Spherical Interpolation for Backward-Compatible Multimodal Representations

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

[1]DINFO, University of Florence [2]MICC – Media Integration and Communication Center [3]University of Trento Contact: simone.ricci@unifi.it

NEURAL INFORMATION PROCESSING SYSTEMS

We **rotate new-model queries** into the old model's space (orthogonal Procrustes), then we select a **point on the spherical arc** between the old and the aligned new query to obtain the final query.

A **single weight**, chosen on a support set, makes the upgrade **backward-compatible in 85 of 90** evaluations (45 with alignment alone).

No retraining, no re-indexing.

We characterize **when** a point inside the arc beats both endpoints.

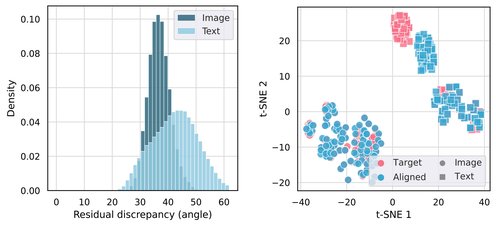
Motivation

Background:

- VLMs embed images and text on a **shared unit hypersphere**.
- Re-indexing a gallery is **costly**, or *impossible* (privacy, retention).

Post-hoc Alignment:

- Train the new model **freely**, then align it with a light map.
- An **orthogonal map** keeps the image–text geometry, but a **residual angle** remains between the two representations:



Residual angle after alignment, SigLIP2 → CLIP ViT-B/32 (Flickr30k).

Method

1. **Align.** Fit an orthogonal map (new ➡ old) on 12.6K CC3M pairs, in *closed form*:

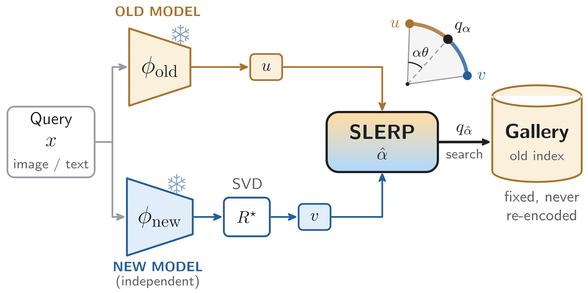
$$R^\star = \arg\min_{R^\top R = I} \left\| \bar{V} R - U \right\|_F^2 = PQ^\top, \qquad \bar{V}^\top U = P\Sigma Q^\top$$

2. **Interpolate.** Move along the geodesic from the old to the aligned new query.

$$u = \phi_{\text{old}}(x), \qquad v = \phi_{\text{new}}(x)\, R^\star, \qquad \theta = \angle(u, v)$$

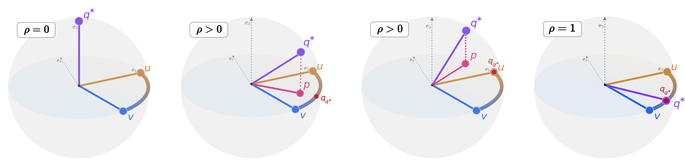
$$q_\alpha = \frac{\sin\big((1-\alpha)\theta\big)}{\sin\theta}\, u + \frac{\sin(\alpha\theta)}{\sin\theta}\, v$$

3. **Retrieve.** Search the **unchanged** old gallery, one interpolation weight for all queries.



Geometric Characterization

The intuition. The new model is **better overall**, but its aligned query is still *slightly off* in the old model space.



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

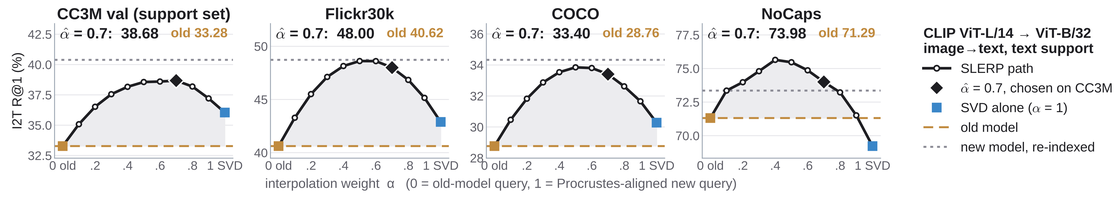
Project the ideal query onto the interpolation plane of the old and the aligned new query. If the projection falls **on the arc** (d), an **interpolated** query is closest to the ideal query. If it falls before or beyond the arc (b, c), one of the two endpoints is the best possible query.

Theorem 1. An interior query strictly beats both endpoints **iff** the ideal query projects **inside the arc**, as in (d):

$$G(\alpha^*) < \min\{G(0), G(1)\} \iff \frac{p}{\rho} \in \operatorname{relint}\left(\widehat{uv}\right)$$

Does it happen in practice? Yes, on Flickr30k:

99.87%	1,408	21.43%
of queries: relevant caption closest inside the arc	queries retrieved correctly only inside the arc	of queries get a strictly better rank inside the arc



Along the arc: Recall@1 peaks **inside the arc**, above both endpoints, on every benchmark. The weight chosen on **CC3M** transfers to all three test sets.

Compatibility Results

45 ➡ 85 /90	40.62 ➡ 48.00	91.1%	45.46 ➡ 58.63
compatible evaluations, SVD ➡ SVD + SLERP	Flickr30k I2T R@1, 0 items re-encoded	of the full re-indexing gain, gallery untouched	SigLIP2 ➡ SigLIP1: compatibility restored

Cross-modal backward-compatible retrieval, Recall@1 (text-only support)

Upgrade (new → old)	Method	Flickr30k I2T@1	Flickr30k T2I@1	COCO I2T@1	COCO T2I@1	NoCaps I2T@1	NoCaps T2I@1
CLIP ViT-L/14 → CLIP ViT-B/32	Old model	40.62	21.73	28.76	14.47	71.29	45.24
	XBT (training)	42.47 ✓	22.38 ✓	30.73 ✓	**15.55** ✓	**75.02** ✓	**48.02** ✓
	SVD	42.89 ✓	21.49 ✗	30.27 ✓	14.03 ✗	69.22 ✗	43.92 ✗
	+SLERP ($\hat{\alpha}$)	**48.00** ✓	**23.33** ✓	**33.40** ✓	15.26 ✓	73.98 ✓	46.59 ✓
SigLIP2 → SigLIP1	Old model	58.09	39.40	46.99	30.88	86.20	64.18
	SVD	45.46 ✗	**46.28** ✓	36.30 ✗	30.53 ✗	75.00 ✗	64.64 ✓
	+SLERP ($\hat{\alpha}$)	**58.63** ✓	45.99 ✓	**46.76** ✗	**32.50** ✓	**85.87** ✗	**66.59** ✓
SigLIP2 → CLIP ViT-B/32	Old model	40.62	21.73	28.76	14.47	71.29	45.24
	SVD	42.80 ✓	23.56 ✓	32.33 ✓	15.62 ✓	72.16 ✓	46.41 ✓
	+SLERP ($\hat{\alpha}$)	**51.01** ✓	**25.79** ✓	**37.35** ✓	**17.18** ✓	**78.56** ✓	**50.59** ✓
SigLIP1 → CLIP ViT-H/14	Old model	59.38	43.07	43.50	28.56	84.27	63.53
	SVD	60.50 ✓	29.91 ✗	41.82 ✗	23.18 ✗	81.40 ✗	56.13 ✗
	+SLERP ($\hat{\alpha}$)	**64.81** ✓	**43.38** ✓	**46.56** ✓	**28.79** ✓	**85.89** ✓	**63.89** ✓

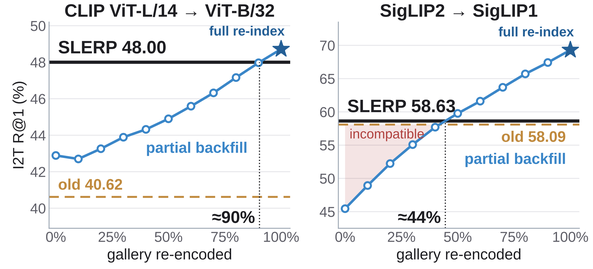
✓ backward compatible: new→old beats old→old. One weight selected on CC3M, fixed across datasets.

No training needed:

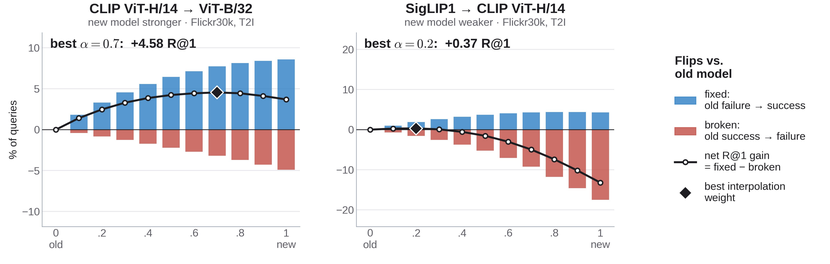
SLERP beats training-based XBT on **5/6** metrics (H/14 ➡ B/32) and **3/6** (L/14 ➡ B/32).

Migration Period

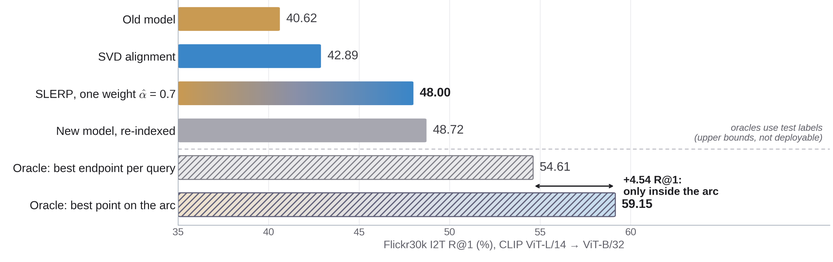
- **Gallery untouched:** backfilling needs **90%** of captions re-encoded to match SLERP.
- **Cheap:** query-side only, **≈0.3 ms** extra with parallel encoders.



Churn Analysis



Per-Query Oracle



Takeaways

- **One SVD, one weight:** align the new queries, then interpolate them with the old ones.
- **Choose the weight once** on a small support set: it transfers across benchmarks.
- **Upgrade now, re-index later:** the old gallery stays usable during the migration.
- **Limits:** both encoders run at query time; gains are bounded by the two models.

GITHUB

WEBSITE