* = p + w_\perp$. Let $\rho := \|p\|$. Then $\rho \in [0, 1]$ and $\|w_\perp\|^2 = 1 - \rho^2$. If $\rho > 0$, there exists a unique $\psi \in (-\pi, \pi]$ such that*

$$p = \rho(\cos\psi\, e_1 + \sin\psi\, e_2). \tag{9}$$

Here, ρ measures the magnitude of the component of q^* in the interpolation plane, while ψ gives its angular coordinate from u. The effect of interpolation is therefore determined by the position of the normalized in-plane projection relative to the arc $[0, \theta]$.

Theorem 1 (Angular retrieval gap along the SLERP arc; proof in Appendix E). *Let $q_\alpha = \mathrm{slerp}(u, v; \alpha)$ with $\theta = \angle(u, v) \in (0, \pi)$, and let ρ and ψ be defined as in Lemma 1 with respect to $e_1 = u$ and $e_2 = (v - \cos\theta\, u)/\sin\theta$. Then:*

(i) If $\rho > 0$, then

$$\langle q_\alpha, q^*\rangle = \rho\cos(\alpha\theta - \psi), \qquad G(\alpha) = \arccos\big(\rho\cos(\alpha\theta - \psi)\big). \tag{10}$$

(ii) If $\rho = 0$, then

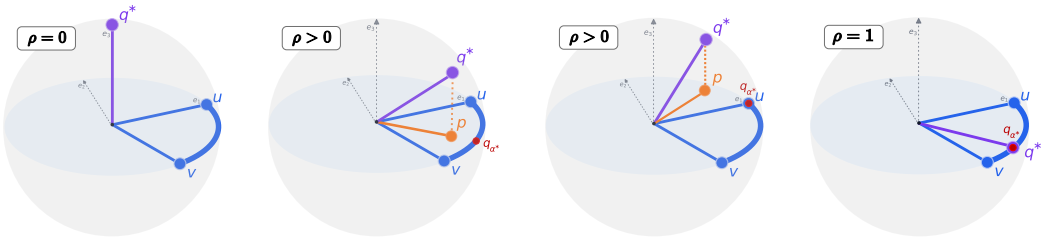
$$G(\alpha) \equiv \pi/2, \qquad \forall\alpha \in [0, 1]. \tag{11}$$

(iii) If $\rho > 0$, the minimizers of $G(\alpha)$ over $\alpha \in [0, 1]$ coincide with the minimizers of the circular distance $d_{\mathrm{circ}}(\alpha\theta, \psi)$ over $\alpha \in [0, 1]$, where

$$d_{\mathrm{circ}}(\tau, \psi) := \min_{k\in\mathbb{Z}} |\tau - \psi - 2\pi k|. \tag{12}$$

Equivalently, every minimizer satisfies

$$\alpha^* = \frac{\tau^*}{\theta}, \qquad \tau^* \in \arg\min_{\tau\in[0,\theta]} d_{\mathrm{circ}}(\tau, \psi). \tag{13}$$



(a) $p = 0$. (b) p lies before the arc. (c) p lies beyond the arc. (d) p lies on the arc.

Figure 2: Geometry of SLERP in the interpolation plane. The old query u and aligned new query v define a minor geodesic arc. Let p be the projection of the retrieval-optimal direction q^* onto this plane, and let $\rho = \|p\|$. The effect of interpolation is determined by the normalized in-plane projection p/ρ when $\rho > 0$. A strict interior improvement occurs exactly when p/ρ lies in the relative interior of the arc; when $\rho = 0$, the angular gap is constant along the arc.

For $\rho > 0$, a strict interior improvement

$$G(\alpha^*) < \min\{G(0), G(1)\} \tag{14}$$

holds if and only if the normalized in-plane projection of q^* lies strictly in the relative interior of the minor arc from u to v.

Theorem 1 shows that the endpoint separation θ alone does not determine whether interpolation yields an interior angular improvement. The decisive quantity is the position of the normalized in-plane projection of q^* relative to the minor arc. If this direction lies in the relative interior of the arc, the geodesic contains an interior query closer to q^* than either endpoint; otherwise, the optimum along the arc is attained at an endpoint. Appendix F connects this angular characterization to Recall@K by instantiating q^* as a top-K-margin-optimal direction. A margin perturbation bound then specifies when angular proximity to this direction locally certifies Recall@K success. Thus, Theorem 1 characterizes when interpolation improves the angular retrieval gap, while the margin result determines when this improvement is sufficient for the discrete top-K retrieval event.

4 Experimental Results

We evaluate whether post-alignment spherical interpolation improves backward-compatible retrieval after mapping the new model into the old-model coordinate system with orthogonal Procrustes alignment. Because the alignment is obtained in closed form from a singular value decomposition (Sec. 3), we refer to it as SVD throughout this paper. The main experiments focus on cross-modal retrieval with a fixed old-model gallery, while Appendix G reports an auxiliary zero-shot classification evaluation with fixed old-model text prototypes.

4.1 Experimental Setup

We evaluate the proposed approach across CLIP, SigLIP1, and SigLIP2 vision-language models. To cover both same-family and cross-family upgrades, we use OpenAI CLIP ViT-B/32 and ViT-L/14 [1], CLIP ViT-H/14 [62], SigLIP1 ViT-SO400M-14 [3], and SigLIP2 ViT-SO400M-14 [5]. These produce embeddings of 512 (CLIP ViT-B/32), 768 (CLIP ViT-L/14), 1024 (CLIP ViT-H/14), and 1152 (SigLIP1/2) dimensions, covering equal- and unequal-dimensional model pairs. Including SigLIP-family models allows us to assess whether the residual post-alignment discrepancy and the gains from spherical interpolation extend beyond CLIP-family representation geometry to models trained with different objectives. All checkpoints are obtained from the OpenCLIP repository [63].

For each old–new model pair, we estimate the Procrustes map from 12,637 available image–text pairs from the CC3M [64] validation split, which serves as the alignment support set. We consider three support-modality choices: text-only, image-only, and joint image–text support, where the joint setting stacks embeddings from both modalities before solving the Procrustes problem. Motivated by [35], these variants test whether a single modality is sufficient to estimate the orthogonal map and whether combining modalities yields more stable cross-modal transfer.

Table 1: Cross-modal backward-compatible retrieval for same-family model upgrades. For each dataset, the α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The weight $\hat{\alpha}$ is selected on CC3M and fixed across all test datasets, whereas $\alpha^\star$ is selected directly on each test dataset and is reported only as an oracle upper bound. Checkmarks indicate backward compatibility; bold and underlined values denote the best and second-best $\text{new} \to \text{old}$ results, respectively.

		Flickr30k			COCO			NoCaps		
Method	Sup.	α	I2T@1	T2I@1	α	I2T@1	T2I@1	α	I2T@1	T2I@1
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*										
► Old model (CLIP ViT-B/32)	–	–	40.62	21.73	–	28.76	14.47	–	71.29	45.24
XBT [51]	–	–	42.47 ✓	22.38 ✓	–	30.73 ✓	**15.55** ✓	–	75.02 ✓	**48.02** ✓
SVD	T	–	42.89 ✓	21.49 ×	–	30.27 ✓	14.03 ×	–	69.22 ×	43.92 ×
+SLERP ($\hat{\alpha}$)	T	0.7/0.5	48.00 ✓	23.33 ✓	0.7/0.5	33.40 ✓	15.26 ✓	0.7/0.5	73.98 ✓	46.59 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.5	**48.61** ✓	23.33 ✓	0.5/0.4	**33.84** ✓	15.27 ✓	0.4/0.4	**75.62** ✓	46.65 ✓
SVD	I	–	41.17 ✓	19.99 ×	–	28.71 ×	13.11 ×	–	65.96 ×	42.16 ×
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	46.30 ✓	23.00 ✓	0.7/0.4	31.77 ✓	15.14 ✓	0.7/0.4	71.40 ✓	46.49 ✓
+SLERP ($\alpha^\star$)	I	0.5/0.4	47.39 ✓	23.00 ✓	0.5/0.3	32.51 ✓	15.15 ✓	0.4/0.3	73.78 ✓	46.51 ✓
SVD	I+T	–	42.44 ✓	21.80 ✓	–	29.95 ✓	14.16 ×	–	69.13 ×	44.41 ×
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.4	47.09 ✓	**23.44** ✓	0.7/0.4	33.05 ✓	15.30 ✓	0.7/0.4	73.56 ✓	46.98 ✓
+SLERP ($\alpha^\star$)	I+T	0.5/0.5	47.77 ✓	**23.44** ✓	0.6/0.5	33.32 ✓	15.35 ✓	0.5/0.4	74.78 ✓	46.98 ✓
► New model (CLIP ViT-L/14)	–	–	48.72	28.27	–	34.33	18.68	–	73.36	47.84
SigLIP2 → SigLIP1 *(same model family)*										
► Old model (SigLIP1)	–	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18
SVD	T	–	45.46 ×	46.28 ✓	–	36.30 ×	30.53 ×	–	75.00 ×	64.64 ✓
+SLERP ($\hat{\alpha}$)	T	0.3/0.4	58.63 ✓	45.99 ✓	0.3/0.4	46.76 ×	32.50 ✓	0.3/0.4	85.87 ×	66.59 ✓
+SLERP ($\alpha^\star$)	T	0.2/0.7	**58.70** ✓	**47.59** ✓	0.1/0.5	47.18 ✓	32.56 ✓	0.1/0.5	86.49 ✓	66.72 ✓
SVD	I	–	50.66 ×	41.78 ✓	–	40.96 ×	28.18 ×	–	80.00 ×	62.69 ×
+SLERP ($\hat{\alpha}$)	I	0.2/0.3	58.59 ✓	43.84 ✓	0.2/0.3	**47.30** ✓	32.02 ✓	0.2/0.3	**86.69** ✓	66.08 ✓
+SLERP ($\alpha^\star$)	I	0.3/0.6	58.61 ✓	45.12 ✓	0.2/0.4	**47.30** ✓	32.06 ✓	0.2/0.5	**86.69** ✓	66.38 ✓
SVD	I+T	–	50.19 ×	46.39 ✓	–	39.41 ×	30.78 ×	–	77.93 ×	64.98 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.2/0.5	58.62 ✓	46.87 ✓	0.2/0.5	47.24 ✓	**32.67** ✓	0.2/0.5	86.49 ✓	**67.02** ✓
+SLERP ($\alpha^\star$)	I+T	0.2/0.7	58.62 ✓	47.58 ✓	0.2/0.5	47.24 ✓	**32.67** ✓	0.1/0.6	86.60 ✓	**67.02** ✓
► New model (SigLIP2)	–	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82

We evaluate backward-compatible retrieval, comparing the training-based XBT baseline [51], three Procrustes support-modality variants, and SLERP applied after each alignment. Old- and new-model performance are included as reference points. For SLERP, $\alpha = 0$ corresponds to the old-model query, while $\alpha = 1$ corresponds to the SVD-aligned new-model query, i.e., the SVD baseline. For each old–new model pair, the Procrustes map is estimated and the interpolation weight $\hat{\alpha}$ is selected on the same CC3M alignment support set. Separate weights are selected for image-to-text (I2T) and text-to-image (T2I) retrieval by maximizing CC3M Recall@1 over $\alpha \in \{0, 0.1, \ldots, 1.0\}$. These model-pair- and retrieval-direction-specific weights are then fixed for evaluation on three cross-modal retrieval benchmarks, following the large-scale protocol of XBT [51]: the full Flickr30k dataset [65] (the union of the Karpathy splits [66]), and the validation splits of COCO 2014 [67] and NoCaps [68]. We additionally report the benchmark-specific oracle weight $\alpha^\star$, selected directly on each test benchmark solely as an upper-bound reference.

4.2 Evaluation Protocol

Following [10, 51], we consider an updated system empirically backward-compatible if, given a performance metric M, the new-to-old performance strictly exceeds the old-to-old performance, i.e.,

$$M_{\text{new}\to\text{old}} > M_{\text{old}\to\text{old}}. \tag{15}$$

Here, $M_{\text{old}\to\text{old}}$ evaluates old-model queries against the fixed old-model gallery, whereas $M_{\text{new}\to\text{old}}$ evaluates the updated query representations against the same old-model gallery. We use Recall@K, with $K \in \{1, 5, 10\}$, as the retrieval metric M and report results for both image-to-text and text-to-image retrieval.

Inspired by [22, 13], we further analyze positive and negative retrieval flips relative to the old-to-old system. For each query $i \in \mathcal{Q}$, let $c_i^{\text{old}\to\text{old}} \in \{0, 1\}$ and $c_i^{\text{new}\to\text{old}} \in \{0, 1\}$ indicate whether retrieval is successful under the corresponding system, where success means that at least one relevant gallery item appears among the top-K retrieved items. The positive-flip rate (PFR) and negative-flip rate

Table 2: Cross-modal backward-compatible retrieval for cross-family model upgrades. For each dataset, the α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The weight $\hat{\alpha}$ is selected on CC3M and fixed across all test datasets, whereas $\alpha^\star$ is selected directly on each test dataset and is reported only as an oracle upper bound. Checkmarks indicate backward compatibility; bold and underlined values denote the best and second-best new $\rightarrow$ old results, respectively.

		Flickr30k			COCO			NoCaps		
Method	Sup.	α	I2T@1	T2I@1	α	I2T@1	T2I@1	α	I2T@1	T2I@1
SigLIP2 → CLIP ViT-B/32 *(cross-family)*										
► Old model (CLIP ViT-B/32)	–	–	40.62	21.73	–	28.76	14.47	–	71.29	45.24
SVD	T	–	42.80 ✓	23.56 ✓	–	32.33 ✓	15.62 ✓	–	72.16 ✓	46.41 ✓
+SLERP ($\hat{\alpha}$)	T	0.6/0.5	51.01 ✓	25.79 ✓	0.6/0.5	**37.35** ✓	17.18 ✓	0.6/0.5	78.56 ✓	50.59 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.6	51.34 ✓	25.83 ✓	0.6/0.6	**37.35** ✓	17.27 ✓	0.5/0.6	**78.80** ✓	50.61 ✓
SVD	I	–	48.01 ✓	23.41 ✓	–	32.28 ✓	15.20 ✓	–	72.42 ✓	46.47 ✓
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	51.33 ✓	28.57 ✓	0.7/0.4	34.86 ✓	18.79 ✓	0.7/0.4	76.18 ✓	53.86 ✓
+SLERP ($\alpha^\star$)	I	0.6/0.5	**51.40** ✓	**28.88** ✓	0.6/0.5	35.05 ✓	**18.97** ✓	0.5/0.5	77.62 ✓	**54.37** ✓
SVD	I+T	–	36.25 ×	23.93 ✓	–	28.16 ×	15.71 ✓	–	66.24 ×	48.66 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.4/0.5	47.97 ✓	27.09 ✓	0.4/0.5	34.81 ✓	17.82 ✓	0.4/0.5	77.40 ✓	53.04 ✓
+SLERP ($\alpha^\star$)	I+T	0.4/0.6	47.97 ✓	27.10 ✓	0.5/0.6	34.86 ✓	17.84 ✓	0.4/0.6	77.40 ✓	53.23 ✓
► New model (SigLIP2)	–	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*										
► Old model (CLIP ViT-H/14)	–	–	59.38	43.07	–	43.50	28.56	–	84.27	63.53
SVD	T	–	60.50 ✓	29.91 ×	–	41.82 ×	23.18 ×	–	81.40 ×	56.13 ×
+SLERP ($\hat{\alpha}$)	T	0.5/0.1	64.81 ✓	43.38 ✓	0.5/0.1	**46.56** ✓	28.79 ✓	0.5/0.1	85.89 ✓	63.89 ✓
+SLERP ($\alpha^\star$)	T	0.6/0.2	**65.10** ✓	43.44 ✓	0.5/0.2	**46.56** ✓	28.88 ✓	0.4/0.2	**86.00** ✓	63.96 ✓
SVD	I	–	48.93 ×	29.13 ×	–	26.43 ×	23.21 ×	–	70.31 ×	55.05 ×
+SLERP ($\hat{\alpha}$)	I	0.0/0.3	59.38 ×	43.25 ✓	0.0/0.3	43.50 ×	**29.04** ✓	0.0/0.3	84.27 ×	63.80 ✓
+SLERP ($\alpha^\star$)	I	0.4/0.2	61.19 ✓	**43.50** ✓	0.2/0.3	44.00 ✓	**29.04** ✓	0.2/0.2	84.93 ✓	63.91 ✓
SVD	I+T	–	58.54 ×	30.03 ×	–	34.47 ×	23.58 ×	–	77.40 ×	56.60 ×
+SLERP ($\hat{\alpha}$)	I+T	0.1/0.1	61.09 ✓	43.36 ✓	0.1/0.1	44.33 ✓	28.80 ✓	0.1/0.1	85.04 ✓	63.87 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.2	64.74 ✓	43.43 ✓	0.4/0.3	45.65 ✓	28.93 ✓	0.3/0.3	85.76 ✓	**64.00** ✓
► New model (SigLIP1)	–	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18

(NFR) are defined as

$$\mathrm{PFR} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{1}\left[c_i^{\text{old}\rightarrow\text{old}} = 0,\ c_i^{\text{new}\rightarrow\text{old}} = 1\right], \tag{16}$$

$$\mathrm{NFR} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{1}\left[c_i^{\text{old}\rightarrow\text{old}} = 1,\ c_i^{\text{new}\rightarrow\text{old}} = 0\right]. \tag{17}$$

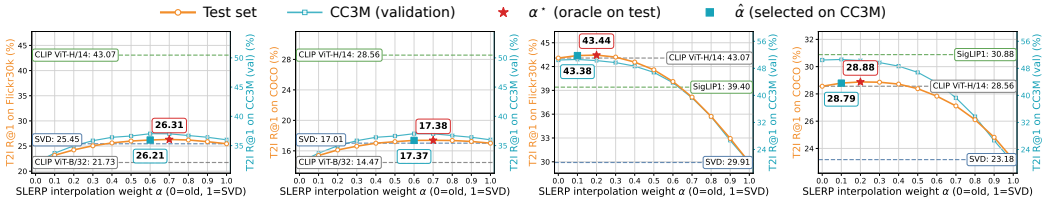
Thus, positive flips correspond to queries that become successful after the upgrade, whereas negative flips correspond to queries that become unsuccessful. Since Recall@K is the average of these binary success indicators, its change relative to the old-to-old system satisfies:

$$\Delta\mathrm{R@}K = \mathrm{PFR} - \mathrm{NFR}. \tag{18}$$

4.3 Compatibility Results

Tables 1 and 2 report cross-modal compatible retrieval results for same-family and cross-family model pairs, respectively. Across 90 evaluations—five model pairs, three support modalities, three datasets, and both retrieval directions—CC3M-selected SLERP satisfies the Recall@1 compatibility criterion in Eq. 15 in 85 cases, compared with 45 for SVD alone. The additional model-pair results included in this aggregate, together with the full Recall@1/5/10 results, are reported in Appendix I. SVD thus provides an effective post-hoc alignment but does not consistently satisfy the compatibility criterion, particularly for cross-family pairs. This is consistent with the residual angular discrepancy in Fig. 1: Procrustes alignment maps new representations into the old-model coordinate system but does not fully resolve the cross-model mismatch.

For same-family upgrades, our approach remains competitive with available XBT results without any specific compatibility training. In the harder SigLIP1 → CLIP ViT-H/14 setting, where the new model is not uniformly stronger than the old one, SLERP still achieves compatibility in most cases. These results suggest that the gains arise from combining old and SVD-aligned new representations,



(a) CLIP ViT-H/14 → CLIP ViT-B/32 (b) SigLIP1 → CLIP ViT-H/14

Figure 3: Text-to-image Recall@1 along the SLERP interpolation path for Flickr30k (left) and COCO (right). The image gallery is encoded by the old model, while text queries interpolate between the old-model query ($\alpha = 0$) and the SVD-aligned new-model query ($\alpha = 1$). Markers indicate the CC3M-selected $\hat{\alpha}$ and the dataset-specific oracle $\alpha^\star$; dashed lines report the old-model, new-model, and SVD performance.

rather than solely from a stronger new encoder. Fig. 3 illustrates this behavior for T2I Recall@1, with performance often maximized at an interior interpolation weight. The CC3M-selected $\hat{\alpha}$ closely tracks the dataset-specific oracle $\alpha^\star$, showing that a single support-set-selected weight transfers to Flickr30k, COCO, and NoCaps without test-set tuning.

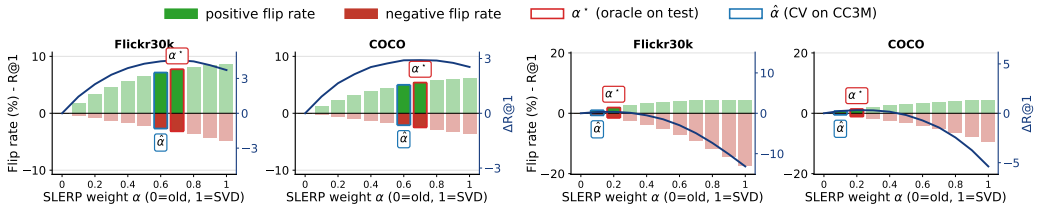
Appendix H reports the corresponding I2T curves, which show the same pattern. Across different support set modalities, text-only alignment is the most stable choice overall. Image-only support, as in [35], and joint support remain competitive in individual cases, but text embeddings appear to provide cleaner semantic anchors for estimating the old–new alignment. Moreover, Appendix M provides a complementary per-query oracle analysis on Flickr30k I2T retrieval for CLIP ViT-L/14 → CLIP ViT-B/32. Among queries achieving Recall@1 at some evaluated interpolation weight, 97.65% are assigned an interior oracle weight. The oracle uses query-level test labels and is therefore non-deployable, it provides observable evidence consistent with the geometric characterization in Sec. 3: retrieval-favorable query representations frequently occur in the interior of the interpolation arc rather than at either endpoint. Appendix J further shows that SLERP often improves over the new model even when re-indexing is allowed, indicating that the benefit of interpolation is not limited to the compatibility-only setting.

Additionally, in Appendix K we evaluate the setting in which a small subset from the target test distribution is available for estimating the SLERP weight. The results show that even a modest subset is sufficient to select a reliable dataset-specific interpolation weight. We further show in Appendix N that the gains are not specific to SLERP: a fixed normalized midpoint, support-set selected NLERP, and score interpolation yield comparable average improvements. We use SLERP for its constant-angular-speed parameterization, which gives the interpolation weight a consistent geometric interpretation across queries.

Flip analysis. Fig. 4 decomposes the T2I Recall@1 change along the interpolation path into positive and negative flips relative to old-to-old retrieval. Since $\Delta\mathrm{R@1} = \mathrm{PFR} - \mathrm{NFR}$, the optimal interpolation weight is determined by the balance between the two rates rather than by positive flips alone. Moving from the old-model query toward the aligned new-model query initially corrects more old-model failures than it introduces regressions; beyond the optimum, negative flips grow faster than positive ones and the gain decreases. For CLIP ViT-H/14 → CLIP ViT-B/32, where SVD alignment alone already exceeds the old model, the gain peaks close to the SVD endpoint ($\alpha^\star = 0.7$). For SigLIP1 → CLIP ViT-H/14, where the SVD-aligned query induces a high negative-flip rate, the gain peaks close to the old-model endpoint ($\alpha^\star = 0.2$), and interpolation retains most old-model successes while still recovering part of the new-model gains. In both cases, the CC3M-selected $\hat{\alpha}$ lies near the peak. Appendix L extends the analysis to I2T retrieval and NoCaps dataset.

5 Conclusion

We studied the problem of upgrading contrastive vision-language retrieval systems without re-encoding the deployed gallery. Motivated by the residual angular discrepancy left by orthogonal post-hoc alignment, we proposed a query-side approach that first maps new-model queries into the old-model representation space through orthogonal Procrustes alignment and then interpolates them with



(a) CLIP ViT-H/14 → CLIP ViT-B/32 (b) SigLIP1 → CLIP ViT-H/14

Figure 4: Flip-rate trade-off for T2I Recall@1. Bars show positive and negative flip rates relative to the old-to-old evaluation. The blue curve reports ΔR@1. Here $\alpha = 0$ corresponds to old-model queries and $\alpha = 1$ to SVD-aligned new-model queries. Red and blue markers denote the dataset-specific oracle $\alpha^\star$ and the CC3M-selected weight $\hat{\alpha}$, respectively.

the corresponding old-model queries along the spherical geodesic. Our analysis characterizes when this interpolation path contains an interior direction closer to an idealized retrieval-optimal direction than either endpoint, and connects this angular improvement to Recall@K through a local margin-based sufficient condition. Empirically, we validated the method across CLIP, SigLIP1, and SigLIP2 upgrades, covering same-family and cross-family pairs and Flickr30k, COCO, and NoCaps retrieval benchmarks. SLERP improves over SVD alignment in nearly all settings, often restores empirical backward-compatibility, and frequently matches or surpasses training-based compatibility baselines while requiring no compatibility training. The support-set selected interpolation weights transfer reliably across datasets, and additional evaluations show that the same mechanism also extends to re-indexed retrieval and zero-shot classification. Overall, these results show that residual post-alignment geometry contains useful retrieval signal and that post-alignment interpolation provides a simple and effective mechanism for backward-compatible vision-language model upgrades.

Limitations. SLERP also involves practical trade-offs, discussed in Appendix O: the interpolation weight must be selected on a support set, query-side inference requires both endpoint embeddings, and performance remains bounded by the quality and complementarity of the available representations. In our experiments, these trade-offs are favorable, suggesting that spherical interpolation provides a simple, robust approach for upgrading contrastive vision-language retrieval systems while preserving backward-compatibility. Moreover, it provides a compatibility mechanism during model migration, allowing the deployed gallery to remain usable while full re-indexing is deferred or performed offline, as described in Appendix P.

Acknowledgments and Disclosure of Funding

Funding: This work was partially supported by research funds from the Department of Information Engineering, University of Florence, and by the EU Horizon Europe projects ELIAS (No. 101120237) and ELLIOT (No. 101214398), by the FIS project GUIDANCE (No. FIS2023-03251).
Competing interests: The authors declare no competing interests.

References

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

[2] 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*, pages 4904–4916. PMLR, 2021.

[3] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[4] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[5] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[7] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

[8] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *CoRR*, abs/1702.08734, 2017.

[10] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6367–6376. IEEE, 2020.

[11] Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, and Feng Zhou. Learning compatible embeddings. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9919–9928. IEEE, 2021.

[12] Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19386–19395, 2022.

[13] Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Fanzi Wu, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Towards universal backward-compatible representation learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1615–1621. IJCAI, 2022.

[15] Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. Stationary representations: Optimally approximating compatibility and implications for improved model replacements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[18] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[19] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[20] Colin Raffel. Building machine learning models like open source software. *Communications of the ACM*, 66(2):38–40, 2023.

[21] Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *Transactions on Machine Learning Research*, 2025.

[22] Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14299–14308, 2021.

[23] Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A model update strategy for compatible LLM evolution. In *EMNLP (Findings)*, pages 7320–7332. Association for Computational Linguistics, 2024.

[24] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11, 2019.

[25] Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Mitigating negative flips via margin preserving training. In *AAAI*, pages 8721–8730. AAAI Press, 2026.

[26] Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. A stationary (and therefore compatible) representation is all you need. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[27] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Feature Extraction: Modern Questions and Challenges*, pages 196–212. PMLR, 2015.

[28] Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. Bt^ 2: Backward-compatible training with basis transformation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11195–11204. IEEE, 2023.

[29] Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Backward-compatible aligned representations via an orthogonal transformation layer. In *ECCV Workshops (17)*, volume 15639 of *Lecture Notes in Computer Science*, pages 451–464. Springer, 2024.

[30] Simone Ricci, Niccolò Biondi, Federico Pernici, Ioannis Patras, and Alberto Del Bimbo. λ-orthogonality regularization for compatible representation learning. In *NeurIPS*, 2025.

[31] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[32] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

[33] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55394–55414. Curran Associates, Inc., 2023.

[34] Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. In *NeurIPS*, 2024.

[35] Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas Garg. Canonicalizing multimodal contrastive representation learning. *CoRR*, abs/2602.17584, 2026.

[36] Lirong Wu, Zicheng Liu, Jun Xia, Zelin Zang, Siyuan Li, and Stan Z Li. Generalized clustering and multi-manifold learning with geometric structure preservation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 139–147, 2022.

[37] A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into plato's cave: Examining cross-modal representational. *arXiv preprint arXiv:2604.18572*, 2026.

[38] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pages 245–254. ACM, 1985.

[39] Shengsen Wu, Liang Chen, Yihang Lou, Yan Bai, Tao Bai, Minghua Deng, and Ling-Yu Duan. Neighborhood consensus contrastive learning for backward-compatible representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2722–2730, 2022.

[40] Binjie Zhang, Yixiao Ge, Yantao Shen, Yu Li, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Hot-refresh model upgrades with regression-free compatible training in image retrieval. In *International Conference on Learning Representations*, 2022.

[41] Niccolo Biondi, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo. Cores: Compatible representations via stationarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[42] Niccolo Biondi, Federico Pernici, Matteo Bruni, Daniele Mugnai, and Alberto Del Bimbo. Cl2r: Compatible lifelong learning representations. *ACM Transactions on Multimedia Computing, Communications and Applications*, 18(2s):1–22, 2023.

[43] Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Regular polytope networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[44] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[45] Binjie Zhang, Shupeng Su, Yixiao Ge, Xuyuan Xu, Yexin Wang, Chun Yuan, Mike Zheng Shou, and Ying Shan. Darwinian model upgrades: Model evolving with selective compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3393–3400, 2023.

[46] Ngoc Bui, Menglin Yang, Runjin Chen, Leonardo Neves, Mingxuan Ju, Rex Ying, Neil Shah, and Tong Zhao. Learning along the arrow of time: Hyperbolic geometry for backward-compatible representation learning. In *Forty-second International Conference on Machine Learning*, 2025.

[47] Zikun Zhou, Yushuai Sun, Wenjie Pei, Xin Li, and Yaowei Wang. Prototype perturbation for relaxing alignment constraints in backward-compatible learning. *IEEE Transactions on Multimedia*, 2026.

[48] Anonymous. Hyperspherical simplex representations from softmax outputs and logits are inherently backward-compatible, 2025. Under review at Transactions on Machine Learning Research.

[49] Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L2r: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 183–192, 2023.

[50] Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *CVPR*, pages 16614–16623. IEEE, 2024.

[51] Young Kyun Jang and Ser-nam Lim. Towards cross-modal backward-compatible representation learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1783–1792, 2025.

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

[53] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2018.

[54] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. Pmlr, 2022.

[55] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

[56] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

[57] Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754, 2023.

[58] Zhixu Tao, Ian Mason, Sanjeev R. Kulkarni, and Xavier Boix. Task arithmetic through the lens of one-shot federated learning. *Trans. Mach. Learn. Res.*, 2025.

[59] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. In *European conference on computer vision*, pages 239–254. Springer, 2024.

[60] Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. WARP: On the benefits of weight averaged rewarded policies. *CoRR*, abs/2406.16768, 2024.

[61] Erik B. Dam, Martin Koch, and Martin Lillholm. Quaternions, interpolation and animation. Technical Report DIKU-TR-98/5, Department of Computer Science, University of Copenhagen, 1998.

[62] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[63] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[64] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[65] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

[66] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[67] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[68] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956. IEEE, 2019.

[69] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[70] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

[71] Chang Wang and Sridhar Mahadevan. A general framework for manifold alignment. In *AAAI fall symposium: manifold learning and its applications*, pages 79–86, 2009.

[72] Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *ICML*, ACM International Conference Proceeding Series, pages 1120–1127. ACM, 2008.

[73] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

[74] Akshit Achara, Tatiana Gaintseva, Mateo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. *arXiv preprint arXiv:2602.06205*, 2026.

[75] Simone Ricci, Tiberio Uricchio, and Alberto Del Bimbo. Meta-learning advisor networks for long-tail and noisy labels in social image classification. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(5s):169:1–169:23, 2023.

[76] Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035*, 2024.

[77] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[78] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[79] Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *ICLR*, 2025.

[80] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in CLIP via modality inversion. In *ICLR*, 2025.

[81] Simone Magistri, Dipam Goswami, Marco Mistretta, Bartlomiej Twardowski, Joost van de Weijer, and Andrew D. Bagdanov. Isoclip: Decomposing CLIP projectors for efficient intra-modal alignment. In *CVPR*, pages 29315–29324. Computer Vision Foundation, 2026.

[82] Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. Infonce induces gaussian distribution. In *International Conference on Learning Representations (ICLR)*, 2026.

[83] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[84] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[85] Andrei N Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Sov Dok*, 4:1035–1038, 1963.

[86] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[87] Seonguk Seo, Mustafa Gokhan Uzunbas, Bohyung Han, Sara Cao, and Ser-Nam Lim. Metric compatible training for online backfilling in large-scale retrieval. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1537–1545. IEEE, 2025.

Appendix Contents

A Additional Related Work **17**

B Extension to Unequal Embedding Dimensions via Rectangular Procrustes **17**

C Relationship Between SLERP and NLERP **20**

D Decomposition of the Retrieval-Optimal Direction **21**

E Proof of Theorem 1 **22**

F Connection Between SLERP and Recall@K **24**

F.1 Recall@K with Multiple Relevant Items 24

F.2 Top-K Margin 25

F.3 Retrieval-Relevant Optimal Direction 25

F.4 Lipschitz Stability of the Top-K Margin 26

F.5 Local Certification of Recall@K 26

G Cross-Modal Zero-shot Classification **27**

H Additional SLERP Arc Results **27**

I Detailed Retrieval Performance **27**

J Re-indexing **32**

K Target-Set Budget Sensitivity **33**

L Additional Flip Analyses **34**

M Per-Query Oracle Analysis **34**

N Alignment and Endpoint-Combination Baselines **36**

N.1 Alignment baselines 37

N.2 Endpoint-combination baselines 37

N.3 Discussion 38

O Limitations **39**

P Deployment Cost and Migration-Period Analysis **39**

P.1 Query-Side Cost and Re-indexing Trade-off 40

P.2 Migration Period with Partial Backfilling 40

A Additional Related Work

Post-hoc alignment and cross-model correspondence. A complementary line of work studies whether the latent spaces of independently trained models can be aligned after training, without modifying the original model weights. Classical manifold alignment and Procrustes methods provide the foundation for this approach by seeking simple correspondences between representation spaces, either from paired anchors or by preserving the geometric structure of the underlying data manifolds [69, 70, 71, 72, 73]. The central assumption is that independently trained models may encode similar semantic structure in different coordinate systems; consequently, a lightweight post-hoc transformation can often make their representations comparable. Recent work supports this view in latent representation spaces, showing that useful cross-model correspondences can be recovered through semantic latent translation and latent functional maps [33, 34]. These results suggest that exact equality between representation spaces is not required for interoperability—a transformation that preserves the task-relevant structure shared across models can be sufficient. This perspective is consistent with broader evidence that independently trained models can develop similar representation geometry across architectures and modalities [27, 32]. However, such convergence is generally approximate rather than exact, and recent results indicate that cross-modal agreement may weaken at realistic scales and in many-to-many settings [37]. Most closely related to our setting, [35] shows that independently trained contrastive VLMs can often be approximately canonicalized by a single orthogonal map shared across modalities, estimated from a small anchor set and without retraining either model. Recent work further extends this viewpoint beyond pairwise alignment by constructing a shared orthogonal reference space across multiple models and applying a retrieval-oriented correction [74]. These works support the assumption that a new-model query can be mapped into the coordinate system of the deployed gallery model with post-hoc orthogonal alignment. Our work focuses on the angular discrepancy after this alignment step: even in a common coordinate system, embeddings of the same input may differ due to architecture, training data (e.g., label noise and long-tailed class distributions [75]), optimization, or inductive bias, and may therefore induce different rankings over the same fixed gallery. We study this residual discrepancy directly by characterizing when spherical interpolation between the old-model query and the aligned new-model query improves retrieval.

A related line of work combines independently trained embedding models directly in parameter space. In particular, [76] studies model merging for general text embeddings by searching over combinations of task vectors, including SLERP-based interpolation; unlike our setting, this operates in model-parameter space and assumes models with the same architecture, whereas we interpolate query representations after cross-model alignment.

Geometry of contrastive vision-language spaces. Our analysis builds on work showing that contrastive VLMs learn a normalized image–text embedding space in which cross-modal comparisons are performed by cosine similarity, but whose geometry is not fully homogeneous or modality-invariant [1, 77]. Prior work shows that image and text embeddings can occupy separated regions of the representation space, leading to modality gaps [78, 79]. Other studies identify intra-modal misalignment within individual encoders [80], as well as spectral decompositions into approximately isotropic shared components and anisotropic modality-specific directions [81]. Recent work further suggests that InfoNCE-trained representations may exhibit approximately Gaussian embedding distributions [82]. Together, these findings indicate that global comparability does not imply full geometric equivalence: a common coordinate system may preserve cross-modal similarity while still retaining model-specific or modality-specific structure. This perspective helps explain why an alignment estimated from one modality can transfer to the other [35], while also leaving residual angular discrepancies after alignment. In our setting, this residual angular structure is precisely the object of interest, and it motivates interpolation between the old-model query and the aligned new-model query.

B Extension to Unequal Embedding Dimensions via Rectangular Procrustes

We describe how the alignment obtained by solving the Procrustes problem used in the main text extends when the old model and the new model have different embedding dimensions. Let the old model, which defines the deployed gallery space, produce normalized embeddings in $\mathbb{R}^{d_{\text{old}}}$, and let the new model produce normalized embeddings in $\mathbb{R}^{d_{\text{new}}}$. Given an alignment support set, denote by

$$U \in \mathbb{R}^{N_a \times d_{\text{old}}}, \qquad \bar{V} \in \mathbb{R}^{N_a \times d_{\text{new}}} \tag{19}$$

the row-stacked old-model and new-model embeddings, respectively. The goal is to construct an orthogonal alignment map

$$R \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}} \tag{20}$$

such that $\bar{V}R$ gives new-model embeddings expressed in the old-model coordinate system. We refer to $\bar{V}R$ as the new aligned representation.

The construction is based on the support-set cross-covariance matrix

$$C := \bar{V}^\top U \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}}. \tag{21}$$

When $d_{\text{new}} = d_{\text{old}}$, this reduces to the square orthogonal Procrustes problem used in the main text. When $d_{\text{new}} \neq d_{\text{old}}$, we use a semi-orthogonal Procrustes factor obtained by maximizing the paired agreement between old-model and new-model embeddings of the support set:

$$R^\star \in \arg\max_{R \in \mathcal{O}_{d_{\text{new}}, d_{\text{old}}}} \operatorname{tr}(R^\top C), \tag{22}$$

where

$$\mathcal{O}_{d_{\text{new}}, d_{\text{old}}} := \begin{cases} \{R \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}} : RR^\top = I_{d_{\text{new}}}\}, & d_{\text{new}} \leq d_{\text{old}}, \\ \{R \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}} : R^\top R = I_{d_{\text{old}}}\}, & d_{\text{new}} \geq d_{\text{old}}. \end{cases} \tag{23}$$

Thus, the alignment is always a partial isometry from the new-model representation space to the old-model gallery space. The two unequal-dimensional cases have different geometric interpretations.

First consider the case $d_{\text{new}} < d_{\text{old}}$. Since the new-model representation has lower dimension than the old-model representation, it can be embedded isometrically into the old-model gallery space. In this case, the trace maximization above is equivalent to the rectangular least-squares Procrustes problem

$$R^\star = \arg\min_{RR^\top = I_{d_{\text{new}}}} \|\bar{V}R - U\|_F^2, \tag{24}$$

because the constraint $RR^\top = I_{d_{\text{new}}}$ makes $\|\bar{V}R\|_F^2 = \|\bar{V}\|_F^2$ independent of R.

Let

$$C = \bar{V}^\top U = P\Sigma Q^\top \tag{25}$$

be a thin singular value decomposition, with

$$P \in \mathbb{R}^{d_{\text{new}} \times d_{\text{new}}}, \qquad Q \in \mathbb{R}^{d_{\text{old}} \times d_{\text{new}}}. \tag{26}$$

The rectangular Procrustes solution is

$$R^\star = PQ^\top \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}}. \tag{27}$$

Because

$$R^\star R^{\star\top} = PQ^\top QP^\top = I_{d_{\text{new}}}, \tag{28}$$

the map preserves all inner products within the new-model space. Indeed, for any $x, y \in \mathbb{R}^{d_{\text{new}}}$,

$$\langle xR^\star, yR^\star \rangle = xR^\star R^{\star\top} y^\top = xy^\top = \langle x, y \rangle. \tag{29}$$

Thus, when $d_{\text{new}} < d_{\text{old}}$, the new aligned representation is an isometric embedding of the new-model representation into the old-model gallery space. Norms, angles, inner products, and cosine similarities between new-model embeddings are preserved exactly.

We now consider the case $d_{\text{new}} > d_{\text{old}}$. In this setting, the new-model representation has higher dimension than the old-model gallery space, so no linear map from $\mathbb{R}^{d_{\text{new}}}$ to $\mathbb{R}^{d_{\text{old}}}$ can preserve all inner products. The rectangular Procrustes map should therefore be interpreted as a partial isometry: it extracts the d_{old}-dimensional component of the new-model representation that is maximally aligned with the old-model gallery space.

Write the full singular value decomposition of the cross-covariance as

$$C = \bar{V}^\top U = [P\ N] \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^\top, \tag{30}$$

where

$$P \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}}, \qquad N \in \mathbb{R}^{d_{\text{new}} \times (d_{\text{new}} - d_{\text{old}})}, \qquad Q \in \mathbb{R}^{d_{\text{old}} \times d_{\text{old}}}. \tag{31}$$

Here, the columns of P span the d_{old}-dimensional subspace of the new-model representation that is aligned with the old-model gallery space, while the columns of N span its orthogonal complement. The rectangular Procrustes map is

$$R^\star = PQ^\top \in \mathbb{R}^{d_{\text{new}} \times d_{\text{old}}}. \tag{32}$$

It satisfies

$$R^{\star\top} R^\star = I_{d_{\text{old}}}, \qquad R^\star R^{\star\top} = PP^\top, \qquad PP^\top + NN^\top = I_{d_{\text{new}}}. \tag{33}$$

Therefore, any new-model embedding $\bar{v} \in \mathbb{R}^{d_{\text{new}}}$ admits the orthogonal decomposition

$$\bar{v} = \bar{v}PP^\top + \bar{v}NN^\top. \tag{34}$$

The first term is the component of $\bar{v}$ that is visible after alignment to the old-model gallery space, while the second term is the residual component that lies outside the old model-aligned subspace.

Equivalently, one can represent the new-model embedding in an augmented space as

$$\tilde{v} := (\bar{v}R^\star, \bar{v}N) \in \mathbb{R}^{d_{\text{old}}} \times \mathbb{R}^{d_{\text{new}} - d_{\text{old}}}, \tag{35}$$

and represent each old-model gallery embedding $u \in \mathbb{R}^{d_{\text{old}}}$ as

$$\tilde{u} := (u, 0). \tag{36}$$

This augmented representation preserves the full geometry of the new-model space. Indeed, for any new-model embeddings $\bar{v}, \bar{w} \in \mathbb{R}^{d_{\text{new}}}$,

$$\langle \tilde{v}, \tilde{w} \rangle = \langle \bar{v}R^\star, \bar{w}R^\star \rangle + \langle \bar{v}N, \bar{w}N \rangle = \bar{v}(PP^\top + NN^\top)\bar{w}^\top = \langle \bar{v}, \bar{w} \rangle. \tag{37}$$

At the same time, the residual component is inert for retrieval against the deployed old-model gallery, since

$$\langle \tilde{v}, \tilde{u} \rangle = \langle \bar{v}R^\star, u \rangle. \tag{38}$$

Thus, when $d_{\text{new}} > d_{\text{old}}$, the projected vector $\bar{v}R^\star$ is the only component of the new aligned representation that affects cross-model retrieval in the old-model gallery space. The residual block $\bar{v}N$ preserves the remaining new-model geometry, but it has zero interaction with old-model gallery embeddings represented as $(u, 0)$.

In practice, retrieval can be performed directly in the old-model gallery space using the projected new aligned embedding $\bar{v}R^\star$. When $d_{\text{new}} > d_{\text{old}}$, this projection is not generally unit-norm:

$$\|\bar{v}R^\star\|_2^2 = \bar{v}PP^\top\bar{v}^\top \le \|\bar{v}\|_2^2. \tag{39}$$

Multiplying all scores for a fixed query by a positive scalar does not change the induced ranking, so this norm reduction does not affect nearest-neighbor retrieval by itself. However, the SLERP analysis requires both endpoints to lie on the old-model unit sphere. We therefore normalize the projected new aligned query before interpolation:

$$v(x) := \frac{\phi_{\text{new}}(x)R^\star}{\|\phi_{\text{new}}(x)R^\star\|_2}, \qquad \text{provided } \|\phi_{\text{new}}(x)R^\star\|_2 > 0. \tag{40}$$

For $d_{\text{new}} \le d_{\text{old}}$, the alignment is norm-preserving, so this normalization is redundant. With this convention, the old-model query

$$u(x) := \phi_{\text{old}}(x) \in \mathbb{S}^{d_{\text{old}}-1} \tag{41}$$

and the new aligned-model query

$$v(x) \in \mathbb{S}^{d_{\text{old}}-1} \tag{42}$$

both lie on the same unit sphere in the old-model coordinate system. Consequently, the SLERP characterization in the main text applies without modification.

C Relationship Between SLERP and NLERP

This appendix clarifies the relationship between spherical linear interpolation (SLERP) and normalized linear interpolation (NLERP) in our setting. Both methods can interpolate between the same two normalized query endpoints: the old-model query embedding u and the SVD-aligned new-model query embedding v, with $\theta = \angle(u, v) \in (0, \pi)$. The difference is not the set of directions that can be reached, but the way in which the interpolation path is parameterized.

Recall that SLERP is defined as

$$q_\alpha^{\mathrm{slerp}} = \frac{\sin((1-\alpha)\theta)}{\sin\theta} u + \frac{\sin(\alpha\theta)}{\sin\theta} v, \qquad \alpha \in [0, 1]. \tag{43}$$

By contrast, NLERP first takes a Euclidean convex combination of the two endpoints and then renormalizes it:

$$q_\beta^{\mathrm{nlerp}} = \frac{(1-\beta)u + \beta v}{\|(1-\beta)u + \beta v\|_2}, \qquad \beta \in [0, 1]. \tag{44}$$

Thus, NLERP also remains on the unit sphere, but its interpolation weight does not correspond to constant angular motion along the arc.

To make the relationship explicit, use the same orthonormal basis as in Sec. 3,

$$e_1 = u, \qquad e_2 = \frac{v - \cos\theta\, u}{\sin\theta}, \tag{45}$$

so that $v = \cos\theta\, e_1 + \sin\theta\, e_2$. The unnormalized NLERP numerator can then be written as

$$(1-\beta)u + \beta v = \big(1 - \beta + \beta\cos\theta\big)e_1 + \beta\sin\theta\, e_2. \tag{46}$$

Therefore, after normalization, the NLERP point has angular coordinate

$$\tau(\beta) = \mathrm{atan2}(\beta\sin\theta,\, 1 - \beta + \beta\cos\theta) \in [0, \theta], \tag{47}$$

and

$$q_\beta^{\mathrm{nlerp}} = \cos(\tau(\beta))e_1 + \sin(\tau(\beta))e_2 = q_{\tau(\beta)/\theta}^{\mathrm{slerp}}. \tag{48}$$

Moreover,

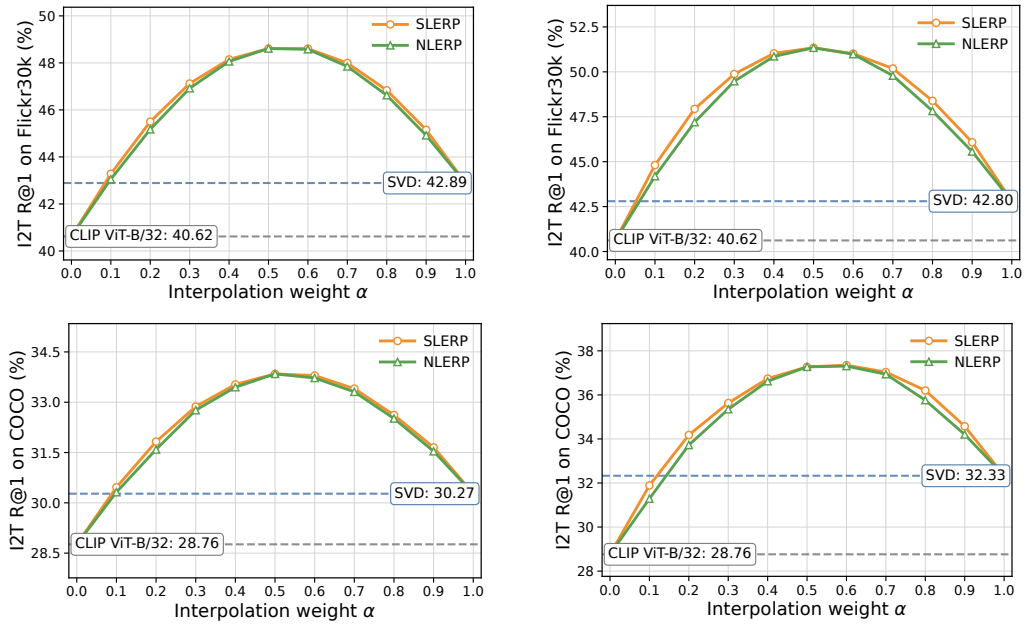
$$\frac{d\tau}{d\beta} = \frac{\sin\theta}{\|(1-\beta)u + \beta v\|_2^2} > 0, \tag{49}$$

so $\tau(\beta)$ is strictly increasing on $[0, 1]$. Consequently, NLERP traces exactly the same minor geodesic arc as SLERP, but under a monotone reparameterization. Conversely, the inverse mapping from a SLERP weight α to the NLERP weight that reaches the same direction is

$$\beta(\alpha) = \frac{\sin(\alpha\theta)}{\sin((1-\alpha)\theta) + \sin(\alpha\theta)}. \tag{50}$$

This equivalence has an immediate implication for continuous weight selection. For any retrieval objective that depends only on the resulting query direction, optimizing over $\alpha \in [0, 1]$ with SLERP and optimizing over $\beta \in [0, 1]$ with NLERP search over the same set of unit vectors. Thus, in the continuous setting, the two methods have the same optimal directions. The distinction is instead geometric and practical: SLERP moves at constant angular speed, since $\angle(u, q_\alpha^{\mathrm{slerp}}) = \alpha\theta$, whereas NLERP moves non-uniformly along the same arc. Using SLERP therefore makes the interpolation weight directly interpretable as a normalized angular displacement from the old-model query toward the SVD-aligned new-model query.

Fig. 5 empirically compares SLERP and NLERP on Flickr30k and COCO retrieval after SVD alignment. The two curves are nearly identical across the interpolation path and achieve their best performance in the same interior region of the arc. The small visible differences are consistent with the reparameterization above: a uniform grid in the NLERP weight β is not a uniform grid in angular position, and therefore does not sample exactly the same directions as a uniform grid in the SLERP weight α. Overall, this comparison supports our use of SLERP not because it accesses a different family of query representations, but because it provides a geometrically canonical parameterization of the same interpolation arc.



(a) CLIP ViT-L/14 → CLIP ViT-B/32.

(b) SigLIP2 → CLIP ViT-B/32.

Figure 5: Empirical comparison of SLERP and NLERP for text-only image-to-text retrieval after SVD alignment. The columns correspond to the two source-target model pairs: CLIP ViT-L/14 → CLIP ViT-B/32 on the left and SigLIP2 → CLIP ViT-B/32 on the right. The rows correspond to the two retrieval datasets: Flickr30k on the top row and COCO on the bottom row. Across Flickr30k and COCO, and for both source-model pairs, the two interpolation schemes trace nearly identical retrieval curves and attain their best performance in the same interior region of the interpolation path. The small discrepancies arise because a uniform grid in the NLERP weight does not correspond to a uniform grid in angular position along the shared geodesic arc. This supports using SLERP as a geometrically canonical parameterization rather than as a different family of query representations.

D Decomposition of the Retrieval-Optimal Direction

Lemma 1 (Decomposition relative to the interpolation plane)**.** *Let $q^* \in \mathbb{S}^{d-1}$. There exist unique vectors $p \in \operatorname{span}(u, v)$ and $w_\perp \perp \operatorname{span}(u, v)$ such that $q^* = p + w_\perp$. Let $\rho := \|p\|$. Then $\rho \in [0, 1]$ and $\|w_\perp\|^2 = 1 - \rho^2$. If $\rho > 0$, there exists a unique $\psi \in (-\pi, \pi]$ such that*

$$p = \rho(\cos\psi\, e_1 + \sin\psi\, e_2). \tag{51}$$

Proof. Let $\mathcal{P} := \operatorname{span}(u, v)$. Since $\mathcal{P}$ is a linear subspace of $\mathbb{R}^d$, the orthogonal projection of q^* onto $\mathcal{P}$ is unique. Denote this projection by p and define

$$w_\perp := q^* - p. \tag{52}$$

Then $p \in \mathcal{P}$, $w_\perp \perp \mathcal{P}$, and

$$q^* = p + w_\perp. \tag{53}$$

The uniqueness of the orthogonal projection also gives the uniqueness of this decomposition.

Because p and $w_\perp$ are orthogonal, Pythagoras gives

$$\|q^*\|^2 = \|p\|^2 + \|w_\perp\|^2. \tag{54}$$

Since $q^* \in \mathbb{S}^{d-1}$, we have $\|q^*\| = 1$. Therefore, with $\rho := \|p\|$, it follows that $\rho \in [0, 1]$ and

$$\|w_\perp\|^2 = 1 - \rho^2. \tag{55}$$

If $\rho > 0$, then p/ρ is a unit vector in the two-dimensional plane $\mathcal{P}$. Since (e_1, e_2) is an orthonormal basis of $\mathcal{P}$, there exist unique coefficients $a, b \in \mathbb{R}$ such that

$$\frac{p}{\rho} = ae_1 + be_2, \qquad a^2 + b^2 = 1. \tag{56}$$

Hence there exists a unique angle $\psi \in (-\pi, \pi]$ such that $a = \cos\psi$ and $b = \sin\psi$. Thus,

$$p = \rho(\cos\psi\, e_1 + \sin\psi\, e_2), \tag{57}$$

which proves the claim. □

E Proof of Theorem 1

Theorem 1 (Angular retrieval gap along the SLERP arc). *Let $q_\alpha = \mathrm{slerp}(u, v; \alpha)$ with $\theta = \angle(u, v) \in (0, \pi)$, and let ρ and ψ be defined as in Lemma 1 with respect to the canonical basis $e_1 = u$ and $e_2 = (v - \cos\theta\, u)/\sin\theta$. Then:*

(i) If $\rho > 0$, then $\langle q_\alpha, q^* \rangle = \rho\cos(\alpha\theta - \psi)$, $G(\alpha) = \arccos\big(\rho\cos(\alpha\theta - \psi)\big)$.

(ii) If $\rho = 0$, then $G(\alpha) \equiv \pi/2$, $\forall \alpha \in [0, 1]$.

(iii) If $\rho > 0$, the minimizers of $G(\alpha)$ over $\alpha \in [0, 1]$ coincide with the minimizers of the circular distance $d_{\mathrm{circ}}(\alpha\theta, \psi)$ over $\alpha \in [0, 1]$, where $d_{\mathrm{circ}}(\tau, \psi) := \min_{k\in\mathbb{Z}} |\tau - \psi - 2\pi k|$. Equivalently, every minimizer satisfies

$$\alpha^* = \frac{\tau^*}{\theta}, \qquad \tau^* \in \arg\min_{\tau\in[0,\theta]} d_{\mathrm{circ}}(\tau, \psi). \tag{58}$$

For $\rho > 0$, a strict interior improvement $G(\alpha^*) < \min\{G(0), G(1)\}$ holds if and only if the normalized in-plane projection of q^* lies strictly in the relative interior of the minor arc from u to v.

Proof. Since $\theta = \angle(u, v) \in (0, \pi)$, the vectors u and v are distinct and non-antipodal, and hence span a two-dimensional interpolation plane. Define the orthonormal basis

$$e_1 := u, \qquad e_2 := \frac{v - \langle u, v\rangle\, u}{\|v - \langle u, v\rangle\, u\|} = \frac{v - \cos\theta\, u}{\sin\theta}. \tag{59}$$

Then

$$u = e_1, \qquad v = \cos\theta\, e_1 + \sin\theta\, e_2. \tag{60}$$

Using the definition of SLERP,

$$q_\alpha = \frac{\sin((1-\alpha)\theta)}{\sin\theta} e_1 + \frac{\sin(\alpha\theta)}{\sin\theta}\big(\cos\theta\, e_1 + \sin\theta\, e_2\big). \tag{61}$$

Since

$$\sin((1-\alpha)\theta) = \sin\theta\cos(\alpha\theta) - \cos\theta\sin(\alpha\theta), \tag{62}$$

the coefficient of e_1 becomes $\cos(\alpha\theta)$, and the coefficient of e_2 becomes $\sin(\alpha\theta)$. Therefore

$$q_\alpha = \cos(\alpha\theta)e_1 + \sin(\alpha\theta)e_2. \tag{A}$$

Assume first that $\rho > 0$. By Lemma 1,

$$q^* = \rho(\cos\psi\, e_1 + \sin\psi\, e_2) + w_\perp, \qquad w_\perp \perp \mathrm{span}(u, v). \tag{63}$$

Since $q_\alpha \in \mathrm{span}(u, v)$, the orthogonal component $w_\perp$ does not contribute to the inner product. Hence, using (A),

$$\begin{aligned}\langle q_\alpha, q^* \rangle &= \langle \cos(\alpha\theta)e_1 + \sin(\alpha\theta)e_2, \rho(\cos\psi\, e_1 + \sin\psi\, e_2) + w_\perp \rangle \\ &= \rho\cos(\alpha\theta)\cos\psi + \rho\sin(\alpha\theta)\sin\psi \\ &= \rho\cos(\alpha\theta - \psi).\end{aligned}$$

Since $G(\alpha) = \arccos(\langle q_\alpha, q^* \rangle)$, this proves (i).

If $\rho = 0$, then $q^* = w_\perp$ is orthogonal to $\mathrm{span}(u, v)$. Since every q_α lies in this plane,

$$\langle q_\alpha, q^* \rangle = 0, \qquad \forall \alpha \in [0, 1]. \tag{64}$$

Thus

$$G(\alpha) = \arccos(0) = \frac{\pi}{2}, \qquad \forall \alpha \in [0, 1], \tag{65}$$

which proves (ii).

We now prove (iii), again assuming $\rho > 0$. Since $\arccos$ is strictly decreasing on $[-1, 1]$, minimizing $G(\alpha)$ is equivalent to maximizing

$$\cos(\alpha\theta - \psi). \tag{66}$$

For any $\tau \in \mathbb{R}$, the circular distance

$$d_{\mathrm{circ}}(\tau, \psi) = \min_{k\in\mathbb{Z}} |\tau - \psi - 2\pi k| = \arccos\big(\cos(\tau - \psi)\big), \tag{67}$$

where the second equality holds because $\arccos(\cos(\cdot))$ computes the minimal unsigned angular distance to the nearest multiple of 2π. Therefore maximizing $\cos(\alpha\theta - \psi)$ is equivalent to minimizing $d_{\mathrm{circ}}(\alpha\theta, \psi)$. Setting $\tau = \alpha\theta$, so that $\tau \in [0, \theta]$, every minimizer has the form

$$\alpha^* = \frac{\tau^*}{\theta}, \qquad \tau^* \in \arg\min_{\tau\in[0,\theta]} d_{\mathrm{circ}}(\tau, \psi). \tag{68}$$

This proves (iii).

It remains to characterize when the minimum is attained by a point that strictly improves over both endpoints. Let

$$\Gamma := \{\cos\tau\, e_1 + \sin\tau\, e_2 : \tau \in [0, \theta]\} \tag{69}$$

be the minor arc from u to v. Since $\rho > 0$, the normalized in-plane projection of q^* is

$$\frac{p}{\|p\|} = \cos\psi\, e_1 + \sin\psi\, e_2. \tag{70}$$

Because $\theta < \pi$, the map

$$\tau \mapsto \cos\tau\, e_1 + \sin\tau\, e_2 \tag{71}$$

is injective on $[0, \theta]$. Hence

$$\frac{p}{\|p\|} \in \mathrm{relint}(\Gamma) \tag{72}$$

if and only if there exists an integer $k \in \mathbb{Z}$ such that

$$\tau_0 := \psi + 2\pi k \in (0, \theta). \tag{B}$$

Suppose first that $p/\|p\| \in \mathrm{relint}(\Gamma)$. Then (B) holds for some $\tau_0 \in (0, \theta)$, and

$$\cos(\tau_0 - \psi) = 1. \tag{73}$$

Thus τ_0 maximizes $\tau \mapsto \cos(\tau - \psi)$ over $[0, \theta]$. Moreover, since τ_0 lies strictly inside the interval and $\theta < \pi$, neither endpoint can attain the same value. Therefore

$$\cos(0 - \psi) < 1, \qquad \cos(\theta - \psi) < 1. \tag{74}$$

Let $\alpha^* = \tau_0/\theta \in (0, 1)$. Using $\rho > 0$ and the strict monotonicity of $\arccos$, we obtain

$$G(\alpha^*) = \arccos(\rho) < \min\{G(0), G(1)\}. \tag{75}$$

Hence a strict interior improvement holds.

Conversely, suppose that

$$\frac{p}{\|p\|} \notin \mathrm{relint}(\Gamma). \tag{76}$$

Then no point of the form $\psi + 2\pi k$ lies in $(0, \theta)$. Define

$$f(\tau) := \cos(\tau - \psi), \qquad \tau \in [0, \theta]. \tag{77}$$

Any interior maximizer of f must satisfy

$$f'(\tau) = -\sin(\tau - \psi) = 0, \tag{78}$$

and hence $\tau = \psi + k\pi$ for some $k \in \mathbb{Z}$. Points of the form $\psi + 2\pi k$ are maxima of the cosine, whereas points of the form $\psi + (2k + 1)\pi$ are minima. Since, by assumption, no point $\psi + 2\pi k$ lies in the open interval $(0, \theta)$, f has no interior maximizer on $(0, \theta)$. Therefore its maximum over the compact interval $[0, \theta]$ is attained at an endpoint:

$$f(\tau) \le \max\{f(0), f(\theta)\}, \qquad \forall \tau \in [0, \theta]. \tag{79}$$

Equivalently, for every $\alpha \in [0,1]$,

$$\cos(\alpha\theta - \psi) \le \max\{\cos(0-\psi), \cos(\theta-\psi)\}. \tag{80}$$

Multiplying by $\rho > 0$ and using again that $\arccos$ is strictly decreasing, we obtain

$$G(\alpha) \ge \min\{G(0), G(1)\}, \qquad \forall \alpha \in [0,1]. \tag{81}$$

Thus no interior point can strictly improve over both endpoints.

Combining the two directions, a strict interior improvement occurs if and only if the normalized in-plane projection $p/\|p\|$ lies strictly in the relative interior of the minor arc from u to v. □

F Connection Between SLERP and Recall@K

This appendix provides the metric-level details connecting the geometric SLERP characterization in Sec. 3 to Recall@K. We first define Recall@K for queries with multiple relevant gallery items, then introduce a top-K margin, and finally prove that angular proximity to a positive-margin direction certifies top-K retrieval success.

F.1 Recall@K with Multiple Relevant Items

Let the deployed gallery be

$$\mathcal{G} = \{y_j\}_{j=1}^{N_g}, \qquad g_j := \phi_{\text{old}}(y_j) \in \mathbb{S}^{d-1}. \tag{82}$$

For a unit query direction $q \in \mathbb{S}^{d-1}$, the score of gallery item y_j is

$$s_j(q) := \langle q, g_j \rangle. \tag{83}$$

Since all embeddings are normalized, this is the cosine similarity used for ranking.

For a query $x \in \mathcal{Q}$, let

$$P(x) \subseteq \{1, \dots, N_g\} \tag{84}$$

denote the set of relevant gallery indices, and let

$$N(x) := \{1, \dots, N_g\} \setminus P(x) \tag{85}$$

denote the set of negative indices. We assume $P(x) \neq \emptyset$ and $|N(x)| \ge K$. The set $P(x)$ may contain multiple relevant items, as in image-to-text retrieval with multiple captions per image or in retrieval benchmarks where several gallery items are valid matches for the same query.

Let $T_K(q) \subseteq \{1, \dots, N_g\}$ be the set of indices of the K highest-scoring gallery items under $\{s_j(q)\}_{j=1}^{N_g}$, with ties resolved by a fixed deterministic rule. The single-query Recall@K event is

$$\text{Recall@}K(x; q) := \mathbf{1}\{T_K(q) \cap P(x) \neq \emptyset\}. \tag{86}$$

Thus, with multiple relevant items, Recall@K is a hit indicator: it is equal to one whenever at least one relevant gallery item appears among the top K retrieved items.

Since the query direction is input-dependent, dataset-level Recall@K is defined for a query rule $h : \mathcal{Q} \to \mathbb{S}^{d-1}$:

$$\text{Recall@}K(h) := \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \text{Recall@}K(x; h(x)). \tag{87}$$

For SLERP with fixed interpolation weight α, the query rule is $h_\alpha(x) = q_\alpha(x)$, and therefore

$$\text{Recall@}K(\alpha) := \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \text{Recall@}K(x; q_\alpha(x)). \tag{88}$$

F.2 Top-K Margin

For a query x, define the best relevant score as

$$s_x^+(q) := \max_{j \in P(x)} s_j(q). \tag{89}$$

Let $\tau_{K,x}^-(q)$ denote the K-th largest value among the negative scores

$$\{s_j(q) : j \in N(x)\}. \tag{90}$$

The top-K retrieval margin is

$$m_{K,x}(q) := s_x^+(q) - \tau_{K,x}^-(q). \tag{91}$$

Proposition 1 (Top-K margin and Recall@K)**.** *For any query x and unit query direction q,*

$$m_{K,x}(q) > 0 \quad \Longrightarrow \quad \mathrm{Recall@}K(x; q) = 1, \tag{92}$$

and

$$m_{K,x}(q) < 0 \quad \Longrightarrow \quad \mathrm{Recall@}K(x; q) = 0. \tag{93}$$

When $m_{K,x}(q) = 0$, the outcome depends on the deterministic tie-breaking rule. Consequently, away from boundary ties, positivity of $m_{K,x}(q)$ is equivalent to Recall@K success.

Proof. If $m_{K,x}(q) > 0$, then

$$s_x^+(q) > \tau_{K,x}^-(q). \tag{94}$$

Thus the best relevant item scores strictly above the K-th largest negative score. Hence at most $K-1$ negatives can score above the best relevant item, so at least one relevant item must appear among the top K gallery items. Therefore

$$\mathrm{Recall@}K(x; q) = 1. \tag{95}$$

If $m_{K,x}(q) < 0$, then

$$s_x^+(q) < \tau_{K,x}^-(q). \tag{96}$$

Thus the K-th largest negative score is strictly above the score of every relevant item. Hence at least K negatives outrank all relevant items, so no relevant item can appear in the top K. Therefore

$$\mathrm{Recall@}K(x; q) = 0. \tag{97}$$

If $m_{K,x}(q) = 0$, the best relevant score coincides with the K-th largest negative score. In this boundary case, whether a relevant item is included in $T_K(q)$ depends on the fixed tie-breaking rule. □

F.3 Retrieval-Relevant Optimal Direction

The geometric analysis in the main text is formulated with respect to a optimal direction $q^* \in \mathbb{S}^{d-1}$. To specialize this direction to Recall@K, we choose a direction that maximizes the top-K margin:

$$q_K^\star(x) \in \arg\max_{q \in \mathbb{S}^{d-1}} m_{K,x}(q). \tag{98}$$

Such a maximizer exists. Indeed, each score $s_j(q) = \langle q, g_j \rangle$ is continuous in q. The maximum over finitely many relevant scores is continuous, and the K-th order statistic over finitely many negative scores is also continuous. Hence $m_{K,x}$ is continuous. Since $\mathbb{S}^{d-1}$ is compact, the maximum is attained.

The maximizer $q_K^\star(x)$ need not be unique; any selected maximizer can serve as a retrieval-relevant optimal direction. With this choice, define the angular gap

$$G_{K,x}(\alpha) := \angle(q_\alpha(x), q_K^\star(x)). \tag{99}$$

For each fixed query x, Theorem 1 applies with

$$u = u(x), \qquad v = v(x), \qquad q^* = q_K^\star(x). \tag{100}$$

Therefore, the SLERP arc contains an interior point closer to the top-K-margin-optimal direction than either endpoint exactly when the normalized in-plane projection of $q_K^\star(x)$ lies strictly in the relative interior of the minor arc joining $u(x)$ and $v(x)$.

F.4 Lipschitz Stability of the Top-K Margin

Lemma 2 (Top-K margin perturbation bound)**.** *For every query $x \in \mathcal{Q}$, every $K \geq 1$, and every pair of unit query directions $q, q' \in \mathbb{S}^{d-1}$,*

$$|m_{K,x}(q) - m_{K,x}(q')| \leq 2\|q - q'\| = 4 \sin\left(\frac{\angle(q, q')}{2}\right) \leq 2\,\angle(q, q'). \tag{101}$$

Proof. For every gallery item $g_j \in \mathbb{S}^{d-1}$,

$$|s_j(q) - s_j(q')| = |\langle q - q', g_j \rangle| \leq \|q - q'\|\,\|g_j\| = \|q - q'\|. \tag{102}$$

Thus every score changes by at most $\|q - q'\|$.

The maximum over relevant scores preserves this bound, so

$$|s_x^+(q) - s_x^+(q')| \leq \|q - q'\|. \tag{103}$$

Similarly, if every negative score changes by at most δ, then the K-th largest negative score changes by at most δ. Taking $\delta = \|q - q'\|$, we obtain

$$|\tau_{K,x}^-(q) - \tau_{K,x}^-(q')| \leq \|q - q'\|. \tag{104}$$

Therefore

$$|m_{K,x}(q) - m_{K,x}(q')| \leq |s_x^+(q) - s_x^+(q')| + |\tau_{K,x}^-(q) - \tau_{K,x}^-(q')| \leq 2\|q - q'\|. \tag{105}$$

Finally, for unit vectors,

$$\|q - q'\| = 2 \sin\left(\frac{\angle(q, q')}{2}\right) \leq \angle(q, q'), \tag{106}$$

which proves the result. □

F.5 Local Certification of Recall@K

Corollary 1 (Local Recall@K certification)**.** *Let*

$$m_{K,x}^\star := m_{K,x}(q_K^\star(x)). \tag{107}$$

Assume $m_{K,x}^\star > 0$. Then $0 < m_{K,x}^\star \leq 2$, and for any SLERP query $q_\alpha(x)$,

$$G_{K,x}(\alpha) < 2 \arcsin\left(\frac{m_{K,x}^\star}{4}\right) \quad \Longrightarrow \quad \text{Recall@}K(x; q_\alpha(x)) = 1. \tag{108}$$

A simpler sufficient condition is

$$G_{K,x}(\alpha) < \frac{m_{K,x}^\star}{2}. \tag{109}$$

Proof. Since all scores are cosine similarities, they lie in $[-1, 1]$. Hence the difference between any relevant score and any negative score is at most 2, so

$$0 < m_{K,x}^\star \leq 2. \tag{110}$$

By Lemma 2, applied with $q = q_\alpha(x)$ and $q' = q_K^\star(x)$,

$$m_{K,x}(q_\alpha(x)) \geq m_{K,x}^\star - 4 \sin\left(\frac{\angle(q_\alpha(x), q_K^\star(x))}{2}\right). \tag{111}$$

Using $G_{K,x}(\alpha) = \angle(q_\alpha(x), q_K^\star(x))$, we get

$$m_{K,x}(q_\alpha(x)) \geq m_{K,x}^\star - 4 \sin\left(\frac{G_{K,x}(\alpha)}{2}\right). \tag{112}$$

Therefore, if

$$G_{K,x}(\alpha) < 2 \arcsin\left(\frac{m_{K,x}^\star}{4}\right), \tag{113}$$

then

$$4 \sin\left(\frac{G_{K,x}(\alpha)}{2}\right) < m^{\star}_{K,x}, \tag{114}$$

and hence

$$m_{K,x}(q_\alpha(x)) > 0. \tag{115}$$

By Proposition 1, this implies

$$\text{Recall@}K(x; q_\alpha(x)) = 1. \tag{116}$$

The simpler condition follows from the coarser inequality

$$|m_{K,x}(q) - m_{K,x}(q')| \leq 2\angle(q, q'). \tag{117}$$

Thus

$$m_{K,x}(q_\alpha(x)) \geq m^{\star}_{K,x} - 2G_{K,x}(\alpha), \tag{118}$$

so $G_{K,x}(\alpha) < m^{\star}_{K,x}/2$ also implies $m_{K,x}(q_\alpha(x)) > 0$. □

G Cross-Modal Zero-shot Classification

We evaluate zero-shot classification as an auxiliary evaluation setting. Class text prototypes are encoded by the old-model text encoder using the template ''`A photo of a <class_name>`'' and kept fixed. Image embeddings are obtained from the new model either directly via SVD alignment or by applying SLERP to the SVD-aligned embeddings, then compared against the fixed old-model prototypes. Thus, only the image side changes, while the text-prototype side remains in the old-model space.

For classification, the Procrustes alignment is estimated on ImageNet-1K [83] validation set using text-only support. The SLERP weight $\hat{\alpha}$ is selected on ImageNet validation top-1 accuracy and then fixed for evaluation on Cars, Pets, Flowers, Aircraft, DTD, EuroSAT, Food101, SUN397, Caltech101, and UCF101. For SLERP, $\alpha = 0$ corresponds to old-model image embeddings, while $\alpha = 1$ corresponds to SVD-aligned new-model image embeddings. We also report $\alpha^{\star}$, the dataset-specific oracle weight, as an upper-bound reference.

Tab. 3 shows that SVD alignment alone recovers a meaningful fraction of the old-model classifier's accuracy, but remains lower on average. Applying SLERP to the SVD-aligned embeddings consistently improves this endpoint, often closing the gap to, or even surpassing, the old-model classifier. This trend mirrors the retrieval results: SVD provides a useful but approximate alignment between the two model representations, while SLERP yields a more effective representation in the old-model space.

H Additional SLERP Arc Results

Figures 6 and 7 extend the SLERP analysis to image-to-text Recall@1 on Flickr30k, COCO, and NoCaps. They complement the text-to-image curves (see Fig. 3) in the main paper and use the same convention: $\alpha = 0$ is the old-model query and $\alpha = 1$ is the SVD-aligned new-model query.

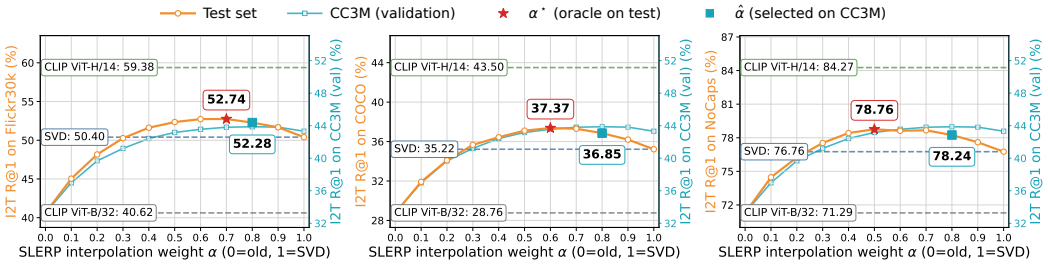
The curves show the same behavior observed in Sec. 4.3. For both CLIP ViT-H/14 $\rightarrow$ CLIP ViT-B/32 and SigLIP1 $\rightarrow$ CLIP ViT-H/14, the best test performance is typically reached at an interior interpolation weight rather than at either endpoint. The CC3M validation set curve also follows the test-set curve closely, and the selected $\hat{\alpha}$ is near the dataset-specific oracle $\alpha^{\star}$. Thus, the benefit of SLERP is not restricted to text-to-image retrieval: the same support-set selected interpolation also transfers to image-to-text retrieval across datasets and model families.

I Detailed Retrieval Performance

Tabs. 4, 5, and 6 provide the full Recall@1/5/10 metrics for the retrieval settings of the experimental results in Sec. 4.3. For each dataset, we report both image-to-text (I2T) and text-to-image (T2I) retrieval, for all support modalities used to estimate the Procrustes alignment. We compare SVD alignment alone, XBT (when available), and SLERP applied to the SVD-aligned embeddings with

Table 3: Zero-shot classification accuracy (%) with fixed old-model text prototypes. For SLERP, the support-set-selected weight $\hat{\alpha}$ is selected on the ImageNet-1K validation set and fixed across target datasets, whereas $\alpha^\star$ is selected directly on each target dataset and is reported only as an oracle upper bound. Bold and underlined values denote the best and second-best classification results, respectively, excluding the old- and new-model reference rows.

Method	Cars	Pets	Flowers	Aircraft	DTD	EuroSAT	Food101	SUN397	Caltech	UCF101	*Average*
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*											
▶ **Old model (CLIP ViT-B/32)**	60.18	87.46	66.50	18.96	44.15	45.21	80.42	62.06	91.36	63.57	61.99
SVD	25.59	87.05	34.51	14.28	40.19	40.78	72.20	54.08	90.22	59.95	51.89
+SLERP ($\hat{\alpha}$)	55.35	89.62	60.45	20.07	47.28	53.16	85.05	64.51	93.18	67.04	63.57
+SLERP ($\alpha^\star$)	**62.23**	**90.05**	**67.97**	**20.94**	**47.34**	**53.88**	**85.44**	**65.29**	**93.39**	**67.27**	**65.38**
▶ **New model (CLIP ViT-L/14)**	76.91	93.46	79.46	32.64	53.01	60.28	90.91	67.68	95.17	74.97	72.45
SigLIP2 → SigLIP1 *(same model family)*											
▶ **Old model (SigLIP1)**	88.30	95.28	91.72	59.89	62.17	60.36	93.65	75.58	98.38	81.79	80.71
SVD	31.70	85.47	69.43	16.17	45.15	42.31	84.00	64.73	97.48	71.45	60.79
+SLERP ($\hat{\alpha}$)	87.27	94.99	90.17	58.93	61.05	58.05	**93.79**	75.80	98.34	81.15	79.95
+SLERP ($\alpha^\star$)	**88.30**	**95.39**	**91.72**	**60.34**	**62.17**	**60.36**	**93.79**	**76.04**	**98.54**	**81.79**	**80.84**
▶ **New model (SigLIP2)**	95.37	96.08	89.36	71.59	65.78	50.68	94.21	75.59	98.46	81.79	81.89
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*											
▶ **Old model (CLIP ViT-H/14)**	93.12	94.55	80.63	42.42	62.77	68.51	90.30	75.09	97.24	78.27	78.29
SVD	41.08	92.83	49.33	16.02	43.32	48.84	81.64	61.82	96.96	64.71	59.66
+SLERP ($\hat{\alpha}$)	90.55	94.74	75.03	37.44	59.99	66.69	90.73	73.65	**98.42**	77.53	76.48
+SLERP ($\alpha^\star$)	**93.15**	**95.07**	**80.63**	**43.44**	**63.24**	**69.72**	**91.11**	**75.51**	**98.42**	**79.25**	**78.95**
▶ **New model (SigLIP1)**	88.30	95.28	91.72	59.89	62.17	60.36	93.65	75.58	98.38	81.79	80.71



(a) Flickr30k (b) COCO (c) NoCaps

Figure 6: **Image-to-text Recall@1 along the SLERP path for CLIP ViT-H/14 → CLIP ViT-B/32.** The text gallery is encoded by the old model, while image queries interpolate between the old-model query and the SVD-aligned new-model query. The case $\alpha = 0$ corresponds to old-model queries, while $\alpha = 1$ corresponds to SVD-aligned new-model queries. Markers show the CC3M-selected $\hat{\alpha}$ and the dataset-specific test oracle $\alpha^\star$; dashed lines show old-model, SVD, and new-model reference performance.

both the support-set selected weight $\hat{\alpha}$ and the dataset-specific oracle $\alpha^\star$. The tables also include the CLIP ViT-H/14 → CLIP ViT-B/32 setting, which has been adopted by [51].

The detailed results confirm the trend of Tables 1 and 2. SVD provides a strong new-to-old alignment with no model retraining or gradient-based optimization, but it does not consistently achieve compatibility across model pairs, datasets, and support modalities. SLERP improves the corresponding SVD endpoint in nearly all settings, and the gains extend beyond Recall@1 to Recall@5 and Recall@10. This shows that interpolation not only improves the top-ranked item, but also produces a more compatible ranking over the old-model gallery.

The same pattern holds across same-family and cross-family pairs. In the CLIP same-family settings, SLERP is consistently competitive with the training-based XBT baseline while requiring no compatibility training. In the cross-family settings, including the difficult SigLIP1 → CLIP ViT-H/14 case, SLERP often turns incompatible SVD rows into compatible ones. Across support

Table 4: **Cross-modal backward-compatible retrieval on Flickr30k.** The α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The support-set-selected weight $\hat{\alpha}$ is selected on CC3M, whereas $\alpha^\star$ is selected directly on Flickr30k and is reported only as an oracle upper bound. Checkmarks indicate backward compatibility; bold and underlined values denote the best and second-best new → old results, respectively.

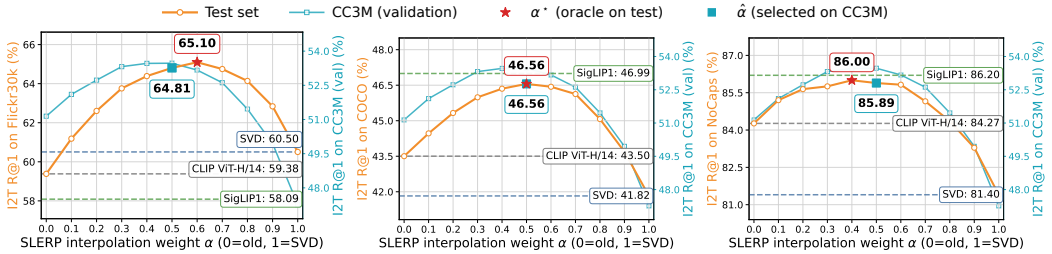
			I2T			T2I		
Method	Sup.	α	R@1	R@5	R@10	R@1	R@5	R@10
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	40.62	64.75	73.72	21.73	41.56	51.10
XBT [51]	–	–	42.47 ✓	66.04 ✓	74.87 ✓	22.38 ✓	42.71 ✓	52.41 ✓
SVD	T	–	42.89 ✓	66.79 ✓	75.77 ✓	21.49 ×	41.35 ×	50.92 ×
+SLERP ($\hat{\alpha}$)	T	0.7/0.5	48.00 ✓	71.84 ✓	80.34 ✓	23.33 ✓	43.77 ✓	53.29 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.5	**48.61** ✓	**72.64** ✓	**80.83** ✓	23.33 ✓	43.77 ✓	53.29 ✓
SVD	I	–	41.17 ✓	65.64 ✓	74.91 ✓	19.99 ×	39.25 ×	48.75 ×
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	46.30 ✓	70.62 ✓	79.18 ✓	23.00 ✓	43.38 ✓	53.01 ✓
+SLERP ($\alpha^\star$)	I	0.5/0.4	47.39 ✓	71.51 ✓	79.93 ✓	23.00 ✓	43.38 ✓	53.01 ✓
SVD	I+T	–	42.44 ✓	66.22 ✓	75.21 ✓	21.80 ✓	41.86 ✓	51.52 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.4	47.09 ✓	70.77 ✓	79.64 ✓	**23.44** ✓	43.99 ✓	53.53 ✓
+SLERP ($\alpha^\star$)	I+T	0.5/0.5	47.77 ✓	72.01 ✓	80.25 ✓	**23.44** ✓	**44.08** ✓	**53.61** ✓
▶ New model (CLIP ViT-L/14)	–	–	48.72	72.93	81.22	28.27	49.63	59.09
CLIP ViT-H/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	40.62	64.75	73.72	21.73	41.56	51.10
XBT [51]	–	–	41.09 ✓	65.18 ✓	74.32 ✓	23.70 ✓	44.47 ✓	54.19 ✓
SVD	T	–	50.40 ✓	74.33 ✓	82.23 ✓	25.45 ✓	46.81 ✓	56.45 ✓
+SLERP ($\hat{\alpha}$)	T	0.8/0.6	52.28 ✓	76.03 ✓	83.58 ✓	26.21 ✓	47.74 ✓	57.40 ✓
+SLERP ($\alpha^\star$)	T	0.7/0.7	**52.74** ✓	**76.28** ✓	**83.90** ✓	26.31 ✓	47.77 ✓	57.43 ✓
SVD	I	–	48.47 ✓	72.18 ✓	80.56 ✓	26.49 ✓	47.90 ✓	57.51 ✓
+SLERP ($\hat{\alpha}$)	I	0.6/0.6	50.95 ✓	74.55 ✓	82.39 ✓	**28.34** ✓	**50.15** ✓	**59.69** ✓
+SLERP ($\alpha^\star$)	I	0.6/0.6	50.95 ✓	74.55 ✓	82.39 ✓	**28.34** ✓	**50.15** ✓	**59.69** ✓
SVD	I+T	–	50.06 ✓	73.42 ✓	81.36 ✓	26.58 ✓	48.00 ✓	57.63 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.5/0.7	52.03 ✓	75.14 ✓	83.02 ✓	27.43 ✓	49.14 ✓	58.78 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.7	52.19 ✓	75.44 ✓	83.11 ✓	27.43 ✓	49.14 ✓	58.78 ✓
▶ New model (CLIP ViT-H/14)	–	–	59.38	82.49	88.97	43.07	65.99	74.31
SigLIP2 → SigLIP1 *(same model family)*								
▶ Old model (SigLIP1)	–	–	58.09	81.69	88.63	39.40	62.46	71.13
SVD	T	–	45.46 ×	71.93 ×	81.06 ×	46.28 ✓	68.99 ✓	76.97 ✓
+SLERP ($\hat{\alpha}$)	T	0.3/0.4	58.63 ✓	82.35 ✓	88.99 ✓	45.99 ✓	69.07 ✓	77.16 ✓
+SLERP ($\alpha^\star$)	T	0.2/0.7	**58.70** ✓	**82.40** ✓	**89.07** ✓	**47.59** ✓	70.25 ✓	77.99 ✓
SVD	I	–	50.66 ×	76.19 ×	84.43 ×	41.78 ✓	65.03 ✓	73.61 ✓
+SLERP ($\hat{\alpha}$)	I	0.2/0.3	58.59 ✓	82.26 ✓	88.96 ✓	43.84 ✓	67.27 ✓	75.66 ✓
+SLERP ($\alpha^\star$)	I	0.3/0.6	58.61 ✓	82.28 ✓	89.01 ✓	45.12 ✓	68.41 ✓	76.66 ✓
SVD	I+T	–	50.19 ×	76.22 ×	84.43 ×	46.39 ✓	69.18 ✓	77.19 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.2/0.5	58.62 ✓	82.18 ✓	88.96 ✓	46.87 ✓	69.87 ✓	77.83 ✓
+SLERP ($\alpha^\star$)	I+T	0.2/0.7	58.62 ✓	82.18 ✓	88.96 ✓	47.58 ✓	**70.36** ✓	**78.19** ✓
▶ New model (SigLIP2)	–	–	69.31	88.50	93.26	51.29	72.97	80.20
SigLIP2 → CLIP ViT-B/32 *(cross-family)*								
▶ Old model (CLIP ViT-B/32)	–	–	40.62	64.75	73.72	21.73	41.56	51.10
SVD	T	–	42.80 ✓	67.79 ✓	77.23 ✓	23.56 ✓	43.89 ✓	53.57 ✓
+SLERP ($\hat{\alpha}$)	T	0.6/0.5	51.01 ✓	75.34 ✓	**83.47** ✓	25.79 ✓	47.20 ✓	56.75 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.6	51.34 ✓	**75.43** ✓	**83.47** ✓	25.83 ✓	47.21 ✓	56.78 ✓
SVD	I	–	48.01 ✓	71.85 ✓	79.99 ✓	23.41 ✓	44.39 ✓	54.10 ✓
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	51.33 ✓	74.99 ✓	82.73 ✓	28.57 ✓	50.66 ✓	60.29 ✓
+SLERP ($\alpha^\star$)	I	0.6/0.5	**51.40** ✓	75.08 ✓	82.89 ✓	**28.88** ✓	**50.99** ✓	**60.70** ✓
SVD	I+T	–	36.25 ×	57.82 ×	66.80 ×	23.93 ✓	45.06 ✓	54.87 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.4/0.5	47.97 ✓	71.34 ✓	79.83 ✓	27.09 ✓	48.73 ✓	58.48 ✓
+SLERP ($\alpha^\star$)	I+T	0.4/0.6	47.97 ✓	71.34 ✓	79.83 ✓	27.10 ✓	48.80 ✓	58.50 ✓
▶ New model (SigLIP2)	–	–	69.31	88.50	93.26	51.29	72.97	80.20
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*								
▶ Old model (CLIP ViT-H/14)	–	–	59.38	82.49	88.97	43.07	65.99	74.31
SVD	T	–	60.50 ✓	82.90 ✓	89.17 ✓	29.91 ×	52.61 ×	62.07 ×
+SLERP ($\hat{\alpha}$)	T	0.5/0.1	64.81 ✓	86.50 ✓	91.84 ✓	43.38 ✓	66.34 ✓	74.56 ✓
+SLERP ($\alpha^\star$)	T	0.6/0.2	**65.10** ✓	**86.60** ✓	**91.88** ✓	43.44 ✓	66.40 ✓	74.69 ✓
SVD	I	–	48.93 ×	75.55 ×	84.23 ×	29.13 ×	51.81 ×	61.25 ×
+SLERP ($\hat{\alpha}$)	I	0.0/0.3	59.38 ×	82.49 ×	88.97 ×	43.25 ✓	66.22 ✓	74.63 ✓
+SLERP ($\alpha^\star$)	I	0.4/0.2	61.19 ✓	84.16 ✓	90.51 ✓	**43.50** ✓	66.40 ✓	74.72 ✓
SVD	I+T	–	58.54 ×	81.21 ×	87.88 ×	30.03 ×	52.66 ×	62.24 ×
+SLERP ($\hat{\alpha}$)	I+T	0.1/0.1	61.09 ✓	83.80 ✓	89.93 ✓	43.36 ✓	66.34 ✓	74.60 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.2	64.74 ✓	86.08 ✓	91.55 ✓	43.43 ✓	**66.41** ✓	**74.76** ✓
▶ New model (SigLIP1)	–	–	58.09	81.69	88.63	39.40	62.46	71.13

Table 5: **Cross-modal backward-compatible retrieval on COCO.** The α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The support-set-selected weight $\hat{\alpha}$ is selected on CC3M, whereas $\alpha^\star$ is selected directly on COCO and is reported only as an oracle upper bound. Checkmarks indicate backward compatibility; bold and underlined values denote the best and second-best new → old results, respectively.

			I2T			T2I		
Method	Sup.	α	R@1	R@5	R@10	R@1	R@5	R@10
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	28.76	50.34	59.86	14.47	30.22	38.92
XBT [51]	–	–	30.73 ✓	52.58 ✓	62.30 ✓	**15.55** ✓	**32.27** ✓	**41.34** ✓
SVD	T	–	30.27 ✓	51.35 ✓	61.07 ✓	14.03 ×	29.71 ×	38.33 ×
+SLERP ($\hat{\alpha}$)	T	0.7/0.5	33.40 ✓	55.44 ✓	64.80 ✓	15.26 ✓	31.57 ✓	40.34 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.4	**33.84** ✓	**55.84** ✓	**65.12** ✓	15.27 ✓	31.56 ✓	40.36 ✓
SVD	I	–	28.71 ×	49.27 ×	58.58 ×	13.11 ×	28.44 ×	36.80 ×
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	31.77 ✓	53.46 ✓	62.90 ✓	15.14 ✓	31.39 ✓	40.21 ✓
+SLERP ($\alpha^\star$)	I	0.5/0.3	32.51 ✓	54.42 ✓	63.69 ✓	15.15 ✓	31.42 ✓	40.19 ✓
SVD	I+T	–	29.95 ✓	50.99 ✓	60.34 ✓	14.16 ×	30.17 ×	38.75 ×
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.4	33.05 ✓	54.79 ✓	63.98 ✓	15.30 ✓	31.70 ✓	40.57 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.5	33.32 ✓	55.25 ✓	64.42 ✓	15.35 ✓	31.74 ✓	40.61 ✓
▶ New model (CLIP ViT-L/14)	–	–	34.33	56.09	65.38	18.68	35.85	44.40
CLIP ViT-H/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	28.76	50.34	59.86	14.47	30.22	38.92
XBT [51]	–	–	30.33 ✓	51.28 ✓	59.97 ✓	15.82 ✓	32.79 ✓	41.84 ✓
SVD	T	–	35.22 ✓	57.40 ✓	66.69 ✓	17.01 ✓	34.49 ✓	43.61 ✓
+SLERP ($\hat{\alpha}$)	T	0.8/0.6	36.85 ✓	59.18 ✓	68.52 ✓	17.37 ✓	34.96 ✓	44.17 ✓
+SLERP ($\alpha^\star$)	T	0.6/0.7	**37.37** ✓	**59.75** ✓	**68.85** ✓	17.38 ✓	35.03 ✓	44.27 ✓
SVD	I	–	33.67 ✓	55.54 ✓	64.91 ✓	17.13 ✓	34.66 ✓	43.84 ✓
+SLERP ($\hat{\alpha}$)	I	0.6/0.6	35.33 ✓	57.91 ✓	66.99 ✓	**18.35** ✓	**36.51** ✓	**45.73** ✓
+SLERP ($\alpha^\star$)	I	0.6/0.6	35.33 ✓	57.91 ✓	66.99 ✓	**18.35** ✓	**36.51** ✓	**45.73** ✓
SVD	I+T	–	35.50 ✓	57.72 ✓	66.74 ✓	17.35 ✓	35.01 ✓	44.15 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.5/0.7	36.71 ✓	59.26 ✓	68.37 ✓	17.90 ✓	35.83 ✓	45.05 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.7	37.11 ✓	59.43 ✓	68.63 ✓	17.90 ✓	35.83 ✓	45.05 ✓
▶ New model (CLIP ViT-H/14)	–	–	43.50	66.34	74.76	28.56	49.50	58.50
SigLIP2 → SigLIP1 *(same model family)*								
▶ Old model (SigLIP1)	–	–	46.99	70.23	78.19	30.88	52.11	61.05
SVD	T	–	36.30 ×	59.91 ×	69.34 ×	30.53 ×	51.66 ×	60.27 ×
+SLERP ($\hat{\alpha}$)	T	0.3/0.4	46.76 ×	70.19 ×	78.32 ✓	32.50 ✓	53.87 ✓	62.53 ✓
+SLERP ($\alpha^\star$)	T	0.1/0.5	47.18 ✓	70.39 ✓	**78.55** ✓	32.56 ✓	53.83 ✓	62.52 ✓
SVD	I	–	40.96 ×	64.64 ×	73.55 ×	28.18 ×	49.17 ×	58.26 ×
+SLERP ($\hat{\alpha}$)	I	0.2/0.3	**47.30** ✓	**70.62** ✓	78.47 ✓	32.02 ✓	53.46 ✓	62.28 ✓
+SLERP ($\alpha^\star$)	I	0.2/0.4	**47.30** ✓	**70.62** ✓	78.47 ✓	32.06 ✓	53.48 ✓	62.23 ✓
SVD	I+T	–	39.41 ×	63.45 ×	72.41 ×	30.78 ×	51.95 ×	60.77 ×
+SLERP ($\hat{\alpha}$)	I+T	0.2/0.5	47.24 ✓	70.47 ✓	78.47 ✓	**32.67** ✓	**54.12** ✓	**62.75** ✓
+SLERP ($\alpha^\star$)	I+T	0.2/0.5	47.24 ✓	70.47 ✓	78.47 ✓	**32.67** ✓	**54.12** ✓	**62.75** ✓
▶ New model (SigLIP2)	–	–	51.06	73.27	80.73	35.04	56.49	64.96
SigLIP2 → CLIP ViT-B/32 *(cross-family)*								
▶ Old model (CLIP ViT-B/32)	–	–	28.76	50.34	59.86	14.47	30.22	38.92
SVD	T	–	32.33 ✓	54.63 ✓	64.31 ✓	15.62 ✓	32.39 ✓	41.27 ✓
+SLERP ($\hat{\alpha}$)	T	0.6/0.5	**37.35** ✓	**60.36** ✓	**69.34** ✓	17.18 ✓	34.75 ✓	43.99 ✓
+SLERP ($\alpha^\star$)	T	0.6/0.6	**37.35** ✓	**60.36** ✓	**69.34** ✓	17.27 ✓	34.83 ✓	44.05 ✓
SVD	I	–	32.28 ✓	54.30 ✓	63.53 ✓	15.20 ✓	31.77 ✓	40.68 ✓
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	34.86 ✓	57.49 ✓	66.59 ✓	18.79 ✓	37.30 ✓	46.77 ✓
+SLERP ($\alpha^\star$)	I	0.6/0.5	35.05 ✓	57.84 ✓	66.78 ✓	**18.97** ✓	**37.66** ✓	**46.98** ✓
SVD	I+T	–	28.16 ×	49.46 ×	59.02 ×	15.71 ✓	32.77 ✓	41.88 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.4/0.5	34.81 ✓	57.37 ✓	66.57 ✓	17.82 ✓	35.92 ✓	45.23 ✓
+SLERP ($\alpha^\star$)	I+T	0.5/0.6	34.86 ✓	57.38 ✓	66.60 ✓	17.84 ✓	36.00 ✓	45.34 ✓
▶ New model (SigLIP2)	–	–	51.06	73.27	80.73	35.04	56.49	64.96
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*								
▶ Old model (CLIP ViT-H/14)	–	–	43.50	66.34	74.76	28.56	49.50	58.50
SVD	T	–	41.82 ×	64.87 ×	73.80 ×	23.18 ×	43.28 ×	52.61 ×
+SLERP ($\hat{\alpha}$)	T	0.5/0.1	**46.56** ✓	**69.50** ✓	**77.46** ✓	28.79 ✓	49.78 ✓	58.81 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.2	**46.56** ✓	**69.50** ✓	**77.46** ✓	28.88 ✓	49.94 ✓	59.00 ✓
SVD	I	–	26.43 ×	45.13 ×	53.39 ×	23.21 ×	43.11 ×	52.32 ×
+SLERP ($\hat{\alpha}$)	I	0.0/0.3	43.50 ×	66.34 ×	74.76 ×	**29.04** ✓	**50.12** ✓	**59.21** ✓
+SLERP ($\alpha^\star$)	I	0.2/0.3	44.00 ✓	67.02 ✓	75.42 ✓	**29.04** ✓	**50.12** ✓	**59.21** ✓
SVD	I+T	–	34.47 ×	55.69 ×	64.66 ×	23.58 ×	43.83 ×	53.12 ×
+SLERP ($\hat{\alpha}$)	I+T	0.1/0.1	44.33 ✓	67.38 ✓	75.58 ✓	28.80 ✓	49.80 ✓	58.83 ✓
+SLERP ($\alpha^\star$)	I+T	0.4/0.3	45.65 ✓	68.65 ✓	76.89 ✓	28.93 ✓	50.07 ✓	59.19 ✓
▶ New model (SigLIP1)	–	–	46.99	70.23	78.19	30.88	52.11	61.05

Table 6: **Cross-modal backward-compatible retrieval on NoCaps.** The α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The support-set-selected weight $\hat{\alpha}$ is selected on CC3M, whereas $\alpha^\star$ is selected directly on NoCaps and is reported only as an oracle upper bound. Checkmarks indicate backward compatibility; bold and underlined values denote the best and second-best new → old results, respectively.

			I2T			T2I		
Method	Sup.	α	R@1	R@5	R@10	R@1	R@5	R@10
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	71.29	91.93	96.20	45.24	74.97	84.51
XBT [51]	–	–	75.02 ✓	93.27 ✓	97.31 ✓	**48.02** ✓	**79.00** ✓	**88.21** ✓
SVD	T	–	69.22 ×	91.18 ×	96.04 ×	43.92 ×	74.22 ×	83.96 ×
+SLERP ($\hat{\alpha}$)	T	0.7/0.5	73.98 ✓	93.33 ✓	96.91 ✓	46.59 ✓	76.47 ✓	85.54 ✓
+SLERP ($\alpha^\star$)	T	0.4/0.4	**75.62** ✓	**93.96** ✓	**97.47** ✓	46.65 ✓	76.52 ✓	85.56 ✓
SVD	I	–	65.96 ×	89.24 ×	94.91 ×	42.16 ×	72.72 ×	82.77 ×
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	71.40 ✓	92.71 ✓	96.87 ✓	46.49 ✓	76.24 ✓	85.64 ✓
+SLERP ($\alpha^\star$)	I	0.4/0.3	73.78 ✓	93.31 ✓	97.36 ✓	46.51 ✓	76.22 ✓	85.56 ✓
SVD	I+T	–	69.13 ×	90.89 ×	96.18 ×	44.41 ×	74.95 ×	84.64 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.4	73.56 ✓	93.20 ✓	96.96 ✓	46.98 ✓	76.90 ✓	85.86 ✓
+SLERP ($\alpha^\star$)	I+T	0.5/0.4	74.78 ✓	93.69 ✓	97.29 ✓	46.98 ✓	76.90 ✓	85.86 ✓
▶ New model (CLIP ViT-L/14)	–	–	73.36	93.42	97.38	47.84	76.55	85.11
CLIP ViT-H/14 → CLIP ViT-B/32 *(same model family)*								
▶ Old model (CLIP ViT-B/32)	–	–	71.29	91.93	96.20	45.24	74.97	84.51
XBT [51]	–	–	76.62 ✓	94.11 ✓	97.58 ✓	51.14 ✓	80.82 ✓	88.93 ✓
SVD	T	–	76.76 ✓	93.56 ✓	97.33 ✓	50.38 ✓	79.64 ✓	87.76 ✓
+SLERP ($\hat{\alpha}$)	T	0.8/0.6	78.24 ✓	94.22 ✓	97.84 ✓	51.09 ✓	80.24 ✓	88.33 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.7	**78.76** ✓	**94.84** ✓	**97.98** ✓	51.27 ✓	80.40 ✓	88.36 ✓
SVD	I	–	75.04 ✓	93.58 ✓	97.07 ✓	50.62 ✓	80.38 ✓	88.57 ✓
+SLERP ($\hat{\alpha}$)	I	0.6/0.6	77.51 ✓	94.53 ✓	97.73 ✓	**52.79** ✓	**81.86** ✓	**89.56** ✓
+SLERP ($\alpha^\star$)	I	0.6/0.6	77.51 ✓	94.53 ✓	97.73 ✓	**52.79** ✓	**81.86** ✓	**89.56** ✓
SVD	I+T	–	76.40 ✓	93.73 ✓	97.20 ✓	51.58 ✓	80.79 ✓	88.83 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.5/0.7	78.42 ✓	94.49 ✓	97.76 ✓	52.55 ✓	81.60 ✓	89.34 ✓
+SLERP ($\alpha^\star$)	I+T	0.6/0.7	78.58 ✓	94.64 ✓	97.84 ✓	52.55 ✓	81.60 ✓	89.34 ✓
▶ New model (CLIP ViT-H/14)	–	–	84.27	97.24	98.98	63.53	87.65	92.93
SigLIP2 → SigLIP1 *(same model family)*								
▶ Old model (SigLIP1)	–	–	86.20	97.78	99.44	64.18	87.62	92.86
SVD	T	–	75.00 ×	93.84 ×	97.49 ×	64.64 ✓	87.86 ✓	93.10 ✓
+SLERP ($\hat{\alpha}$)	T	0.3/0.4	85.87 ×	97.60 ×	**99.31** ×	66.59 ✓	89.07 ✓	94.02 ✓
+SLERP ($\alpha^\star$)	T	0.1/0.5	86.49 ✓	97.78 ×	99.22 ×	66.72 ✓	89.13 ✓	94.05 ✓
SVD	I	–	80.00 ×	96.16 ×	98.51 ×	62.69 ×	86.95 ×	92.64 ×
+SLERP ($\hat{\alpha}$)	I	0.2/0.3	**86.69** ✓	**97.84** ✓	**99.31** ×	66.08 ✓	88.97 ✓	93.94 ✓
+SLERP ($\alpha^\star$)	I	0.2/0.5	**86.69** ✓	**97.84** ✓	**99.31** ×	66.38 ✓	89.12 ✓	94.07 ✓
SVD	I+T	–	77.93 ×	95.16 ×	98.18 ×	64.98 ✓	88.35 ✓	93.42 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.2/0.5	86.49 ✓	97.73 ×	99.29 ×	**67.02** ✓	**89.41** ✓	94.20 ✓
+SLERP ($\alpha^\star$)	I+T	0.1/0.6	86.60 ✓	97.76 ×	**99.31** ×	**67.02** ✓	89.38 ✓	**94.23** ✓
▶ New model (SigLIP2)	–	–	89.18	98.64	99.56	69.82	90.85	95.10
SigLIP2 → CLIP ViT-B/32 *(cross-family)*								
▶ Old model (CLIP ViT-B/32)	–	–	71.29	91.93	96.20	45.24	74.97	84.51
SVD	T	–	72.16 ✓	92.69 ✓	96.82 ✓	46.41 ✓	75.39 ✓	84.29 ×
+SLERP ($\hat{\alpha}$)	T	0.6/0.5	78.56 ✓	**95.33** ✓	**98.33** ✓	50.59 ✓	79.65 ✓	87.64 ✓
+SLERP ($\alpha^\star$)	T	0.5/0.6	**78.80** ✓	95.24 ✓	98.31 ✓	50.61 ✓	79.60 ✓	87.51 ✓
SVD	I	–	72.42 ✓	92.71 ✓	97.04 ✓	46.47 ✓	75.94 ✓	85.20 ✓
+SLERP ($\hat{\alpha}$)	I	0.7/0.4	76.18 ✓	94.31 ✓	97.82 ✓	53.86 ✓	82.62 ✓	90.09 ✓
+SLERP ($\alpha^\star$)	I	0.5/0.5	77.62 ✓	94.42 ✓	97.82 ✓	**54.37** ✓	**82.88** ✓	**90.31** ✓
SVD	I+T	–	66.24 ×	88.96 ×	94.20 ×	48.66 ✓	78.19 ✓	87.19 ✓
+SLERP ($\hat{\alpha}$)	I+T	0.4/0.5	77.40 ✓	94.38 ✓	97.67 ✓	53.04 ✓	81.99 ✓	89.60 ✓
+SLERP ($\alpha^\star$)	I+T	0.4/0.6	77.40 ✓	94.38 ✓	97.67 ✓	53.23 ✓	82.16 ✓	89.76 ✓
▶ New model (SigLIP2)	–	–	89.18	98.64	99.56	69.82	90.85	95.10
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*								
▶ Old model (CLIP ViT-H/14)	–	–	84.27	97.24	98.98	63.53	87.65	92.93
SVD	T	–	81.40 ×	96.69 ×	98.80 ×	56.13 ×	82.11 ×	89.27 ×
+SLERP ($\hat{\alpha}$)	T	0.5/0.1	85.89 ✓	**97.98** ✓	**99.38** ✓	63.89 ✓	87.82 ✓	93.12 ✓
+SLERP ($\alpha^\star$)	T	0.4/0.2	**86.00** ✓	97.96 ✓	99.31 ✓	63.96 ✓	87.89 ✓	93.22 ✓
SVD	I	–	70.31 ×	92.96 ×	97.13 ×	55.05 ×	81.70 ×	89.21 ×
+SLERP ($\hat{\alpha}$)	I	0.0/0.3	84.27 ×	97.24 ×	98.98 ×	63.80 ✓	87.89 ✓	**93.35** ✓
+SLERP ($\alpha^\star$)	I	0.2/0.2	84.93 ✓	97.27 ✓	99.13 ✓	63.91 ✓	87.87 ✓	93.30 ✓
SVD	I+T	–	77.40 ×	95.29 ×	98.02 ×	56.60 ×	82.68 ×	89.72 ×
+SLERP ($\hat{\alpha}$)	I+T	0.1/0.1	85.04 ✓	97.36 ✓	99.16 ✓	63.87 ✓	87.83 ✓	93.14 ✓
+SLERP ($\alpha^\star$)	I+T	0.3/0.3	85.76 ✓	97.47 ✓	99.27 ✓	**64.00** ✓	**87.94** ✓	93.27 ✓
▶ New model (SigLIP1)	–	–	86.20	97.78	99.44	64.18	87.62	92.86



(a) Flickr30k (b) COCO (c) NoCaps

Figure 7: **Image-to-text Recall@1 along the SLERP path for SigLIP1 → CLIP ViT-H/14.** The text gallery is encoded by the old model, while image queries interpolate between the old-model query and the SVD-aligned new-model query. The case $\alpha = 0$ corresponds to old-model queries, while $\alpha = 1$ corresponds to SVD-aligned new-model queries. Markers show the CC3M-selected $\hat{\alpha}$ and the dataset-specific test oracle $\alpha^\star$; dashed lines show old-model, SVD, and new-model reference performance.

Table 7: Reindexed-gallery retrieval for same-family model upgrades. For each dataset, the α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The support-set-selected weight $\hat{\alpha}$ is selected on CC3M and fixed across test datasets, whereas $\alpha^\star$ is selected directly on each test dataset and is reported only as an oracle upper bound. Bold and underlined values denote the best and second-best results, respectively, excluding the old- and new-model reference rows.

Method	Sup.	Flickr30k			COCO			NoCaps		
		α	I2T@1	T2I@1	α	I2T@1	T2I@1	α	I2T@1	T2I@1
CLIP ViT-L/14 → CLIP ViT-B/32 *(same model family)*										
► Old model (CLIP ViT-B/32)	–	–	40.62	21.73	–	28.76	14.47	–	71.29	45.24
XBT [51]	–	–	43.50	**39.59**	–	33.32	**27.61**	–	77.22	**63.48**
SVD	T	–	48.72	28.27	–	34.33	18.68	–	73.36	47.84
+SLERP ($\hat{\alpha}$)	T	0.7/0.7	52.96	31.57	0.7/0.7	36.79	20.36	0.7/0.7	76.47	51.34
+SLERP ($\alpha^\star$)	T	0.6/0.7	**53.23**	31.57	0.7/0.7	36.79	20.36	0.5/0.6	77.09	51.81
SVD	I	–	48.72	28.27	–	34.33	18.68	–	73.36	47.84
+SLERP ($\hat{\alpha}$)	I	0.9/0.7	49.51	31.46	0.9/0.7	34.79	20.23	0.9/0.7	73.71	51.25
+SLERP ($\alpha^\star$)	I	0.9/0.7	49.51	31.46	0.9/0.7	34.79	20.23	0.8/0.6	73.93	51.60
SVD	I+T	–	48.72	28.27	–	34.33	18.68	–	73.36	47.84
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.7	53.16	31.04	0.7/0.7	**36.80**	20.12	0.7/0.7	76.84	51.22
+SLERP ($\alpha^\star$)	I+T	0.7/0.6	53.16	31.04	0.7/0.7	**36.80**	20.12	0.5/0.5	**77.47**	51.60
► New model (CLIP ViT-L/14)	–	–	48.72	28.27	–	34.33	18.68	–	73.36	47.84
SigLIP2 → SigLIP1 *(same model family)*										
► Old model (SigLIP1)	–	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18
SVD	T	–	69.31	51.29	–	51.06	35.04	–	**89.18**	69.82
+SLERP ($\hat{\alpha}$)	T	0.5/0.5	69.26	50.97	0.5/0.5	52.05	35.16	0.5/0.5	88.98	69.51
+SLERP ($\alpha^\star$)	T	0.8/0.8	**71.21**	52.46	0.7/0.7	**52.56**	35.71	1.0/0.8	**89.18**	70.41
SVD	I	–	69.31	51.29	–	51.06	35.04	–	**89.18**	69.82
+SLERP ($\hat{\alpha}$)	I	0.6/0.5	67.05	50.58	0.6/0.5	49.04	35.28	0.6/0.5	86.62	69.79
+SLERP ($\alpha^\star$)	I	0.9/0.8	69.64	52.36	1.0/0.7	51.06	35.86	1.0/0.8	**89.18**	70.48
SVD	I+T	–	69.31	51.29	–	51.06	35.04	–	**89.18**	69.82
+SLERP ($\hat{\alpha}$)	I+T	0.6/0.5	70.38	51.05	0.6/0.5	52.39	35.53	0.6/0.5	88.93	69.83
+SLERP ($\alpha^\star$)	I+T	0.8/0.8	71.05	**52.48**	0.7/0.7	52.48	**35.94**	1.0/0.7	**89.18**	**70.60**
► New model (SigLIP2)	–	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82

modalities, text-only support remains the most stable overall, while image-only and joint support can be competitive for specific retrieval directions (I2T or T2I).

J Re-indexing

Tabs. 7 and 8 report the re-indexing setting, where the gallery is allowed to be re-indexed. This setting differs from the compatibility evaluation in Eq. 15, since the gallery is no longer fixed in the old-model space. Instead, both query and gallery embeddings can be interpolated using the same

Table 8: Reindexed-gallery retrieval for cross-family model upgrades. For each dataset, the α column reports the interpolation weights for I2T/T2I retrieval. The Sup. column indicates the support modality used to estimate the Procrustes map: text-only (T), image-only (I), or joint image–text (I+T). The support-set-selected weight $\hat{\alpha}$ is selected on CC3M and fixed across all test datasets, whereas $\alpha^\star$ is selected directly on each test dataset and is reported only as an oracle upper bound. Bold and underlined values denote the best and second-best results, respectively, excluding the old- and new-model reference rows.

		Flickr30k			COCO			NoCaps		
Method	Sup.	α	I2T@1	T2I@1	α	I2T@1	T2I@1	α	I2T@1	T2I@1
SigLIP2 → CLIP ViT-B/32 *(cross-family)*										
► Old model (CLIP ViT-B/32)	–	–	40.62	21.73	–	28.76	14.47	–	71.29	45.24
SVD	T	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82
+SLERP ($\hat{\alpha}$)	T	0.8/0.9	70.37	51.72	0.8/0.9	51.67	35.33	0.8/0.9	**89.36**	70.34
+SLERP ($\alpha^\star$)	T	0.8/0.9	70.37	51.72	0.9/0.9	**51.91**	35.33	0.8/0.9	**89.36**	70.34
SVD	I	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82
+SLERP ($\hat{\alpha}$)	I	0.9/0.8	68.33	51.61	0.9/0.8	50.34	35.23	0.9/0.8	88.09	70.55
+SLERP ($\alpha^\star$)	I	1.0/0.9	69.31	**52.01**	1.0/0.9	51.06	**35.57**	1.0/0.9	89.18	**70.61**
SVD	I+T	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82
+SLERP ($\hat{\alpha}$)	I+T	0.8/0.8	**71.16**	51.59	0.8/0.8	51.69	35.12	0.8/0.8	89.18	70.46
+SLERP ($\alpha^\star$)	I+T	0.8/0.9	**71.16**	51.86	0.9/0.9	51.84	35.41	1.0/0.8	89.18	70.46
► New model (SigLIP2)	–	–	69.31	51.29	–	51.06	35.04	–	89.18	69.82
SigLIP1 → CLIP ViT-H/14 *(cross-family; new model not uniformly stronger than old model)*										
► Old model (CLIP ViT-H/14)	–	–	59.38	43.07	–	43.50	28.56	–	84.27	63.53
SVD	T	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18
+SLERP ($\hat{\alpha}$)	T	0.5/0.6	64.56	47.62	0.5/0.6	48.03	33.02	0.5/0.6	86.69	**67.42**
+SLERP ($\alpha^\star$)	T	0.4/0.5	64.68	**48.04**	0.7/0.7	48.63	**33.06**	0.9/0.6	**87.27**	**67.42**
SVD	I	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18
+SLERP ($\hat{\alpha}$)	I	0.9/0.6	60.12	46.98	0.9/0.6	47.52	32.89	0.9/0.6	86.11	67.31
+SLERP ($\alpha^\star$)	I	0.6/0.5	61.34	47.40	0.9/0.7	47.52	32.96	1.0/0.6	86.20	67.31
SVD	I+T	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18
+SLERP ($\hat{\alpha}$)	I+T	0.7/0.7	64.32	46.22	0.7/0.7	**49.27**	32.95	0.7/0.7	86.62	67.19
+SLERP ($\alpha^\star$)	I+T	0.4/0.5	**65.64**	47.49	0.7/0.7	**49.27**	32.95	0.8/0.6	87.07	67.31
► New model (SigLIP1)	–	–	58.09	39.40	–	46.99	30.88	–	86.20	64.18

SLERP weight: $\alpha = 0$ corresponds to the old-model representations, whereas $\alpha = 1$ corresponds to the SVD-aligned new-model representations.

In this setting, the SVD endpoint is equivalent to evaluating the new model directly. Because the orthogonal Procrustes map is applied to both query and gallery embeddings, it preserves all pairwise inner products and therefore leaves the retrieval ranking unchanged relative to the new model's original representation space. Consequently, the SVD results coincide with the new-model results.

SLERP often improves over this endpoint, indicating that interpolation can be beneficial even when re-indexing is allowed. These gains are particularly notable because SLERP requires no model retraining and, in several cases, matches or exceeds XBT, which requires compatibility training. Overall, these results show that SLERP not only restores compatibility with an existing old-model index, but can also produce stronger interpolated representations when re-indexing is permitted.

K Target-Set Budget Sensitivity

Fig. 8 examines how much target-distribution data is required to estimate a dataset-specific SLERP weight. For each target-set fraction, we subsample the target benchmark using three random seeds and compute Recall@1 along the SLERP path. The solid curves show the mean Recall@1 across seeds, and the shaded regions indicate one standard deviation.

Across both model pairs, even small target subsets identify the same high-performing region of the SLERP curve. Reducing the target-set budget increases the variance of the Recall@1 estimates, but the maximizer remains stable and close to that obtained using the full target set. This suggests that a reliable dataset-specific interpolation weight can be estimated from a modest subset of the target distribution, without requiring evaluation on the full benchmark.

This analysis is separate from the protocol used in the main results, where $\hat{\alpha}$ is selected once on the CC3M validation set and then fixed for evaluation on Flickr30k, COCO, and NoCaps. Its purpose

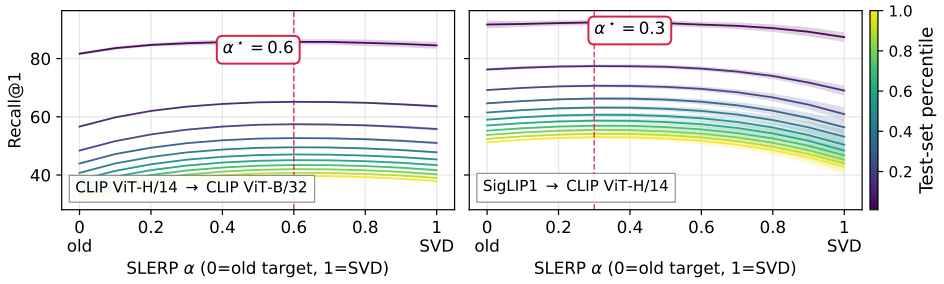
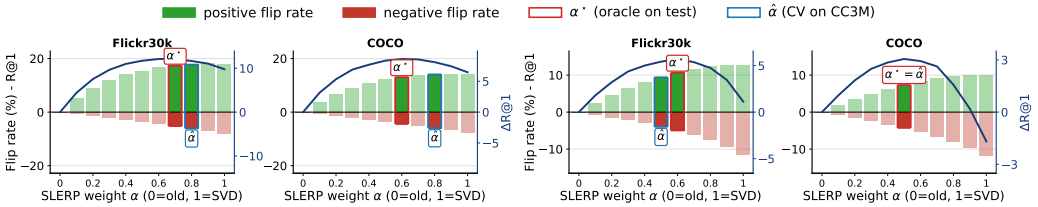


Figure 8: Recall@1 along the SLERP path under different target-set fractions. Solid curves report the average over three random seeds; shaded regions denote $\pm 1\sigma$ across seeds.



(a) CLIP ViT-H/14 → CLIP ViT-B/32 (b) SigLIP1 → CLIP ViT-H/14

Figure 9: **I2T Recall@1 flip-rate trade-off on Flickr30k and COCO.** Bars show positive and negative flip rates relative to the old-to-old evaluation. The blue curve reports $\Delta\mathrm{R@1} = \mathrm{PFR} - \mathrm{NFR}$. Here $\alpha = 0$ corresponds to old-model queries and $\alpha = 1$ to SVD-aligned new-model queries. Red and blue markers denote the dataset-specific oracle $\alpha^\star$ and the CC3M-selected weight $\hat{\alpha}$, respectively.

is to assess a complementary scenario in which a small subset from the deployment distribution is available. In this setting, the results indicate that such a subset is sufficient to estimate a reliable dataset-specific SLERP weight.

L Additional Flip Analyses

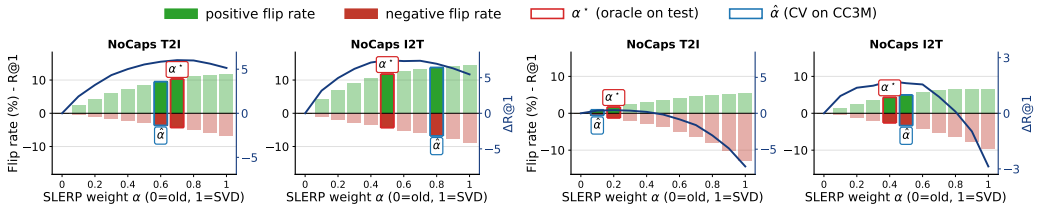
Figures 9 and 10 extend the flip analysis from Sec. 4.3. Fig. 9 presents I2T Recall@1 flips on Flickr30k and COCO for CLIP ViT-H/14 → CLIP ViT-B/32 and SigLIP1 → CLIP ViT-H/14. Fig. 10 presents the corresponding NoCaps analysis for both T2I and I2T retrieval. In all plots, $\alpha = 0$ corresponds to the old-model query representation, whereas $\alpha = 1$ corresponds to the SVD-aligned new-model query representation.

These additional results are consistent with the trend observed in Fig. 4. The optimal interpolation weight is governed by the balance between positive and negative flips, rather than by the positive-flip rate alone. Across retrieval directions and datasets, SLERP improves Recall@1 when it corrects more previously incorrect queries than it causes previously correct queries to fail.

M Per-Query Oracle Analysis

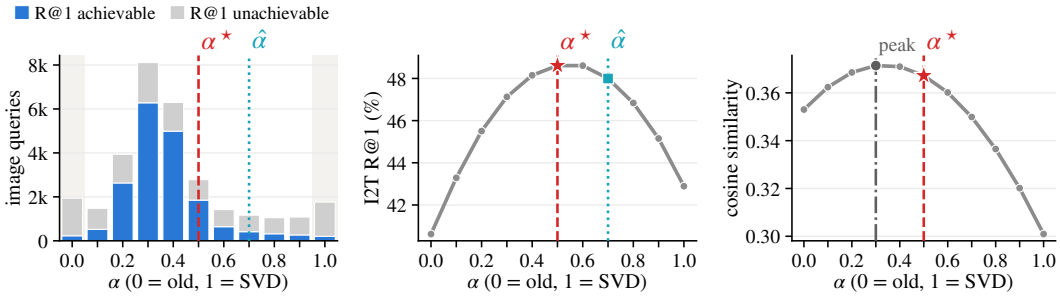
The geometric characterization in Sec. 3 is formulated with respect to an idealized retrieval-optimal direction q^*, which is not observable in practice. We therefore complement it with an empirical analysis at the level of individual queries, asking whether interior points of the evaluated SLERP arc provide retrieval benefits that neither endpoint provides. This analysis uses query-level test labels and is intended only as an oracle upper bound; it does not constitute a deployable weight-selection procedure.

We consider Flickr30k image-to-text retrieval for CLIP ViT-L/14 → CLIP ViT-B/32 with text-only Procrustes alignment, evaluated on the grid $\mathcal{A} = \{0, 0.1, \ldots, 1\}$, where $\alpha = 0$ is the old-model query and $\alpha = 1$ is the SVD-aligned new-model query. For each query $i \in \mathcal{Q}$ and weight $\alpha \in \mathcal{A}$, let $c_i(\alpha) \in \{0, 1\}$ indicate whether a relevant caption is retrieved at rank one, and let $s_i(\alpha)$ denote



(a) CLIP ViT-H/14 → CLIP ViT-B/32 (b) SigLIP1 → CLIP ViT-H/14

Figure 10: **NoCaps Recall@1 flip-rate trade-off.** We report both T2I and I2T flip rates for the same old–new pairs used in the main paper. Bars show positive and negative flip rates relative to the old-to-old evaluation, and the blue curve reports $\Delta\mathrm{R@1} = \mathrm{PFR} - \mathrm{NFR}$. Here $\alpha = 0$ corresponds to old-model queries and $\alpha = 1$ to SVD-aligned new-model queries. Red and blue markers denote the dataset-specific oracle $\alpha^\star$ and the CC3M-selected weight $\hat{\alpha}$, respectively.



(a) Per-query oracle weights. (b) Fixed-weight Recall@1. (c) Relevant-item similarity.

Figure 11: Per-query oracle analysis on Flickr30k image-to-text retrieval for CLIP ViT-L/14 → CLIP ViT-B/32 with text-only alignment. **(a)** Distribution of per-query oracle interpolation weights, where each query is assigned the weight attaining the best rank of its highest-ranked relevant item, with ties broken by the highest relevant-item similarity and then by proximity to $\alpha = 0.5$. Queries are divided into those for which Recall@1 is achievable at some interpolation weight (blue, $n = 18{,}344$) and those for which it is not (gray, $n = 12{,}670$); shaded columns mark the two endpoints of the interpolation path. **(b)** Recall@1 obtained with a single fixed interpolation weight applied to all queries, maximized at the dataset-level optimum $\alpha^\star = 0.5$. **(c)** Cosine similarity between the interpolated query and its highest-scoring relevant item at each weight, averaged over Recall@1-achievable queries. Red dashed and cyan dotted lines mark the dataset oracle $\alpha^\star = 0.5$ and the CC3M-selected weight $\hat{\alpha} = 0.7$, respectively. Because most successful queries attain rank one over many weights, the oracle distribution follows the relevant-item similarity and peaks at $\alpha = 0.3$, rather than at the Recall@1-optimal fixed weight $\alpha^\star$.

the cosine similarity between the interpolated query and its highest-scoring relevant caption. We distinguish two quantities: $c_i(\alpha)$ measures retrieval success and depends on all gallery items, whereas $s_i(\alpha)$ measures proximity to the relevant item alone.

The per-query oracle Recall@1 selects the best weight independently for each query,

$$\mathrm{R@1}_{\mathrm{PQ}} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \max_{\alpha \in \mathcal{A}} c_i(\alpha). \tag{119}$$

Among the 31,014 Flickr30k image queries, 18,344 achieve Recall@1 for at least one weight, giving $\mathrm{R@1}_{\mathrm{PQ}} = 59.15\%$. To isolate the contribution of the interior of the arc, we compare it with an endpoint-only oracle that, for each query, may choose only between the old-model query and the SVD-aligned query, $\max\{c_i(0), c_i(1)\}$. This endpoint-only oracle reaches 54.61% (16,936 queries). The remaining 1,408 queries achieve Recall@1 exclusively at interior weights, so interior points account for a gain of 4.54 Recall@1 points that no per-query choice between the two endpoints can recover. Beyond rank one, for 6,647 queries (21.43%) the best relevant-item rank attained at an interior weight is strictly better than the best rank attained at either endpoint.

As Recall@1 is a thresholded event, most successful queries remain successful over a wide range of weights: 96.57% of the Recall@1-achievable queries are retrieved at rank one for more than one weight (8.53 of the 11 weights on average), including $8{,}873$ queries retrieved at rank one for every weight. For such queries, Recall@1 alone does not identify a preferred position along the arc.

The relevant-item similarity $s_i(\alpha)$ provides a direct observable counterpart of Theorem 1. Instantiating q^* as the normalized embedding of a relevant caption g, Theorem 1 gives $\langle q_\alpha, g \rangle = \rho \cos(\alpha\theta - \psi)$. Therefore, if the similarity to g at some interior grid weight strictly exceeds its values at both endpoints, the in-plane projection of g lies in the relative interior of the minor arc from u to v. We observe this for 99.87% of all queries: for nearly every query, the similarity to at least one relevant caption is maximized strictly inside the arc. Averaged over Recall@1-achievable queries, $s_i(\alpha)$ peaks at $\alpha = 0.3$, whereas the fixed-weight Recall@1 peaks at $\alpha^\star = 0.5$ (Figs. 11b and 11c), since retrieval success also depends on how the similarities to non-relevant items change along the arc.

To assign a single oracle weight to each query, we select the weight with the best relevant-item rank; ties are broken by the highest relevant-item similarity $s_i(\alpha)$, and any remaining tie in favor of the weight closest to $\alpha = 0.5$. Fig. 11a shows the resulting distribution. Because most successful queries are tied in rank over many weights, the similarity criterion determines their oracle weight, and the distribution follows the relevant-item similarity: it peaks at $\alpha = 0.3$ rather than at $\alpha^\star = 0.5$ or at the CC3M-selected $\hat{\alpha} = 0.7$. Under this rule, $17{,}912$ of the $18{,}344$ Recall@1-achievable queries (97.65%) and 88.1% of all queries are assigned an interior weight. These fractions reflect where the relevant-item similarity is maximized among rank-optimal weights and should not be read as the fraction of queries that require interpolation to succeed, which is given by the $1{,}408$ interior-only queries above.

Tab. 9 summarizes the resulting upper bounds. The per-query oracle reaches 59.15% I2T Recall@1, compared with 40.62% for the old model, 42.89% for SVD alignment alone, 48.00% for the fixed CC3M-selected weight, and 48.61% for the dataset-level oracle weight. Part of this gain comes from choosing, per query, between the two endpoints (54.61%); the remaining 4.54 points are attainable only at interior weights. The per-query oracle also exceeds the 48.72% obtained by fully re-indexing with CLIP ViT-L/14; this comparison should not be interpreted as a deployable advantage, since the oracle uses query-level test labels to select a different weight for each query.

Table 9: Flickr30k I2T Recall@1 (%) for CLIP ViT-L/14 $\rightarrow$ ViT-B/32 with text-only alignment.

Method	R@1
▶ **Old model**	40.62
SVD	42.89
SLERP ($\hat{\alpha} = 0.7$)	48.00
SLERP ($\alpha^\star = 0.5$)	48.61
▶ **New model**	48.72
Endpoints only (per-query oracle)	54.61
SLERP (per-query oracle)	**59.15**

This experiment does not recover q^* itself. It tests two observable counterparts of Theorem 1. First, instantiating q^* as a relevant item, the interior condition of the theorem holds for 99.87% of queries, indicating that the residual direction between the old-model and aligned new-model queries consistently moves the query closer to its relevant items. Second, at the level of the discrete retrieval event, interior weights yield Recall@1 successes unavailable at either endpoint for $1{,}408$ queries and strictly improve the best relevant-item rank for 21.43% of queries. Whether a higher relevant-item similarity translates into a retrieval gain depends on the margin to non-relevant items, consistent with the margin-based certification in Appendix F. Finally, the gap between the per-query oracle and the fixed support-set-selected weight indicates that the preferred interpolation position varies across queries.

N Alignment and Endpoint-Combination Baselines

To identify the source of the improvements observed in the main experiments, we decompose the proposed approach into its three components: the alignment map, the operator that combines the two query representations, and the selection of the interpolation position. We compare orthogonal Procrustes with non-isometric alignment maps, and SLERP with alternative ways of combining the old-model query u and the aligned new-model query v. The old-model gallery is always left unchanged throughout the analysis.

All results cover the five model pairs, three retrieval datasets, three alignment-support modalities, and both retrieval directions of Sec. 4, i.e., 45 configurations and 90 compatibility evaluations. Every alignment map, regularization parameter, and interpolation parameter is estimated or selected only on the CC3M validation split used as alignment support set, separately for I2T and T2I Recall@1, and then used for evaluation on Flickr30k, COCO, and NoCaps. Interpolation coefficients are selected over the grid $\{0, 0.1, \dots, 1\}$. Compatibility is assessed with the empirical protocol of Eq. 15.

N.1 Alignment baselines

Let $\bar{V} \in \mathbb{R}^{N_a \times d_{\text{new}}}$ and $U \in \mathbb{R}^{N_a \times d_{\text{old}}}$ be the support-set embeddings of Sec. 3. Each map below replaces the orthogonal Procrustes map $R^\star$ used in our approach; the resulting query is normalized before retrieval, and no interpolation is applied.

Affine alignment. An unconstrained linear map with intercept is fitted by least squares,

$$(A^\star, b^\star) = \arg\min_{A,b} \left\|\bar{V}A + \mathbf{1}b^\top - U\right\|_F^2, \tag{120}$$

and the normalized aligned query is $v = (\bar{v}A^\star + b^{\star\top})/\|\bar{v}A^\star + b^{\star\top}\|_2$.

Ridge alignment [84]. The same map is fitted with Tikhonov regularization [85],

$$(A^\star, b^\star) = \arg\min_{A,b} \left\|\bar{V}A + \mathbf{1}b^\top - U\right\|_F^2 + \lambda\|A\|_F^2, \tag{121}$$

with $\lambda \in \{10^{-6}, 10^{-4}\}$ selected directly on the support set.

One-sided CCA. The new-model embeddings are projected onto d_{old} canonical directions with covariance regularization $\epsilon \in \{10^{-6}, 10^{-4}\}$, while the old-model side is kept in its original coordinates so that the gallery remains unchanged. This is a special case of standard CCA [86]. With an untransformed target, the canonical projection followed by the least-squares map to the old coordinates reduces to regularized regression; one-sided CCA therefore coincides with ridge alignment, and indeed produces identical results in all 45 configurations.

Whitened Procrustes. Both embedding sets are centered and whitened with ϵ-regularized covariances before solving the orthogonal Procrustes problem, and the aligned query is mapped back to the old-model coordinates and normalized; $\epsilon \in \{10^{-4}, 10^{-2}, 10^{-1}, 1\}$ is selected on the support set.

N.2 Endpoint-combination baselines

All combination methods use the same orthogonally aligned endpoint v obtained with Procrustes.

Fixed normalized midpoint. The simplest combination uses no weight selection:

$$q_{\text{mid}} = \frac{u + v}{\|u + v\|_2} = \text{slerp}(u, v; 0.5), \tag{122}$$

where the second equality holds for unit, non-antipodal endpoints. This baseline isolates the effect of combining the old and aligned-new representations from the effect of validation-based position selection.

Normalized linear interpolation. We obtain

$$q_\beta^{\text{nlerp}} = \frac{(1-\beta)u + \beta v}{\|(1-\beta)u + \beta v\|_2}, \qquad \beta \in [0, 1], \tag{123}$$

with β selected on the support set using the same protocol as the SLERP weight. As shown in Appendix C, NLERP and SLERP traverse the same minor geodesic between u and v under a monotone reparameterization: SLERP is linear in angular displacement, whereas the angular position associated with an NLERP coefficient depends on the query-specific endpoint angle.

Score interpolation and score ensembling. For an old-model gallery embedding g_j, interpolating the scores of the two endpoints gives

$$s_j^{\text{score}}(\beta) = (1-\beta)\langle u, g_j\rangle + \beta\langle v, g_j\rangle = \big\langle (1-\beta)u + \beta v,\, g_j\big\rangle. \tag{124}$$

The ensemble of old-model scores and aligned-new-model scores is the same computation. For a fixed query, the normalization in Eq. 123 rescales all gallery scores by the same positive constant, so score interpolation, score ensembling, and NLERP at the same coefficient induce identical rankings.

Query-adaptive NLERP. To test whether a query-dependent position helps, the coefficient is predicted from the endpoint agreement of each query,

$$\beta_i = \sigma\big(b_m + s_m\langle u_i, v_i\rangle\big), \tag{125}$$

where σ is the logistic function and (b_m, s_m) are selected separately for image and text queries on CC3M over $b_m \in \{-3, -1.5, 0, 1.5, 3\}$ and $s_m \in \{-3, 0, 3, 6\}$. The rule uses no test labels: at inference it only requires the cosine between the two endpoints, which is available at negligible cost.

N.3 Discussion

Tab. 10 shows that orthogonal Procrustes is the strongest alignment method among the evaluated ones. The non-isometric maps, although more flexible on the support set, generalize substantially worse to the target benchmarks: affine alignment loses 15.35 mean Recall@1 points relative to Procrustes, ridge regression and one-sided CCA lose 11.32, and whitened Procrustes loses 1.83, with compatibility dropping from $45/90$ to 11, 14, 14, and $37/90$, respectively. Because these maps rescale or shear the new-model representation, they distort the inner-product geometry that retrieval relies on, whereas an isometry preserves it. The table also shows that alignment alone is not sufficient: SVD remains below the unchanged old-model query by 1.64 I2T and 0.49 T2I Recall@1 points on average, which is why it satisfies compatibility in only half of the evaluations.

Table 10: Aggregate Recall@1 ablation over 45 configurations (90 compatibility evaluations) reported in Tabs. 4, 5, and 6. Δ_{SVD} and Δ_{old} are averaged over both retrieval directions; Comp. counts evaluations satisfying Eq. 15.

Method	I2T	T2I	Δ_{SVD}	Δ_{old}	Comp.
▶ **Old-model query**	53.36	34.26	+1.06	0.00	–
Alignment only					
SVD (Procrustes)	51.72	33.77	0.00	−1.06	45/90
Affine	31.69	23.10	−15.35	−16.42	11/90
Ridge	33.96	28.89	−11.32	−12.38	14/90
One-sided CCA	33.96	28.89	−11.32	−12.38	14/90
Whitened Procrustes	49.06	32.78	−1.83	−2.89	37/90
SVD + endpoint combination					
Midpoint	57.76	37.23	+4.75	+3.68	72/90
NLERP	57.73	37.38	+4.81	+3.74	84/90
Score interp.	57.73	37.38	+4.81	+3.74	84/90
Adaptive NLERP	**57.86**	**37.42**	**+4.89**	**+3.83**	**88/90**
SLERP	57.72	37.38	+4.80	+3.74	85/90

Replacing the SVD endpoint by the fixed normalized midpoint raises mean I2T and T2I Recall@1 from 51.72 and 33.77 to 57.76 and 37.23, a mean improvement of 4.75 points over SVD. Relative to the unchanged old-model query, it improves I2T and T2I Recall@1 by 4.40 and 2.97 points, and it increases compatibility from $45/90$ to $72/90$. Since this baseline has no tunable weight and uses no validation data, this gain is attributable to the combination of the two endpoints.

Support-set selected interpolation substantially improves compatibility robustness. SLERP improves mean Recall@1 over SVD by 4.80 points, only 0.05 more than the midpoint, while raising compatibility from $72/90$ to $84/90$; relative to the old-model query, it improves I2T and T2I Recall@1 by 4.36 and 3.12 points. Support-set selected NLERP reaches 4.81 points and $84/90$, and score interpolation matches it, as expected from their ranking equivalence.[3] Query-adaptive NLERP gives the best results, 4.89 points and $88/90$, but improves mean Recall@1 by only 0.09 points over SLERP while requiring four additional calibrated parameters; its higher compatibility nevertheless indicates that query-dependent positions are a promising direction, consistent with the per-query oracle analysis in Appendix M.

The ablation separates the two sources of improvement. Combining the old-model query with the orthogonally aligned new-model query accounts for most of the mean Recall@1 improvement, while

[3] In one configuration the selected coefficient coincides with the old-model endpoint, where floating-point ties change the outcome of a single query; we count it as non-compatible, consistent with NLERP.

validation-based position selection makes these improvements consistently satisfy the backward-compatibility criterion across model pairs, datasets, support modalities, and retrieval directions. Similar improvements are observed with SLERP, NLERP, and score interpolation, indicating that they arise from exploiting the residual post-alignment geometry rather than from a particular combination operator. We adopt SLERP as the canonical parameterization because $\angle(u, q_\alpha^{\mathrm{slerp}}) = \alpha \angle(u, v)$, so α represents the same fraction of the angular displacement from the old-model query toward the aligned new-model query for every query. This constant-angular-speed property is precisely what the characterization in Sec. 3 relies on. Under NLERP or score interpolation, by contrast, the angular position associated with a fixed coefficient depends on the query-specific angle between the endpoints.

O Limitations

Selection of the interpolation weight. SLERP requires selecting an interpolation weight along the geodesic arc between the old-model query embedding and the SVD-aligned new-model query embedding. The optimal value is not known a priori and must be estimated from data, either from the deployment distribution or from a separate support set. In the main experiments, we avoid target-set tuning by selecting $\hat{\alpha}$ on the CC3M alignment support set and applying the same value to Flickr30k, COCO, and NoCaps. The target-set budget analysis in Appendix K shows that, when a small labeled subset from the deployment distribution is available, a reliable interpolation weight can be estimated from only a modest number of examples. At the same time, the retrieval results in Sec. 4 and the zero-shot classification results in Appendix G show that SLERP is robust to this choice: support-set selected weights transfer well across target datasets and remain close to the best dataset-specific performance.

Query-side computation. SLERP requires no compatibility training and leaves the deployed gallery unchanged, but it needs both endpoint representations at inference time: the old-model query embedding and the SVD-aligned new-model query embedding. Compared with using a single encoder, SLERP therefore adds query-side computation and requires both encoders, or suitable cached representations, to be available during deployment. This overhead is independent of the gallery size: the existing gallery remains indexed in the old-model space, and only incoming queries require the additional computation. A full model upgrade, by contrast, requires re-encoding and re-indexing the entire gallery before the new model can be used directly. In practice, SLERP is best viewed as a compatibility mechanism for the migration period (Appendix P) following a model update, keeping the old index usable while full gallery re-indexing proceeds offline.

Dependence on endpoint quality. Our method requires neither model retraining nor gradient-based optimization: it does not update either encoder, learn dataset-specific correction layers, or optimize compatibility losses on the target benchmarks. Its only fitted components are the closed-form Procrustes map and the interpolation weight, both estimated on the alignment support set. Consequently, its performance depends on the quality and complementarity of the old-model query embedding, the new-model query embedding, and the estimated Procrustes map. If both endpoint representations are weak for a deployment domain, interpolation alone cannot introduce new task-specific information. The same property that limits the method also supports its transfer: because nothing is fitted to a specific target benchmark, the same support-set-selected weight carries over across retrieval benchmarks and to zero-shot classification with fixed old-model text prototypes.

P Deployment Cost and Migration-Period Analysis

This appendix quantifies the practical cost of the proposed approach. We first analyze the query-side computation and latency required by SLERP queries against the old-model gallery, and relate them to the cost of full gallery re-indexing. We then evaluate SLERP during the migration period, in which the gallery is progressively re-encoded with the new aligned model, and compare it with partial gallery backfilling. Unless otherwise stated, we consider the CLIP ViT-L/14 $\rightarrow$ CLIP ViT-B/32 upgrade with text-only Procrustes support.

Table 11: Per-query serving cost for CLIP ViT-L/14 $\rightarrow$ CLIP ViT-B/32 with a gallery of 10^7 items on one NVIDIA A100. Encoder compute is taken from the OpenCLIP profiles and excludes the Procrustes projection ($\approx$ 0.79 MFLOPs) and the search. Latency is the sum of independently measured median (p50) encoder and IVF-PQ search components; in the two-GPU setting, the old and new encoders are executed concurrently on separate devices.

			Encoder compute (GFLOPs)		Latency (ms)	
Serving configuration	Gallery	Query encoders	I2T	T2I	I2T	T2I
Old model	ϕ_{old}	ϕ_{old}	8.82	5.96	19.93	19.95
Full re-indexing	ϕ_{new}	ϕ_{new}	162.03	13.30	25.59	20.40
SLERP, one GPU	ϕ_{old}	$\phi_{\text{old}} + \phi_{\text{new}}$	170.85	19.26	31.04	25.76
SLERP, two GPUs	ϕ_{old}	$\phi_{\text{old}} + \phi_{\text{new}}$	170.85	19.26	25.87	20.70
SLERP, cached old query	ϕ_{old}	ϕ_{new}	162.03	13.30	25.87	20.68

P.1 Query-Side Cost and Re-indexing Trade-off

SLERP requires both the old-model query embedding u and the SVD-aligned new-model query embedding v for every incoming query. Its main cost is therefore the evaluation of the two encoders, whereas the remaining operations are negligible. According to the OpenCLIP profiles, the ViT-B/32 image and text encoders require 8.82 and 5.96 GFLOPs per input, respectively, while the corresponding ViT-L/14 encoders require 162.03 and 13.30 GFLOPs. By comparison, the 768×512 Procrustes projection requires $393{,}216$ multiply-accumulate operations, or approximately 0.79 MFLOPs, and the spherical interpolation operation is linear in the embedding dimension. The gallery, its index, and the complexity of approximate nearest-neighbor search are left unchanged, since each query remains a single d_{old}-dimensional unit vector searched against the old-model index. Aggregated over 10^5 queries and a gallery of 10^7 items, the encoder compute of SLERP amounts to 17.09 PFLOPs for I2T and 1.93 PFLOPs for T2I retrieval, compared with 0.88 and 0.60 PFLOPs for old-model queries. Full re-indexing followed by new-model queries instead requires 149.20 PFLOPs and 1.62 EFLOPs, respectively, the latter being dominated by the re-encoding of the image gallery.

Tab. 11 reports the corresponding query latency, measured on an NVIDIA A100 with an IVF-PQ index built over 10^7 synthetic gallery vectors. When the two encoders are evaluated sequentially on a single GPU, SLERP increases the latency by approximately 5.4 ms relative to the fully re-indexed system. Since u and v are computed independently, the two encoders can be executed concurrently, reducing the encoding latency from $\phi_{\text{old}} + \phi_{\text{new}}$ to $\max(\phi_{\text{old}}, \phi_{\text{new}})$. With the encoders placed on separate GPUs, or when the old-model query embedding is cached, the gap to the fully re-indexed system reduces to approximately 0.3 ms. In all configurations, the projection and interpolation stages contribute less than 1 ms, and the overhead is dominated by encoder inference.

P.2 Migration Period with Partial Backfilling

In practice, large galleries are rarely re-encoded at once; instead, they are backfilled progressively while the system remains in service [12, 13]. We therefore compare SLERP with partial gallery backfilling for image-to-text retrieval on Flickr30k. In the partial-backfill baseline, a random fraction $f \in \{0\%, 10\%, \ldots, 100\%\}$ of the captions is re-encoded with the new model, while the remaining captions retain their old-model embeddings, and queries are computed with the aligned new-model encoder. Consequently, $f = 0\%$ recovers the SVD baseline with a fixed old-model gallery, and $f = 100\%$ recovers full re-indexing. SLERP, instead, uses the CC3M-selected $\hat{\alpha}$ of Tab. 1 and leaves the gallery entirely unchanged. We report the results of the analysis in Fig. 12 for CLIP ViT-L/14 $\rightarrow$ CLIP ViT-B/32 and SigLIP2 $\rightarrow$ SigLIP1 model pairs.

For CLIP ViT-L/14 $\rightarrow$ CLIP ViT-B/32, SLERP matches or exceeds partial backfilling at every evaluated fraction below full re-indexing. Without re-encoding any gallery item, SLERP reaches 48.00 Recall@1, whereas partial backfilling requires re-encoding 90% of the text gallery ($139{,}563$ captions) to reach 47.98. Relative to the old model, SLERP thus recovers 91.1% of the improvement obtained by full re-indexing (48.72). For SigLIP2 $\rightarrow$ SigLIP1, the SVD-aligned query falls far below the old model (45.46 vs. 58.09), and partial backfilling exceeds the old model only once 50% of the text gallery has been re-encoded. SLERP, in contrast, satisfies the compatibility criterion (58.63)

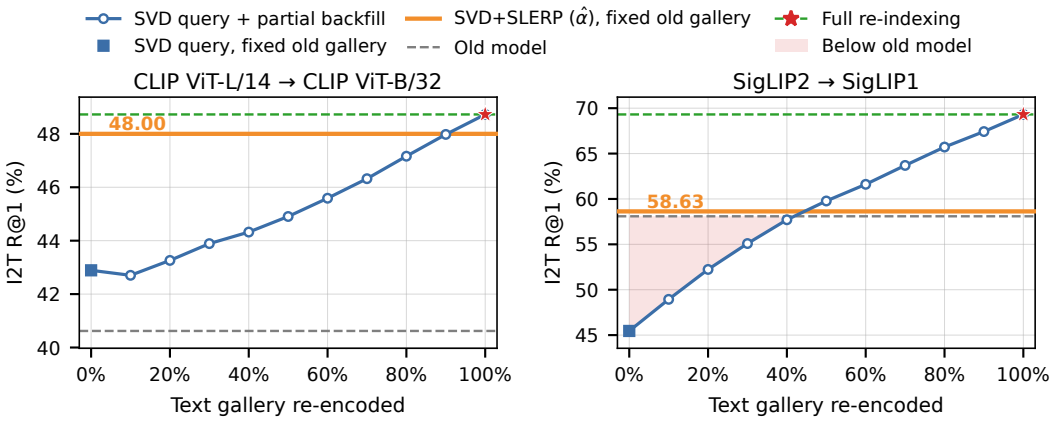


Figure 12: Image-to-text Recall@1 during partial backfilling of the text gallery on Flickr30k, with text-only Procrustes support. The partial-backfill baseline issues SVD-aligned new-model queries against a mixed gallery in which an increasing fraction of old-model caption embeddings is replaced by new-model embeddings. At 0%, the baseline coincides with the SVD query on the fixed old-model gallery (square), and at 100% with full re-indexing (star). SVD+SLERP uses the CC3M-selected $\hat{\alpha}$ and leaves the gallery unchanged. Shaded regions mark backfilling fractions at which the baseline falls below the old-model system and thus violates Eq. 15.

with an unchanged gallery, and partial backfilling surpasses it only once between 40% and 50% of the captions have been re-encoded.

As backfilling progresses, direct retrieval with the new model becomes preferable, and once re-indexing is complete, the old encoder can be retired. Combining SLERP queries for non-backfilled items with new-model queries for backfilled items is a natural extension of this analysis, also supported by [87].

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 accurately describe the paper's contributions and scope: a training-free query-side SLERP method after orthogonal Procrustes alignment for backward-compatible VLM retrieval. The stated theoretical claims are developed in Sec. 3 and Appendix F, and the empirical claims are supported by experiments across CLIP, SigLIP, and SigLIP2 on Flickr30k, COCO, and NoCaps in Sec. 4.

   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: The paper includes an introduction to the limitations discussion in Sec. 5 with a dedicated Appendix O, covering interpolation-weight selection, the need to compute both old and new query embeddings at inference time, and the dependence of the training-free method on the quality of the endpoint representations and Procrustes alignment.

   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: The paper states the assumptions for the geometric analysis in Sec. 3, including normalized embeddings and the non-degenerate condition. The main theoretical results are demonstrated in Appendix E, and Appendices D and F.

   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: The paper uses closed-form alignment and interpolation procedures, and provides the details needed to reproduce the main experiments in Sec. 4.1 and Sec. 4.2. These include the public model checkpoints, datasets, alignment support set, support-modality choices, Procrustes/SVD alignment procedure, SLERP grid search for $\hat{\alpha}$, baselines, metrics, and compatibility criterion. The main results in Sec. 4.3 are further supported by detailed Recall@1/5/10 results and additional analyses in Appendices I, G, J, and K.

   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: [No]

   Justification: The experiments use publicly available datasets and model checkpoints, and the paper describes the closed-form Procrustes/SVD alignment and SLERP evaluation protocol in Sec. 4.1 and Sec. 4.2. A minimal reference implementation of the proposed method is publicly available at https://github.com/miccunifi/SLERP_backward_compatibility. However, exact run commands and full reproduction scripts for the proposed method and baselines are not yet included; they will be released in the same repository.

   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: The paper specifies the experimental setting in Sec. 4.1 and Sec. 4.2, including the evaluated model pairs, public checkpoints, datasets, CC3M alignment/validation split, support-modality choices, baselines, Recall@K metrics, compatibility criterion, and the SLERP grid used to select $\hat{\alpha}$. Since the proposed method is training-free and uses closed-form Procrustes/SVD alignment, no optimizer or training hyperparameters are required for the main method.

   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: [No]

   Justification: Most of the main experiments are deterministic evaluations based on frozen pretrained models and closed-form Procrustes/SVD alignment, so the paper reports point estimates for the main retrieval and classification results. The only analysis involving randomness is the target-set budget sensitivity study in Appendix K, where target subsets are sampled with three random seeds and the paper reports the mean Recall@1 curve with shaded $\pm 1\sigma$ bands (see Fig. 8).

   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: [No]

   Justification: The paper reports the experimental protocol, datasets, model checkpoints, and evaluation metrics, but it does not currently provide detailed compute-resource information such as GPU/CPU type or memory.

   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 research is conducted according to the NeurIPS Code of Ethics.

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: [No]

Justification: The paper focuses on a technical method for backward-compatible vision-language retrieval and discusses practical limitations in Appendix O, but it does not separately discuss both positive and negative societal impacts. Potential impacts include reducing the cost of model upgrades for retrieval systems, while possible risks include improving retrieval systems used in sensitive applications such as surveillance or biased content search.

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: The paper does not release new pretrained models, image generators, scraped datasets, or other assets with high misuse risk. Our work uses existing public pretrained vision-language models and benchmark datasets.

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: [No]

    Justification: The paper uses existing assets, including public pretrained model checkpoints and benchmark datasets, and credits their original sources in Sec. 4.1. However, the current version does not explicitly list the licenses or terms of use for each dataset, model checkpoint, or code asset used in the experiments.

    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: [N/A]

    Justification: The paper does not introduce or release new assets.

    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 does not involve crowdsourcing nor research with 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 does not involve crowdsourcing nor research with human subjects.

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: The use of LLMs was limited to writing, editing, or formatting.

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.