You CAN Teach an Old Model New Tricks: Domain Adaptation via Complementary Subspace Expansion

Donghoon Han SungHyun Moon SeungJae Lee
Dnotitia Inc.
{dhk1349, sunghyun, seungjae.lee}@dnotitia.com

Abstract

Adapting a deployed retrieval encoder can improve domain quality while degrading general retrieval and making existing corpus embeddings incompatible. We introduce **Complementary Subspace Expansion (CoSE)**, which preserves the frozen encoder's output and appends a complementary channel trained with both blended and standalone retrieval losses. An inference-time blending weight controls the domain–general trade-off, with a frozen endpoint that preserves every legacy score. Across four domains and three seeds, CoSE trains 6.4M parameters and keeps all reported general benchmark means across three training seeds within 4.8 percentage points of Frozen at equal channel weighting, while validation-selected weights improve domain retrieval over Frozen. Complementary-channel truncation and partial branching provide storage and encoding-cost trade-offs. Serving benchmarks and comparisons with two-table alternatives characterize the deployment trade-off: CoSE retains an auditable legacy-score endpoint within one expanded representation, at the cost of additional compute and corpus backfill.

1 Introduction

Vision-language embedding models underpin large-scale multimodal retrieval, where queries and documents are mapped to a shared vector space and compared against an index built from massive corpora. In deployment, this index is often the largest asset in the serving stack: re-embedding a billion-document corpus can take weeks of GPU time. Adapting a deployed model to a new domain therefore incurs a second cost beyond fine-tuning: index migration.

Current adaptation methods force a trade-off. PEFT methods such as LoRA Hu et al. [2022a] and DoRA Liu et al. [2024b] modify the base encoder's weights, invalidating old indices and often degrading general retrieval. BCT Shen et al. [2020], Hu et al. [2022b], Pan et al. [2023] relaxes compatibility into an approximate distillation objective. L2C Chi et al. [2025], the closest architectural neighbor, combines a frozen encoder with a parallel network but still changes the final feature, breaking exact index compatibility. Keeping the base frozen preserves general retrieval, but provides no domain adaptation.

We ask whether the old model can learn new tricks *without being touched*—whether backward compatibility can be guaranteed by construction. We propose **Complementary Subspace Expansion (CoSE)**, a side-branch LoRA architecture in which the base encoder stays frozen while a jointly trained side branch produces a complementary embedding $\mathbf{z}_{\text{new}}$. The final retrieval embedding is the concatenation $\mathbf{z}_{\text{final}} = [\mathbf{z}_{\text{old}} \,\|\, \mathbf{z}_{\text{new}}]$, where $\mathbf{z}_{\text{old}}$ is structurally preserved from the frozen encoder's output on the same input. At the frozen endpoint, CoSE preserves every query–document score against the legacy corpus. We call this exact backward compatibility (BC); using the complementary channel requires additional corpus embeddings. The complementary channel $\mathbf{z}_{\text{new}}$ adds domain signal through contrastive training, while a blending scalar α shifts weight between $\mathbf{z}_{\text{old}}$ and $\mathbf{z}_{\text{new}}$ at inference without retraining.

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

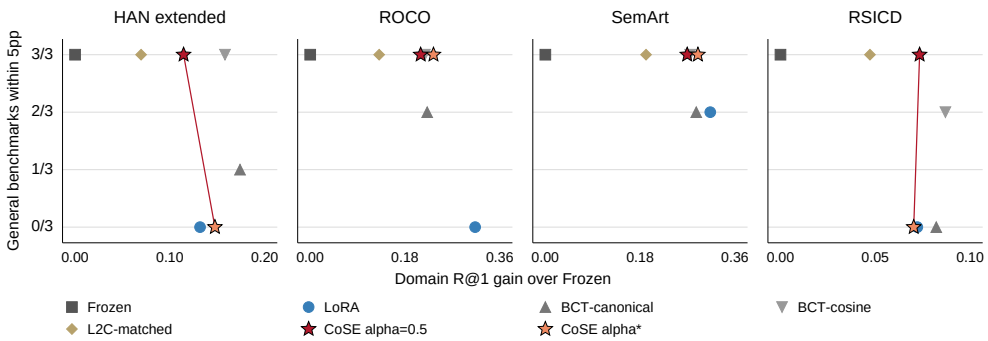


Figure 1: **Domain quality and general preservation after full corpus migration.** x: domain R@1 gain over Frozen, avg(i2t,t2i); y: number of general benchmarks within 5pp of Frozen. Points use Table 1 results and each method's populated corpus representation; SemArt/RSICD LoRA general coverage uses seed 0, while its domain gain averages three seeds. The red segment connects CoSE's fixed and validation-selected blends from the same checkpoint family; it is not a measured intermediate trajectory.

Figure 1 summarizes fully migrated domain quality and general-preservation coverage under a 5pp budget. CoSE's fixed and validation-selected blends expose different operating points; the experiments separately evaluate legacy-score preservation, deployment alternatives, and serving cost.

Contributions.

- **Domain adaptation with an unchanged legacy channel.** CoSE preserves the frozen encoder's output as a coordinate block and learns a complementary retrieval channel. The frozen endpoint preserves legacy scores, while blended retrieval adds domain signal after corpus backfill.
- **Joint training with controllable deployment costs.** Contrastive supervision of both the blended and complementary representations supports different inference-time operating points. Optional Matryoshka supervision on HAN and partial branching provide storage and encoding-cost trade-offs.
- **Evaluation of quality, preservation, and deployment cost.** Four-domain experiments quantify domain retrieval and general retention, while serving benchmarks and two-table comparisons establish the benefits and limits of preserving the legacy channel within an expanded representation.

Deployment scope. CoSE does not eliminate corpus backfill: domain gains require computing the complementary channel for indexed documents. We recommend serving the frozen endpoint until backfill completes, then switching to the selected blend. After migration, the frozen channel remains available and the blending weight controls the domain–general trade-off without retraining.

2 Related Work

Backward-compatible representation learning. BCT Shen et al. [2020] and variants (adversarial Pan et al. [2023], aligner-based Hu et al. [2022b], forward-compatible Ramanujan et al. [2022], efficient-update Jaeckle et al. [2023]) treat compatibility as a distillation target. CoSE instead preserves the base's outputs *structurally* as the first d coordinates of $F_\theta = [\mathbf{z}_{\text{old}}; \mathbf{z}_{\text{new}}]$ by expanding the score channel rather than perturbing within it (Sec. 3.1).

PEFT with weight-space constraints. LoRA Hu et al. [2022a] and DoRA Liu et al. [2024b] provide parameter-efficient weight adaptation, while OPLoRA Xiong and Xie [2026] and MiLoRA Wang et al. [2025] constrain the update subspace. These methods do not enforce exact equality with the frozen encoder's outputs. CoSE's complementarity is analyzed in *output* space between $\mathbf{z}_{\text{old}}$ and $\mathbf{z}_{\text{new}}$; it does not require an explicit orthogonality constraint.

Frozen-backbone complementary-network architectures — closest comparable. Side-Tuning Zhang et al. [2020] and LST Sung et al. [2022] established the frozen-backbone-plus-side pattern. L2C Chi et al. [2025] is closest: a frozen CLIP Radford et al. [2021] encoder plus a parallel network combined via "reverted attention," $\tilde{\mathbf{I}}(x) = \mathbf{I}(x) + \mathrm{CP}^{\mathrm{RT}}(x)$. CoSE differs on three axes: (i) L2C's additive form modifies the final feature, breaking BC; CoSE's concatenation keeps $\mathbf{z}_{\mathrm{old}}$ verbatim; (ii) L2C is classification-focused, CoSE is retrieval-formulated; (iii) CoSE uses in-backbone LoRA, enabling partial branching. We compare against parameter-matched and full-capacity retrieval-adapted L2C in Sec. 4.2.

Multimodal embedding adaptation. CLIP-family models Radford et al. [2021], Tschannen et al. [2025], Chuang et al. [2025], domain-specialized descendants Zhang et al. [2023], Liu et al. [2024a], and VLM-derived embedders Jiang et al. [2024], Lin et al. [2024], Zhang et al. [2024] adapt via full or projector fine-tuning; none target BC. CoSE's side-branch recipe is applicable on top of any such base.

CLIP adapter family. CLIP-Adapter Gao et al. [2024] and Tip-Adapter Zhang et al. [2021] add small trainable modules on a frozen CLIP for few-shot classification. They share CoSE's frozen-backbone premise but are designed for classification-time feature/logit adaptation rather than deployed retrieval embedding expansion.

Continual learning, Matryoshka, and InfoNCE. Forgetting mitigation via parameter regularization (EWC Kirkpatrick et al. [2017]) or output distillation (LwF Li and Hoiem [2018]) reduces but does not eliminate drift; CoSE achieves exact base preservation architecturally. We adopt Matryoshka Representation Learning Kusupati et al. [2022] for $\mathbf{z}_{\mathrm{new}}$ truncation; partial branching is algorithmically similar to top-layer adapters Houlsby et al. [2019] but motivated by serving cost. Contrastive training uses InfoNCE Oord et al. [2018] with in-batch negatives; we use no momentum queue He et al. [2020]; small-batch quality remains a limitation (App. A.9).

3 Methodology

3.1 Problem Formulation: Exact Compatibility as a Score Constraint

Let $f_0 : \mathcal{X} \to \mathbb{R}^d$ denote a frozen vision-language embedding encoder and

$$\mathbf{z}_{\mathrm{old}}(x) = \frac{f_0(x)}{\|f_0(x)\|_2}$$

the L2-normalized legacy embedding. A deployed corpus $\mathcal{D} = \{d_i\}_{i=1}^N$ has the fixed index $\mathcal{I}_0 = \{\mathbf{z}_{\mathrm{old}}(d_i)\}_{i=1}^N$. Given paired domain data $\mathcal{T}_{\mathrm{dom}}$, the target is domain-adapted retrieval with (i) an exactly auditable frozen endpoint, (ii) improved domain quality, and (iii) general-retrieval preservation. CoSE satisfies the first requirement by expanding the score channel instead of perturbing the deployed one.

Definition 1 (Exact backward compatibility of legacy scores)**.** Let $Z_{\mathrm{old}} \in \mathbb{R}^{N \times d}$ have rows $\mathbf{z}_{\mathrm{old}}(d_i)^\top$. A query encoder $g : \mathcal{X} \to \mathbb{R}^d$ preserves every legacy score if

$$Z_{\mathrm{old}} g(q) = Z_{\mathrm{old}} \mathbf{z}_{\mathrm{old}}(q) \quad \text{for every } q. \tag{1}$$

If the legacy document matrix has full column rank, preserving every legacy score within the original embedding channel forces the adapted query embedding to equal the frozen embedding: $Z_{\mathrm{old}}(g(q) - \mathbf{z}_{\mathrm{old}}(q)) = 0$ implies $g(q) = \mathbf{z}_{\mathrm{old}}(q)$. CoSE therefore retains the legacy channel and learns an additional one. This observation concerns exact score equality, which is stronger than preserving rankings.

3.2 Complementary Expansion and Architecture

CoSE represents each input by

$$F_\theta(x) = \begin{bmatrix} \mathbf{z}_{\mathrm{old}}(x) \\ \mathbf{z}_{\mathrm{new}}(x;\theta) \end{bmatrix} \in \mathbb{R}^{d+m},$$

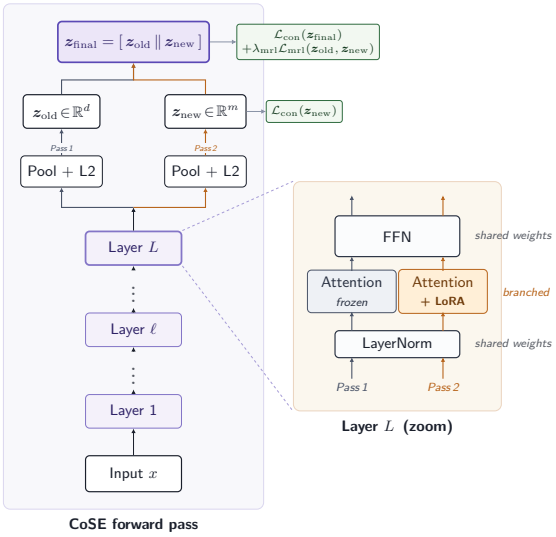


Figure 2: CoSE architecture. Pass 1 runs the frozen encoder and produces the preserved legacy embedding $\mathbf{z}_{\text{old}}$. Pass 2 is a trainable side-branch LoRA pathway that produces the complementary embedding $\mathbf{z}_{\text{new}}$. The final representation is $F_\theta = [\mathbf{z}_{\text{old}} ; \mathbf{z}_{\text{new}}]$. Shared weights act separately on each pathway; MRL supervises truncated blends and is used only on HAN in the main experiments.

where $\mathbf{z}_{\text{new}}(x;\theta) \in \mathbb{R}^m$ is trainable and domain-specialized.

If a fixed full-rank linear readout recovers the frozen embedding from an expanded representation, an invertible coordinate transformation expresses that representation as a frozen block plus a complementary block. CoSE implements this form directly, making the legacy readout a coordinate slice. Formal statements and proofs are in Appendix H.

The reference implementation uses one forward over a duplicated batch $[x; x]$. The backbone weights are frozen, and hooks on attention projections add LoRA deltas only to the second half. Each half computes its own hidden states from the input; the complementary half does not consume the first half's intermediate states. The first half produces $\mathbf{z}_{\text{old}}$ and the second produces $\mathbf{z}_{\text{new}}$, each with last-token pooling and L2 normalization. The pooled $\mathbf{z}_{\text{old}}$ is detached in training. Thus the first half is independent of the LoRA parameters.

The default width is $m = d$. For Qwen3-VL-Embedding-2B ($L = 28$, $d = 2048$), rank-16 LoRA on $Q/K/V/O$ adds 6.4M trainable parameters (0.30% of the base). Complementary coordinates can be truncated and renormalized at deployment. An optional Matryoshka loss supports this use on HAN; it is not used by the main-table ROCO, SemArt, or RSICD checkpoints.

Lemma 1 (Index preservation by construction)**.** *For every input x and parameter value θ, $F_\theta(x)_{1:d} = \mathbf{z}_{\text{old}}(x)$. Therefore, for any query q and legacy document d_i,*

$$\langle F_\theta(q)_{1:d}, \mathbf{z}_{\text{old}}(d_i)\rangle = \langle \mathbf{z}_{\text{old}}(q), \mathbf{z}_{\text{old}}(d_i)\rangle,$$

which is exactly the frozen-model score. The operating point using only the first d coordinates recovers the frozen ranking over the legacy index.

Proof. The first pathway uses frozen weights and contains no trainable operation from the side branch. The stop-gradient makes this separation explicit in the computation graph. Hence the first d coordinates are independent of θ and equal $\mathbf{z}_{\text{old}}(x)$ up to kernel-level floating-point determinism. □

At the frozen endpoint, the legacy index remains valid without rewriting. For adapted retrieval with nonzero contribution from $\mathbf{z}_{\text{new}}$, corpus documents additionally store complementary coordinates, while the legacy $\mathbf{z}_{\text{old}}$ entries remain unchanged. The guarantee concerns underlying similarity scores. ANN indices built over different representations may return different candidates; retaining the unchanged legacy ANN index preserves the original search path for rollback.

3.3 Retrieval Scores and Operating Points

For a query-document pair (q, d), define

$$r_0(q,d) = \langle \mathbf{z}_{\text{old}}(q), \mathbf{z}_{\text{old}}(d) \rangle, \qquad r_1(q,d) = \langle \mathbf{z}_{\text{new}}(q), \mathbf{z}_{\text{new}}(d) \rangle.$$

CoSE uses

$$s_\alpha(q,d) = \alpha r_0(q,d) + (1-\alpha) r_1(q,d), \qquad \alpha \in [0,1]. \tag{2}$$

The weighted concatenation $\phi_\alpha(x) = [\sqrt{\alpha}\,\mathbf{z}_{\text{old}}(x); \sqrt{1-\alpha}\,\mathbf{z}_{\text{new}}(x)]$ induces s_α and a positive-semidefinite kernel for every $\alpha \in [0,1]$. At $\alpha = 1$, it reduces to the frozen score; at $\alpha = 0.5$, it equals cosine similarity of the unweighted, branch-normalized concatenation.

We report $\alpha = 0.5$ as a tuning-free reference and select α^* on domain validation i2t R@1 over a fixed grid. If general retention is required, the same sweep can instead maximize domain quality subject to a specified general-degradation budget; this need not select the domain-optimal weight. Validation selection and general preservation are separate criteria: a domain-validation optimum need not meet a general-quality budget. We therefore evaluate the fixed weight directly, rather than infer its preservation from the stability of a selected weight.

Changing the blend with fixed stored vectors. The symmetric representation ϕ_α defines the score, but need not be the storage convention for every weight. For example, keeping $D(d) = \phi_{1/2}(d)$ fixed and using

$$Q_\alpha(q) = [\sqrt{2}\alpha \mathbf{z}_{\text{old}}(q); \sqrt{2}(1-\alpha)\mathbf{z}_{\text{new}}(q)]$$

gives $\langle Q_\alpha(q), D(d) \rangle = s_\alpha(q,d)$ without rewriting documents. This is an inner-product construction, with no further query normalization. It establishes score equivalence, not invariant approximate-search behavior: changing the query weights can change HNSW traversal, recall, and cost. Our serving-cost measurements use the fixed $\alpha = 0.5$ operating point; we do not measure a query-only reweighting cost or recall sweep. A separate frozen-endpoint ANN diagnostic is described in Appendix G. Exact legacy scores at $\alpha = 1$ therefore remain distinct from reproducing the legacy ANN candidate set.

3.4 Training Objective

For a minibatch $\mathcal{B} = \{(u_i, v_i)\}_{i=1}^B$ of ordered query–positive pairs, the reported implementation uses the directional contrastive loss

$$\mathcal{L}_{\text{NCE}}(s) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s(u_i,v_i)/\tau)}{2\exp(s(u_i,v_i)/\tau) + \sum_{j \in \mathcal{N}_i} \exp(s(u_i,v_j)/\tau)},$$

where $\mathcal{N}_i$ excludes the diagonal and other pairs with the same image ID. The factor two reflects the positive logit and retained in-batch diagonal; the former is the target (Appendix G). Training is directional: image-to-text on ROCO, SemArt, and RSICD and the sampled HAN pair types (Appendix C). Evaluation reports both directions. All CoSE contrastive terms use this implemented loss. It is not standard InfoNCE up to a constant: writing the positive exponential as P and the negative sum as N, the difference is $\log(1 + P/(P+N))$, which depends on the embeddings. Baseline objectives are distinguished in Appendix B.

We define the contrastive training recipe as

$$\mathcal{L}_{\text{CoSE}} = \mathcal{L}_{\text{NCE}}(s_{1/2}) + \lambda_{\text{new}}\mathcal{L}_{\text{NCE}}(r_1) + \lambda_{\text{mrl}}\mathcal{L}_{\text{MRL}}(\mathbf{z}_{\text{old}}, \mathbf{z}_{\text{new}}). \tag{3}$$

The main-table checkpoints minimize $\mathcal{L}_{\text{CoSE}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}$ with $\lambda_{\text{orth}} = 0.1$; the paired loss controls use $\lambda_{\text{orth}} = 0$. Appendix A.1 defines the penalty. The first term trains the default unweighted concatenation; the second term anchors the standalone complementary endpoint; the $\alpha = 1$ endpoint is inherited from the frozen encoder.

For Matryoshka training, let $\mathcal{M} = \{256, 512, 1024, 2048\}$ be the reference truncation set, with $m' \le m$, and define

$$\mathbf{z}_{\text{new}}^{(m')}(x) = \text{norm}(\mathbf{z}_{\text{new}}(x)_{1:m'}), \qquad r_1^{(m')}(q,d) = \langle \mathbf{z}_{\text{new}}^{(m')}(q), \mathbf{z}_{\text{new}}^{(m')}(d) \rangle.$$

Then

$$\mathcal{L}_{\text{MRL}}(\mathbf{z}_{\text{old}}, \mathbf{z}_{\text{new}}) = \frac{1}{|\mathcal{M}|}\sum_{m' \in \mathcal{M}} \mathcal{L}_{\text{NCE}}(\tfrac{1}{2}r_0 + \tfrac{1}{2}r_1^{(m')}).$$

The MRL term averages losses on the blended, branch-normalized concatenation $[\mathbf{z}_{\text{old}}; \mathbf{z}_{\text{new}}^{(m')}]$, including full width. It does not supervise $\mathbf{z}_{\text{new}}$ alone. We use $\lambda_{\text{new}} = 1$, $\tau = 0.02$, and $\lambda_{\text{mrl}} = 0.5$ for HAN and 0 for ROCO/SemArt/RSICD. MRL is an optional trade-off for truncated blended deployment; its measured benefit is limited to HAN (Section 5.2). Training configurations are given in Appendix G.

3.5 Partial Branching and Serving Cost

Full CoSE runs both the frozen pathway and the complementary pathway, giving approximately $2\times$ base forward cost and $d+m$ stored coordinates per document. Let $K \in \{0, \ldots, L\}$ denote the number of shared prefix layers. The frozen pathway still runs all L layers and produces $\mathbf{z}_{\text{old}}$ exactly, while the complementary pathway starts from the detached hidden state after layer K and runs the remaining $L - K$ layers with LoRA adapters. The analytical base-equivalent cost $C(K) = 1 + (L - K)/L$ assumes the prefix is computed once. The duplicated-batch reference recomputes it and does not realize this saving. Output-equivalent prefix sharing realizes this saving in the tested implementation (Appendix A.11). Since the frozen pathway is unchanged for every K, Lemma 1 holds regardless of branching depth. Storage is $d + m$ at full width and $d + m'$ under Matryoshka truncation with $m' \leq m$.

4 Experiments

4.1 Setup

Backbone. Qwen3-VL-Embedding-2B Li et al. [2026] (L=28, d=2048, last-token pool). Additional experiments include an 8B diagnostic, a 2B batch-size control, and transfer to SigLIP2-so400m (§4.5). **Datasets.** Four domains spanning broadcast/cultural, medical, art, and satellite imagery: HAN extended[1], ROCO Pelka et al. [2018], SemArt Garcia and Vogiatzis [2018], RSICD Lu et al. [2018]. General-retrieval eval: Flickr30K Plummer et al. [2015], COCO 2017 val Lin et al. [2014] (1000 pairs each), ImageNet-1k Deng et al. [2009] zero-shot (5000 images), all seed 42. Per-domain split sizes and protocol in Appendix C. **Training.** The CoSE checkpoints use batch 64 with MRL weight 0.5 for HAN, and batch 32 without MRL for ROCO/SemArt/RSICD. They use lr 2×10^{-4}, 200-step warmup and cosine decay over 8000 steps, $\tau = 0.02$, and rank-16 LoRA on all $Q/K/V/O$ projections (6.42M trainable). The main-table checkpoints include $\lambda_{\text{orth}} = 0.1$; paired ablations use zero. LoRA and BCT use the same rank and reference schedule; actual batch settings are distinguished in Appendix B. L2C has parameter-matched and full-capacity variants.

Evaluation. Domain retrieval uses the same 1000 sampled image records across methods (sampling seed 0), with one selected caption and one matching positive per record. Caption rules are in Appendix C. Fully migrated retrieval uses each method's populated corpus representation; cross retrieval against an unchanged frozen corpus is evaluated separately. R@1 in the main text; R@5/R@10 and per-direction i2t/t2i breakdowns are reported in Appendix D. CoSE α swept on val over 12 values; $\alpha^* = \arg\max_\alpha$ val-i2t-R@1 (ties choose the smallest α); test reports both α=0.5 and α^*. **Baselines.** *Frozen* (unadapted base), *LoRA* (standard PEFT, matched rank), *BCT-canonical* Shen et al. [2020] (symmetric cross-encoded InfoNCE), *BCT-cosine* (feature-alignment distillation), and *L2C* Chi et al. [2025] (reverted attention, retrieval-adapted). We additionally evaluate a full-capacity retrieval-adapted L2C, a Frozen+LoRA score blend, and a two-table system retaining separate Frozen and LoRA embeddings. Full definitions in Appendix B.

4.2 Main Domain Performance

Table 1 reports test-split R@1 on four domains.

At $\alpha = 0.5$, CoSE improves domain retrieval on all four domains while all general mean metrics remain within 4.8pp of Frozen. Domain-validation selection yields another operating point, trailing the numerical domain leader by $2.7/8.0/2.3/1.7$pp on HAN/ROCO/SemArt/RSICD. Domain-only selection exceeds the 5pp general budget on HAN and RSICD for every seed and on ROCO for one of three seeds, although ROCO's mean satisfies it. L2C-matched also satisfies the 5pp budget on

[1]Extended HAN Moon et al. [2025]; base on AIHub (`https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71866`); extended captions release pending review.

Table 1: **Fully migrated retrieval and general preservation.** Every method uses its corresponding populated corpus representation. Domain and COCO/Flickr values are avg(i2t,t2i) R@1; ImageNet is zero-shot Top-1. Domain values are mean±sample std across three training seeds (0/42/123). General evaluation uses shared samples across methods (1000 retrieval pairs, 5000 ImageNet images). General values are three-seed means except SemArt/RSICD LoRA, reported for seed 0 only; their domain values remain three-seed means. The *endpoint* mark for CoSE means exact legacy scores are available at $\alpha = 1$, not that scores at the displayed α equal Frozen; $\approx$ denotes approximate compatibility training for single-encoder baselines. Grayscale marks degradation from Frozen: at most 3pp, 3–10pp, and over 10pp. CoSE α^* is selected per seed on validation i2t R@1.

Trained on	Method	Legacy-score endpoint	Domain R@1 avg(i2t,t2i)	INet Top-1	COCO avg	Flickr avg
Frozen	(no adaptation)	✓	0.320 / 0.142 / 0.204 / 0.088	0.644	0.687	0.880
HAN extended	LoRA	×	0.450 ± 0.007	0.524	0.507	0.776
	BCT-canonical	≈	0.492 ± 0.001	0.542	0.571	0.858
	BCT-cosine	≈	0.476 ± 0.003	0.623	0.647	0.888
	L2C-matched	×	0.388 ± 0.005	0.625	0.682	0.865
	CoSE α=0.5	✓	0.433 ± 0.006	0.635	0.683	0.881
	CoSE $\alpha^* \approx 0.12$	✓	0.465 ± 0.005	0.531	0.507	0.757
ROCO	LoRA	×	0.457 ± 0.007	0.491	0.568	0.767
	BCT-canonical	≈	0.365 ± 0.003	0.593	0.645	0.844
	BCT-cosine	≈	0.361 ± 0.001	0.622	0.660	0.857
	L2C-matched	×	0.273 ± 0.007	0.637	0.683	0.883
	CoSE α=0.5	✓	0.353 ± 0.002	0.638	0.684	0.893
	CoSE $\alpha^* \approx 0.25$	✓	0.377 ± 0.007	0.600	0.655	0.865
SemArt	LoRA	×	0.513 ± 0.003	0.622	0.615	0.850
	BCT-canonical	≈	0.487 ± 0.001	0.602	0.624	0.844
	BCT-cosine	≈	0.476 ± 0.007	0.630	0.669	0.869
	L2C-matched	×	0.393 ± 0.003	0.633	0.669	0.866
	CoSE α=0.5	✓	0.470 ± 0.004	0.650	0.685	0.889
	CoSE $\alpha^* \approx 0.30$	✓	0.490 ± 0.007	0.629	0.656	0.870
RSICD	LoRA	×	0.157 ± 0.011	0.287	0.453	0.645
	BCT-canonical	≈	0.167 ± 0.011	0.449	0.567	0.763
	BCT-cosine	≈	0.172 ± 0.006	0.511	0.637	0.831
	L2C-matched	×	0.133 ± 0.005	0.631	0.679	0.876
	CoSE α=0.5	✓	0.159 ± 0.004	0.596	0.675	0.865
	CoSE $\alpha^* \approx 0.23$	✓	0.156 ± 0.006	0.497	0.586	0.789

all four domains at the level of benchmark means, and BCT-cosine on three. CoSE at $\alpha = 0.5$ has higher domain R@1 than L2C-matched on all four domains; general retention alone is not unique to CoSE. The fixed and tuned weights therefore serve different operating objectives. Seed variation is conditional on the training protocol (Appendix G); gaps are interpreted descriptively.

Full-capacity L2C. With approximately 155M trainable parameters, L2C-full obtains domain i2t R@1 of $0.358/0.453/0.142$ on ROCO/SemArt/RSICD, compared with $0.355/0.474/0.165$ for CoSE at $\alpha = 0.5$ (CoSE: three-seed means; L2C-full: seed 0 only). Its COCO avg(i2t,t2i) R@1 is 1.2–2.1pp below Frozen; Flickr and ImageNet also remain within 5pp. The domain i2t comparisons are separate from Table 1's bidirectional averages.

4.3 Deployment Alternatives and Migration

A two-table alternative retains Frozen corpus embeddings and builds a separate LoRA table. With known query domains, hard routing selects one encoder and searches its table over the full corpus. It exceeds CoSE at fixed $\alpha = 0.5$ on all three tested mixed-traffic domains (Table 2). CoSE therefore targets deployments valuing an expanded representation and a domain–general blend without a per-query routing decision, rather than universally higher retrieval quality.

During partial corpus backfill, blending covered and uncovered documents creates coverage bias. At 25% coverage, HAN R@1 falls from 0.341 to 0.270 and ROCO from 0.151 to 0.134; all four domains show excess top-1 selection of covered documents. We therefore serve the frozen endpoint until backfill completes, then switch atomically. Blue-green migration can provide interim continuity for other methods too; CoSE retains the frozen score channel inside the final model after migration.

Table 2: Mixed-traffic text-to-image R@1 (mean±sample std, three seeds) on the union of domain and Flickr documents, with domain-test and Flickr text queries. Hard routing assumes known query domains. SemArt/RSICD LoRA seeds 42/123 are retrained replicas, distinct from the main-table runs (Appendix G). This evaluation differs from the ROCO-only serving benchmark.

System	ROCO	SemArt	RSICD
Two-table, known-domain routing	0.672 ± 0.004	0.693 ± 0.004	0.515 ± 0.003
CoSE, fixed $\alpha = 0.5$	0.623 ± 0.003	0.680 ± 0.003	0.510 ± 0.003

Table 3: ROCO HNSW search work and projected storage, excluding any additional legacy index retained for ANN rollback (Appendix F). Payload is a proxy, not measured memory traffic or encoder throughput. The 2304-dimensional CoSE row truncates only the complementary channel; full-width domain/general performance must not be attributed to this configuration.

Index	Dim	ndis/query	Payload	Size (100k)
Frozen	2048	797	6.5 MB	846 MB
LoRA	2048	1049	8.6 MB	846 MB
Two-table, hard routing	2048/table	1049	8.6 MB	1692 MB
CoSE full	4096	532	8.7 MB	1666 MB
CoSE $[2048 \| 256]$	2304	837	7.7 MB	949 MB

4.4 Serving Cost

We benchmark HNSW on 65k distinct ROCO documents at ANN recall@$10 \geq 0.95$. Table 3 reports distance computations per query (ndis), an evaluated-vector payload proxy (ndis $\times$ dimension $\times 4$ bytes), and index size projected to 100k documents. The two-table row stores both tables but reports search work on its selected LoRA path; it is not a mixed-traffic average or a matched quality–cost Pareto comparison with Table 2.

The benchmark uses inner-product search. Each branch is L2-normalized independently; the compact row truncates the complementary branch to 256 coordinates and renormalizes it before concatenation. Documents and queries are both stored or queried as $[\mathbf{z}_{\text{old}}; \mathbf{z}_{\text{new}}]/\sqrt{2}$, so the measured blend is $\alpha = 0.5$. A separate HNSW graph is built for each representation. Replacing these stored weights by another symmetric ϕ_α requires rebuilding that representation's index; the query-only construction in Section 3.3 avoids this rewrite but is not part of this cost benchmark. The reported recall and search work apply only to these measured configurations.

On an isolated H200, batch-one text-encoder p50 is $16.6/16.7/27.9$ms for Frozen/merged LoRA/CoSE. CoSE costs $1.67\times$ Frozen at batch one and 2.1–$2.2\times$ at batches 32–64 ($\text{QPS}_{\text{Frozen}}/\text{QPS}_{\text{CoSE}}$, using external queries). Encoder, preprocessing, and ANN costs are measured separately; full procedures and image-query results are in Appendix G.

At $\alpha = 0.5$, the $[2048 \| 256]$ configuration obtains ROCO domain i2t R@1 of 0.338 (full width 0.355, Frozen 0.132; three seeds). Its domain/Flickr/COCO averages are $0.342/0.883/0.675$ and ImageNet Top-1 is 0.618, all general means within 5pp of Frozen. These test-quality measurements and the 65k-train-corpus ANN benchmark share a configuration, not an evaluation task. In a separate endpoint diagnostic, a full-width graph built over unweighted $[\mathbf{z}_{\text{old}}; \mathbf{z}_{\text{new}}]$ documents is queried with $[\mathbf{z}_{\text{old}}; 0]$. Its top-10 set overlap with a legacy graph is 98.4% (Appendix G), despite equal underlying frozen scores. This is not a reuse of Table 3's graph; retaining the legacy index preserves the original rollback search path.

Storage with exact ANN rollback. Table 3 counts the active index. Retaining the unchanged Frozen HNSW graph additionally costs 846 MB at the projected 100k-document scale. The compact and full CoSE configurations then require 1795 MB and 2512 MB in total, respectively. This optional retained graph preserves the original approximate-search path; keeping the frozen coordinates alone preserves scores but does not force the expanded graph to visit the same candidates. Thus the 6.25–25% vector-width overhead available with short complementary channels is neither an end-to-end latency bound nor the total storage cost of retaining an additional rollback index.

Table 4: ROCO test i2t R@1 at validation-selected α^*, seed 0. All rows use the same validation-selection and test-evaluation protocol. The matched-batch comparison is single-seed.

Model	Training batch	R@1
CoSE 2B	32	0.397
CoSE 2B	8	0.334
CoSE 8B	8	0.384

Table 5: Paired HAN loss controls (mean±sample std, three seeds). Reference uses standalone weight 1, MRL weight 0.5, and $\lambda_{\text{orth}} = 0$; the main-table checkpoint additionally includes $\lambda_{\text{orth}} = 0.1$. Overlap is agreement between the two branches' top-1 retrieval results.

Variant	Default i2t (α=0.5)	α=0.1 i2t	$\mathbf{z}_{\text{old}}$/$\mathbf{z}_{\text{new}}$ overlap@1
Reference	0.420 ± 0.003	0.503 ± 0.010	$25.3 \pm 0.2\%$
w/o standalone	0.439 ± 0.004	0.187 ± 0.004	$2.7 \pm 0.3\%$
w/o MRL	0.411 ± 0.002	0.515 ± 0.008	$27.5 \pm 0.9\%$

4.5 Scale and Backbone Transfer

At matched batch size eight, 8B exceeds 2B by 5.0pp (Table 4). Reducing the 2B batch from 32 to eight lowers R@1 by 6.3pp, consistent with batch size contributing to the apparent regression in the unmatched comparison. These single-seed comparisons neither isolate the cause nor establish universal scaling. The validation-selected weights for 2B batch 32, 2B batch eight, and 8B batch eight are 0.25, 0.15, and 0.25, respectively.

On SigLIP2-so400m Tschannen et al. [2025], a variant with an additional trainable projection MLP uses 7.96M LoRA and 5.31M projection parameters. CoSE at $\alpha = 0.5$ improves seed-0 i2t R@1 from 0.113 to 0.321 on ROCO and 0.059 to 0.147 on RSICD. On ROCO only, $\alpha = 0.9$ keeps Flickr and COCO within 5pp of Frozen; that diagnostic uses a different domain subset. RSICD general retention and ImageNet transfer were not evaluated. These results extend the evidence beyond Qwen3-VL without establishing joint attainment of those gains and preservation at $\alpha = 0.5$.

5 Analysis and Ablations

We focus the main text on two ablations that support the paper's core claims directly. Additional analyses (orthogonality geometry, $\mathbf{z}_{\text{new}}$ standalone performance, rank ablation, batch-size ablation, post-hoc Frozen+LoRA score-ensemble) are in Appendix A.

5.1 Joint Training versus Post-hoc Score Blending

A natural control blends Frozen with a separately trained LoRA encoder using the same score form and validation grid as CoSE. Domain-only tuning selects weights near the LoRA endpoint on three domains, but interior blends can improve general retention. Inference-time blending is therefore not unique to CoSE, and post-hoc blends can reach overlapping domain–general trade-off regions. On SemArt the validation-selected ensemble obtains domain R@1 0.518 versus CoSE's 0.490, with all three seeds satisfying the general budget; the SemArt/RSICD ensemble includes retrained LoRA seeds 42/123. CoSE's joint training determines the complementary channel available throughout its operating range; the following ablation tests the role of its supervision directly.

5.2 Loss-Term Ablation

We compare HAN loss terms with three paired seeds (0/42/123), matched initialization and data order, batch 64, 8000 steps, and $\lambda_{\text{orth}} = 0$ (Table 5). Each configuration selects α^* on validation i2t R@1.

Removing standalone supervision collapses low-α retrieval: mean i2t R@1 at $\alpha = 0.1$ falls from 0.503 to 0.187. Although midpoint i2t improves from 0.420 to 0.439, bidirectional domain retrieval decreases. At $\alpha = 0.5$, COCO and Flickr means fall 5.6 and 5.5pp below Frozen; each seed exceeds

the general-preservation budget. MRL remains in this control but supervises blended representations, not the standalone channel.

MRL improves HAN blended retrieval at widths $512/1024$ by 1.0–2.2pp per paired seed, at a 1.6pp mean cost to validation-selected full-width retrieval. Its benefit is specific to these truncated blends; the ROCO control does not establish the same gain (Appendix A.10).

6 Discussion and Conclusion

Cost and deployment boundary. Exact legacy-score preservation does not require doubling vector width: retaining the full frozen channel and adding 128–512 complementary coordinates gives 6.25–25% width overhead for $d = 2048$. This preserves the frozen endpoint, not necessarily full-width domain quality. Query encoding remains more expensive: full branching costs 2.1–$2.2\times$ Frozen at saturated batches. A verified prefix-sharing implementation costs $1.58\times$ at half depth; the duplicated-batch reference does not obtain that reduction. Known-domain two-table routing has higher mixed-traffic quality in our tested settings, while CoSE offers a domain–general blend within one expanded representation.

Quality and scope. CoSE is domain-competitive rather than domain-best. Domain-optimal weights need not satisfy the general-preservation budget, and partial backfill does not provide safe blended domain gains. The matched-batch 8B comparison and SigLIP2 transfer are single-seed evidence; neither establishes universal scaling or backbone-general performance guarantees. Retaining K domain branches concurrently requires $d + \sum_{k=1}^{K} m_k$ coordinates: truncation reduces the constant, not linear growth. Sequential updates to one domain can replace the previous branch after backfill. Concurrent multi-domain adaptation and adaptive per-query blending remain future work.

Conclusion. CoSE adds domain specialization while preserving the deployed score channel within an expanded representation. Joint training supports a tunable domain–general operating range; truncation and partial branching control its costs. Domain gains require corpus backfill, with frozen service until coverage completes. CoSE is useful when retaining an auditable frozen-score endpoint and controlling general retention within one representation justify the additional compute and storage.

Acknowledgements

This research was supported by the 'Advanced GPU Utilization Support Program' funded by the Government of the Republic of Korea (Ministry of Science and ICT).

Competing interests. The authors declare no additional competing interests beyond their disclosed affiliations.

Author Contributions Donghoon Han established the concept, problem definition, and logical framework of the study, and led the direction, development, and implementation of the methodology. He also led the experiments, the analysis and visualization of results, and the writing of the initial draft, contributing to every stage of the work from conception to completion.

SungHyun Moon contributed to the discussions that validated and refined the core idea and the proposed method, and to the planning of additional experiments. He also led the direction of revisions to the manuscript and the reviewer responses. The two authors jointly wrote the reviewer responses, took part in the author–reviewer discussion, and prepared the resubmission and camera-ready manuscripts.

SeungJae Lee reviewed the final manuscript.

References

Zhixiang Chi, Li Gu, Huan Liu, Ziqiang Wang, Yanan Wu, Yang Wang, and Konstantinos N. Plataniotis. Learning to adapt frozen CLIP for few-shot test-time domain adaptation. In *International Conference on Learning Representations*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/file/a5f2dc873f2971491d34b0e797411d82-Paper-Conference.pdf`.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

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

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

Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multimodal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 676–691, 2018. doi: 10.1007/978-3-030-11012-3_52.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

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*, 2022a.

Weihua Hu, Rajas Bansal, Kaidi Cao, Nikhil Rao, Karthik Subbian, and Jure Leskovec. Learning backward compatible embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3018–3028, 2022b.

Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. *arXiv preprint arXiv:2303.04766*, 2023.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

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

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. 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.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2018.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024.

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

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*, 62:1–16, 2024a.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024b.

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. doi: 10.1109/TGRS.2017.2776321.

SungHyun Moon, Aidyn Zhakatayev, and SeungJae Lee. Han: Korean heritage augmented narrative visual-language description dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 12859–12866, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758229. URL `https://doi.org/10.1145/3746027.3758229`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tan Pan, Furong Xu, Xudong Yang, Sifeng He, Chen Jiang, Qingpei Guo, Feng Qian, Xiaobo Zhang, Yuan Cheng, Lei Yang, et al. Boundary-aware backward-compatible representation via adversarial learning in image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15201–15210, 2023.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in COntext (ROCO): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, volume 11043 of *Lecture Notes in Computer Science*, pages 180–189. Springer, 2018. doi: 10.1007/978-3-030-01364-6_20.

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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

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

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2020.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022.

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.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. Milora: Harnessing minor singular components for parameter-efficient llm finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4823–4836, 2025.

Yifeng Xiong and Xiaohui Xie. OPLoRA: Orthogonal projection LoRA prevents catastrophic forgetting during parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34088–34096, 2026. doi: 10.1609/aaai.v40i40.40703.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European conference on computer vision*, pages 698–714. Springer, 2020.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

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.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

A Additional Ablations

This appendix extends the analyses in Section 5.

A.1 Orthogonality-Penalty Diagnostics

The main-table setting includes an output-space orthogonality penalty $\mathcal{L}_{\text{orth}}$ of weight $\lambda_{\text{orth}} = 0.1$; the paired loss controls omit it. The penalty averages squared cross-correlations of unit-normalized channels. Its weighted contribution is bounded by $\lambda_{\text{orth}}/d^2$, at most 2.4×10^{-8} for $d = 2048$, and is therefore numerically tiny. The following single-run diagnostics and the paired controls are separate experiments.

For full-width training, let $Q_0, Q_1 \in \mathbb{R}^{B\times d}$ contain the unit-normalized frozen and complementary query embeddings as rows. The penalty is

$$\mathcal{L}_{\text{orth}} = \frac{1}{d^2} \left\| \frac{Q_1^\top Q_0}{B} \right\|_F^2 .$$

It acts on query embeddings only, with the frozen channel detached. Since each outer product of unit vectors has Frobenius norm one, the averaged cross-correlation has Frobenius norm at most one; hence $\lambda_{\text{orth}}\mathcal{L}_{\text{orth}} \le \lambda_{\text{orth}}/d^2$.

We define overlap@1 as the fraction of queries for which the frozen and complementary channels retrieve the same top-ranked candidate, computed separately for i2t and t2i. It measures agreement between channels, not correctness against the ground-truth pair.

Diagnostic: λ_{orth}=0.1 vs λ_{orth}=0. Comparing the reported reference checkpoint (λ_{orth}=0.1) against a separate single run with λ_{orth}=0 on 1000 HAN extended val samples:

Metric	λ_{orth}=0.1	λ_{orth}=0	Δ
$\cos(\mathbf{z}_{\text{old}}, \mathbf{z}_{\text{new}})$ image	0.596 ± 0.048	0.590 ± 0.046	-0.006
$\cos(\mathbf{z}_{\text{old}}, \mathbf{z}_{\text{new}})$ text	0.504 ± 0.112	0.492 ± 0.124	-0.012
overlap@1 i2t / t2i	26.0 / 27.6%	26.9 / 26.8%	≤ 0.9 pp

These diagnostics show small observed changes in cosine and overlap when the penalty is removed; they are not pairwise statistical-equivalence tests.

λ_{orth} sweep (200× range). We sweep λ_{orth} across five values on HAN extended test; a sixth point (λ_{orth}=0) was the diagnostic above (val-measured). These are single-run diagnostics, including an unseeded reference; they are not paired seeded controls. The paired experiment is Table 5. For each nonzero-penalty run, α^* is selected by validation i2t R@1 on the 12-point grid; the table reports test performance at that selected weight. The selected weights for $\lambda_{\text{orth}} = 0.01, 0.1, 0.5, 1.0, 2.0$ are $0.15, 0.15, 0.15, 0.15, 0.10$, respectively.

λ_{orth}	0 (val)	0.01	0.1	0.5	1.0	2.0
i2t R@1 (α=0.5)	0.435	0.420	0.419	0.418	0.416	0.415
i2t R@1 (test, α^*)	—	0.516	0.499	0.511	0.512	0.504

The reported nonzero-weight test runs span 0.005 in default i2t R@1 and 0.017 in tuned i2t R@1. The zero-weight column uses validation rather than test and should not be compared as an identical evaluation. These are descriptive diagnostics, not proof that the penalty has no effect for every checkpoint.

Standalone supervision and geometry. Removing $\mathcal{L}_{\text{NCE}}(r_1)$ reduces overlap@1 to 2.4%, whereas removing $\mathcal{L}_{\text{orth}}$ changes overlap by less than 1pp in the reported HAN diagnostic. This supports an association between standalone retrieval supervision and the observed partial-overlap geometry; it does not establish a universal cosine equilibrium or isolate a rank-capacity mechanism.

Recommendation. Equation 3 defines the contrastive objective with optional MRL (HAN weight 0.5; other-domain weight 0). The main-table setting additionally uses $\lambda_{\text{orth}} = 0.1$, and its zero-penalty ablation remains a distinct run. Small diagnostic differences do not establish checkpoint equivalence.

A.2 Discrete α sweep across four domains

We sweep $\alpha \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 1\}$, the same grid used for selection, using the main-table seed-0 checkpoint per domain and independent encoding instances. Figure 3 shows validation i2t R@1 and ImageNet Top-1 on 5000 images. Markers use only validation i2t selection, never test results. Flickr/COCO use 1000 pairs (sampling seed 42); domain subsets use sampling seed 0.

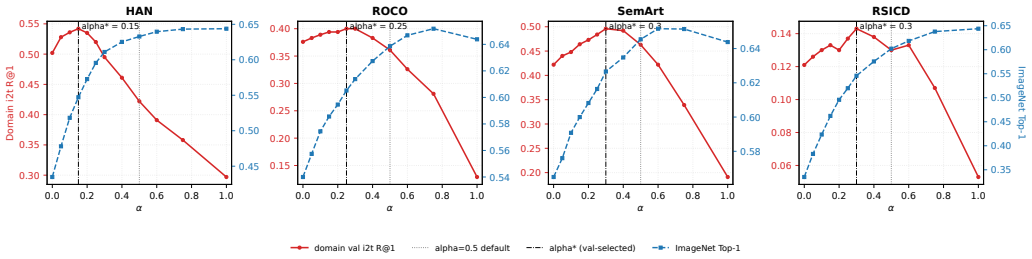


Figure 3: Seed-0 CoSE weight sweeps across the four domains. Solid red: validation i2t R@1; dashed blue: ImageNet Top-1 (5000 images). Vertical guides mark $\alpha = 0.5$ and validation-selected α^*. General quality need not change monotonically, and domain-only selection need not satisfy the preservation budget.

At $\alpha = 1$ all models recover the frozen score endpoint. Interior weights trade domain and general quality; other frozen-plus-adapter blends also support this sweep. Main-table means average three checkpoints, whereas these curves show seed 0 only.

A.3 Per-domain Pareto plot

Figure 1 is derived from the fully migrated, bidirectional domain and general means in Table 1. It counts benchmarks satisfying an operational 5pp budget, not statistical significance. It makes no claim that CoSE is the only system with an exact frozen-score endpoint; Frozen+adapter blends and retained-table systems can also provide one.

A.4 α-swept Pareto: CoSE vs Frozen+LoRA score ensemble

Both CoSE and the Frozen+LoRA ensemble preserve the frozen score at $\alpha = 1$ and require another stored channel to realize adapted retrieval. Table 6 reports fixed and domain-validation-selected ensemble operating points. Domain-only selection near the LoRA endpoint does not establish that all interior blends are redundant; preservation-constrained selection may prefer them. The loss ablation isolates the effect of standalone complementary supervision within CoSE.

Table 6: Frozen+LoRA score ensemble vs CoSE (test split, avg(i2t,t2i) R@1, mean±std across 3 seeds). α^*_{ens} is selected per seed on val i2t R@1 over the same 12-value sweep used for CoSE, breaking ties toward the smallest weight. SemArt/RSICD LoRA seeds 42/123 were retrained; thus their endpoint values need not equal the original main table. Ens. α=0.5 equals the unit-norm post-hoc concat by algebraic identity ($\cos$ of unit-concat $= \frac{1}{2}\cos(\mathbf{z}_{\text{old}}) + \frac{1}{2}\cos(z_{\text{lora}})$).

Domain	Frozen	LoRA	Ens. α=0.5	Ens. α^*_{ens}	CoSE α^*
HAN extended	0.320	0.450	0.409 ± 0.004	0.446 ± 0.010	0.465 ± 0.005
ROCO	0.142	0.457	0.407 ± 0.005	0.458 ± 0.008	0.377 ± 0.007
SemArt	0.204	0.512	0.439 ± 0.008	0.518 ± 0.007	0.490 ± 0.007
RSICD	0.088	0.150	0.161 ± 0.005	0.162 ± 0.003	0.156 ± 0.006

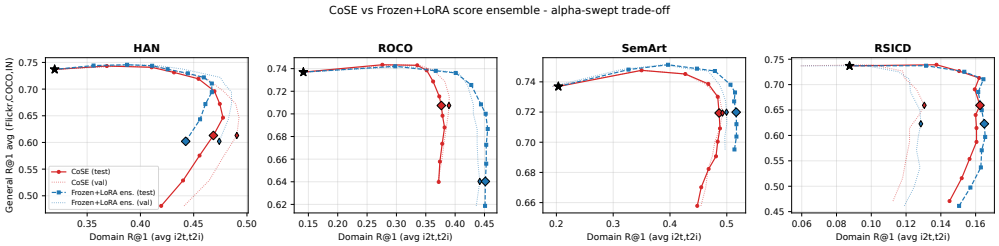


Figure 4: Seed-0 CoSE and Frozen+LoRA trade-off curves across the four domains over the 12-point grid. Solid/dashed curves show test and dotted curves show validation domain avg(i2t,t2i) R@1 on the horizontal axis; the vertical axis averages Flickr/COCO bidirectional R@1 and ImageNet Top-1 (5000 images). Diamonds mark the same validation-i2t-selected weights on both splits; the black star marks Frozen. These are sampled trade-off curves, not a claim that selected points lie on the test Pareto frontier.

A.5 Domain-level BC retrieval: legacy-index migration scenario

Table 1 reports fully migrated retrieval. The following cross-retrieval diagnostic uses an unchanged frozen corpus. It samples 1000 test pairs with evaluation seed equal to the training seed (0/42/123), uses FlashAttention-2, and uses the step-7500 HAN LoRA checkpoint rather than its final checkpoint. These protocol differences explain why self and Frozen references differ from Table 1; the values are not interchangeable.

Table 7: Domain-level BC cross retrieval (avg(i2t,t2i) R@1, mean±sample std across 3 seeds, 1000 pairs; variation combines training initialization and test-subset sampling). *self* = method-Q × method-K (the reference reported for this diagnostic, not numerically identical to Table 1); *cross* = method-Q × frozen-K (legacy-index scenario); *frozen* = frozen self-retrieval anchor. *% cap* reports the fraction of self-gain over frozen recovered by cross retrieval, $(c-f)/(s-f)$. CoSE row reports the Lemma-1 prediction: BC cross against a d-dim frozen corpus reduces to frozen self-retrieval ranking exactly, so % cap is 0 (legacy users see the unchanged frozen ranking) when using the frozen channel at $\alpha = 1$.

Domain	Method	self	cross	frozen	% cap	avg % cap
HAN	LoRA	0.456 ± 0.019	0.390 ± 0.008	0.321	50.8%	
	BCT-canonical	0.494 ± 0.016	0.424 ± 0.003	0.321	59.2%	
	BCT-cosine	0.480 ± 0.007	0.369 ± 0.005	0.321	30.2%	
	L2C-matched	0.388 ± 0.004	0.341 ± 0.002	0.321	28.7%	
ROCO	LoRA	0.452 ± 0.011	0.247 ± 0.006	0.133	35.8%	
	BCT-canonical	0.369 ± 0.010	0.263 ± 0.010	0.133	54.7%	
	BCT-cosine	0.362 ± 0.011	0.205 ± 0.006	0.133	31.5%	
	L2C-matched	0.267 ± 0.012	0.171 ± 0.010	0.133	28.4%	
SemArt	LoRA	0.514 ± 0.011	0.341 ± 0.011	0.203	44.5%	
	BCT-canonical	0.484 ± 0.005	0.388 ± 0.002	0.203	65.8%	
	BCT-cosine	0.475 ± 0.003	0.294 ± 0.007	0.203	33.4%	
	L2C-matched	0.393 ± 0.004	0.260 ± 0.003	0.203	29.7%	
RSICD	LoRA	0.151 ± 0.003	0.108 ± 0.008	0.084	35.5%	
	BCT-canonical	0.160 ± 0.010	0.132 ± 0.009	0.084	63.6%	
	BCT-cosine	0.166 ± 0.008	0.113 ± 0.011	0.084	35.5%	
	L2C-matched	0.135 ± 0.004	0.094 ± 0.002	0.084	19.9%	
Per-method averages across 4 domains						
LoRA						41.7%
BCT-canonical						60.8%
BCT-cosine						32.7%
L2C-matched						26.7%
CoSE (frozen-channel retrieval, $\alpha = 1$)						**0**%

Inversion of the general-BC story. In the separately reported general-benchmark diagnostic, LoRA cross retrieval can exceed LoRA self retrieval (RSICD-trained LoRA on Flickr: 0.733 versus 0.643). These values are retained as that diagnostic's references, not substituted for the differently reported means in Table 1. On domain data, Table 7 reports lower cross than self retrieval. Cross and fully migrated retrieval answer different corpus-state questions.

CoSE's structural BC sidesteps this. By Lemma 1, CoSE queries using the frozen channel at $\alpha = 1$ preserve legacy scores. Domain gains require the populated expanded corpus. Exact pairwise scores do not imply identical candidate sets from different ANN graphs; an unchanged legacy graph preserves the original search path.

A.6 Effect of Detaching the Frozen Output

The reference detaches the pooled frozen output, not intermediate hidden states. A single-run unseeded diagnostic obtains HAN default test i2t R@1 0.419 with detach and 0.422 without it. Since the base weights are frozen and LoRA deltas are confined to the second batch half, removing this detach does not train the frozen parameters or invalidate the score-preservation argument. It is a graph-separation diagnostic, not evidence that freezing fails without detach.

A.7 What $\mathbf{z}_{\text{new}}$ Learns Standalone

We evaluate seed-0 domain test i2t R@1 with the main-table checkpoints at $\alpha = 1$ (Frozen), 0 (complementary alone), and 0.5 (blend):

Domain	Frozen	$\mathbf{z}_{\text{new}}$ alone	Blend, 0.5
HAN extended	0.298	0.455	0.415
ROCO	0.132	0.390	0.359
SemArt	0.196	0.416	0.467
RSICD	0.084	0.140	0.166

The complementary endpoint supplies domain signal; balancing it with the frozen channel controls general retention. These are i2t, single-seed measurements, whereas Table 1 averages directions and seeds.

A.8 LoRA Rank Ablation

We vary LoRA rank on HAN extended validation at $\alpha = 0.5$:

Rank	Trainable params	HAN extended val i2t R@1
4	1.6M	0.418
16 (default)	**6.4M**	**0.440**
32	12.8M	0.448

Increasing rank from 4 to 16 improves HAN validation i2t R@1 by 0.022; increasing it from 16 to 32 adds 0.008. This within-method ablation measures sensitivity to adapter rank, rather than attributing differences between architectures to parameter count.

A.9 Batch-Size Ablation

Contrastive retrieval often benefits from large batches. On HAN extended val, DDP with cross-GPU all-gather:

Config	i2t R@1	t2i R@1
bs=64 (single GPU)	0.440	0.491
bs=128 (2-GPU DDP)	0.444	0.482
bs=256 (4-GPU DDP)	0.438	0.471

Increasing batch size from 64 to 128 or 256 yields no consistent improvement in the reported HAN validation retrieval metrics.

A.10 $\mathbf{z}_{\text{new}}$ Truncation via Matryoshka

We evaluate optional MRL with paired controls on HAN and ROCO. Each configuration is trained at seeds 0/42/123 with matched initialization and data order, $\lambda_{\text{orth}} = 0$, and MRL weight either 0

or 0.5. MRL averages the four blended losses at complementary widths 256/512/1024/2048. The main-table HAN checkpoints use MRL; ROCO/SemArt/RSICD checkpoints do not.

	HAN MRL	HAN no MRL	ROCO MRL	ROCO no MRL
Full, 0.5	0.432 ± 0.004	0.432 ± 0.004	0.349 ± 0.005	0.360 ± 0.003
Full, validation-selected α^*	0.469 ± 0.008	0.485 ± 0.009	0.371 ± 0.005	0.387 ± 0.004
$m = 256$, 0.5	0.420 ± 0.003	0.415 ± 0.014	0.345 ± 0.004	0.343 ± 0.002
$m = 512$, 0.5	0.444 ± 0.004	0.430 ± 0.003	0.354 ± 0.003	0.347 ± 0.004
$m = 1024$, 0.5	0.450 ± 0.004	0.433 ± 0.002	0.355 ± 0.002	0.353 ± 0.007

Entries are test avg(i2t,t2i) R@1, mean±sample std. For HAN, MRL minus no-MRL paired gains at $m = 512$ are $1.6/1.0/1.6$pp and at $m = 1024$ are $2.2/1.1/2.0$pp; at $m = 256$ the signs differ. The full-width validation-selected mean is 1.6pp lower with MRL. On ROCO, truncation gains are at most 0.7pp at the midpoint, comparable to observed same-seed retraining variation, while full-width means decrease 1.0–1.7pp. Thus the demonstrated MRL benefit concerns HAN truncated blends, not all domains or standalone complementary retrieval.

For the main-table ROCO checkpoint family, the three-seed $[2048\|256]$ configuration at $\alpha = 0.5$ obtains domain avg R@1 0.342, Flickr 0.883, COCO 0.675, and ImageNet Top-1 0.618 (Section 4.4). This result is without MRL. Truncating to $m = 128$ would give 6.25% width overhead but is outside the MRL training set.

A.11 Partial Branching across all four domains

Partial branching restricts adapters to layers K through $L - 1$. Computing the prefix once would give analytical cost $1 + (L - K)/L$; the reference duplicated-batch implementation recomputes it. The quality diagnostics below concern adapter placement, not measured speed. Table below evaluates K=0 (full), K=7 (75% overhead), and K=14 (50% overhead) across all four domains (i2t R@1, val at α=0.5).

Domain	$K = 0$	$K = 7$	$K = 14$	LoRA ref
HAN extended	0.440	0.414	0.405	0.543
ROCO	0.411	0.391	0.376	0.358
SemArt	0.484	**0.484**	0.454	0.497
RSICD	0.166	0.159	0.146	0.154

At $K = 14$, HAN i2t R@1 is 0.035 below full depth; SemArt at $K = 7$ equals its full-depth diagnostic. These adapter-placement diagnostics do not imply free computation. For the $K = 14$ ROCO checkpoint, prefix sharing preserves $\mathbf{z}_{\text{old}}/\mathbf{z}_{\text{new}}$ bit-for-bit on 200 texts and 100 images in bf16, and preserves $\alpha = 0.5$ scores and rankings exactly. The frozen image output also matches an independent Frozen forward bit-for-bit. An isolated H200 text benchmark (50 warmup, 150 queries, three repeats) measures p50 $15.6/21.4/20.8$ms for Frozen/reference/prefix-sharing. At batch 64, reference and prefix-sharing cost $2.08\times$ / $1.58\times$ Frozen. Throughput uses external query counts; the analytical $1.5\times$ is not a measured latency ratio.

A.12 Full HAN extended Partial-Branching Sweep

Full start-layer sweep (HAN extended val, α=0.5):

	K=0 (full)	K=4	K=7	K=14	K=21	LoRA ref
Analytical extra cost	100%	86%	75%	50%	25%	0%
Trainable params	6.4M	5.5M	4.8M	3.2M	1.6M	6.4M
HAN extended i2t R@1	0.440	0.427	0.414	0.405	0.377	0.543
HAN extended t2i R@1	0.491	0.453	0.442	0.431	0.412	0.400

At $K = 21$, HAN validation i2t/t2i R@1 is $0.377/0.412$, compared with $0.440/0.491$ at full depth. The listed compute savings assume a shared-prefix implementation; the duplicated-batch reference does not realize them. Index preservation (Lemma 1) holds at every K.

B Baseline Implementation Details

BCT — two variants. We report two BCT formulations:

1. **Canonical (Shen et al. 2020)** — the main-table version. The loss is symmetric cross-encoded InfoNCE: $\mathcal{L}_{\text{bct}} = \frac{1}{2}[\text{InfoNCE}(f_{\text{new}}(q), f_{\text{old}}(d^+), \text{negs}_{\text{old}}) + \text{InfoNCE}(f_{\text{old}}(q), f_{\text{new}}(d^+), \text{negs}_{\text{new}})]$. This directly targets "new queries retrieve correctly from the old index," the property backward compatibility is defined by.
2. **Cosine-alignment distillation** — a simpler feature-distillation variant: $\mathcal{L}_{\text{bct}}^{\cos} = \mathbb{E}_x\,[1 - \cos(f_{\text{new}}(x), f_{\text{old}}(x))]$. This pulls new outputs toward old outputs in feature space, rather than targeting retrieval-against-old directly.

Both BCT variants use the same 6.42M LoRA configuration and reference schedule. CoSE batches are 64 on HAN and 32 on the other domains; matched parameter counts and schedules do not imply that every baseline uses the same batch size. The run-specific launch configurations for seed-0 BCT-canonical and all twelve L2C-matched runs specify microbatch 16 with four accumulation steps. These launch settings are not independently verified as the effective settings of every final checkpoint; accumulation does not increase the simultaneous contrastive candidate set. These are parameter-count comparisons, not comparisons with identical contrastive candidate sets across every method. The SemArt/RSICD LoRA replicas used for the ensemble and two-table comparisons use batches 32/128, respectively. Table 1 uses the baseline checkpoints from the single-encoder evaluation. Canonical BCT directly supervises cross-encoded retrieval, whereas cosine alignment supervises feature proximity; neither gives CoSE's structural legacy-score equality.

L2C — retrieval adaptation. We adapt L2C Chi et al. [2025] to retrieval with the following combining rule and side-branch configuration:

Element-wise reverted attention (faithful). The official combining computes

$$m = \mathbf{I}_{\text{CLIP}} \odot \mathbf{I}_{\text{Side}}, \quad A = \mathbf{1} - \text{softmax}(m, \text{dim}{=}{-1}), \quad \tilde{\mathbf{I}}_{\text{Side}} = \mathbf{I}_{\text{Side}} \odot A$$

where $\odot$ is element-wise multiplication and softmax is taken over the feature dimension. We use this element-wise rule rather than a matrix-product attention rule. A LayerNorm follows the gated output before combining.

Side branch reads backbone hidden states (adaptation). The official L2C uses a fully-parallel separate encoder (SideNet) processing raw image tokens; adapting that to a decoder-based VLM like Qwen3-VL would require training a separate VLM encoder, an orthogonal design change. For comparison with CoSE's side-branch LoRA, our CPNet reads the backbone's detached final hidden states as input. The combining rule (element-wise reverted attention + addition) follows the official implementation.

Parameter-matched configuration (L2C-matched). Since the official CPNet (3 transformer blocks at backbone hidden dim, ~ 155M trainable) is not parameter-comparable to our 6.4M baselines, we report a parameter-matched variant: CPNet uses one block of width 576, eight attention heads, and an FFN width of $4 \times 576 = 2304$, with learned projections from 2048 to 576 and back, giving ≈ 6.4M trainable parameters. This provides a parameter-count control for comparison with CoSE. We also evaluate a full-capacity variant (155M trainable, three blocks, sixteen heads). The full variant uses width 2048 and FFN width 8192. The matched variant controls parameter count, not all training conditions or the isolated effect of the combining rule; neither retrieval adaptation is the original classification task.

Objectives and candidate sets. Table 8 separates the implementations used for the reported model families. Standard symmetric InfoNCE averages the two directions of a $B \times B$ similarity matrix, with one diagonal positive per direction and no duplicated positive. In contrast, the directional implementation in Section 3.4 has $B + 1$ logits before same-image masking, including two copies of the positive. Gradient accumulation accumulates gradients, not candidates: microbatch 16 with four accumulation steps still provides only 16 candidates per directional standard-InfoNCE query, of which 15 are negatives. Batch sizes below are paired-record counts, not the duplicated input count inside CoSE. The CoSE and SigLIP2 reference trainers backpropagate each minibatch without gradient accumulation. Parameter matching does not control the objective or contrastive candidate set and therefore does not isolate architecture alone.

Table 8: Training-objective mapping. D denotes the directional duplicated-positive loss with same-image masking; S denotes standard symmetric InfoNCE. The objectives and contrastive candidate sets differ across model families.

Model family	Objective and scope
Main CoSE	D on blend and complementary channel; orthogonality weight 0.1. HAN: MRL 0.5, batch 64; other domains: MRL 0, batch 32.
Paired CoSE controls	Same D; orthogonality 0. Standalone and MRL terms vary as specified in the ablation.
LoRA	HAN uses D on the sampled pair types; other domains use S. No CoSE standalone or MRL term.
BCT-canonical	S on adapted pairs plus 0.5 times the mean of two cross-encoded S losses (new/old and old/new).
BCT-cosine	S plus 0.5 times the mean query/positive cosine-alignment penalty.
L2C-matched/full	S on the combined embeddings. No duplicated positive or CoSE MRL term. Matched launch settings: microbatch 16, accumulation 4; effective settings are not verified for every checkpoint.
SigLIP2 CoSE	D on blend and complementary channel; orthogonality 0.1, no MRL; batch 32, 4000 steps, seed 0.

SigLIP2 complementary projection. The SigLIP2-so400m variant applies rank-16 LoRA to all attention $Q/K/V/O$ projections in both towers. Its frozen channel uses the native pooled output. In the trainable pathway, vision hidden states are mean-pooled over tokens, while text uses the last token; each modality has a separate Linear–GELU–Linear projection before branch L2 normalization. The first linear map takes the tower hidden width to the shared embedding width, and the second preserves that width. Both towers and the embedding have width 1152, giving two $1152 \to 1152 \to 1152$ MLPs (5.31M parameters including biases). Training uses AdamW, learning rate 2×10^{-4}, weight decay 0.01, 200-step warmup followed by cosine decay, and temperature 0.02. This projection-equipped variant differs architecturally from the Qwen implementation.

C Dataset Details

Retrieval evaluation protocol. Each domain evaluation samples 1000 image records without replacement and selects one caption per record, using sampling seed 0 for both validation and test subsets. HAN selects one of 11 caption types uniformly, then selects English or Korean with equal probability. RSICD selects one of the available captions uniformly; ROCO and SemArt use the sole available caption. Flickr30K and COCO evaluation use sampling seed 42 and the first caption stored for each sampled image. COCO samples 1000 images from the full 5000-image validation split; ImageNet samples 5000 images from the complete validation set with seed 42. These general evaluation subsets are shared across methods. For these one-to-one retrieval evaluations, the image and selected caption from the same sampled row are the sole positive pair in either direction. Unselected captions are absent from the candidate set; this is not a multi-caption relevance evaluation.

HAN extended. HAN extended is an extended version of the HAN dataset Moon et al. [2025], currently under review for public release. We use a version with $51{,}065$ Korean-broadcast images; each image is annotated with 11 parallel English/Korean caption pairs (4 factual, 6 inferential, 1 comprehensive; 22 captions per image across both languages), yielding roughly 1.1M image-caption pairs. We split deterministically $80/10/10$ (seed 42): train $40{,}852$, val $5{,}106$, test $5{,}107$. For training, we sample four pair types: (A) image↔text (50%), (B) cross-lingual EN↔KO parallel (20%), (C) same-language paraphrase (20%), (D) cross-lingual non-parallel (10%). The base HAN dataset is publicly available on AIHub (`https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71866`).

ROCO (medical radiology). Radiology images with technical captions (e.g. "axial CT showing hypodense lesion in segment VII of the liver"). Native train/validation/test splits (train $65{,}422$, validation $8{,}175$, test $8{,}176$).



"Haeundae Beach blossoms into a communal summer celebration."

"Intelligent robotic arms separate waste, signaling sustainability."

"The workshop becomes a small sanctuary where dust and shared life meet."

Figure 5: **HAN extended** (Korean broadcast, EN shown). Each image carries 11 parallel EN/KO caption pairs; one English caption is shown per image.



"Hydrocephalus is seen on the preoperative axial MRI."

"60-year-old man with relapsed multiple myeloma; CT demonstrates bilateral perirenal masses."

"ERCP shows the mass spreading from the cystic duct to the upper-middle bile duct."

Figure 6: **ROCO** (medical radiology). Images letterboxed to square for display; black padding preserves original content. Captions are single technical narratives.

SemArt (art paintings). 19,244 paintings with long-form art-historical descriptions (artist, technique, subject). Native train/val/test splits.

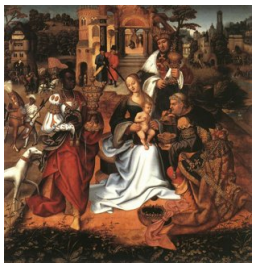
RSICD (remote sensing). $\sim$ 10,000 satellite/aerial images with short factual captions. Native train/val/test splits. Smallest of our domains and a stress test for catastrophic forgetting (RSICD-trained LoRA obtains ImageNet Top-1 of 0.287, seed 0).

General benchmarks. Flickr30K Plummer et al. [2015] and COCO 2017 val Lin et al. [2014]: 1000 image-caption pairs each, sampled with seed 42. ImageNet-1k Deng et al. [2009] zero-shot classification: 5000 images, seed 42. All external to training.

Representative image-caption samples for each domain are shown in Figs. 5 (HAN extended), 6 (ROCO), 7 (SemArt), and 8 (RSICD).

D Full R@1/R@5/R@10 i2t/t2i Breakdowns

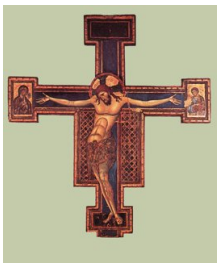
This section gives the complete per-method, per-domain breakdown across all retrieval metrics: R@1, R@5, R@10 for both image-to-text (i2t) and text-to-image (t2i). Tables are split into (i) domain test (Table 9), (ii) Flickr30K (Table 10), (iii) COCO 2017 val (Table 11), and (iv) ImageNet-1k zero-shot (Table 12). Entries report mean $\pm$ sample std across seeds 0, 42, 123, except SemArt/RSICD LoRA general results, which use only the available original seed-0 checkpoints and have no std. All methods use the same general evaluation samples; domain results retain three seeds. For CoSE, two operating points are reported on domain (Table 9): the zero-tune α=0.5 default and the val-selected α^* per seed. For general retrieval (Tables 10–12), CoSE is evaluated at α=0.5 to provide a single comparable column against the non-blending baselines; Fig. 3 shows the seed-0 ImageNet Top-1 curve over the 12-point α grid.



"This panel is the front of the wing of an altarpiece; on the reverse is St Bernard of Clairvaux."



"View of the Sala del Camino (Room of the Hearth), decorated by Vasari personally."

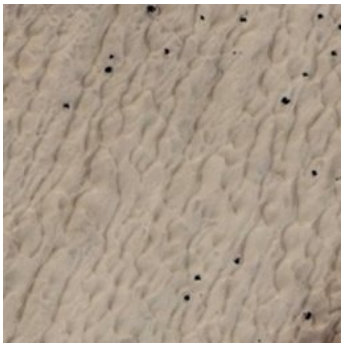


"From the 13th century the Berlinghieri, a family of painters . . ."

Figure 7: **SemArt** (art paintings). Portraits rendered at matched height. Captions are long-form art-historical descriptions (abbreviated here).



"A port with several boats is close to many buildings."



"A piece of sand is like scales of fish in the desert."



"There are school buses stopping in the parking lot next to a grassland and buildings with black roof."

Figure 8: **RSICD** (remote sensing). Square aerial tiles. Each image has five short factual captions.

D.1 Domain Test (matched protocol, 1000 pairs, seed-0 sampler)

Per-method, per-domain breakdown of domain-test retrieval at all three recall cutoffs in both directions, mirroring the unified main-paper Table 1 but at full granularity. Entries are mean±std across the three seeds $(0, 42, 123)$. LoRA leads on i2t R@1 on HAN, ROCO, and SemArt, with a larger ROCO gap; RSICD is numerically close. On t2i, BCT-canonical leads on HAN extended at R@1, with CoSE α=0.5 within 7pp; LoRA leads on ROCO/SemArt; on RSICD the methods cluster within ±2pp.

D.2 Flickr30K (general retrieval, 1000 pairs, seed-42 sampler)

Flickr30K results at all six metrics (R@1/5/10 in both directions) show that CoSE at $\alpha = 0.5$ remains close to Frozen. L2C-matched and BCT-cosine also retain high general retrieval quality. RSICD-trained LoRA has the largest observed drop; its general results use seed 0 only.

D.3 COCO 2017 val (general retrieval, 1000 pairs, seed-42 sampler)

Under the shared COCO evaluation, CoSE at $\alpha = 0.5$ stays within 2pp of Frozen in both R@1 directions. L2C-matched is within 2pp in bidirectional R@1 on all four domains, and BCT-cosine within 5pp. LoRA has larger bidirectional drops (7.2–23.4pp); SemArt/RSICD LoRA values are single-seed observations.

Table 9: Domain test — full R@1/R@5/R@10 in both directions (i2t and t2i). Mean±sample std across 3 seeds, all percentages. Same 1000-pair seed-0 sampler used in Table 1.

		i2t			t2i		
Trained on	Method	R@1	R@5	R@10	R@1	R@5	R@10
HAN extended	LoRA	52.7 ± 0.5	85.1 ± 0.9	92.9 ± 0.5	37.2 ± 1.8	69.3 ± 1.1	80.7 ± 1.2
	BCT-canonical	47.2 ± 0.8	80.2 ± 0.7	90.1 ± 0.5	51.1 ± 0.9	81.9 ± 1.0	89.9 ± 0.8
	BCT-cosine	45.6 ± 0.9	78.2 ± 0.4	89.0 ± 0.6	49.6 ± 0.5	81.5 ± 0.3	89.0 ± 0.8
	L2C-matched	37.3 ± 0.8	68.2 ± 0.7	81.0 ± 0.8	40.4 ± 0.2	70.6 ± 0.7	81.3 ± 0.4
	CoSE α=0.5	41.6 ± 0.4	73.9 ± 0.5	85.6 ± 0.3	44.9 ± 0.8	76.5 ± 0.6	84.9 ± 0.5
	CoSE α^*	50.4 ± 1.1	82.6 ± 0.2	91.6 ± 0.4	42.7 ± 1.3	74.0 ± 0.7	85.0 ± 1.0
ROCO	LoRA	45.6 ± 0.9	74.6 ± 0.4	84.3 ± 0.2	45.7 ± 0.7	74.2 ± 0.6	84.4 ± 0.9
	BCT-canonical	35.9 ± 0.6	63.3 ± 0.6	75.9 ± 0.7	37.1 ± 1.0	66.5 ± 0.3	78.1 ± 0.8
	BCT-cosine	35.5 ± 0.6	63.9 ± 0.5	76.1 ± 0.8	36.6 ± 0.4	66.9 ± 1.4	78.1 ± 1.2
	L2C-matched	26.5 ± 1.0	53.8 ± 0.9	67.8 ± 0.8	28.2 ± 0.4	57.4 ± 0.9	69.9 ± 0.6
	CoSE α=0.5	35.5 ± 0.6	65.1 ± 0.5	77.4 ± 0.6	35.0 ± 1.0	63.8 ± 0.4	75.3 ± 0.2
	CoSE α^*	39.6 ± 1.1	69.8 ± 1.0	82.4 ± 0.3	35.8 ± 0.3	64.7 ± 0.6	75.8 ± 0.4
SemArt	LoRA	51.5 ± 0.4	75.4 ± 0.4	83.5 ± 0.1	51.2 ± 0.4	76.4 ± 0.7	84.2 ± 0.3
	BCT-canonical	47.8 ± 1.0	72.3 ± 0.8	81.7 ± 1.0	49.6 ± 0.9	74.6 ± 0.6	82.5 ± 0.1
	BCT-cosine	45.7 ± 0.9	71.8 ± 0.8	81.0 ± 0.9	49.4 ± 1.0	72.2 ± 1.1	81.6 ± 1.2
	L2C-matched	38.5 ± 0.8	63.6 ± 0.5	74.4 ± 1.0	40.0 ± 1.2	65.1 ± 0.2	75.3 ± 0.6
	CoSE α=0.5	47.4 ± 0.6	70.5 ± 0.3	79.7 ± 0.5	46.6 ± 0.5	70.8 ± 0.9	78.6 ± 0.2
	CoSE α^*	49.4 ± 0.8	74.7 ± 1.0	83.4 ± 0.5	48.6 ± 0.6	73.1 ± 1.0	81.5 ± 0.2
RSICD	LoRA	15.6 ± 0.5	38.7 ± 0.3	53.7 ± 0.4	15.9 ± 1.7	39.5 ± 0.7	53.9 ± 0.2
	BCT-canonical	16.0 ± 1.4	41.7 ± 0.6	58.5 ± 1.2	17.4 ± 0.9	43.4 ± 0.8	60.4 ± 0.9
	BCT-cosine	16.5 ± 1.0	41.3 ± 0.9	58.5 ± 0.8	17.8 ± 0.2	43.6 ± 1.2	59.4 ± 1.5
	L2C-matched	12.5 ± 1.0	36.2 ± 0.6	52.3 ± 1.1	14.2 ± 0.5	38.9 ± 1.3	55.3 ± 0.6
	CoSE α=0.5	16.5 ± 0.1	42.6 ± 0.6	58.6 ± 0.3	15.2 ± 0.6	40.5 ± 0.6	57.9 ± 1.2
	CoSE α^*	16.1 ± 0.7	41.5 ± 1.2	58.9 ± 1.2	15.0 ± 0.8	39.2 ± 1.2	56.0 ± 0.6

Table 10: Flickr30K full breakdown. Shared evaluation samples; mean±sample std (3 seeds), except SemArt/RSICD LoRA (seed 0 only, no std). Frozen reference: i2t $87.4/98.2/99.6$, t2i $88.6/98.5/99.3$.

		i2t			t2i		
Trained on	Method	R@1	R@5	R@10	R@1	R@5	R@10
HAN	LoRA	84.7 ± 0.2	96.6 ± 0.5	98.5 ± 0.1	70.4 ± 1.4	90.9 ± 0.9	95.4 ± 0.5
	BCT-canonical	85.7 ± 0.4	97.3 ± 0.1	98.5 ± 0.3	85.8 ± 1.2	97.1 ± 0.4	98.6 ± 0.2
	BCT-cosine	88.3 ± 0.9	97.9 ± 0.3	99.2 ± 0.3	89.2 ± 0.5	98.3 ± 0.1	99.0 ± 0.2
	L2C-matched	85.7 ± 0.7	97.8 ± 0.3	98.9 ± 0.2	87.2 ± 0.5	97.8 ± 0.2	98.8 ± 0.2
	CoSE α=0.5	87.9 ± 0.6	97.9 ± 0.2	99.4 ± 0.1	88.3 ± 0.2	97.8 ± 0.4	98.9 ± 0.1
ROCO	LoRA	74.6 ± 0.8	92.3 ± 0.9	95.9 ± 0.5	78.7 ± 2.5	94.4 ± 0.6	97.0 ± 0.3
	BCT-canonical	82.7 ± 0.7	96.2 ± 0.4	98.1 ± 0.3	86.0 ± 0.2	97.0 ± 0.5	98.5 ± 0.1
	BCT-cosine	84.4 ± 0.3	97.1 ± 0.3	98.9 ± 0.2	86.9 ± 0.2	97.5 ± 0.3	98.8 ± 0.2
	L2C-matched	87.4 ± 0.3	98.2 ± 0.3	99.7 ± 0.0	89.1 ± 0.2	98.3 ± 0.1	99.1 ± 0.2
	CoSE α=0.5	89.0 ± 0.4	98.2 ± 0.2	99.2 ± 0.1	89.6 ± 0.5	98.8 ± 0.2	99.4 ± 0.1
SemArt	LoRA	84.7	96.4	98.0	85.2	97.3	98.3
	BCT-canonical	83.1 ± 1.2	95.9 ± 0.4	98.1 ± 0.2	85.7 ± 0.9	97.4 ± 0.3	98.7 ± 0.1
	BCT-cosine	85.5 ± 0.4	97.2 ± 0.4	98.7 ± 0.2	88.3 ± 0.3	97.9 ± 0.2	99.0 ± 0.2
	L2C-matched	85.4 ± 0.2	97.7 ± 0.3	99.2 ± 0.2	87.7 ± 0.4	98.0 ± 0.4	99.1 ± 0.2
	CoSE α=0.5	88.1 ± 0.3	98.0 ± 0.2	99.2 ± 0.2	89.7 ± 0.2	98.8 ± 0.2	99.5 ± 0.1
RSICD	LoRA	63.8	82.1	87.2	65.1	85.3	90.3
	BCT-canonical	76.3 ± 1.9	92.0 ± 0.9	95.7 ± 1.1	76.3 ± 1.4	91.9 ± 1.0	94.8 ± 0.9
	BCT-cosine	82.3 ± 2.8	95.7 ± 0.6	97.8 ± 0.7	83.9 ± 3.2	96.4 ± 1.1	97.9 ± 0.5
	L2C-matched	87.4 ± 0.6	98.2 ± 0.2	99.2 ± 0.1	87.8 ± 0.2	97.9 ± 0.2	99.2 ± 0.2
	CoSE α=0.5	86.0 ± 0.6	97.7 ± 0.2	98.9 ± 0.3	87.0 ± 0.5	98.2 ± 0.2	99.3 ± 0.2

D.4 ImageNet-1k zero-shot (5000 images, seed-42 sampler)

ImageNet zero-shot evaluates general visual–text alignment, with a Frozen reference of $64.4/86.8$ (Top-1 / Top-5). LoRA Top-1 drops range from 2.2 to 35.7pp; the SemArt/RSICD values use seed 0 only. CoSE at $\alpha = 0.5$ stays within 4.8pp of Frozen on both metrics. L2C-matched also preserves ImageNet performance, with Top-1 drops of 0.7–1.9pp, at lower domain quality than CoSE in Table 1.

Table 11: COCO 2017 val full breakdown. Shared evaluation samples; mean±sample std (3 seeds), except SemArt/RSICD LoRA (seed 0 only, no std). Frozen reference: i2t $69.6/92.4/96.3$, t2i $67.8/91.3/96.8$.

		i2t			t2i		
Trained on	Method	R@1	R@5	R@10	R@1	R@5	R@10
HAN	LoRA	58.5 ± 1.6	86.1 ± 0.5	93.2 ± 0.5	42.8 ± 0.9	73.7 ± 0.8	85.4 ± 0.6
	BCT-canonical	57.9 ± 1.0	85.1 ± 0.4	93.3 ± 0.5	56.3 ± 1.3	83.8 ± 1.5	92.5 ± 1.3
	BCT-cosine	64.0 ± 1.2	90.1 ± 0.9	95.2 ± 0.2	65.5 ± 0.6	89.5 ± 0.8	95.2 ± 0.3
	L2C-matched	69.0 ± 0.4	91.2 ± 0.3	96.3 ± 0.2	67.4 ± 0.4	90.5 ± 0.2	95.8 ± 0.1
	CoSE α=0.5	68.5 ± 0.7	91.8 ± 0.2	95.8 ± 0.5	68.1 ± 0.2	91.4 ± 0.3	95.7 ± 0.3
ROCO	LoRA	55.7 ± 1.9	81.3 ± 1.7	89.7 ± 1.1	57.9 ± 1.0	84.0 ± 1.4	90.8 ± 0.9
	BCT-canonical	63.8 ± 0.6	87.4 ± 0.4	93.7 ± 1.3	65.2 ± 1.0	89.3 ± 0.6	94.8 ± 0.1
	BCT-cosine	64.0 ± 1.3	88.8 ± 0.4	95.3 ± 0.2	68.0 ± 0.6	90.3 ± 0.2	95.5 ± 0.3
	L2C-matched	69.3 ± 0.4	92.1 ± 0.2	96.8 ± 0.2	67.3 ± 0.3	90.7 ± 0.3	96.6 ± 0.2
	CoSE α=0.5	68.9 ± 0.2	91.5 ± 0.3	96.2 ± 0.2	67.9 ± 0.4	91.1 ± 0.2	96.0 ± 0.2
SemArt	LoRA	61.2	89.3	95.4	61.7	88.8	94.5
	BCT-canonical	61.3 ± 0.2	89.5 ± 0.6	94.9 ± 0.6	63.5 ± 0.9	89.1 ± 0.6	94.3 ± 0.2
	BCT-cosine	65.7 ± 1.5	91.2 ± 0.4	96.4 ± 0.3	68.0 ± 0.5	90.5 ± 0.2	95.6 ± 0.1
	L2C-matched	68.1 ± 0.3	91.3 ± 0.1	96.3 ± 0.1	65.6 ± 0.3	90.3 ± 0.4	95.8 ± 0.3
	CoSE α=0.5	68.3 ± 0.3	92.4 ± 0.3	96.7 ± 0.3	68.7 ± 0.6	91.4 ± 0.2	96.5 ± 0.2
RSICD	LoRA	45.2	72.4	82.1	45.4	74.2	83.9
	BCT-canonical	56.5 ± 0.6	83.6 ± 0.6	91.4 ± 0.5	56.9 ± 0.9	82.9 ± 1.4	90.5 ± 0.9
	BCT-cosine	63.8 ± 3.2	88.8 ± 3.0	94.4 ± 1.5	63.7 ± 3.7	88.2 ± 1.0	94.3 ± 0.4
	L2C-matched	69.4 ± 0.1	91.8 ± 0.2	96.4 ± 0.1	66.5 ± 0.1	89.8 ± 0.5	95.6 ± 0.2
	CoSE α=0.5	68.0 ± 0.7	91.6 ± 0.6	95.6 ± 0.5	66.9 ± 0.4	90.0 ± 0.6	95.3 ± 0.3

Table 12: ImageNet zero-shot Top-1 / Top-5. Shared evaluation samples; mean±sample std (3 seeds), except SemArt/RSICD LoRA (seed 0 only, no std). Frozen reference: $64.4/86.8$.

Trained on	Method	Top-1	Top-5
HAN	LoRA	52.4 ± 1.2	80.2 ± 0.7
	BCT-canonical	54.2 ± 2.5	82.0 ± 0.9
	BCT-cosine	62.3 ± 1.6	86.3 ± 1.0
	L2C-matched	62.5 ± 0.6	85.7 ± 0.6
	CoSE α=0.5	63.5 ± 0.3	86.5 ± 0.1
ROCO	LoRA	49.1 ± 1.4	73.4 ± 1.3
	BCT-canonical	59.3 ± 2.2	82.3 ± 1.8
	BCT-cosine	62.2 ± 1.5	85.0 ± 1.1
	L2C-matched	63.7 ± 0.1	86.8 ± 0.1
	CoSE α=0.5	63.8 ± 0.1	86.4 ± 0.1
SemArt	LoRA	62.2	84.7
	BCT-canonical	60.2 ± 0.6	83.4 ± 0.7
	BCT-cosine	63.0 ± 1.1	85.3 ± 0.7
	L2C-matched	63.3 ± 0.1	86.0 ± 0.1
	CoSE α=0.5	65.0 ± 0.4	87.2 ± 0.1
RSICD	LoRA	28.7	54.2
	BCT-canonical	44.9 ± 2.0	71.1 ± 1.4
	BCT-cosine	51.1 ± 4.8	77.6 ± 4.0
	L2C-matched	63.1 ± 0.3	85.7 ± 0.2
	CoSE α=0.5	59.6 ± 0.7	83.4 ± 0.7

E Feature Drift: CoSE vs BCT vs LoRA

Setup. For each method and domain, we measure the cosine similarity between the method's output embedding and the frozen base encoder's embedding on the same input, averaged over 200 val-split images (Table 13). For CoSE we measure the $\mathbf{z}_{\text{old}}$ slice (first d coordinates of $\mathbf{z}_{\text{final}}$); for LoRA / BCT-canonical / BCT-cosine we measure the full post-adapter output. CoSE's structural-preservation claim implies zero drift up to kernel determinism; BCT's distillation target implies small-but-nonzero drift; LoRA has no alignment loss and is expected to drift substantially.

Takeaways. CoSE's reported mean cosine drift rounds to zero under bf16/SDPA independent forwards; this is not an FP32 experiment or a maximum-coordinate-error measurement. BCT-cosine, which explicitly minimizes $1 - \cos(f_{\text{new}}, f_{\text{frozen}})$ during training, drifts least among the adapted baselines (0.06–0.14) but is still many orders of magnitude larger than the observed cosine-rounding

Table 13: Per-domain output-embedding drift relative to the frozen base. Drift $= 1 - \cos(f_{\text{method}}(x), f_{\text{frozen}}(x))$, averaged over 200 val images per domain, seed 0. CoSE mean drift rounds to 0 under bf16/SDPA independent forwards; sample std is approximately 7×10^{-8}. Cosine arithmetic is subject to floating-point rounding.

Domain	CoSE (ours)	LoRA	BCT-canonical	BCT-cosine
HAN extended	$\mathbf{0.00 \times 10^{0}}$	3.58×10^{-1}	4.18×10^{-1}	9.72×10^{-2}
ROCO	$\mathbf{0.00 \times 10^{0}}$	3.03×10^{-1}	5.11×10^{-1}	6.14×10^{-2}
SemArt	$\mathbf{0.00 \times 10^{0}}$	4.47×10^{-1}	5.30×10^{-1}	9.31×10^{-2}
RSICD	$\mathbf{0.00 \times 10^{0}}$	7.23×10^{-1}	5.82×10^{-1}	1.39×10^{-1}

scale. BCT-canonical (symmetric cross-encoded InfoNCE, Shen et al. 2020) does not enforce feature-space proximity directly, and its drift is actually larger than LoRA on three of four domains (0.42–0.58 vs LoRA's 0.30–0.45); canonical BCT targets retrieval-via-old-index consistency, not embedding proximity. LoRA drifts 0.30–0.72 with no alignment target — particularly severe on small-data RSICD where it catastrophically moves away from the frozen output. The quantitative gap between "structural preservation" (CoSE, 0) and "approximate preservation" (BCT-cosine, $\sim 10^{-1}$) is what the paper refers to as a guarantee by construction versus a distillation approximation.

F Deployment Scenarios

CoSE preserves the old coordinates while computing a new complementary channel. Domain gains require backfilling that channel; uninterrupted frozen service during backfill can also be provided by blue-green migration with other methods.

Frozen service and rollback. At $\alpha = 1$, the complementary branch can be bypassed and queries can use the original legacy index. Scores are preserved up to execution-level numerical determinism. Switching the weight to one on a different ANN graph does not guarantee identical candidates; retaining the original graph preserves the original search path and adds storage during the rollback window. Without storage sharing, adding the projected 846MB legacy index to the compact/full expanded indices in Table 3 gives 1795/2512MB at 100k documents. These sums illustrate additional rollback storage; they are not separate measurements.

Backfill followed by an atomic switch. A rolling job populates $\mathbf{z}_{\text{new}}$ while serving continues with the frozen pathway. Mixed scoring of covered and uncovered documents introduces coverage bias. Ten coverage draws per operating point yielded the following diagnostic at 25% coverage; the baselines are specific to that experiment, not Table 1's sampler references.

	HAN	ROCO	SemArt	RSICD
Frozen R@1	0.341	0.151	0.211	0.091
R@1 at 25% coverage	0.270	0.134	0.215	0.084
Excess covered top-1 share	0.52	0.71	0.44	0.65

Excess share is $P(\text{top-1} \in \text{covered}) - 0.25$. Rank fusion and isotonic calibration reduced but did not eliminate this bias. We recommend frozen service until coverage is complete, then an atomic switch to a common selected weight.

Steady-state cost. Full-width storage is $d + m$ coordinates per document, or $d + m'$ when only the complementary channel is truncated. This preserves frozen scores, not necessarily full-width adapted quality. Query computation is approximately $1 + (L - K)/L$ base forwards with a shared prefix of K layers. Computing complementary coordinates for an existing document can require recomputing hidden states unless they were cached; reusing its stored final frozen vector does not by itself supply those hidden states. ANN search work is measured separately in Table 3.

G Implementation and Reproducibility

The main checkpoints use PyTorch and bf16. The domain/general, serving, and paired-control evaluations use the default SDPA backend; the separate cross-retrieval diagnostic explicitly uses FlashAttention-2. AdamW uses lr 2×10^{-4}, weight decay 0.01, 200-step warmup and cosine decay through 8000 steps. Split seed is 42; domain evaluation sampling seed is 0 and general sampling seed is 42, except the explicitly labeled cross-retrieval diagnostic. Final-step checkpoints are used for the main evaluation.

Contrastive implementation. The loss in Section 3.4 is directional. A positive similarity is appended to the in-batch logits, whose diagonal is also retained; hence its denominator contains two copies of the positive exponential. Only the appended logit is the cross-entropy target. Off-diagonal pairs with the same image ID are masked. This applies to the blended, standalone, and MRL losses of the reported CoSE checkpoints. The duplicated-positive objective has a different gradient from standard InfoNCE.

Training configurations. The main-table CoSE models use batch 64 and MRL weight 0.5 on HAN, and batch 32 without MRL on ROCO/SemArt/RSICD; all include orthogonality weight 0.1. HAN paired loss controls and ROCO paired MRL controls use orthogonality weight 0. For the ensemble and two-table comparisons, SemArt/RSICD LoRA seeds 42/123 use retrained replicas; these are distinct from the corresponding single-encoder baseline checkpoints in Table 1.

Serving protocol. Isolated H200 encoder measurements use fresh system instances, bf16/SDPA, no-grad evaluation, external query counts, 100 warmup queries, 300 batch-one queries, and three repeats. Text p50/p95 is 16.6/17.3ms for Frozen, 16.7/17.6ms for merged LoRA, and 27.9/29.1ms for CoSE. Image p50/p95 is 31.0/37.1, 29.8/33.3, and 42.5/52.2ms, respectively. CPU tokenization takes 3.7–4.0ms for text and 8.9–9.8ms for images. At text batch 64, QPS is 985 for Frozen and approximately 446 for CoSE (2.21 times cost); peak memory is 5.4GB for Frozen/merged LoRA and 6.6GB for CoSE. The ANN benchmark uses one CPU thread, 65k ROCO train images, 1000 test text queries, HNSW $M = 32$, efConstruction 200, and efSearch 64/64/32/64 for Frozen/LoRA/CoSE full/compact. Recall@10 against each representation's exact search is 0.953/0.982/0.956/0.981 and search p50 is 0.357/0.469/0.414/0.418ms. These are separate encoder and search measurements, not summed end-to-end quantiles.

Frozen-endpoint ANN diagnostic. Separately from Table 3, we build an inner-product HNSW graph over 65k ROCO documents stored as $[\mathbf{z}_{\text{old}}; \mathbf{z}_{\text{new}}] \in \mathbb{R}^{4096}$, with each branch L2-normalized and no subsequent concatenation normalization. This graph is queried with $[\mathbf{z}_{\text{old}}; 0]$ for 1000 text queries, without rebuilding it at $\alpha = 1$. A separately built legacy graph stores and queries only the 2048-dimensional $\mathbf{z}_{\text{old}}$. Both use $M = 32$, efConstruction 200, and efSearch 256. The mean query-wise set intersection $|I^{10}_{\text{expanded}}(q) \cap I^{10}_{\text{legacy}}(q)|/10$ is 0.984; this is set overlap, not the fraction of identical lists. Recall against exact frozen top-10 is 0.982 for expanded and 0.989 for legacy. The expanded graph is built specifically for this diagnostic, not reused from the $\alpha = 0.5$ cost benchmark, and no full-α cost or recall sweep is implied.

Conditional seed variation. ROCO/SemArt/RSICD training uses a fixed batch sampler (seed 42), while training seeds 0/42/123 vary LoRA initialization and torch RNG. Their sample std excludes data-order variation and does not estimate kernel nondeterminism from repeated runs. ROCO same-seed retraining comparisons moved midpoint R@1 by 0.5–0.8pp; these two pairs differ in a numerically tiny penalty and do not isolate the source of variation. The corresponding HAN data-order sensitivity and repeated-run distribution have not been established.

Numerical preservation. The score-preservation lemma is structural. The independent-instance 2B bf16/SDPA cosine diagnostic uses 200 images per domain and has sample variation of approximately 10^{-7}. Separately, the $K = 14$ shared-prefix ROCO checkpoint matches both the reference and independent Frozen image outputs bit-for-bit on 100 images. Neither observation is a universal FP32 claim or a guarantee of equal candidates from distinct ANN graphs.

G.1 Measured Training Memory

We measure training memory with the actual trainer in train mode, bf16, and gradient checkpointing enabled. GPU-wide peak usage from `nvidia-smi` includes reserved memory.

Configuration	Batch	Peak (GB)
HAN, paired reference (MRL + standalone)	64	78.9
HAN, without standalone or without MRL	64	70.0
ROCO, MRL on/off control	32	35.3

These GPU-wide usage peaks apply only to the listed configurations and are not allocator-only peaks. Smaller microbatches were not evaluated. Gradient accumulation does not reproduce a larger in-batch contrastive negative set.

H Supporting Compatibility Results

The main text gives the score definition and architectural guarantee. Here we record the supporting linear-algebra statements and proofs.

Theorem 1 (Same-channel exact compatibility is trivial under full-rank legacy indices)**.** *If $g : \mathcal{X} \to \mathbb{R}^d$ satisfies $Z_{\mathrm{old}} g(q) = Z_{\mathrm{old}} \mathbf{z}_{\mathrm{old}}(q)$ for every query q, then*

$$g(q) - \mathbf{z}_{\mathrm{old}}(q) \in \mathrm{Null}(Z_{\mathrm{old}}).$$

In particular, if $\mathrm{rank}(Z_{\mathrm{old}}) = d$, then $g(q) = \mathbf{z}_{\mathrm{old}}(q)$ for every q.

Proof. Subtracting the two score vectors gives $Z_{\mathrm{old}}(g(q) - \mathbf{z}_{\mathrm{old}}(q)) = 0$. Hence the offset lies in $\mathrm{Null}(Z_{\mathrm{old}})$. Full column rank gives $\mathrm{Null}(Z_{\mathrm{old}}) = \{0\}$. □

Theorem 2 (Direct-sum normal form for auditable exact compatibility)**.** *Let $F : \mathcal{X} \to \mathbb{R}^{d+m}$ be an expanded representation. Suppose there exists a fixed rank-d linear readout $P : \mathbb{R}^{d+m} \to \mathbb{R}^d$ such that*

$$PF(x) = \mathbf{z}_{\mathrm{old}}(x) \qquad \forall x \in \mathcal{X}.$$

Then, up to an invertible linear change of coordinates,

$$F(x) = \begin{bmatrix} \mathbf{z}_{\mathrm{old}}(x) \\ h(x) \end{bmatrix}$$

for some $h : \mathcal{X} \to \mathbb{R}^m$.

Proof. Since P has rank d, the rows of P can be completed to a basis of the dual space. Hence there exists $R : \mathbb{R}^{d+m} \to \mathbb{R}^m$ such that $T(v) = [Pv; Rv]$ is invertible. Applying T gives $TF(x) = [PF(x); RF(x)] = [\mathbf{z}_{\mathrm{old}}(x); h(x)]$ with $h(x) = RF(x)$. □

Proposition 1 (CoSE defines a positive-semidefinite retrieval kernel family)**.** *For every $\alpha \in [0, 1]$,*

$$k_\alpha(x, y) = \alpha \langle \mathbf{z}_{\mathrm{old}}(x), \mathbf{z}_{\mathrm{old}}(y) \rangle + (1 - \alpha) \langle \mathbf{z}_{\mathrm{new}}(x), \mathbf{z}_{\mathrm{new}}(y) \rangle$$

is positive semidefinite and equals the inner product induced by

$$\phi_\alpha(x) = \begin{bmatrix} \sqrt{\alpha}\, \mathbf{z}_{\mathrm{old}}(x) \\ \sqrt{1 - \alpha}\, \mathbf{z}_{\mathrm{new}}(x) \end{bmatrix}, \qquad k_\alpha(x, y) = \langle \phi_\alpha(x), \phi_\alpha(y) \rangle.$$

At $\alpha = 1$, k_1 is the frozen encoder kernel exactly.

Proof. Inner-product kernels are positive semidefinite, and nonnegative weighted sums preserve positive semidefiniteness. The representation by ϕ_α follows by direct expansion. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction distinguish exact legacy-score preservation at the frozen endpoint from domain quality and general retention at other weights. Table 1 and Sections 4.3–4.5 report operating points, deployment alternatives, serving cost, and single-seed controls. No unique dominance or safe mixed-coverage gain is claimed.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 covers encoding and storage costs, domain-quality gaps, partial-backfill bias, ANN rollback scope, single-seed scale/transfer evidence, and linear storage growth for concurrent branches. Dataset-release limitations are stated in the setup.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The main text states the score definition and index-preservation lemma with its architectural argument. Appendix H gives the supporting same-channel, direct-sum, and kernel statements and proofs with explicit assumptions. Score equality is distinguished from approximate-search agreement.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [No]

Justification: The manuscript describes the core architecture, objective, reference penalty setting, datasets, and evaluation rules. Checkpoint-to-objective mapping and diagnostic sampling differences are specified in Appendix B and the implementation section. Public reproducibility is limited by the absence of an accessible checkpoint package and by restrictions on the HAN extended captions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The paper does not provide a public training/evaluation-code or adapter-checkpoint access URL. Public benchmarks are cited; HAN extended captions are subject to provider review.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [No]

   Justification: The main text reports the reference schedule, validation selection, and primary samplers. The appendix specifies actual CoSE batch sizes and MRL usage, validation selection, diagnostic sampler differences, paired controls, and retrained baseline settings. The paper does not fully specify all baseline configurations or provide public reproduction access.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: The main domain table reports mean and sample standard deviation across three training seeds; the domain and general appendices use the same sample-standard-deviation convention. SemArt/RSICD LoRA general results are explicitly reported for seed 0 only. Loss controls are paired across three seeds. Non-HAN CoSE seed variation is conditional on fixed data order; observed same-seed retraining variation is reported separately. L2C-full and scale/backbone controls are labeled single-seed. Preservation budgets and grayscale bins are descriptive thresholds, not pairwise significance or equivalence tests.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The reference trials use an H200 with the schedule in the setup; an 8000-step 2B run takes approximately six hours. Appendix G.1 reports actual train-mode peaks of 78.9 GB for HAN reference batch 64 and 35.3 GB for ROCO batch 32. Isolated encoding benchmarks distinguish batch-one latency, throughput, and prefix sharing. Smaller-GPU trainability and complete project-compute accounting are not provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses publicly available pretrained models (Qwen3-VL-Embedding) and standard benchmarks; HAN extended is licensed via AIHub. No human-subjects

research, no scraped private data, and no models with foreseeable misuse risk beyond the standard image-text retrieval setting.

Guidelines:

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

10. **Broader impacts**

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

Answer: [No]

Justification: Exact frozen-score preservation can reduce behavioral uncertainty during retrieval migration, but domain adaptation still requires corpus backfill and compute. Retrieval systems can inherit biases and expose sensitive content; preserving frozen scores does not remove those risks. A dedicated discussion of societal impacts and mitigations is not provided.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: We release no new pretrained generative models or scraped datasets that pose misuse risk. CoSE adapters, when released, are domain-specific embedding modules that inherit the base encoder's risk profile unchanged.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [No]

    Justification: Creators and datasets are cited, including Qwen3-VL and SigLIP2. A complete per-asset license/version table and verification of all release terms are not included; HAN extended-caption release remains separately restricted.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not provide a public access path for new code or adapter checkpoints; HAN extended captions remain subject to provider review.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A]

Justification: No crowdsourcing or human-subjects research is conducted. All training and evaluation data come from existing public or licensed datasets.

Guidelines:

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

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

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

Answer: [N/A]

Justification: No human-subjects research is conducted in this work. The HAN dataset's underlying captions were collected by a third-party data provider under their own consent procedures; we use the released data under that platform's license.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not used as a methodologically important, original, or non-standard component of the core method. Per NeurIPS policy, declaration is not required for writing/editing assistance.

Guidelines:

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