LEARN THE (REGISTER) GAP: MAKING EFFECTIVE SPECIALIST TEXT-TO-IMAGE RETRIEVAL WITH GENERALIST ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision-language encoders often underperform on text-to-image retrieval in specialist collections, even in domains well represented in web-scale pretraining. We show that two small linear maps, one for captions and one for images, trained on the collection's own caption-image pairs with the encoder frozen, recover much of this loss. The maps improve recall in 57 of 58 encoder-collection pairs spanning medicine, remote sensing, biodiversity, art, science, fashion and news, by up to 34 points at R@10, and more than doubling it in eight of them. The frozen encoder's own ranking of the catalog can predict the gain before training, and maps learned on one collection transfer only to collections that share its conventions of description and depiction. The failure is therefore a register gap, not a domain gap: the specialist features are present, but specialist ways of describing and depicting leave captions and images misaligned. The maps recover half or more of LoRA's gain at a fraction of its cost, and add to low-rank LoRA when applied on top. Applied to a caption's phrases and an image's patches, the same maps improve retrieval further and ground phrases in images without box annotations.

1 INTRODUCTION

Contrastive vision-language encoders (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) are the default index for text-to-image retrieval. They work well for web-scale search, but underperform on specialist collections such as herbaria, museum archives, and satellite imagery. The communities that curate these datasets develop their own conventions for photographing, captioning, and searching. These differences negatively affect the retrieval accuracy in specialist collections using general-purpose and even sometimes domain specific encoders.

The drop could mean that the images or text differ so much that a generalist encoder does not extract the necessary features for the search. That is a domain gap, and it calls for retraining or fine-tuning. We argue that for many specialist collections the generalist features are sufficient, and the text and image features are only misaligned by those collection-specific conventions. Linguists call such conventions a register:

> "A communication situation that recurs regularly in a society... will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations." (Ferguson, 1994, p. 20)

Register is a standard notion in linguistics (Halliday et al., 1964; Halliday, 1978; Biber & Finegan, 1994), and computational linguistics studies the closely related sublanguages of specialist domains (Harris, 1968; Kittredge & Lehrberger, 1982). A register can shape how captions are written, how images are made, or both. A professional caption often says what an image means rather than what it shows, and specialist images (an axial CT, an overhead tile, a specimen sheet) look unlike web photographs. Figure 1 shows both failures with one encoder, holding one side fixed each time. The two images of a man playing piano have near identical image embeddings, yet there is a noticeable difference in the way the caption is written. The same can be said of the aerial tile on the right side of Figure 1 which has near identical captions but drastically different visual perspectives because one comes from a dataset of aerial tiles.

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

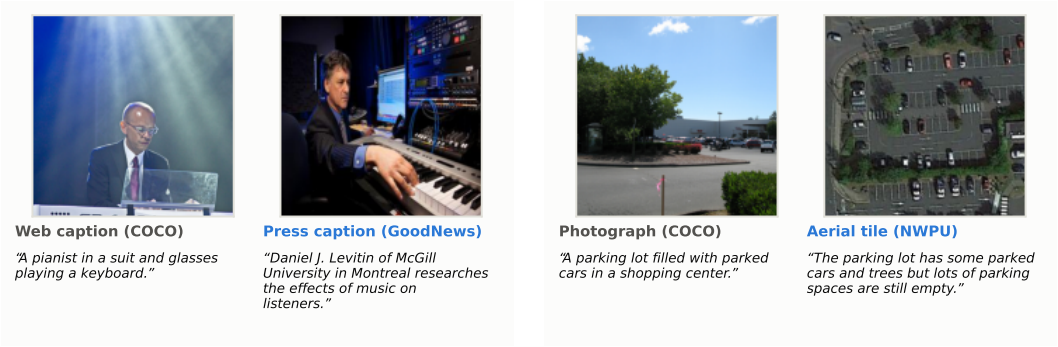


Figure 1: **The register gap.** Inside a specialist dataset, captions and images may take different forms than in general datasets such as COCO. Left: for very similar images, the GoodNews dataset has structured captions that go beyond the visual content. Right: similar captions may belong to images with very different structure, as in the NWPU aerial data.

We call this the *register gap*: a collection-specific joint misarrangement of query and gallery embeddings under a frozen encoder. The correction works by learning a metric specific to each collection, and linear maps learned on one collection help another only when the two share a register. The contribution of this paper is to explore and characterize ways to repair this register gap, specifically:

1. **Simple, efficient correction.** Demonstrating that linear transforms of the text and image features significantly improves specialist image retrieval.
2. **Ability to predict improvement.** We show that simple statistics of the generic encoder predicts the gain across many domains and encoders (§4.1), and that the maps rescale space semantic directions most important for the register.
3. **Extended transformations to patch tokens and caption phrases**. This further improves performance and contributes to better visual grounding in the specialist domain (§4.3, §4.4).
4. **Characterize performance across models.** We document variation in performance across models and show that register correction even helps the most powerful new embedding models.

2 Related Work

Training vision-language models on domain-specific image–text pairs has produced specialized models for medical imaging (Sellergren et al., 2025; Zhang et al., 2023), remote sensing (Liu et al., 2024; Zhang et al., 2024), fashion (Chia et al., 2022; Marqo, 2024), and biodiversity (Stevens et al., 2024; Gu et al., 2025). Linear transformations have been used to align word embeddings across languages (Mikolov et al., 2013; Xing et al., 2015; Conneau et al., 2018), for image–text retrieval through canonical correlation analysis (Rasiwasia et al., 2010), and to connect frozen vision and language models (Merullo et al., 2023). CLIP-Adapter (Gao et al., 2024) learns a small residual network over frozen CLIP features, while Tip-Adapter (Zhang et al., 2022) augments classification scores using a cache of labeled examples. CoOp (Zhou et al., 2022) instead learns text prompts for classification. We compare our linear maps with retrieval adaptations of CLIP-Adapter and Tip-Adapter, and with LoRA (Hu et al., 2022), which learns low-rank weight updates within the encoders rather than transforming their output features.

The modality gap describes the separation between image and text embeddings in contrastively trained models (Liang et al., 2022). Explanations include mismatched pairs and learned temperature (Yaras et al., 2024), dimensional collapse (Fahim et al., 2024), differences in the information conveyed by images and their captions (Schrodi et al., 2025), and the contrastive loss itself (Mager et al., 2026). The gap can be reduced during training (Eslami & de Melo, 2025; Yaras et al., 2024; Mager et al., 2026), but post-training corrections give mixed benefits: removing selected dimensions yields modest gains (Schrodi et al., 2025), while shared-subspace projection can close the gap without improving downstream performance (Yi et al., 2025). Related work identifies misalignment within each modality that limits image-to-image and text-to-text retrieval (Mistretta et al., 2025).

Our register gap concerns the alignment of matching image–caption pairs within a collection, rather than the overall separation between modalities.

Retrieval can also be affected by hubness, where some gallery items appear among the nearest neighbors of many queries (Radovanović et al., 2010). Cross-domain Similarity Local Scaling (CSLS) (Conneau et al., 2018) reduces scores for embeddings with high average similarity to their nearest neighbors in the other space. Querybank Normalisation (QB-Norm) (Bogolin et al., 2022) normalizes scores against a reference set of queries, penalizing gallery items that score highly across many queries. We compare with both to test whether reducing hubness alone accounts for the gains from our learned maps.

Fine-grained image–text alignment is also studied through weakly supervised grounding, where models learn to associate words or phrases with image regions using only image–caption pairs (Huang et al., 2021; Xu et al., 2022; Yao et al., 2022; Mukhoti et al., 2023; Bica et al., 2024). Our phrase–patch extension uses the same image-level supervision to learn linear maps of frozen phrase and patch features. We compare with weakly supervised phrase-localization methods (Akbari et al., 2019; Gupta et al., 2020; Arbelle et al., 2021; Shaharabany et al., 2022; Shaharabany & Wolf, 2023; He et al., 2024) on Flickr30k Entities, using the pointing game to measure whether the highest-scoring image location falls within the annotated bounding box for a phrase.

3 METHODS

We consider the text-based image retrieval setting, where a frozen vision–language encoder maps text captions and images to unit vectors in $\mathbb{R}^d$. For a collection of paired captions and images $\{(q_i, g_i)\}_{i=1}^N$, the unmodified encoder ranks gallery images by $q^\top g$. We split each collection into training, validation, and test sets. The paired training examples provide supervision for the transformations below, validation data are used for model selection, and captions for test images from a held-out gallery are used as queries. We report R@10 and R@50, the larger cutoff for retrieval is increasingly relevant as more expensive re-ranking is becoming very effective (Liu et al., 2025; Li et al., 2026).

3.1 LINEAR FEATURE CORRECTION

We correct the frozen embeddings rather than the encoder: a residual affine transformation (a linear *'map'*) per modality, initialized to the identity, that can reweight and realign the features the encoder already produces but cannot add new ones.

$$\phi_t(q) = \text{normalize}(q + W_t q + b_t), \qquad \phi_i(g) = \text{normalize}(g + W_i g + b_i), \tag{1}$$

where $W_t, W_i \in \mathbb{R}^{d\times d}$. All map parameters are initialized to zero, so the initial model exactly reproduces retrieval with the frozen encoder. For a batch of B caption–image pairs, we train the maps with a text-to-image InfoNCE loss,

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\phi_t(q_i)^\top \phi_i(g_i)/\tau_{\text{NCE}}\big)}{\sum_{j=1}^{B} \exp(\phi_t(q_i)^\top \phi_i(g_j)/\tau_{\text{NCE}})}, \qquad \tau_{\text{NCE}} = 0.05. \tag{2}$$

Thus, the only supervision needed to adapt a collection is the caption–image pairing already present in its catalog.

Map variants. Unless otherwise noted, we learn separate text and image maps. We also consider three constrained variants. A *tied* model shares the same affine map between modalities; this form is useful for the decomposition in Appendix F, since a shared orthogonal rotation cannot change retrieval scores. A *text-only* model leaves the image-side mapping as the identity, allowing a fixed gallery index to be adapted entirely at query time. The corresponding *image-only* model leaves the text side unchanged and isolates adaptation of the gallery features.

3.2 PHRASE- AND PATCH-LEVEL SCORING

A single pooled embedding can obscure useful correspondence when a caption mentions several concepts or when the relevant content occupies only a small part of an image. We therefore also

define a score that retains structure on both sides. Let $p_1, \ldots, p_P$ denote phrases extracted from caption q (noun phrases for the image–text collections and verb phrases for CityFlow-NL), and let $e_1, \ldots, e_J$ denote image patch tokens. Each phrase is embedded with the frozen text encoder. Each patch token is passed through the encoder's own pooling head so that it lies in the same joint embedding space as the global image representation. We then apply the same learned maps in Eq. 1 to the phrase and patch features.

The resulting retrieval score combines the global image–caption similarity with a soft best-patch match for each phrase:

$$s(q, g) = \phi_t(q)^\top \phi_i(g) + \lambda \sum_{k=1}^{P} w_k \left[\tau_s \log \sum_{j=1}^{J} \exp\big(\phi_t(p_k)^\top \phi_i(e_j)/\tau_s\big) - \tau_s \log J \right]. \quad (3)$$

The bracketed term is a smooth maximum over image patches; subtracting $\tau_s \log J$ normalizes for the number of patches. With uniform $w_k = 1/P$, each phrase contributes equally. We also consider a learned phrase gate, $w_k = \mathrm{softmax}_k\big(u^\top \phi_t(p_k)\big)$, where $u \in \mathbb{R}^d$ is initialized to zero so that training begins with uniform weights. The global/local weight λ and set-score temperature τ_s are learned, with $\tau_s \geq 0.01$. For these experiments, Eq. 3 replaces the pooled cosine similarity in the same contrastive training objective used above.

For grounding evaluation, a phrase is assigned to the patch with maximum $\phi_t(p_k)^\top \phi_i(e_j)$. Bounding boxes are used only to evaluate this assignment, not to train the maps (§4.4).

3.3 Near-miss diagnostic

Our correction can only reorganize information that is already present in the frozen embedding space. This suggests a simple way to ask whether a collection is likely to benefit before learning any maps: are the correct images already close to being retrieved, or are they buried much deeper in the gallery? A large number of near misses indicates that the frozen encoder has captured enough of the relevant structure to place true pairs near the useful retrieval cutoff, even if their ordering is not yet good enough.

We measure this directly from the frozen training embeddings. For each caption in the training catalog, we rank all catalog images and record the rank ρ_i of its paired image. We define the near-miss mass at cutoff k as

$$N_k = \Pr_i[k < \rho_i \leq 3k]. \quad (4)$$

Thus, N_k is the fraction of true pairs that fall outside the retrieval cutoff but remain nearby in the ranking. We use $k = 50$ throughout. Because the statistic uses only the frozen training embeddings, it can be computed before fitting any correction; it is intended as a diagnostic of how much potentially recoverable mass lies near the operating cutoff, rather than as an additional learned predictor.

3.4 Experimental setup

Collections and encoders. To separate collection-specific alignment from the quality of the underlying representation, we evaluate both general-purpose encoders and, where available, encoders trained specifically for the corresponding domain. Our main evaluation includes eleven collections spanning seven specialist settings: ROCOv2 (Rückert et al., 2024) and SkinCAP (Zhou et al., 2024) for medicine; NWPU-Captions (Cheng et al., 2022) and RSICD (Lu et al., 2018) for remote sensing; Fashion200k (Han et al., 2017) and FACAD (Yang et al., 2020) for fashion; SemArt (Garcia & Vogiatzis, 2018) for art; SciMMIR (Wu et al., 2024) for scientific figures; GoodNews (Biten et al., 2019) for news; and TreeOfLife-10M captions (Stevens et al., 2024) for biodiversity. We use COCO with the Karpathy splits (Lin et al., 2014; Karpathy & Fei-Fei, 2015) as a general web-caption control (Table 4).

We evaluate four general-purpose encoders on every collection: SigLIP2 (Tschannen et al., 2025), SigLIP-so400m (Zhai et al., 2023), CLIP ViT-L/14 trained on LAION-2B (Radford et al., 2021; Schuhmann et al., 2022), and MetaCLIP-2 (Chuang et al., 2025). Where available, we also evaluate encoders trained for the corresponding specialist domain: MedSigLIP (Sellergren et al., 2025) and BiomedCLIP (Zhang et al., 2023), RemoteCLIP (Liu et al., 2024) and GeoRSCLIP (Zhang et al.,

Table 1: **Linear correction across collections.** For the default SigLIP2 encoder (left), raw and corrected R@50, the gain from two learned maps, and the pre-training near-miss mass N_{50}. The right block gives the best corrected encoder for each collection together with its raw and corrected recall. Rows are sorted by the default encoder's gain; COCO is the general web-caption control.

	SigLIP2 (default)				best encoder after correction		
Collection	raw	corrected	ΔR@50	N_{50}	encoder	raw	corrected
NWPU	0.670	0.952	+0.281	0.214	SigLIP2	0.670	0.952
SkinCAP	0.656	0.846	+0.190	0.181	SigLIP-so400m	0.666	0.874
SemArt	0.556	0.735	+0.180	0.167	MetaCLIP-2	0.895	0.945
GoodNews	0.583	0.728	+0.145	0.103	MetaCLIP-2	0.974	0.979
SciMMIR	0.551	0.696	+0.145	0.129	SigLIP-so400m	0.767	0.878
TreeOfLife	0.550	0.683	+0.133	0.166	MetaCLIP-2	0.668	0.786
RSICD	0.804	0.916	+0.112	0.157	RemoteCLIP	0.915	0.943
ROCOv2	0.494	0.581	+0.087	0.200	BiomedCLIP	0.890	0.884
Fashion200k	0.954	0.996	+0.042	0.036	Marqo-FashionSigLIP	0.977	0.997
FACAD	0.988	0.995	+0.007	0.008	SigLIP2	0.988	0.995
COCO	0.957	0.964	+0.006	0.033	SigLIP-so400m	0.955	0.972

2024), FashionCLIP (Chia et al., 2022) and Marqo-FashionSigLIP (Marqo, 2024), and BioCLIP (Stevens et al., 2024) and BioCLIP-2 (Gu et al., 2025). This yields 58 collection–encoder combinations. Split construction, caption selection, and collection-specific exceptions are described in Appendix A.

The phrase-level and grounding experiments additionally use CityFlow-NL (Feng et al., 2021) and Flickr30k Entities (Plummer et al., 2015). Dataset-specific evaluation protocols are described with those experiments.

Training and evaluation. We optimize the maps with AdamW using learning rate 10^{-4}, weight decay 10^{-4}, batch size 256, and 30 epochs. We select the epoch with the highest validation R@50 and average test performance over two random seeds. We report R@10, R@50, and the change ΔR@k relative to retrieval with the frozen encoder. Before adaptation, we verify each embedding cache by reproducing the corresponding published baseline performance to four decimal places (Appendix B).

Baselines. The baselines test whether the same retrieval failures can be addressed without learning a collection-specific linear metric, or whether the gains simply come from adding adaptation capacity. Training-free corrections include CSLS (Conneau et al., 2018), QB-Norm (Bogolin et al., 2022), query centroid shift, symmetric centering, and the exact symmetric $\lambda = \frac{1}{2}$ modality-gap correction of Liang et al. (2022). Their hyperparameters are selected on the same validation split as our maps. For learned adaptation, we translate CLIP-Adapter (Gao et al., 2024) and Tip-Adapter(-F) (Zhang et al., 2022) to retrieval and train them on the same collection pairs. We also compare with rank-16 LoRA (Hu et al., 2022) fine-tuning of both encoder towers. Full baseline protocols are given in Appendix C.

4 Results

We evaluate the learned maps first on pooled caption and image embeddings across 58 collection–encoder pairs. We then compare them with score corrections, feature adapters, and LoRA fine-tuning, before turning to the phrase–patch extension and the grounding experiments.

4.1 Linear correction across collections and encoders

Table 1 and Figure 2 summarize the 58 collection–encoder pairs; Appendix Table 13 gives the full results. The maps improve R@50 in 57 of 58 cases, with a mean gain of 0.11 and a maximum gain of 0.29. At R@10, the largest gain is 0.34. The one decrease is BiomedCLIP on ROCOv2: the frozen specialist encoder already reaches 0.89 R@50, and correction changes it by -0.005. NWPU

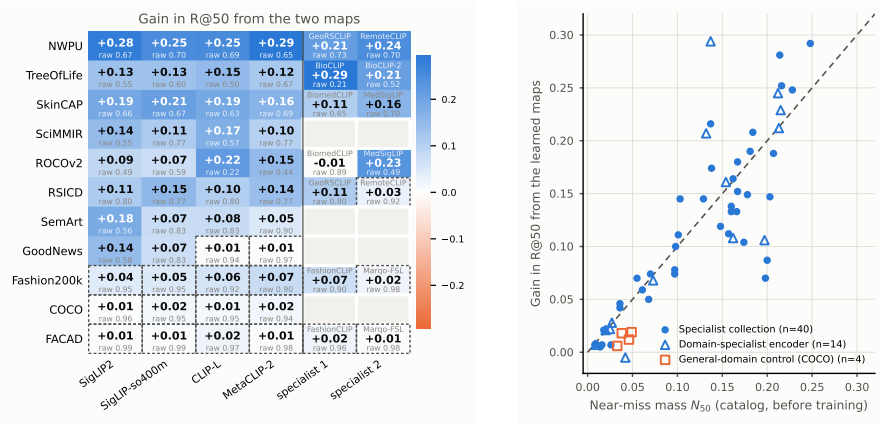


Figure 2: **Results across collections and encoders.** (left) Change in R@50 after learning separate text and image maps for all 58 collection–encoder pairs; raw R@50 is shown beneath each gain. Dashed entries begin above 0.90 R@50. The maps improve 57 of 58 pairs. (right) Near-miss mass N_{50} computed before training versus the realized ΔR@50; the dashed line is $y = x$ and is not fitted to the data.

is the opposite extreme. All six encoders, including RemoteCLIP and GeoRSCLIP, gain between 0.21 and 0.29 at R@50. On COCO, the web-caption control, the gain is at most 0.02 for all four encoders.

Separate text and image maps are slightly better than a tied map. They improve on the tied variant in 51 of 58 pairs, but by only 0.007 R@50 on average (Appendix Table 13). We therefore use two maps for the main experiments and examine that small difference in §4.5. The maps are also cheap to train: median training time across the 58 pairs is 5 seconds, the maximum is 22 seconds on one RTX 3080, and GPU memory remains below 0.8 GB.

Near-miss diagnostic. Figure 2 compares N_{50}, computed before map training, with the eventual change in R@50. The parameter-free prediction $\Delta R@50 = N_{50}$ gives $R^2 = 0.74$ and MAE 0.03 over all 58 pairs. A line fit on ten collections and evaluated on the held-out eleventh gives $R^2 = 0.70$ and MAE 0.033 (Appendix Table 6). The two ends of the plot are informative: COCO has little probability mass just outside the cutoff and little gain, whereas NWPU has a large near-miss band under every encoder and large gains under every encoder.

N_{50} reliably targets the ranking region from which the majority of rescued positives originate (typically within a small multiple of the cutoff). Analysis of the rank transport dynamics and the metric's robustness to different cutoff boundaries ($2k$ through $4k$) is detailed in Appendix J.

Cross-collection transfer. To test whether the maps are learning a global correction, we apply the map learned on each collection to every other collection. Off the diagonal, the mean change in R@50 is -0.03 over 110 transfers, and three quarters of the transfers hurt by more than 0.01. The only transfer above $+0.03$ is RSICD→NWPU; both contain aerial imagery described with similar scene-level captions. A map learned on COCO has essentially no effect elsewhere (Appendix G, Figure 8). The correction is therefore not a global shift shared across collections.

4.2 Comparison with score corrections and adaptation

Table 2 compares our maps with training-free score corrections, learned feature adapters, and LoRA fine-tuning. Baseline details and hyperparameters are given in Appendix C, and Tables 10 and 11 give the training-free and adapter baselines for every encoder.

Hubness and modality-gap corrections. CSLS (Conneau et al., 2018) and QB-Norm (Bogolin et al., 2022) help on most specialist collections, but recover much less than the learned maps: taking the better of the two gives a median 24% of the maps' gain. Hubness is therefore part of the re-

Table 2: **Alternative corrections and adaptation strategies.** Change in R@50 relative to the frozen encoder. Training-free corrections and feature adapters use SigLIP2; the bottom block compares the learned maps with rank-16 LoRA on CLIP ViT-L/14. All methods use the same validation split; LoRA results use one seed.

	NWPU	SkinCAP	SemArt	GoodNews	SciMMIR	TreeOfLife	RSICD	ROCOv2	Fashion200k	FACAD	COCO
raw R@50	0.670	0.656	0.556	0.583	0.551	0.550	0.804	0.494	0.954	0.988	0.957
two linear maps	**+.281**	**+.190**	**+.180**	**+.145**	**+.145**	**+.133**	**+.112**	+.087	**+.042**	**+.007**	+.006
CSLS (k=10)	+.076	+.049	+.047	+.020	+.033	+.015	+.017	+.021	+.007	+.002	+.000
QB-Norm (β=20)	+.042	+.069	+.033	+.024	+.016	+.019	+.023	-.003	+.001	+.000	-.000
symmetric centering	+.037	+.057	+.044	+.026	+.032	+.017	+.018	-.013	-.005	+.000	-.010
Liang shift ($\lambda=\frac{1}{2}$)	-.048	-.139	-.078	-.138	-.208	-.082	-.083	-.091	-.040	-.020	-.046
query centroid shift	-.070	-.157	-.117	-.190	-.271	-.112	-.106	-.115	-.085	-.037	-.078
Tip-Adapter(-F)	+.011	-.069	-.035	-.073	-.088	-.031	-.037	-.036	-.004	-.004	-.017
CLIP-Adapter (text)	+.221	+.147	+.158	+.109	+.073	+.112	+.107	+.047	+.033	+.007	+.003
CLIP-Adapter (image)	+.258	+.178	+.133	+.095	+.076	+.111	+.099	+.089	+.039	+.006	+.007
CLIP-Adapter (both)	+.280	+.180	+.176	+.138	+.128	+.132	+.109	**+.113**	+.042	+.007	**+.007**
CLIP ViT-L/14											
raw R@50	0.688	0.632	—	0.945	—	0.499	—	0.216	—	—	—
two linear maps	+.250	+.192	—	+.006	—	+.148	—	+.178	—	—	—
LoRA (r=16)	+.276	+.213	—	+.012	—	+.159	—	+.318	—	—	—
maps on LoRA	+.281	+.216	—	+.014	—	+.165	—	+.325	—	—	—

trieval problem, but does not account for most of the improvement. Corrections aimed at the global modality gap do worse. Query centroid shift and the symmetric $\lambda = \frac{1}{2}$ construction of Liang et al. (2022) decrease recall on nearly every collection. Symmetric centering helps on seven of eleven collections, but only modestly.

Feature adapters. CLIP-Adapter (Gao et al., 2024), implemented as a residual bottleneck MLP and trained on the same pairs, is close to the linear maps on most collections. The difference is within about 0.02 R@50 on ten of eleven collections; ROCOv2 is the clear exception. A larger nonlinear transform is therefore not necessary to obtain most of the gain.

Applying this approach to frontier embeddings models Gemini Embedding 2 (Shanbhogue et al., 2026) yields similarly strong, predictable improvements without requiring weight access. Furthermore, scaling the correction to larger residual MLPs provides negligible additional benefit over simple linear maps (Appendix Table 9).

LoRA fine-tuning. LoRA (Hu et al., 2022) changes the encoder itself rather than only its output features. It generally gives a larger gain, but at substantially higher training cost. Across the five CLIP ViT-L/14 comparisons, the maps recover a median 90% of LoRA's R@50 improvement while training in seconds rather than 32–80 minutes (Appendix Table 12). ROCOv2 is the largest gap: the maps recover 56% of the LoRA improvement. On GoodNews they recover about half, although both methods change R@50 there by very little. Depending on the fine-tuning set, LoRA also reduces COCO R@50 by as much as 0.043; the maps leave the frozen encoder unchanged.

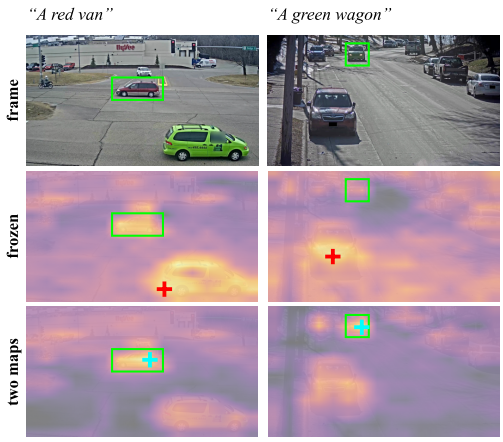
Training maps after LoRA gives a small additional improvement on every collection tested, between 0.002 and 0.007 R@50. With only 1,000 training pairs, the maps match LoRA on SkinCAP and TreeOfLife. The two methods are therefore correcting partly different things: fine-tuning changes the representation, while the maps can still improve the alignment of its output features for a particular collection.

4.3 Phrase- and patch-level retrieval

Table 3 compares global pooled features with the phrase–patch score from §3.2. Using the set score with the frozen encoder adds only 1–9 points of R@10. Training the maps with the phrase–patch score has a much larger effect: without the phrase gate, R@10 improves by 17–30 points over the frozen encoder and by 7–13 points over maps trained only on pooled embeddings. Adding the phrase gate increases the largest improvements to 32 and 14 points, respectively. The difference is largest for SkinCAP and SemArt, whose captions contain about 19 noun phrases on average; RSICD captions contain about four.

Table 3: Phrase- and patch-level scoring improves retrieval beyond pooled embeddings. Text-to-image recall using SigLIP2 so400m/16-384, mean of two seeds. "Set score, no maps" uses the frozen encoder and selects only the global/local weight and set-score temperature on validation data. "Pooled two maps" uses global embeddings only; the final two rows train and score with phrase and patch features.

	SkinCAP (600)			RSICD (1,093)			SemArt (1,069)		
Method	R@1	R@10	R@50	R@1	R@10	R@50	R@1	R@10	R@50
Frozen encoder	0.108	0.375	0.653	0.114	0.437	0.810	0.157	0.359	0.558
Set score, no maps	0.117	0.387	0.663	0.123	0.472	0.828	0.200	0.444	0.626
Pooled two maps	0.207	0.574	0.858	0.145	0.538	0.908	0.202	0.512	0.728
Two maps over sets	0.241	0.677	0.911	0.194	0.604	0.917	0.338	0.639	0.845
Sets + phrase gate	0.275	0.698	0.928	0.196	0.591	0.914	0.347	0.652	0.854



	CityFlow-NL hit (%)		Flickr30k
method	random	new cameras	pointing (%)
chance (box area)	2.6	2.0	27.7
center patch			50.7
frozen encoder	32.8	26.9	67.0
pooled two maps	15.6	14.7	68.6
two maps over sets	61.2	59.5	78.1
sets + phrase gate	**63.3**	**59.7**	81.6
MG			61.7
GbS			74.5
WWbL			75.6
InfoGround			76.7
g++			80.0
SelfEQ			**84.1**

Figure 3: **Grounding without box supervision.** The TIPSv2-B/14 (Cao et al., 2026) maps are trained from caption–image pairs only; boxes are used for evaluation. Left: CityFlow-NL examples (Feng et al., 2021); heat maps show phrase–patch similarity and the cross marks the highest-scoring patch. Right: CityFlow-NL subject-phrase localization on the random split and on cameras unseen during training, and pointing-game accuracy on Flickr30k Entities. CityFlow-NL uses all three captions per test frame. Flickr30k evaluates 14,481 phrases on 1,000 images.

Appendix I evaluates the set-based models under published protocols for SemArt, RSICD, and SkinCAP. Despite differences in training regimes and available query context, our lightweight corrections remain highly competitive with existing domain-specific benchmarks and more customized architectures, even beating them in the case of semart versus ContextNet.

4.4 Grounding without box supervision

For grounding, we assign each phrase to its highest-scoring image patch and count a hit when the patch center falls inside the annotated box. On CityFlow-NL, the frozen encoder localizes the subject vehicle in 32.8% of queries on the random split and 26.9% on cameras not seen during training. Maps trained with the pooled objective actually reduce those numbers to 15.6% and 14.7%, even though they improve global retrieval. Training with the phrase–patch objective raises them to 61.2% and 59.5%; the phrase gate gives 63.3% on the random split and 59.7% on new cameras. Across the 316 CityFlow-NL test frames, the set-trained maps correct 124 cases that were wrong with the frozen encoder and introduce 29 new errors. In the examples in Figure 3, the frozen encoder points to vehicles of the wrong color, while the corrected model selects the described vehicle.

The phrase gate gives the visible vehicle phrase roughly twice the weight of motion phrases. On Flickr30k Entities, the frozen encoder reaches 67.0% pointing-game accuracy and the set-based maps with the phrase gate reach 81.6%. Among the published weakly supervised methods listed in Figure 3 (Akbari et al., 2019; Gupta et al., 2020; Arbelle et al., 2021; Shaharabany et al., 2022;

Shaharabany & Wolf, 2023), only SelfEQ (He et al., 2024) is higher (caveats in Appendix I). The maps themselves are still trained without bounding boxes.

4.5 Systemic analysis

We find two key properties by analyzing the learned transformations. Appendix E.1 shows one query's ranking before and after correction.

Map geometry. We use polar decomposition to separate a tied linear transformation into an orthogonal rotation and a symmetric stretch. A rotation applied to both modalities leaves every score unchanged, so only the stretch and the bias can change retrieval. Across the 14 collection-encoder pairs we factorized, the stretch alone accounts for a median 98% of the tied map's R@50 gain. The improvement therefore comes almost entirely from stretching or shrinking directions already present in the frozen embedding space, rather than rotating them. Two maps add a small relative rotation between the modalities, which accounts for their small edge over one tied map (Appendix E.2).

Encoder spread. Raw zero-shot recall conflates the underlying quality of an encoder's features with how well those features happen to align with a specific collection's register. While specific encoders have substantial variation in raw performance, our maps shrink this max-min R@50 spread by 16 to 83% on 10 of 11 collections. This shows that once corrected for the target register, generalist and specialist encoders perform much more comparably on most collections (Appendix E.3).

5 Discussion and limitations

The reduction in encoder spread following correction §4.5 is super informative for model selection: raw zero-shot recall does not reliably indicate which encoder will perform best after collection-specific alignment. Furthermore the N_{50} provides an easy and surprisingly accurate estimate of the recall recoverable from a frozen representation. However, correction does not render all underlying representations perfectly equivalent; as seen with SkinCAP, certain specialist models (e.g., Biomed-CLIP) may still benefit less than others.

The method also does not require access to the encoder weights. This matters for models such as Gemini Embedding 2 that are available only through an API: the same maps can be learned from the returned embeddings, and the linear maps perform similarly to substantially larger residual MLPs (Appendix Table 9). At deployment time, the two-map model requires the gallery embeddings to be transformed once. If the gallery index cannot be changed, the text-only variant can instead be applied to each incoming query, with the loss in performance quantified in Appendix H.

Limitations. The method assumes access to paired captions and images from the target collection. The training-set-size experiments in Appendix D show that the benefit decreases as fewer pairs are available. The near-miss statistic also depends on gallery size: N_{50} should be computed using a catalog that is representative of the gallery size at which the system will be used. Its MAE of 0.033 is useful for estimating whether a collection is likely to benefit from correction, but not for distinguishing between encoders whose expected gains are very close. Finally, our experiments use self-gallery text-to-image retrieval with one positive image per query and English captions. We have not evaluated multi-positive retrieval or multilingual queries.

6 Conclusion

Poor retrieval on a specialist collection does not necessarily indicate that a generalist vision-language encoder lacks the relevant features. Instead, these features are often present but poorly aligned due to a register gap—a mismatch in the collection-specific conventions of description and depiction. By applying lightweight linear maps to frozen embeddings, we correct this misalignment, often recovering a median 90% heavy fine-tuning (like LoRA) benefits at a fraction of the computational cost and often complimenting fine-tuning by adding to its gains. Extending this approach to phrase and patch features further demonstrates that robust spatial and semantic correspondences exist within these frozen models, requiring only the proper alignment to be utilized effectively.

REPRODUCIBILITY STATEMENT

The code, configuration files, and results are released as a single repository at `https://anonymous.4open.science/r/register-gap-reproduction-306B`. To facilitate reproducibility, our pipeline can be executed at two levels:

From embeddings. The foundational unit of our pipeline is the embedding cache. We provide one file per collection-encoder pair. This contains the caption and image embeddings, split assignments, raw captions, and pair IDs. Every quantitative result in this paper derives exclusively from these caches, including the the linear maps (`ltg.maps.linear`), the MLP comparisons (`ltg.maps.mlp`), all baselines, factorizations, and the near-miss and transfer analyses. A single script regenerates every table and figure directly from the output JSON files. Using the provided caches (13 GB total for all 58 configurations), the full experimental grid trains in roughly half an hour on a single consumer GPU (Appendix B). All subsequent analyses run comfortable on a CPU.

From raw data. The caches can also be rebuilt entirely from scratch. Data is from public releases of the image collections. The dataset split rules are detailed in Appendix A (using seed 0 and one pairs file per collection), the encoder checkpoints specified in the configuration, and our provided embedding script. For datasets with redistribution restrictions (e.g., CityFlow-NL), we release the pair lists and bounding boxes; raw frames must be obtained directly from the source.

Hyperparameters are detailed in §3.4, and random seeds, per-seed metrics, and wall-clock execution times were logged in the released result files.

ETHICS STATEMENT

The collections are public research datasets, used under their licenses. ROCOv2 and SkinCAP contain de-identified medical images, and GoodNews contains news photographs of real people. CityFlow-NL contains traffic-camera video. The frames in Figure 3 come from the CityFlow-NL dataset of the NVIDIA AI City Challenge (Feng et al., 2021) and are shown under the NVIDIA dataset license, which permits a minimal portion in research papers without unredacted faces or license plates; we blurred the license plates and the one visible driver.

AI USE STATEMENT

We used Anthropic's Claude, a generative AI assistant, throughout this work. It wrote, ran and debugged experiment and figure code, helped design the ablation and decomposition experiments in dialogue with the authors, drafted and edited parts of the text, and searched for references and checked their metadata, which the authors then verified against the sources. The authors reviewed the code and the text, and every reported number comes from the authors' pipeline and verified embedding caches. We take full responsibility for the content of this paper, including text, claims and artifacts produced with generative AI.

REFERENCES

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12468–12478, 2019. doi: 10.1109/CVPR.2019.01276.

Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, Chun-Fu Chen, Alex Bronstein, Kate Saenko, Shimon Ullman, Raja Giryes, Rogerio Feris, and Leonid Karlinsky. Detector-free weakly supervised grounding by separation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1781–1792, 2021. doi: 10.1109/ICCV48922.2021.00182.

Douglas Biber and Edward Finegan (eds.). *Sociolinguistic Perspectives on Register*. Oxford University Press, 1994.

Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrovic. Improving fine-grained understanding in image-text pre-training. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3974–3995, 2024.

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2604.12012.

Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12, 2022.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Scott Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, volume 38, pp. 53541–53568, 2025. doi: 10.52202/085713-1601.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in CLIP. In *International Conference on Learning Representations (ICLR)*, 2025.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. CityFlow-NL: Tracking and retrieval of vehicles at city scale by natural language descriptions. *arXiv preprint arXiv:2101.04741*, 2021.

Charles A. Ferguson. Dialect, register, and genre: Working assumptions about conventionalization. In Douglas Biber and Edward Finegan (eds.), *Sociolinguistic Perspectives on Register*, pp. 15–30. Oxford University Press, New York, 1994.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132:581–595, 2024. arXiv:2110.04544.

Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multimodal retrieval. In *European Conference on Computer Vision Workshops (ECCVW)*, 2018.

Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pp. 25–33, 2019. doi: 10.1145/3323873.3325028.

Jianyang Gu, Sam Stevens, Elizabeth Campolongo, Matthew Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113880–113913, 2025. doi: 10.52202/085713-3436.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision (ECCV)*, pp. 752–768, 2020. doi: 10.1007/978-3-030-58580-8_44.

M. A. K. Halliday. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London, 1978.

M. A. K. Halliday, Angus McIntosh, and Peter Strevens. *The Linguistic Sciences and Language Teaching*. Longmans, London, 1964.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zellig S. Harris. *Mathematical Structures of Language*. Interscience (Wiley), New York, 1968.

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C. Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13095–13105, 2024. doi: 10.1109/CVPR52733.2024.01244.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931, 2021. doi: 10.1109/ICCV48922.2021.00391.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Richard Kittredge and John Lehrberger (eds.). *Sublanguage: Studies of Language in Restricted Semantic Domains*. de Gruyter, Berlin, 1982.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 (ECCV)*, 2014.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large multimodal model as your advanced retrieval assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025, 2025. doi: 10.1109/CVPR52734.2025.00380.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2018.

Fabian Mager, Hiba Nassar, and Lars Kai Hansen. On the modality gap and the contrastive loss in multi-modal representation learning. *arXiv preprint arXiv:2607.10698*, 2026.

Marqo. Marqo-FashionSigLIP. `https://huggingface.co/Marqo/marqo-fashionSigLIP`, 2024.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H. S. Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19413–19423, 2023. doi: 10.1109/CVPR52729.2023.01860.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015. doi: 10.1109/ICCV.2015.303.

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 *International Conference on Machine Learning (ICML)*, 2021.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia (MM)*, pp. 251–260, 2010. doi: 10.1145/1873951.1873987.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. ROCOv2: Radiology objects in COntext version 2, an updated multimodal image dataset. *Scientific Data*, 11, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2022.

Andrew Sellergren et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Tal Shaharabany and Lior Wolf. Similarity maps for self-training weakly-supervised phrase grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6925–6934, 2023. doi: 10.1109/CVPR52729.2023.00669.

Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 28222–28237, 2022. doi: 10.52202/068431-2046.

Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, et al. Gemini Embedding 2: A native multimodal embedding model from Gemini. *arXiv preprint arXiv:2605.27295*, 2026.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL*, 2024.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1006–1011, 2015. doi: 10.3115/v1/N15-1104.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18113–18123, 2022. doi: 10.1109/CVPR52688.2022.01760.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12681–12690, 2025. doi: 10.1109/ICCV51701.2025.01178.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *European Conference on Computer Vision (ECCV)*, 2020.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2022.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

Lingjie Yi, Raphael Douady, and Chao Chen. Decipher the modality gap in multimodal contrastive learning: From convergent representations to pairwise alignment. *arXiv preprint arXiv:2510.03268*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022. arXiv:2111.03930.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yan Zhang, Zhong Ji, Changxu Meng, Yanwei Pang, and Jungong Han. iEBAKER: Improved remote sensing image-text retrieval framework via eliminate before align and keyword explicit reasoning. *arXiv preprint arXiv:2504.05644*, 2025.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Juexiao Zhou, Liyuan Sun, Yan Xu, Wenbin Liu, Shawn Afzal, Bin Wang, Yuetan Ye, et al. SkinCAP: A multi-modal dermatology dataset annotated with rich medical captions. *arXiv preprint arXiv:2405.18004*, 2024.

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

APPENDIX OVERVIEW

The following appendices provide complete details to reproduce our experiments and further analyze the learned register corrections. Appendices A through C detail the datasets, split constructions, and baseline hyperparameter protocols. Appendix D extends the training-set size analysis. Appendices E and F provide the qualitative examples, geometric proofs, and encoder spread data supporting the systemic analysis in Section 4.5. Finally, Appendices G through K contain extended ablation studies (cross-collection transfer, single-side mapping) and the full result tables for all 58 collection-encoder pairs.

A DATA PREPARATION AND SPLIT CONSTRUCTION

Every collection is reduced to a list of (caption, image) pairs with a train / val / test split; the test split is its own gallery. Sizes are read from the result reports (Table 4); the per-collection notes below record how each pair list was built and every exception a reader would need to reproduce the numbers.

Table 4: Collections in the grid with split sizes as used. Sizes are identical across encoders except where noted in the text.

Collection	Domain	Caption register	Image register	train	val	test
ROCOv2 (Rückert et al., 2024)	medicine	PubMed figure captions	radiology	59,962	9,903	9,927
SkinCAP (Zhou et al., 2024)	dermatology	clinical descriptions	lesion photos	2,800	598	598
NWPU-Captions (Cheng et al., 2022)	remote sensing	scene sentences	aerial tiles	25,199	3,150	3,150
RSICD (Lu et al., 2018)	remote sensing	scene sentences	aerial tiles	8,734	1,094	1,093
Fashion200k (Han et al., 2017)	fashion	garment descriptions	catalogue photos	50,000	4,999	4,999
FACAD (Yang et al., 2020)	fashion	marketing copy	catalogue photos	8,754	1,874	1,874
SemArt (Garcia & Vogiatzis, 2018)	art	curatorial prose	paintings	19,244	1,069	1,069
SciMMIR (Wu et al., 2024)	science	figure captions	figures/charts	16,184	3,467	3,467
GoodNews (Biten et al., 2019)	news	article captions	news photos	49,998	5,000	4,999
TreeOfLife-10M (Stevens et al., 2024)	biodiversity	taxon captions	specimen photos	8,400	1,793	1,793
COCO (Karpathy) (Lin et al., 2014; Karpathy & Fei-Fei, 2015)	web (control)	visual captions	photographs	20,000	3,000	5,000‡

Common rules. Pairs are (caption, image) with target = identity: the i-th caption's single ground-truth image is the i-th gallery image. Splits are drawn once per collection with seed 0 and cached, so every encoder sees the same pairs. Where a collection has no native split, or its native split is unsuitable for self-gallery retrieval, we draw an image-disjoint stratified split with target sizes 50,000 / 5,000 / 5,000 (train / val / test). When the pool is smaller than the sum of the targets, val and test are each capped at 15% of the pool and train takes the remainder, which is why SkinCAP, FACAD, SciMMIR and TreeOfLife run at their natural sizes (Table 4). Within each stratum the records are shuffled once with the seed and dealt test, then val, then train in proportion, so every stratum is represented in every split. The stratum is the collection's natural category: disease for SkinCAP, product category for Fashion200k and FACAD, figure sub-class for SciMMIR, taxonomic class for TreeOfLife, and a single stratum for GoodNews. NWPU, RSICD, SemArt, ROCOv2 and COCO keep their native splits. Where a training split is larger than a cell's training budget it is subsampled with a seeded permutation of the cached split, so subsampled runs are reproducible from the same cache. Images are materialised once to a local JPEG cache (long side 512 px) and every cell of a collection reads the same files.

Per-collection notes.

- **ROCOv2.** Hugging Face release (`eltorio/ROCOv2-radiology`); the PubMed figure caption is the query and the radiology figure the image. Native train / val / test splits; figures without a caption are dropped.
- **SkinCAP.** Captions are the English polished caption (`caption_zh_polish_en`) from the released CSV, joined to images by file name. Stratified split by disease label from a 4,000-pair pool.
- **NWPU-Captions.** The dataset ships with a per-image split label (560 / 70 / 70 tiles in each of the 45 scene classes), which we use as is. Each tile has five captions; the query is the *longest* by word count (ties to the first), because the longer captions carry the spatial language specific to the overhead view (18.2 words on average against 15.2 for the first caption).

- **RSICD.** Native train / valid / test splits of the Hugging Face release (`arampacha/rsicd`); five captions per tile, and the query is the longest, as for NWPU.
- **Fashion200k.** `Marqo/fashion200k` is one undivided pool of about 202k images in which several photographs of a product share one description. We keep one image per product (72,382 products) so that each caption has a single positive, then draw the stratified split by `category2`.
- **FACAD.** We use the released TEST partition (101,225 images) as our pool, decode the WORDMAP-encoded marketing caption, and keep one image per unique caption (12,504 products), since FACAD stores several near-duplicate images per product under one caption. Stratified split by product category.
- **SemArt.** The curatorial description is the query and the painting the image, using the released split CSVs; rows without a description are dropped.
- **SciMMIR.** `m-a-p/SciMMIR`, figure subset only (`super_class = fig`; rendered tables are excluded because retrieving them is text-in-image reading rather than visual matching), pooled from the released validation and test partitions (23,118 rows; the 59 GB train partition is not used). The query is the `text` field (the figure caption). Stratified split by `sub_class` (architecture / illustration / result).
- **GoodNews.** One pair per article image from `captioning_dataset.json`; the query is the newspaper caption. There is no native split, so we draw the 50,000 / 5,000 / 5,000 split from the 489,237-pair pool with a single stratum; the reported sizes are two pairs smaller because two images failed to decode.
- **TreeOfLife-10M captions.** One image per species: streaming `imageomics/TreeOfLife-10M`, we keep the first image of each species that has a Wikipedia description in `TreeOfLife-10M-Captions`, up to 12,000 species. The query is that Wikipedia description (not a constructed name caption). Stratified split by taxonomic class.
- **COCO (control).** Karpathy splits; one caption per image (the first of the five), the image is its own positive. Train is the first 20,000 pairs of the Karpathy train split, val the first 3,000 of Karpathy val, test the full Karpathy test (footnote ‡). Images are downloaded once from their COCO URLs.

Encoders. Generic: SigLIP2 so400m/16-384 (Tschannen et al., 2025), SigLIP so400m/14-384 (Zhai et al., 2023), CLIP ViT-L/14 LAION-2B (Radford et al., 2021; Schuhmann et al., 2022), MetaCLIP-2 worldwide-huge (Chuang et al., 2025). Domain-specialist: MedSigLIP (Sellergren et al., 2025), BiomedCLIP (Zhang et al., 2023), RemoteCLIP (Liu et al., 2024), GeoRSCLIP (Zhang et al., 2024), FashionCLIP (Chia et al., 2022), Marqo-FashionSigLIP (Marqo, 2024), BioCLIP (Stevens et al., 2024), BioCLIP-2 (Gu et al., 2025). All embeddings are L2-normalized.

B Training, evaluation and verification details

Contrastive Learning Residual maps $A = I + W$ with W, b zero-initialized; in-batch InfoNCE (text→image direction), $\tau = 0.05$; AdamW, lr 10^{-4}, weight decay 10^{-4}; batch 256; 30 epochs. The epoch with the best validation R@50 is kept. Identical for the text-only, image-only, two-map and tied forms; the tied form shares one module for both sides.

Metrics. R@k is the fraction of test captions whose paired image is among the top k by cosine over the full test gallery; ties are broken pessimistically (a tied score counts as ranked above). N_{50} is computed from Eq. 4 on the training catalog. The d' column of Table 13 is the signal-detection discriminability of each true pair against its own distractors, $d' = \langle (s(q_i, g_i) - \mu_i)/\sigma_i \rangle_i$ with μ_i, σ_i over $\{s(q_i, g_j)\}_{j \neq i}$, an earlier predictor retained for completeness (Table 6). Near-duplicate rate is the fraction of embeddings whose maximum cosine to another embedding of the same modality exceeds 0.9, on a subsample of at most 8,000.

Predictive Diagnostics Near-Miss Diagnostic (N_{50}). Because absolute ranks depend on gallery size, a collection should be diagnosed at the same gallery size it will be served at. We compute N_{50} by sampling a subset of the training catalog equal in size to the test gallery. Each sampled caption is scored against every sampled image, and we record the rank of its paired image. N_{50} is simply the fraction of these paired images that land in the rank band (50, 150]. We average this across three random seeds. Test embeddings are never used in this calculation.

Leave-one-collection-out Evaluation. To evaluate how well N_{50} predicts performance on unseen domains, we fit a simple linear model (ΔR@50 $= ax+b$) to the encoder cells of ten collections. We then use this line to predict the gain for the held-out eleventh collection. We pool these out-of-fold predictions to report the R^2 and mean absolute error (MAE).

C BASELINE PROTOCOLS

Here we share very terse details of the parameter we use in computing performance of alternative approaches to improve embeddings.

CSLS (Conneau et al., 2018). $\mathrm{CSLS}(q, g) = 2\cos(q, g) - r_G(q) - r_Q(g)$ with r the mean cosine to the k=10 nearest neighbors in the other modality; $r_G(q)$ is constant per query and rank-invariant, so the gallery-side term $r_Q(g)$, computed against a querybank of training captions, carries the effect.

QB-Norm (Bogolin et al., 2022). Dynamic inverted softmax with β=20: each gallery item's score is normalized by its softmax mass over a querybank of training captions; the activated set is the querybank's top-1 gallery items, and normalization is gated on whether the raw argmax lies in that set. A training-caption querybank is the paper's endorsed choice.

Centroid shift, centering, and the Liang construction. Query centroid shift subtracts the training-caption centroid from queries; centering subtracts each modality's own training centroid, then renormalizes. The Liang $\lambda = \frac{1}{2}$ construction moves each modality half the centroid difference toward the other (Liang et al., 2022). For unit gallery vectors, the shifted similarity expands to $(q - s) \cdot g = q \cdot g - s \cdot g$. Therefore, applying a query centroid shift is mathematically equivalent to taking the raw ranking and applying a fixed, item-specific bias ($-s \cdot g$) to the gallery. Because this bias correlates heavily with gallery centrality (Pearson $r \approx 0.90$ to 0.98), centroid shifting inadvertently strengthens existing hubs rather than mitigating them.

Tip-Adapter (Zhang et al., 2022). Cache keys = training caption embeddings, values = paired image embeddings; affinity $A = \exp(-\beta(1 - q \cdot K^\top))$; score $= q \cdot g + \alpha\,(AV) \cdot g$ with $\alpha \in \{0.25, 0.5, 1, 2\}$, $\beta \in \{1, 3, 5.5, 10\}$ tuned on val R@50. Tip-Adapter-F makes the keys learnable (SGD, lr 10^{-3}, 20 epochs) with α, β from the training-free winner.

CLIP-Adapter (Gao et al., 2024). A $d \to d/4 \to d$ bottleneck with ReLU, output blended with the input by a residual ratio $\in \{0.2, 0.4, 0.6\}$ tuned on val, then normalized; placements text, image, both; our training recipe otherwise.

LoRA (Hu et al., 2022). Rank-16 adapters (α=32, dropout 0.05) on the query, key, value and output projections of every attention block in both encoders of CLIP ViT-L/14 (LAION-2B), 4.3M trainable parameters of 432M with base weights frozen. Text→image in-batch InfoNCE at τ=0.05, batch 64, AdamW 10^{-4}, bf16 autocast with gradient checkpointing, 10 epochs (30 for SkinCAP and the 1,000-pair runs), epoch chosen on validation R@50. The maps are trained on the identical pairs from the fine-tuned or frozen encoder's cached embeddings with the paper recipe. Table 12 lists every run.

D TRAINING-SET SIZE

Figure 4 sweeps the number of training pairs from 100 to the full split on five collections (two maps, two seeds; test gallery fixed). The gain rises roughly log-linearly with the number of pairs and is still rising at the full split on GoodNews and ROCOv2, the two largest catalogs. SkinCAP reaches its full gain, and NWPU three quarters of it, by 2,500 pairs. Because N_{50} is defined on the served gallery, it does not change along the sweep: it states the ceiling the gain approaches.

E ANALYSIS OF THE LEARNED MAPS

This section gives the evidence behind §4.5: one query before and after correction, the geometry of the maps, and the spread between encoders.

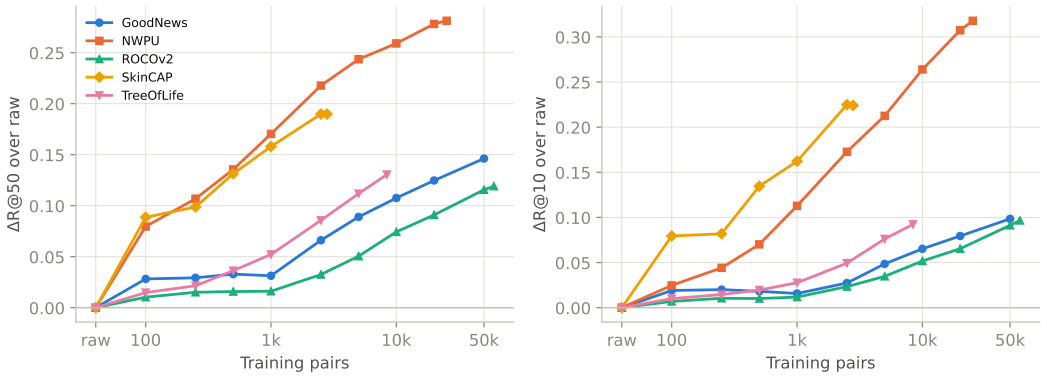


Figure 4: **Training-set size.** Gain at R@50 (left) and R@10 (right) against training pairs for the two-map arm, two seeds; the test gallery is fixed.

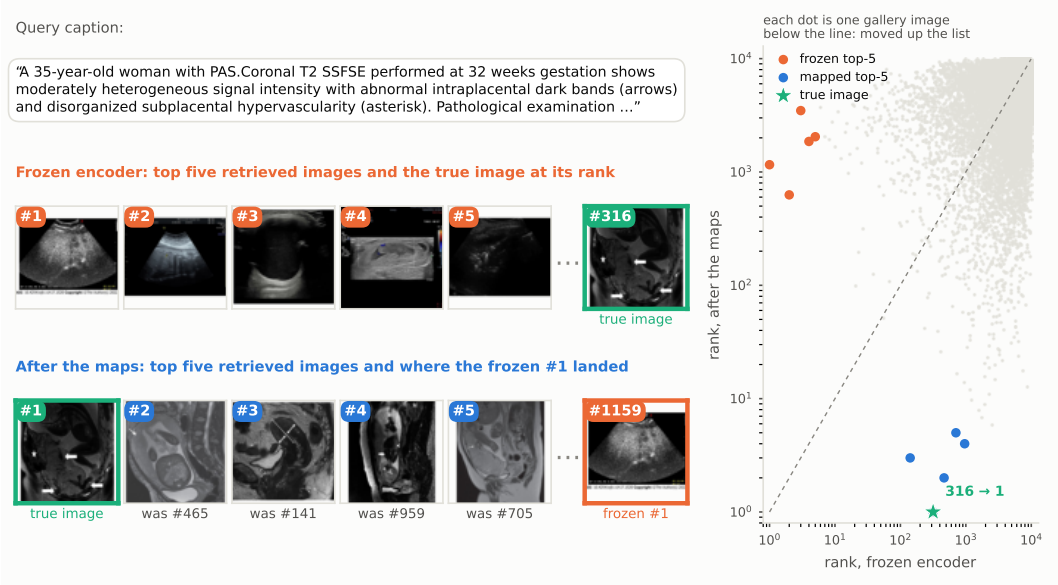


Figure 5: **One ROCOv2 query before and after correction** (SigLIP2, 9,927-image gallery). The frozen encoder retrieves five obstetric ultrasounds and ranks the true fetal MRI at 316. After correction, the true MRI is rank 1; the other four mapped top-five images came from ranks 465, 141, 959, and 705, while the former rank-1 ultrasound falls to 1,159. The scatter plot shows the rank of every gallery image before and after correction.

E.1 One query before and after correction

Figure 5 shows how much the local ranking can change. The query describes a fetal MRI at 32 weeks' gestation. With the frozen encoder, all five top-ranked images are obstetric ultrasounds and the true MRI is at rank 316, although the caption's "T2 SSFSE" and "dark bands" are typical of an MRI report rather than an ultrasound. After correction, the true image moves to rank 1, and the other four images in the top five are pregnancy MRIs that had been at ranks 465, 141, 959, and 705. The former rank-1 ultrasound moves to rank 1,159. The correction does not nudge one positive across a cutoff; it reorders much of the neighborhood.

E.2 Map geometry

For a tied map, polar decomposition separates the learned linear transformation into a rotation and a stretch, $A = RS$, with R orthogonal and S symmetric positive semidefinite (Appendix F). A

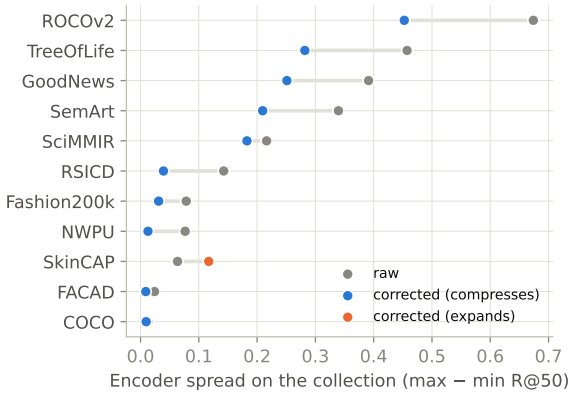


Figure 6: **Encoder spread before and after correction.** Spread is the difference between the highest and lowest R@50 among the encoders evaluated on each collection. It decreases on 10 of 11 collections, by 16 to 83%. NWPU contracts from 0.077 to 0.013; SkinCAP increases from 0.064 to 0.117.

rotation applied to both modalities cannot change any dot product, so retrieval with the tied map can change only through S and the bias. On the 14 collection-encoder pairs with a factorization, the stretch alone recovers a median 98% of the tied map's R@50 gain, and the bias alone contributes essentially nothing. Most of the tied map's effect is therefore a rescaling of directions already present in the frozen embedding: on NWPU, one tied map amplifies 294 directions and suppresses 395 (Figure 7). Two maps add one relative rotation between the modalities (Proposition 2), which accounts for their small edge over the tied map (Appendix F).

E.3 Encoder spread after correction

Figure 6 compares the max-min R@50 across encoders before and after correction. The spread shrinks on ten of eleven collections, by 16 to 83% (Table 8). On NWPU, GeoRSCLIP has the best raw R@50 at 0.727. After correction, SigLIP2 is highest at 0.952, and all six encoders fall within 0.013 of one another. The apparent ordering of encoders on a collection can change substantially after each one is aligned to that collection.

SkinCAP is the exception. Its six encoders begin within a spread of 0.064, but BiomedCLIP gains only 0.106 at R@50, compared with 0.161 to 0.208 for the other encoders, increasing the corrected spread to 0.117. Correction also leaves substantial differences on ROCOv2 and SemArt. Raw performance on these specialist collections therefore reflects both the features available from the encoder and how those features are arranged for the collection. The maps reduce the latter effect, but they do not make the encoders equivalent.

F Decomposition details

Here we explore ways to analyze the learned text and image transformations.

Polar decomposition. Any real square matrix admits a polar decomposition $A = RS$ with R orthogonal and $S = V\Sigma V^\top$ symmetric positive semidefinite. This operation acts as a stretch that rescales space along the orthogonal axes V by the singular values Σ.

Proposition 1 (Rotation is inert in the tied form)**.** *Applying the same orthogonal R to both modalities leaves every retrieval score, hence every ranking, unchanged. A tied map $A = RS$ acts on retrieval only through its stretch S.*

Proposition 2 (Two maps are two stretches and one relative rotation)**.** *With per-side maps $A_t = R_t S_t$ and $A_i = R_i S_i$,*

$$(A_t, A_i) \;\equiv\; (S_t,\; R_{\text{rel}} S_i), \qquad R_{\text{rel}} = R_t^\top R_i, \tag{5}$$

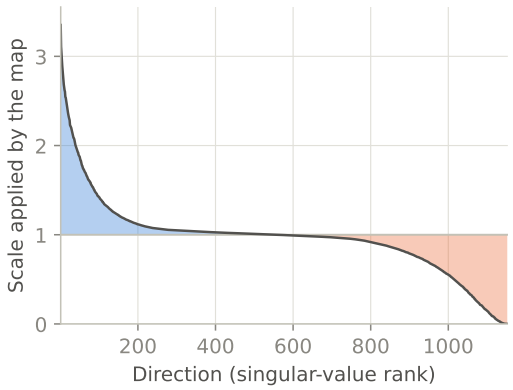


Figure 7: **The correction is rescaling along learned axes.** Singular-value spectrum of one tied map (NWPU, SigLIP2): 294 directions amplified, 395 suppressed, 42 driven below 0.1.

where $\equiv$ denotes equality of every retrieval score. The relative rotation R_{rel} is the entire expressivity that two maps have and a tied map lacks.

Proofs of Props. 1 and 2. For orthogonal R and any q, g: $(Rq) \cdot (Rg) = q^\top R^\top R\, g = q \cdot g$. A common rotation therefore leaves every score unchanged. A tied map $A = RS$ acts on retrieval only through S. For per-side maps $A_t = R_t S_t$ and $A_i = R_i S_i$, we apply the common rotation $R_t^\top$ to both sides. The scores remain unchanged by our first claim. The text map becomes $R_t^\top R_t S_t = S_t$. The image map becomes $R_t^\top R_i S_i = R_{\text{rel}} S_i$. Biases rotate with their maps. Both statements therefore hold for the affine form of Eq. 1.

The Role of the Second Map. Proposition 2 shows that learning two separate maps provides the model with two independent stretches and one relative rotation. Because these components are learned jointly, evaluating them in isolation is misleading. The network relies on them simultaneously to resolve the register gap. Instead, the geometric properties of these components provide a clear path for future analysis. First, we observe that the relative rotation is remarkably small. It averages just 3 to 18 degrees across our experiments. Second, we observe clear asymmetry between the text and image stretches. Future work should investigate the structure of this stretch asymmetry directly. Analyzing how these differing stretches interact with the small relative rotation will clarify exactly how the maps align two distinct semantic spaces.

The sequential split tied $\rightarrow$ two-stretch $\rightarrow$ two maps assigns the interaction between stretch asymmetry and relative rotation to the rotation term; the order-free statement is that removing R_{rel} from the trained pair lands below the tied map in 11 of 14 cells.

G THE CORRECTION IS COLLECTION-SPECIFIC

If the register gap were a property of the encoder rather than of the collection, a map learned anywhere would help everywhere. Figure 8 applies each collection's maps to every other collection: the maps help only the collection they were trained on (diagonal mean $+0.12$) and on average hurt everywhere else (off-diagonal mean -0.03). The one exception is the pair that shares both registers, aerial tiles with scene-sentence captions, where the RSICD maps carry $+0.05$ to NWPU. A map learned on COCO does nothing anywhere.

H WHICH SIDE SHOULD MOVE

Figure 9 analyzes the 26 collection-encoder pairs with single-side arms (Table 14). We plot the performance advantage of mapping the image side instead of the text side against the collection's collapse imbalance. The image side performs better in 19 of the 26 cases. Collections with highly collapsed galleries relative to their captions gain the most from image-side mapping. ROCOv2 under generic encoders clearly demonstrates this trend. Furthermore, the two NWPU configurations

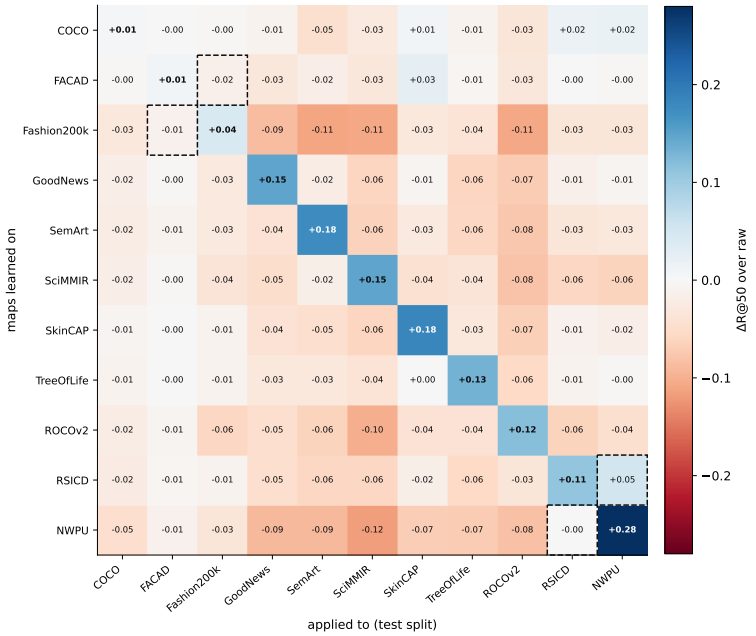


Figure 8: **Maps do not transfer between collections.** Two maps learned on the row collection, applied to the column collection's test split; entries are ΔR@50 over the column's raw encoder (SigLIP2). Dashed cells mark pairs that share a register.

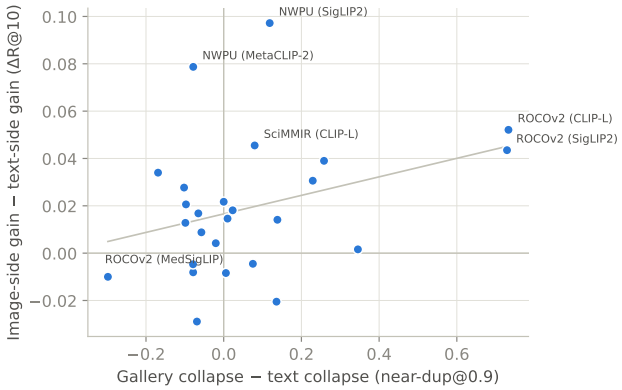


Figure 9: **Which single side wins.** Image-side minus text-side gain (R@10) against collapse imbalance (gallery minus text near-duplicate rate at cosine 0.9), 26 cells. The image side wins in 19; the trend with imbalance is suggestive but not significant ($\rho = 0.29$, $p = 0.16$).

with the largest image-side advantage exhibit near-total gallery collapse. However, the overall rank correlation is weak and lacks statistical significance. We therefore do not recommend using collapse imbalance as a rule to choose which side to map.

I Comparison under published protocols

The results in §4 use a single evaluation protocol. We measure self-gallery text-to-image retrieval on our test split with one positive per query. We then calculate R@k over the full test gallery (§3, Appendix A). Existing literature uses different splits, gallery sizes, caption sets, and cutoffs. A fair comparison requires matching these published protocols exactly.

Table 5: **Evaluation under published benchmark protocols** using the default encoder. To ensure fair comparison, we match established dataset-specific conventions. **SkinCAP:** 600 held-out query captions evaluated against the full 4,000-image gallery, mirroring the Derm1M protocol (Yan et al., 2025) (which uses 3,989 images). **RSICD:** All five captions for each of the 1,093 test images are used as individual queries. The cited RSICD results report R@1, R@5 and R@10 only, so R@50 is left blank. **SemArt:** Evaluated on the Text2Art test split using description-only queries. Published baselines are reported directly from their respective sources: DermLIP from Yan et al. (2025), GeoRSCLIP (Zhang et al., 2024) and iEBAKER-Split from Zhang et al. (2025), and ContextNet (KGM-Author) from Garcia et al. (2019).

protocol	method	R@1	R@5	R@10	R@50
SkinCAP, Derm1M	DermLIP (best published, zero-shot)	–	–	0.213	0.462
	so400m, frozen encoder	0.035	0.105	0.152	0.338
	so400m, pooled two maps	0.078	0.190	0.288	0.538
	so400m, sets + phrase gate	0.088	0.275	0.383	0.657
RSICD, RSITR (T2I)	GeoRSCLIP	0.156	0.412	0.580	–
	iEBAKER-Split (best published)	0.198	0.471	0.634	–
	so400m, frozen encoder	0.101	0.292	0.415	–
	so400m, pooled two maps	0.124	0.363	0.525	–
	so400m, sets + phrase gate	0.176	0.439	0.588	–
SemArt, Text2Art	ContextNet, best row (uses title and author)	0.247	0.477	0.581	–
	so400m, frozen encoder	0.157	0.293	0.359	0.558
	so400m, pooled two maps	0.198	0.409	0.510	0.727
	so400m, sets + phrase gate	0.352	0.563	0.645	0.857

Table 5 re-evaluates our set-based two-map model (§4.3) on three published benchmarks. We test the Text2Art challenge on SemArt (Garcia & Vogiatzis, 2018). We test the SkinCAP evaluation from Derm1M (Yan et al., 2025). We also test the standard RSICD protocol (Zhang et al., 2025). We use the exact maps trained for the main experimental grid (§3.4). We do not perform any additional per-benchmark tuning.

These comparisons evaluate different learning regimes. The Text2Art evaluation uses the standard split, gallery, and metric. Our queries provide only the artwork description. In contrast, the best published model, ContextNet (Garcia et al., 2019), uses the description, title, and author. We found no more recent baselines for this exact protocol.

The SkinCAP evaluation compares our method against DermLIP. DermLIP is evaluated zero-shot after training on one million dermatology image pairs. Our maps instead train directly on the SkinCAP training pairs. We use 600 held-out queries against a 4,000-image gallery. This gallery includes the training images and closely matches the 3,989 gallery size of the original paper.

The RSICD evaluation uses all five captions per image as independent queries. All reported models evaluate on the identical split. Our mapped model matches GeoRSCLIP (Zhang et al., 2024) and trails the current state-of-the-art, iEBAKER.

Grounding. The Flickr30k Entities column of Figure 3 compares pointing-game accuracy on the test split (14,481 phrases on 1,000 images) with published weakly supervised methods, each at its best reported number: MG (Akbari et al., 2019), GbS (Arbelle et al., 2021), WWbL (Shaharabany et al., 2022), InfoGround (Gupta et al., 2020), g++ (Shaharabany & Wolf, 2023) and SelfEQ (He et al., 2024). Three differences matter when reading it: TIPSv2 is a newer encoder than the bases of the published methods, our 16×16 patch grid is coarse enough that the center patch alone scores 50.7%, and InfoGround relies on a detector trained with boxes on Visual Genome.

J Predictor comparison and band sensitivity

Table 6 compares every pre-query predictor under the same leave-one-collection-out protocol, and Table 7 reports the rank transport of the trained maps that fixes the band: the probability of rescue into the top-50 by raw-rank band, and the raw rank of rescued pairs in multiples of the cutoff.

Table 6: Pre-query predictors of the two-map ΔR@50 across all 58 cells: Pearson and Spearman correlation, leave-one-collection-out R^2 and MAE, and the held-out error on the six NWPU cells.

Predictor	r	ρ	LOCO R^2	MAE	NWPU MAE
d' (per-query, train)	-0.75	-0.76	0.48	0.044	0.144
catalog R@10 (train)	-0.87	-0.86	0.68	0.036	0.086
N_{50}, band $\times 1.5$	+0.83	+0.84	0.67	0.032	0.016
N_{50}, band $\times 2$	+0.85	+0.83	0.69	0.031	0.030
N_{50}, band $\times 3$ (used)	+0.85	+0.85	0.68	0.033	0.057
N_{50}, band $\times 4$	+0.85	+0.84	0.67	0.034	0.073
N_{50}, band $\times 6$	+0.85	+0.84	0.64	0.034	0.093
N_{50}, band $\times 8$	+0.84	+0.84	0.62	0.035	0.105
N_{50}, band $\times 3$, val split	+0.88	+0.86	0.72	0.030	0.051

Table 7: Rank transport of the trained two-map arm (seed 0): probability that a test query whose true image had the given raw rank ends inside the corrected top-50, by raw-rank band; and the raw rank of rescued pairs in multiples of the cutoff (median, 90th percentile).

Cell	51–100	101–150	151–200	201–300	301–500	501–1k	>1k	reach p50	p90
COCO (SigLIP2)	0.56	0.27	0.00	0.14	0.00	0.00	0.00	1.3	2.0
GoodNews (SigLIP2)	0.68	0.64	0.54	0.40	0.40	0.34	0.24	4.7	38.8
NWPU (SigLIP2)	0.91	0.88	0.88	0.88	0.87	0.73	0.62	1.9	5.9
NWPU (GeoRSCLIP)	0.89	0.83	0.74	0.84	0.62	0.58	0.30	1.8	4.8
ROCOv2 (SigLIP2)	0.60	0.46	0.40	0.32	0.21	0.14	0.08	2.5	8.9
ROCOv2 (MedSigLIP)	0.75	0.64	0.59	0.50	0.41	0.29	0.14	2.7	9.3
SciMMIR (SigLIP2)	0.68	0.58	0.48	0.40	0.40	0.28	0.20	3.5	19.4
SemArt (SigLIP2)	0.64	0.54	0.64	0.48	0.36	0.27	0.00	3.2	10.3
SkinCAP (SigLIP2)	0.79	0.60	0.59	0.58	0.42	0.22	–	2.6	6.5
TreeOfLife (SigLIP2)	0.61	0.42	0.44	0.34	0.25	0.20	0.13	2.6	9.9
pooled	0.72	0.60	0.54	0.44	0.35	0.25	0.16		

K FULL RESULT TABLES

Table 8: Encoder spread (max−min R@50 across encoders) before and after correction, per collection; positive = compression.

Collection	encoders	raw spread	corrected spread	compression
COCO	4	0.013	0.010	+23%
FACAD	6	0.024	0.009	+63%
Fashion200k	6	0.078	0.031	+60%
GoodNews	4	0.391	0.251	+36%
NWPU	6	0.077	0.013	+83%
ROCOv2	6	0.674	0.452	+33%
RSICD	6	0.143	0.039	+72%
SciMMIR	4	0.216	0.183	+16%
SemArt	4	0.340	0.210	+38%
SkinCAP	6	0.064	0.117	-84%
TreeOfLife	6	0.457	0.282	+38%

Table 9: **Linear two maps against a residual MLP per side on Gemini Embedding 2** (Shanbhogue et al., 2026) (3072-d; nine public collections). Same recipe for all (InfoNCE τ=0.05, AdamW 10^{-4}, batch 256, 30 epochs, val-R@50 selection); mean of three seeds; hidden widths 1024 (fewer parameters than linear) and 4096 (more). R@10 / R@50 on each test gallery.

collection	gallery	frozen	linear two maps	MLP h=1024	MLP h=4096
SkinCAP	600	0.573 / 0.827	0.821 / 0.954	0.826 / 0.964	0.818 / 0.951
SemArt	1,069	0.856 / 0.952	0.915 / 0.975	0.915 / 0.970	0.903 / 0.970
RSICD	1,093	0.507 / 0.851	0.619 / 0.930	0.615 / 0.927	0.620 / 0.924
SciMMIR	3,467	0.909 / 0.955	0.935 / 0.971	0.933 / 0.970	0.933 / 0.970
ROCOv2	9,927	0.491 / 0.718	0.558 / 0.784	0.548 / 0.777	0.548 / 0.780
Fashion200k	4,999	0.826 / 0.958	0.945 / 0.993	0.940 / 0.993	0.938 / 0.992
FACAD	1,874	0.921 / 0.980	0.954 / 0.994	0.949 / 0.995	0.950 / 0.996
COCO (control)	1,000	0.962 / 0.993	0.966 / 0.994	0.963 / 0.993	0.968 / 0.993
Flickr30k (control)	1,000	0.995 / 0.998	0.996 / 0.998	0.996 / 0.999	0.996 / 0.999

Table 10: Training-free baselines, ΔR@50 over raw. CSLS (k=10), QB-Norm (β=20, dynamic inverted softmax, train-caption querybank), symmetric centering, the exact symmetric $\lambda=\frac{1}{2}$ gap-closure construction of Liang et al. (2022), and query centroid shift. Ours = two maps, one per side.

Collection	Encoder	raw R@50	Ours	CSLS	QB-Norm	Center	Liang $\lambda=\frac{1}{2}$	Centroid
COCO	CLIP ViT-L/14	0.951	+0.012	+0.007	+0.004	-0.020	-0.045	-0.075
COCO	SigLIP2	0.957	+0.006	+0.000	-0.000	-0.010	-0.046	-0.078
COCO	MetaCLIP-2	0.945	+0.019	+0.009	+0.007	-0.008	-0.078	-0.122
COCO	SigLIP-so400m	0.955	+0.018	+0.005	-0.001	-0.010	-0.059	-0.097
FACAD	CLIP ViT-L/14	0.973	+0.021	+0.007	+0.010	+0.008	-0.018	-0.032
FACAD	SigLIP2	0.988	+0.007	+0.002	+0.000	+0.000	-0.020	-0.037
FACAD	FashionCLIP	0.964	+0.022	+0.003	+0.004	-0.001	-0.015	-0.032
FACAD	Marqo-FashionSigLIP	0.984	+0.007	-0.001	+0.001	-0.002	-0.007	-0.026
FACAD	MetaCLIP-2	0.983	+0.007	+0.003	-0.002	+0.003	-0.093	-0.164
FACAD	SigLIP-so400m	0.987	+0.008	+0.004	+0.003	+0.004	-0.028	-0.058
Fashion200k	CLIP ViT-L/14	0.918	+0.059	+0.012	-0.002	-0.014	-0.030	-0.053
Fashion200k	SigLIP2	0.954	+0.042	+0.007	+0.001	-0.005	-0.040	-0.085
Fashion200k	FashionCLIP	0.898	+0.068	+0.013	-0.002	-0.016	-0.032	-0.056
Fashion200k	Marqo-FashionSigLIP	0.977	+0.021	+0.004	-0.003	-0.010	-0.017	-0.032
Fashion200k	MetaCLIP-2	0.903	+0.074	+0.020	+0.004	-0.011	-0.091	-0.140
Fashion200k	SigLIP-so400m	0.952	+0.046	+0.010	+0.008	-0.001	-0.044	-0.083
GoodNews	CLIP ViT-L/14	0.945	+0.007	+0.001	-0.002	-0.005	-0.019	-0.044
GoodNews	SigLIP2	0.583	+0.145	+0.020	+0.024	+0.026	-0.138	-0.190
GoodNews	MetaCLIP-2	0.974	+0.005	+0.003	-0.002	-0.004	-0.027	-0.052
GoodNews	SigLIP-so400m	0.832	+0.070	+0.016	+0.018	+0.021	-0.018	-0.054
NWPU	CLIP ViT-L/14	0.688	+0.252	+0.051	+0.037	+0.008	-0.027	-0.043
NWPU	SigLIP2	0.670	+0.281	+0.076	+0.042	+0.037	-0.048	-0.070
NWPU	GeoRSCLIP	0.727	+0.212	+0.020	-0.018	-0.021	-0.176	-0.237
NWPU	MetaCLIP-2	0.651	+0.292	+0.075	+0.023	+0.008	-0.134	-0.183
NWPU	RemoteCLIP	0.698	+0.245	+0.040	+0.044	+0.028	-0.035	-0.054
NWPU	SigLIP-so400m	0.696	+0.248	+0.050	+0.030	+0.017	-0.072	-0.100
ROCOv2	BiomedCLIP	0.890	-0.005	+0.001	-0.004	-0.014	-0.029	-0.042
ROCOv2	CLIP ViT-L/14	0.216	+0.216	+0.021	+0.008	+0.015	-0.054	-0.069
ROCOv2	SigLIP2	0.494	+0.087	+0.021	-0.003	-0.013	-0.091	-0.115
ROCOv2	MedSigLIP	0.487	+0.229	+0.072	+0.084	+0.041	+0.027	+0.022
ROCOv2	MetaCLIP-2	0.439	+0.147	+0.034	+0.018	-0.000	-0.083	-0.105
ROCOv2	SigLIP-so400m	0.590	+0.070	+0.013	-0.009	-0.024	-0.092	-0.115
RSICD	CLIP ViT-L/14	0.800	+0.104	+0.016	+0.028	+0.024	-0.035	-0.061
RSICD	SigLIP2	0.804	+0.112	+0.017	+0.023	+0.018	-0.083	-0.106
RSICD	GeoRSCLIP	0.800	+0.108	-0.010	+0.006	+0.013	-0.191	-0.232
RSICD	MetaCLIP-2	0.772	+0.138	+0.049	+0.035	+0.036	-0.113	-0.174
RSICD	RemoteCLIP	0.915	+0.028	+0.006	+0.014	+0.006	-0.007	-0.021
RSICD	SigLIP-so400m	0.773	+0.152	+0.029	+0.046	+0.037	-0.080	-0.107
SciMMIR	CLIP ViT-L/14	0.571	+0.174	+0.058	+0.053	+0.067	-0.112	-0.172
SciMMIR	SigLIP2	0.551	+0.145	+0.033	+0.016	+0.032	-0.208	-0.271
SciMMIR	MetaCLIP-2	0.767	+0.100	+0.030	+0.013	+0.015	-0.147	-0.201
SciMMIR	SigLIP-so400m	0.767	+0.111	+0.031	+0.011	+0.021	-0.135	-0.204
SemArt	CLIP ViT-L/14	0.832	+0.078	+0.014	+0.002	+0.000	-0.020	-0.036
SemArt	SigLIP2	0.556	+0.180	+0.047	+0.033	+0.044	-0.078	-0.117
SemArt	MetaCLIP-2	0.895	+0.050	+0.003	-0.007	-0.017	-0.045	-0.081
SemArt	SigLIP-so400m	0.829	+0.074	+0.018	+0.014	+0.006	-0.034	-0.070
SkinCAP	BiomedCLIP	0.651	+0.106	+0.032	+0.045	+0.035	-0.005	-0.013
SkinCAP	CLIP ViT-L/14	0.632	+0.188	+0.025	+0.050	+0.038	-0.120	-0.139
SkinCAP	SigLIP2	0.656	+0.190	+0.049	+0.069	+0.057	-0.139	-0.157
SkinCAP	MedSigLIP	0.696	+0.161	+0.065	+0.069	+0.059	+0.017	+0.017
SkinCAP	MetaCLIP-2	0.686	+0.164	+0.060	+0.069	+0.057	-0.132	-0.169
SkinCAP	SigLIP-so400m	0.666	+0.208	+0.042	+0.060	+0.062	-0.084	-0.105
TreeOfLife	BioCLIP	0.210	+0.294	+0.046	+0.095	+0.098	+0.069	+0.045
TreeOfLife	BioCLIP-2	0.521	+0.207	+0.037	+0.051	+0.042	+0.001	-0.007
TreeOfLife	CLIP ViT-L/14	0.499	+0.149	+0.029	+0.031	+0.028	-0.033	-0.050
TreeOfLife	SigLIP2	0.550	+0.133	+0.015	+0.019	+0.017	-0.082	-0.112
TreeOfLife	MetaCLIP-2	0.668	+0.119	+0.013	+0.016	-0.001	-0.085	-0.121
TreeOfLife	SigLIP-so400m	0.602	+0.133	+0.020	+0.026	+0.015	-0.065	-0.109

Table 11: Adapter baselines, ΔR@50 over raw, same recipe and budget. CLIP-Adapter (bottleneck MLP, residual ratio tuned on val) at three placements; Tip-Adapter and Tip-Adapter-F (cache of training pairs; α, β tuned on val).

Collection	Encoder	Ours	CA-text	CA-image	CA-both	Tip	Tip-F
COCO	CLIP ViT-L/14	+0.012	+0.004	+0.010	+0.011	-0.014	-0.014
COCO	SigLIP2	+0.006	+0.003	+0.007	+0.007	-0.017	-0.017
FACAD	SigLIP2	+0.007	+0.007	+0.006	+0.007	-0.004	-0.004
Fashion200k	SigLIP2	+0.042	+0.033	+0.039	+0.042	-0.004	-0.004
GoodNews	SigLIP2	+0.145	+0.109	+0.095	+0.138	-0.073	-0.073
NWPU	CLIP ViT-L/14	+0.252	+0.215	+0.229	+0.252	+0.024	+0.024
NWPU	SigLIP2	+0.281	+0.221	+0.258	+0.280	+0.011	+0.011
ROCOv2	CLIP ViT-L/14	+0.216	+0.119	+0.195	+0.230	-0.031	-0.031
ROCOv2	SigLIP2	+0.087	+0.047	+0.089	+0.113	-0.036	-0.036
RSICD	SigLIP2	+0.112	+0.107	+0.099	+0.109	-0.037	-0.037
SciMMIR	SigLIP2	+0.145	+0.073	+0.076	+0.128	-0.088	-0.088
SemArt	SigLIP2	+0.180	+0.158	+0.133	+0.176	-0.035	-0.036
SkinCAP	CLIP ViT-L/14	+0.188	+0.159	+0.157	+0.183	-0.067	-0.067
SkinCAP	SigLIP2	+0.190	+0.147	+0.178	+0.180	-0.069	-0.069
TreeOfLife	SigLIP2	+0.133	+0.112	+0.111	+0.132	-0.031	-0.031

Table 12: LoRA fine-tuning against the learned maps on CLIP ViT-L/14 (LAION-2B), test R@10/R@50. LoRA: rank 16 on the q/k/v/out projections of both towers (4.3M trainable parameters), text→image InfoNCE at batch 64, bf16, epoch chosen on validation R@50. Maps: the paper recipe on the same training pairs (batch 256). Cost is training wall-clock on one RTX 3080 and peak GPU memory; the last column is the change in COCO-5k R@10/R@50 of the fine-tuned encoder (the maps leave it unchanged). One seed.

Collection	pairs	raw	maps	LoRA	maps on LoRA	maps (s)	LoRA (min)	GB	COCO Δ
NWPU	25,199	0.282/0.688	0.575/0.938	0.668/0.964	0.683/0.969	4.7	80	4.5	-0.038/-0.016
NWPU	1,000	0.282/0.688	0.364/0.827	0.465/0.877	0.476/0.880	1.4	26	4.5	-0.003/-0.004
ROCOv2	20,000	0.101/0.216	0.191/0.394	0.301/0.534	0.306/0.541	12.2	71	4.5	-0.007/-0.011
ROCOv2	1,000	0.101/0.216	0.117/0.269	0.154/0.314	0.160/0.321	10.4	43	4.5	+0.001/-0.003
GoodNews	20,000	0.879/0.945	0.879/0.951	0.894/0.957	0.895/0.959	5.7	66	4.5	-0.052/-0.022
GoodNews	1,000	0.879/0.945	0.882/0.948	0.890/0.954	0.889/0.954	3.3	31	4.5	+0.009/+0.002
TreeOfLife	8,400	0.297/0.499	0.387/0.647	0.397/0.658	0.396/0.664	1.6	32	4.5	-0.087/-0.043
TreeOfLife	1,000	0.297/0.499	0.342/0.581	0.343/0.599	0.355/0.609	0.6	29	4.5	-0.007/-0.004
SkinCAP	2,800	0.321/0.632	0.525/0.824	0.592/0.845	0.612/0.848	0.7	35	4.5	+0.004/-0.003
SkinCAP	1,000	0.321/0.632	0.480/0.796	0.525/0.784	0.518/0.799	0.3	19	4.5	+0.007/-0.002

Table 13: Full grid, 58 dataset×encoder cells. Raw and corrected (two maps, one per side) R@10 / R@50 with the across-seed standard deviation of corrected R@50; gains at R@50 for the two-map and tied-map arms; our predictor near-miss mass N_{50} from the training catalog. Two seeds, val-R@50 model selection.

Collection	Encoder	n_{train}	raw R@10	raw R@50	R@10	R@50	sd	ΔR@50	tied ΔR@50	N_{50}
COCO	CLIP ViT-L/14	20000	0.781	0.951	0.804	0.963	0.000	+0.012	+0.009	0.046
COCO	SigLIP2	20000	0.817	0.957	0.832	0.964	0.001	+0.006	+0.006	0.033
COCO	MetaCLIP-2	20000	0.782	0.945	0.815	0.964	0.000	+0.019	+0.014	0.049
COCO	SigLIP-so400m	20000	0.820	0.955	0.846	0.972	0.001	+0.018	+0.017	0.038
FACAD	CLIP ViT-L/14	8754	0.890	0.973	0.942	0.993	0.001	+0.021	+0.018	0.018
FACAD	SigLIP2	8754	0.940	0.988	0.969	0.995	0.000	+0.007	+0.007	0.008
FACAD	FashionCLIP	8754	0.870	0.964	0.905	0.986	0.001	+0.022	+0.020	0.025
FACAD	Marqo-FashionSigLIP	8754	0.929	0.984	0.947	0.991	0.001	+0.007	+0.005	0.009
FACAD	MetaCLIP-2	8754	0.909	0.983	0.937	0.990	0.001	+0.007	+0.006	0.016
FACAD	SigLIP-so400m	8754	0.942	0.987	0.966	0.995	0.000	+0.008	+0.008	0.009
Fashion200k	CLIP ViT-L/14	50000	0.740	0.918	0.883	0.978	0.000	+0.059	+0.058	0.061
Fashion200k	SigLIP2	50000	0.805	0.954	0.950	0.996	0.000	+0.042	+0.041	0.036
Fashion200k	FashionCLIP	50000	0.705	0.898	0.843	0.966	0.001	+0.068	+0.061	0.073
Fashion200k	Marqo-FashionSigLIP	50000	0.877	0.977	0.954	0.997	0.000	+0.021	+0.019	0.020
Fashion200k	MetaCLIP-2	50000	0.704	0.903	0.877	0.976	0.000	+0.074	+0.069	0.070
Fashion200k	SigLIP-so400m	50000	0.811	0.952	0.950	0.997	0.000	+0.046	+0.044	0.036
GoodNews	CLIP ViT-L/14	49998	0.879	0.945	0.884	0.952	0.000	+0.007	+0.006	0.026
GoodNews	SigLIP2	49998	0.440	0.583	0.539	0.728	0.002	+0.145	+0.131	0.103
GoodNews	MetaCLIP-2	49998	0.932	0.974	0.940	0.979	0.000	+0.005	+0.006	0.014
GoodNews	SigLIP-so400m	49998	0.729	0.832	0.804	0.902	0.000	+0.070	+0.064	0.055
NWPU	CLIP ViT-L/14	25199	0.282	0.688	0.581	0.940	0.000	+0.252	+0.238	0.216
NWPU	SigLIP2	25199	0.278	0.670	0.593	0.952	0.002	+0.281	+0.276	0.214
NWPU	GeoRSCLIP	25199	0.285	0.727	0.568	0.939	0.001	+0.212	+0.202	0.213
NWPU	MetaCLIP-2	25199	0.235	0.651	0.566	0.943	0.001	+0.292	+0.283	0.248
NWPU	RemoteCLIP	25199	0.281	0.698	0.587	0.944	0.002	+0.245	+0.235	0.212
NWPU	SigLIP-so400m	25199	0.267	0.696	0.604	0.944	0.000	+0.248	+0.244	0.228
ROCOv2	BiomedCLIP	59962	0.687	0.890	0.670	0.884	0.000	-0.005	-0.008	0.042
ROCOv2	CLIP ViT-L/14	59962	0.101	0.216	0.222	0.432	0.001	+0.216	+0.186	0.137
ROCOv2	SigLIP2	20000	0.257	0.494	0.325	0.581	0.001	+0.087	+0.083	0.200
ROCOv2	MedSigLIP	59962	0.225	0.487	0.454	0.716	0.000	+0.229	+0.223	0.215
ROCOv2	MetaCLIP-2	59962	0.231	0.439	0.335	0.586	0.001	+0.147	+0.137	0.203
ROCOv2	SigLIP-so400m	59962	0.337	0.590	0.399	0.660	0.002	+0.070	+0.060	0.198
RSICD	CLIP ViT-L/14	8734	0.363	0.800	0.506	0.904	0.000	+0.104	+0.101	0.174
RSICD	SigLIP2	8734	0.437	0.804	0.563	0.916	0.001	+0.112	+0.114	0.157
RSICD	GeoRSCLIP	8734	0.409	0.800	0.510	0.908	0.001	+0.108	+0.100	0.162
RSICD	MetaCLIP-2	8734	0.387	0.772	0.504	0.910	0.002	+0.138	+0.141	0.160
RSICD	RemoteCLIP	8734	0.540	0.915	0.584	0.943	0.000	+0.028	+0.024	0.027
RSICD	SigLIP-so400m	8734	0.416	0.773	0.552	0.925	0.002	+0.152	+0.151	0.167
SciMMIR	CLIP ViT-L/14	16184	0.404	0.571	0.559	0.745	0.000	+0.174	+0.153	0.138
SciMMIR	SigLIP2	16184	0.382	0.551	0.510	0.696	0.002	+0.145	+0.125	0.129
SciMMIR	MetaCLIP-2	16184	0.613	0.767	0.729	0.866	0.001	+0.100	+0.090	0.098
SciMMIR	SigLIP-so400m	16184	0.608	0.767	0.743	0.878	0.001	+0.111	+0.096	0.101
SemArt	CLIP ViT-L/14	19244	0.629	0.832	0.733	0.909	0.001	+0.078	+0.070	0.097
SemArt	SigLIP2	19244	0.358	0.556	0.517	0.735	0.004	+0.180	+0.170	0.167
SemArt	MetaCLIP-2	19244	0.734	0.895	0.792	0.945	0.003	+0.050	+0.041	0.068
SemArt	SigLIP-so400m	19244	0.640	0.829	0.749	0.903	0.000	+0.074	+0.066	0.097
SkinCAP	BiomedCLIP	2800	0.376	0.651	0.455	0.757	0.007	+0.106	+0.098	0.197
SkinCAP	CLIP ViT-L/14	2800	0.321	0.632	0.523	0.820	0.003	+0.188	+0.180	0.207
SkinCAP	SigLIP2	2800	0.391	0.656	0.612	0.846	0.004	+0.190	+0.185	0.181
SkinCAP	MedSigLIP	2800	0.428	0.696	0.635	0.857	0.004	+0.161	+0.163	0.154
SkinCAP	MetaCLIP-2	2800	0.433	0.686	0.622	0.850	0.008	+0.164	+0.176	0.162
SkinCAP	SigLIP-so400m	2800	0.348	0.666	0.590	0.874	0.001	+0.208	+0.193	0.184
TreeOfLife	BioCLIP	8400	0.100	0.210	0.271	0.504	0.004	+0.294	+0.252	0.137
TreeOfLife	BioCLIP-2	8400	0.381	0.521	0.504	0.729	0.000	+0.207	+0.197	0.132
TreeOfLife	CLIP ViT-L/14	8400	0.297	0.499	0.390	0.648	0.001	+0.149	+0.140	0.178
TreeOfLife	SigLIP2	8400	0.327	0.550	0.417	0.683	0.002	+0.133	+0.130	0.166
TreeOfLife	MetaCLIP-2	8400	0.457	0.668	0.536	0.786	0.000	+0.119	+0.109	0.148
TreeOfLife	SigLIP-so400m	8400	0.397	0.602	0.483	0.735	0.000	+0.133	+0.135	0.160

Table 14: All four arms for every cell where the single-side arms were trained (the 14 original cells plus each collection's best-gain encoder). Gains over raw at R@50 and R@10.

Collection	Encoder	raw R@50	text_{50}	image_{50}	both_{50}	tied_{50}	text_{10}	image_{10}	both_{10}	tied_{10}
COCO	CLIP ViT-L/14	0.951	+0.007	+0.013	+0.012	+0.009	+0.005	+0.026	+0.023	+0.014
COCO	SigLIP2	0.957	+0.003	+0.009	+0.006	+0.006	+0.004	+0.022	+0.015	+0.011
FACAD	SigLIP2	0.988	+0.007	+0.006	+0.007	+0.007	+0.025	+0.020	+0.029	+0.024
FACAD	FashionCLIP	0.964	+0.020	+0.020	+0.022	+0.020	+0.021	+0.038	+0.035	+0.027
Fashion200k	SigLIP2	0.954	+0.033	+0.039	+0.042	+0.041	+0.113	+0.126	+0.145	+0.132
Fashion200k	MetaCLIP-2	0.903	+0.051	+0.068	+0.074	+0.069	+0.114	+0.148	+0.173	+0.154
GoodNews	SigLIP2	0.583	+0.115	+0.099	+0.145	+0.131	+0.067	+0.059	+0.099	+0.092
NWPU	CLIP ViT-L/14	0.688	+0.220	+0.227	+0.252	+0.238	+0.197	+0.225	+0.299	+0.260
NWPU	SigLIP2	0.670	+0.228	+0.261	+0.281	+0.276	+0.171	+0.268	+0.315	+0.308
NWPU	GeoRSCLIP	0.727	+0.182	+0.168	+0.212	+0.202	+0.189	+0.160	+0.283	+0.256
NWPU	MetaCLIP-2	0.651	+0.235	+0.271	+0.292	+0.283	+0.186	+0.264	+0.332	+0.302
NWPU	RemoteCLIP	0.698	+0.205	+0.218	+0.245	+0.235	+0.204	+0.213	+0.306	+0.269
ROCOv2	CLIP ViT-L/14	0.216	+0.125	+0.193	+0.216	+0.186	+0.049	+0.102	+0.121	+0.098
ROCOv2	SigLIP2	0.494	+0.038	+0.077	+0.087	+0.083	+0.017	+0.061	+0.067	+0.062
ROCOv2	MedSigLIP	0.487	+0.203	+0.193	+0.229	+0.223	+0.187	+0.177	+0.229	+0.220
RSICD	SigLIP2	0.804	+0.116	+0.100	+0.112	+0.114	+0.086	+0.117	+0.125	+0.116
RSICD	SigLIP-so400m	0.773	+0.150	+0.135	+0.152	+0.151	+0.106	+0.120	+0.135	+0.133
SciMMIR	CLIP ViT-L/14	0.571	+0.129	+0.155	+0.174	+0.153	+0.098	+0.143	+0.155	+0.139
SciMMIR	SigLIP2	0.551	+0.103	+0.099	+0.145	+0.125	+0.065	+0.079	+0.129	+0.099
SemArt	CLIP ViT-L/14	0.832	+0.068	+0.067	+0.078	+0.070	+0.082	+0.074	+0.104	+0.089
SemArt	SigLIP2	0.556	+0.167	+0.144	+0.180	+0.170	+0.128	+0.108	+0.159	+0.134
SkinCAP	CLIP ViT-L/14	0.632	+0.181	+0.177	+0.188	+0.180	+0.181	+0.202	+0.202	+0.197
SkinCAP	SigLIP2	0.656	+0.171	+0.190	+0.190	+0.185	+0.174	+0.213	+0.221	+0.214
SkinCAP	SigLIP-so400m	0.666	+0.190	+0.185	+0.208	+0.193	+0.221	+0.222	+0.242	+0.246
TreeOfLife	BioCLIP	0.210	+0.267	+0.265	+0.294	+0.252	+0.154	+0.149	+0.171	+0.144
TreeOfLife	SigLIP2	0.550	+0.121	+0.115	+0.133	+0.130	+0.072	+0.077	+0.089	+0.076