COMPRESS REDUNDANCY, PREDICT WHAT IS MISSING: BEYOND GLOBAL POOLING FOR SCENE SKETCH RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Sketch-to-photo retrieval enables users to query images with free-hand drawings, but its central challenge is matching sparse strokes to visually rich photographs. Existing methods emphasize semantic alignment or local matching, yet pay limited attention to two limitations of single-vector aggregation: similar regions are repeatedly encoded, while spatial layout is lost during compression. We argue that learning a compact representation for discriminative retrieval requires consolidating repeated evidence while preserving the spatial ordering of regions. Accordingly, we encode sketches and photos with a shared pretrained DINOv3 ConvNeXt-B backbone and introduce redundancy-aware balanced semantic compression and masked spatial prediction to learn complementary global, regional, and spatial representations. Under identical protocols on FS-COCO, SketchyCOCO-SL, and QMUL-ShoeV2, our method improves upon the previous best results by 8.6%–32.7% in R@1, 5.9%–10.6% in R@5, and 1.9%–6.3% in R@10 across scene-level and fine-grained object retrieval. This provides a new direction for compact representation learning in sketch-driven cross-modal image retrieval, jointly suppressing redundancy while preserving spatial structure.

1 INTRODUCTION

A free-hand sketch can specify a target's silhouette, local details, and spatial layout with only a few strokes, moving image retrieval from recognizing *which category* to matching *which instance*, and extending the search to complete scenes (Yu et al., 2016; Liu et al., 2020). This direct query form also introduces fundamental difficulty: objects may be omitted, strokes vary markedly with drawing style and abstraction level, and sparse lines must be matched against visually rich photographs (Sain et al., 2021; Chowdhury et al., 2022a; Koley et al., 2024). To bridge this gap, prior studies have leveraged pretrained vision–language models for semantic transfer, established interpretable local correspondences, and explicitly modeled spatial relations among scene objects (Sain et al., 2023a; Lin et al., 2023; Liu et al., 2020; Chowdhury et al., 2022a). More recent work instead emphasizes strong pretraining, shared encoders, and robust objectives for ambiguous and noisy sketches (Demić & Zajc, 2025). These efforts focus mainly on extracting and matching more evidence, however, while paying little attention to how limited representational capacity should be allocated across regions: repeated content is encoded again and again, crowding out scarce capacity, and naive aggregation can erode the spatial structure required to tell one object or scene from another.

This picture echoes efficient coding in biological vision. Early studies found that natural visual signals carry largely predictable redundancy, and that discriminative information concentrates at few structurally salient locations. Attneave (1954) illustrated that connecting a small set of high-curvature contour points can preserve recognizable shape despite substantial removal of contour detail; Barlow (1961) proposed redundancy reduction as a fundamental principle of sensory coding, whereby sensory pathways recode statistically dependent inputs into more efficient representations. Biological vision does not merely discard redundancy. Predictive coding (Rao & Ballard, 1999) shows that high-level representations use context to predict lower-level activity, leaving feedforward pathways to carry mostly unpredicted residuals. Efficient perception thus involves two complementary processes: compressing recurring evidence, and inferring poorly observed evidence from context (Attneave, 1954; Barlow, 1961; Rao & Ballard, 1999).

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

Scene-sketch retrieval faces an analogous problem. Skies, roads, building textures, and recurring objects elicit many semantically similar region features in photographs; if global aggregation encodes them equally, they consume finite capacity over and over and dilute a few discriminative regions. A sketch, in turn, is itself a highly compressed and incomplete expression: much appearance detail is deliberately omitted, yet object positions, contours, and the organization among regions survive. We therefore adopt a view different from aggregating ever more local features: an effective cross-modal representation should reduce repeated encoding of redundant visual evidence while learning to recover discriminative spatial structure from incomplete regional observations. This view motivates both redundancy-aware balanced semantic compression and masked spatial prediction in this paper.

Turning this view into a compact representation learning method for cross-modal retrieval requires addressing two interdependent technical challenges. First, redundancy does not imply unimportance. Two regions with similar responses need not be interchangeable: they may jointly describe one object or differ only in subtle structures that separate instances. The task is thus not selecting the most salient regions but estimating relative redundancy among local cues so that high-frequency patterns do not dominate during compression. Allocation must also stay balanced: when a few regions keep receiving excessive weight, aggregation degenerates into another form of collapse. We therefore need a mechanism that attenuates repeated contributions without sacrificing complementary evidence.

The second challenge is subtler: spatial structure cannot be preserved by storing coordinates. Local appearance is unreliable across modalities—the same entity may evoke very different responses in sketch and photo—so adding positional encodings merely records *where* features occur without explaining why regions co-constitute a scene. We instead seek predictive spatial representations: when part of the visual evidence is missing, the remaining regions and their relative positions should suffice to constrain the semantics of hidden ones. Spatial relations then cease to be passively stored attributes and become explicitly shaped through recovering missing semantics from incomplete observations.

Our contributions are threefold:

- Existing scene-sketch retrieval methods center on cross-modal semantic alignment and rarely consider how visual evidence is organized within a compact representation. Inspired by efficient coding and prediction in biological vision, we learn cross-modal representations from two complementary angles: compressing redundant evidence and inferring missing evidence.
- We instantiate this view as a sketch–photo retrieval framework built on a shared pretrained DINOv3 ConvNeXt-B encoder, with redundancy-aware balanced semantic compression and masked spatial prediction that model complementary semantics and spatial structure respectively, producing compact and discriminative retrieval representations.
- Systematic experiments on public benchmarks under multiple protocols show consistent gains at both scene level and fine-grained instance level, validating the effectiveness of jointly suppressing redundancy while preserving spatial structure when learning compact cross-modal retrieval representations.

2 Related Work

2.1 Sketch-Based Image Retrieval

SBIR has evolved from category-level matching toward instance-level fine-grained retrieval. Early FG-SBIR relied on metric learning to build a shared embedding; later work turned to drawing style, local correspondence, and abstraction level (Yu et al., 2016; Sain et al., 2021; Lin et al., 2023; Koley et al., 2024). Lin et al. (2023) grounded cross-modal matching in interpretable patch correspondences, showing that instance discrimination depends on local evidence beyond global semantics; Koley et al. (2024) showed that the granularity a sketch demands varies with its abstraction. Research has accordingly shifted from bridging modality gaps toward preserving what truly separates instances.

Scene-level SBIR extends the task to multiple objects and their spatial organization. Liu et al. (2020) conducted the first systematic study of scene-level FG-SBIR, jointly describing appearance,

scale, and layout with scene structures; Chowdhury et al. (2022a) observed that free-hand scene sketches are naturally partial and brought in region-level optimal transport with scene adjacency, so retrieval leans on visible partial structure rather than complete observation; FS-COCO moved scene-level research to real free-hand drawings and analyzed object- and scene-level abstraction together with scene salience encoded by stroke order (Chowdhury et al., 2022b); SceneTrilogy studied complementarity among sketches, photos, and text (Chowdhury et al., 2023); recently, Demić & Zajc (2025) revisited scene-level SBIR and showed strong pretraining, a shared encoder, and training designs robust to ambiguity and noise substantially improve performance. These works established local correspondence, spatial relations, and partial observation as central. Rather than introducing another local matching mechanism, we optimize the formation of the fixed-length representation itself: limiting the influence of repeated regions while strengthening spatial structure through missing-semantic prediction.

2.2 Compact Retrieval Representations and Feature Aggregation

Compressing many local observations into one discriminative vector is classical in image retrieval. Classical aggregation work revealed a lasting phenomenon: frequency of local evidence does not imply discriminative value. Democratic Aggregation modulates each local representation's contribution to the final similarity so that recurrent patterns cannot dominate; Generalized Max Pooling showed sum-like aggregation over-amplifies frequent but low-information local representations and weakens rare informative ones (Jégou & Zisserman, 2014; Murray & Perronnin, 2014). Aggregation, in short, is not counting evidence but deciding its relative influence in a finite representation. Deep retrieval folded this idea into end-to-end learning: NetVLAD made residual aggregation toward visual centers learnable (Arandjelović et al., 2016); DOLG stressed complementarity between global and local components for single-vector retrieval (Yang et al., 2021). As large-scale self-supervised models yield denser, more semantic local features, aggregation has taken new forms: SALAD recasts the assignment of local features to cluster centers as Sinkhorn optimal transport, with a dustbin letting uninformative features opt out (Izquierdo & Civera, 2024). Together these show that compact-representation quality depends not only on backbone strength but on how local evidence shares limited capacity. Prior aggregation methods primarily balance contributions among local representations within a single image, whereas scene-sketch retrieval also faces an evidence-density asymmetry between sketches and photos. We therefore incorporate inter-region semantic redundancy into the capacity allocation of the compact cross-modal retrieval representation, limiting the influence of repeated evidence while preserving discriminative complementary information.

2.3 Predictive Representation Learning

Inferring unseen content from partial observation is another major line of representation learning. Masked image modeling hides part of the input and restores it, forcing broader use of context. I-JEPA formulates masked prediction in latent space, predicting high-level representations of hidden regions from visible context rather than reconstructing pixels (Assran et al., 2023); Siamese Image Modeling conditions dense feature prediction on relative positions and scale changes between augmented views, strengthening spatial sensitivity while maintaining semantic alignment (Tao et al., 2023). This suits free-hand scenes naturally: omitted regions admit no unique pixel answer, yet their semantic roles are constrained by the remaining objects and their relations. We do not reconstruct masked appearance; we treat the high-level semantics of missing regions as supervision, constraining prediction with compact spatial memory, visible semantics, and their relative geometry. Incompleteness thereby plays a role beyond traditional partial SBIR: instead of adapting retrieval to incomplete queries, we use constructed incomplete observations to shape the representation itself, turning an input defect into an effective supervisory signal for spatial structure.

3 Method

DINOv3 is a large-scale self-supervised vision model that learns transferable generic features without manual labels (Siméoni et al., 2025). Unlike pretraining centered on image-level semantics, DINOv3 emphasizes dense-feature quality over long-horizon training, so a single backbone supplies stable global semantics alongside spatially resolved local representations. We adopt DINOv3 ConvNeXt-B. ConvNeXt preserves hierarchical spatial structure while absorbing effective princi-

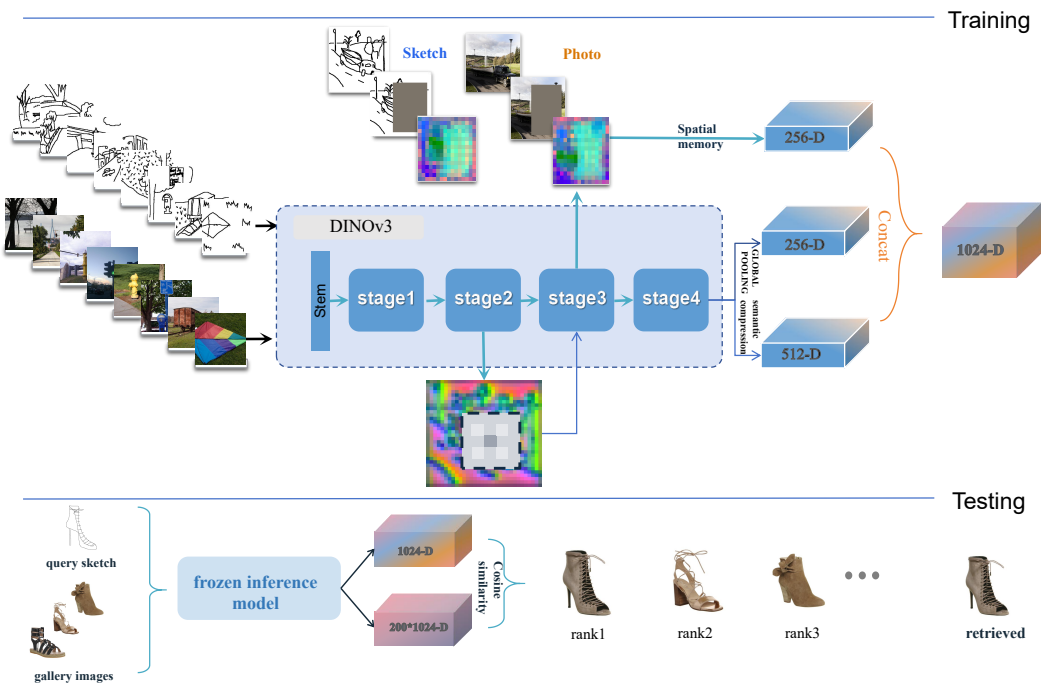

Figure 1: **Overview of the proposed framework.** Training proceeds in two stages with a shared DINOv3 ConvNeXt-B encoder for paired sketches and photos. The first stage combines global semantic alignment, masked-context prediction, and redundancy-aware balanced semantic compression to learn global and complementary semantic representations. With the backbone then frozen, the second stage learns a spatial memory from intermediate features and regularizes it through masked-region prediction. At inference, the encoder, semantic compressor, and spatial-memory projection produce three normalized components, which are concatenated into the retrieval representation used to rank a precomputed gallery by cosine similarity.

ples of modern vision transformers (Liu et al., 2022b); at Base scale its four stages carry 128, 256, 512, and 1024 channels at progressively lower resolutions. This hierarchy matters here: deeper stages provide stable cross-modal semantics while intermediate stages retain spatial detail, allowing semantic compression and spatial relation modeling to rest on one encoder. Figure 1 summarizes the complete training and retrieval pipeline.

3.1 Debiased and Predictive Cross-Modal Alignment

Instance-level contrastive learning typically assumes one exact match per query and pushes every other sample away (van den Oord et al., 2018). Scene-sketch retrieval violates this assumption: distinct scenes share objects, layouts, or contextual semantics, forming partially relevant hard negatives (Demić & Zajc, 2025). We adopt the debiased I-Con objective (Alshammari et al., 2025), replacing strict one-versus-all repulsion with a soft supervision distribution that keeps paired samples dominant while reducing false repulsion of semantically overlapping negatives.

Given a batch of B sketch–photo pairs $\{(S_i, P_i)\}_{i=1}^{B}$ with normalized representations z_i^s and z_i^p, the sketch-to-photo matching distribution is

$$q_{ij} = \frac{\exp(\langle z_i^s, z_j^p \rangle / \tau)}{\sum_{k=1}^{B} \exp(\langle z_i^s, z_k^p \rangle / \tau)}, \tag{1}$$

with I-Con supervision

$$p_{ij} = (1 - \alpha)\, \mathbf{1}[i = j] + \frac{\alpha}{B}, \tag{2}$$

where τ is the temperature and α the debiasing strength. The clean-view objective is

$$\mathcal{L}_{\text{clean}} = \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}(p_i \,\|\, q_i). \tag{3}$$

We set $\tau = 0.07$ and $\alpha = 0.20$, following validated scene-level settings (Demić & Zajc, 2025).

We further require cross-modal correspondence to remain recoverable when local evidence is perturbed. Specifically, we mask a contiguous region of the Stage 2 sketch features and let a light predictor produce $\hat{z}_i^s$ from the remaining context, with the complete photo representation as a stop-gradient target:

$$q_{ij}^{\text{pred}} = \frac{\exp(\langle \hat{z}_i^s, \text{sg}(z_j^p)\rangle/\tau)}{\sum_{k=1}^{B}\exp(\langle \hat{z}_i^s, \text{sg}(z_k^p)\rangle/\tau)}, \qquad \mathcal{L}_{\text{pred}} = \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}(p_i \,\|\, q_i^{\text{pred}}). \tag{4}$$

The final objective is

$$\mathcal{L}_{\text{main}} = \tfrac{1}{2}\mathcal{L}_{\text{clean}} + \tfrac{1}{2}\mathcal{L}_{\text{pred}}. \tag{5}$$

The former stabilizes neighborhoods under complete observation; the latter writes predictability under partial evidence into the same representation space.

3.2 Redundancy-Aware Balanced Semantic Compression

Stage 4 produces semantically rich local tokens. The key is not finding the most salient regions but avoiding double counting: approximate content appearing at several positions should not grow proportionally with its occurrences. We decouple compression into how much capacity each token deserves and where its information is written.

Given T tokens $\{x_i\}_{i=1}^{T}$ with normalized semantic features k_i, we estimate token i's redundancy and capacity mass as

$$d_i = 1 + \sum_{j\neq i}\sigma\left(\frac{k_i^\top k_j - \rho}{\beta}\right), \qquad m_i = \frac{T d_i^{-\gamma}}{\sum_{j=1}^{T} d_j^{-\gamma}}, \quad \sum_{i=1}^{T} m_i = T. \tag{6}$$

Here, ρ thresholds the similarity counted as duplicate, β is a smoothing temperature, and a larger d_i indicates semantics co-expressed by more regions. Thus, repeated evidence shares limited shares rather than being deleted, while distinctive evidence gains larger capacity. Compared with hard selection, this retains complete scene semantics while preventing visual frequency from converting directly into representational weight.

S learnable semantic slots then organize the evidence (Locatello et al., 2020). Token–slot similarities decide placement, and Sinkhorn balancing (Sinkhorn, 1964; Cuturi, 2013; Caron et al., 2020) yields the soft assignment A and slot representation

$$\sum_{s=1}^{S} A_{si} = m_i, \qquad \sum_{i=1}^{T} A_{si} = \frac{T}{S}, \qquad h_s = \frac{\sum_i A_{si} v_i}{\sum_i A_{si}}, \qquad c = [c_1; \ldots; c_S]. \tag{7}$$

The two marginals control two separate things: how much one piece of evidence contributes in total, and that every slot receives comparable capacity. Redundant regions can no longer occupy excess space, nor can the representation collapse onto a few dominant semantics. Here, v_i denotes the value representation, and a low-dimensional projection maps h_s to c_s; their ordered concatenation forms the complementary semantic representation. We use $S = 32$ slots of 16 dimensions each, i.e., a 512-dimensional output. We set $\rho = 0.70$, $\beta = 0.08$, $\gamma = 1$, and use eight Sinkhorn iterations. In short, redundancy decides how much capacity evidence deserves; balanced slotting decides how that capacity is organized—suppressing double counting while preserving complementary structure.

3.3 Spatial Representation Learning from Incomplete Evidence

Beyond compact Stage 4 semantics, Stage 3 features retain finer spatial structure. We therefore build an independent spatial branch on Stage 3 features and constrain it with predictive learning, while keeping the trained shared encoder and semantic compressor frozen.

Given $F_3(I) \in \mathbb{R}^{512\times16\times16}$, light projection and spatial downsampling followed by fixed-order flattening give the spatial memory $m(I) = H(\mathrm{vec}(\tilde{F}_3(I))) \in \mathbb{R}^{256}$. To let spatial memory exploit contextual relations, we introduce region prediction. For both sketch and photo, we mask the same contiguous input-image region R, aligned with the 16×16 Stage 3 grid, and pick an anchor region a among the visible cells. Passing each masked input through the frozen encoder gives a masked memory $\tilde{m}^u$ and anchor feature f_a^u for modality $u \in \{s, p\}$. Conditioned on these features and their shared relative geometry $r = [\Delta x, \Delta y, w_R, h_R]$, where Δx and Δy are the anchor-to-center displacements of the hidden region normalized by the 16×16 grid extent and w_R, h_R are its corresponding normalized width and height, a predictor outputs $\hat{t}_R^u = G(\tilde{m}^u, f_a^u, r)$, the latent semantics of the hidden region. Unlike pixel reconstruction we predict high-level features, consistent with latent-space prediction (Assran et al., 2023). To keep the coarse layout inside the target, its features split into four ordered subregions extracted from the corresponding complete Stage 3 map. Concretely, each $t_R^{u,\mathrm{q}} \in \mathbb{R}^{512}$ for $\mathrm{q} \in \{\mathrm{TL}, \mathrm{TR}, \mathrm{BL}, \mathrm{BR}\}$ is the mean-pooled Stage 3 feature of a non-overlapping quadrant; these abbreviations denote the top-left, top-right, bottom-left, and bottom-right quadrants, listed in raster order. The normalized concatenation is therefore a 2048-dimensional target:

$$\begin{aligned} t_R^u &= [\, t_R^{u,\mathrm{TL}}; t_R^{u,\mathrm{TR}}; t_R^{u,\mathrm{BL}}; t_R^{u,\mathrm{BR}} \,], \\ \mathcal{L}_{\mathrm{sp}}^u &= 1 - \cos(\hat{t}_R^u, t_R^u), \qquad \mathcal{L}_{\mathrm{sp}} = \tfrac{1}{2}(\mathcal{L}_{\mathrm{sp}}^s + \mathcal{L}_{\mathrm{sp}}^p). \end{aligned} \tag{8}$$

We apply this prediction symmetrically to sketches and photos. Sketch and photo spatial memories are additionally aligned by the I-Con objective of Section 3.1, giving

$$\mathcal{L}_{\mathrm{spatial}} = \mathcal{L}_{\mathrm{ICon}}^{\mathrm{sp}} + \lambda\, \mathcal{L}_{\mathrm{sp}}, \qquad \lambda = 0.2. \tag{9}$$

3.4 Unified Retrieval Representation

The three components yield a global semantic representation $g(I)$, the complementary semantic representation $c(I)$, and the spatial memory $m(I)$. Their dimensions are 256, 512, and 256, respectively. After separate normalization we concatenate them with equal block weights and rank a sketch query S against a candidate photo P by

$$\begin{aligned} f(I) &= [\sqrt{w_g}\, \bar{g}(I); \sqrt{w_c}\, \bar{c}(I); \sqrt{w_m}\, \bar{m}(I)], \qquad w_g = w_c = w_m = \tfrac{1}{3}, \\ z(I) &= \mathrm{Norm}(f(I)), \qquad \mathrm{sim}(S, P) = z(S)^\top z(P). \end{aligned} \tag{10}$$

This yields a 1024-dimensional retrieval representation. All experiments retrieve with this unified representation and introduce no query-specific optimization.

4 Experiments

4.1 Experimental Setup

We evaluate on three representative benchmarks covering scene-level and fine-grained instance-level settings.

4.1.1 Datasets

FS-COCO (Chowdhury et al., 2022b) targets scene-level understanding of free-hand sketches, pairing sketches with corresponding photos and scene semantics. Its official Normal and Unseen protocols each contain 7000 training and 3000 test pairs; we additionally construct a user-disjoint Unseen split with the same 7000/3000 sizes. Its complex natural scenes probe retrieval under large domain gaps and multi-object environments.

SketchyCOCO-SL (Gao et al., 2020; Liu et al., 2020) evaluates scene-level fine-grained sketch-to-photo retrieval; following prior protocols we use the 1015/210 train/test split for direct comparison, and additionally report the full SketchyCOCO setting (11265/2816) to test effectiveness at full data scale.

QMUL-ShoeV2 (Yu et al., 2016) is a classic object-level benchmark of shoe sketches and product photos for instance-level matching. Following the ARNet directory split (Jiang et al., 2025), it

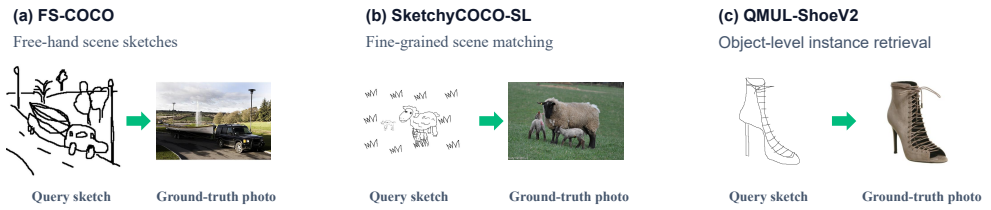

Figure 2: **Benchmark protocols and real paired examples.** FS-COCO, SketchyCOCO-SL, and QMUL-ShoeV2 pair free-hand sketches with ground-truth photos under scene- and object-level retrieval protocols.

contains 5982 training pairs and 666 test sketches; each test sketch is ranked against 200 unique photos.

Together they span complex scenes, fine-grained scene matching, and object-level instance retrieval, testing whether the learned sketch–photo representation preserves cross-modal semantics, instance identity, and spatial layout across abstraction levels and domain gaps. Figure 2 illustrates representative paired queries from these protocols.

4.1.2 METRICS

Following prior sketch-retrieval studies (Yu et al., 2016; Chowdhury et al., 2022b), we report Recall@K (R@K), which checks whether the ground-truth photo appears among the top K results.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1: Comparison with state-of-the-art methods on FS-COCO.

	Normal			Unseen		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Siam-VGG (ECCV'22)	23.3	–	52.6	10.6	–	32.5
SceneTrilogy (CVPR'23)	24.1	–	53.9	–	–	–
SceneDiff (IJCAI'24)	25.2	45.9	55.9	–	–	–
FreestyleRet (ECCV'24)	29.6	–	56.1	–	–	–
Back to the Drawing Board (BMVC'25)	61.9	81.4	87.2	60.0	80.2	86.1
Ours	**67.2**	**86.4**	**91.1**	**67.7**	**86.5**	**91.5**

(a) FS-COCO.

Table 2: Comparison with state-of-the-art methods on SketchyCOCO-SL and QMUL-ShoeV2.

	SketchyCOCO SL				ShoeV2		
Method	R@1	R@5	R@10	Method	R@1	R@5	R@10
SceneDiff (IJCAI'24)	34.3	69.1	81.4	StyleMeUp (CVPR'21)	36.5	68.1	81.8
Partially Does It (CVPR'22)	34.5	–	89.3	Part-SBIR (CVPR'22)	39.9	68.2	82.9
SceneTrilogy (CVPR'23)	38.2	–	87.6	ARNet (AAAI'25)	42.9	73.0	81.7
SceneSketcher-v2 (TIP'22)	68.1	87.6	95.2	NT-SBIR (CVPR'22)	43.7	74.9	–
Back to the Drawing Board (BMVC'25)	70.0	90.0	96.2	EUPS-SBIR (CVPR'23)	44.2	70.8	84.7
Ours	**92.9**	**99.5**	**100.0**	**Ours**	**48.8**	**79.3**	**86.3**

(b) SketchyCOCO-SL and QMUL-ShoeV2.

Notes: Siam-VGG (Chowdhury et al., 2022b); SceneTrilogy (Chowdhury et al., 2023); SceneDiff (Zuo et al., 2024); FreestyleRet (Li et al., 2024); Back to the Drawing Board (Demić & Zajc, 2025); Partially Does It / Part-SBIR (Chowdhury et al., 2022a); SceneSketcher-v2 (Liu et al., 2022a); StyleMeUp (Sain et al., 2021); ARNet (Jiang et al., 2025); NT-SBIR (Bhunia et al., 2022); EUPS-SBIR (Sain et al., 2023b).

FS-COCO's official unseen split does not strictly isolate authors. For a stricter cross-author test, we construct a user-disjoint unseen split (7000/3000) whose training and test sketches come from completely disjoint

authors, verified before training. Under an identical fair protocol (epochs/seed/batch), retraining Back to the Drawing Board (Demić & Zajc, 2025) from official code yields R@1/R@5/R@10 = 59.13/78.40/84.87 versus our 66.07/84.60/89.73; on full SketchyCOCO (11265/2816), we reach 56.50/75.00/81.82 against 47.41/68.18/76.92 for the same fairly retrained baseline.

Table 1 summarizes quantitative comparisons under each dataset's own protocol: our method attains the best reported score for every R@1/R@5/R@10 metric. Relative to the strongest prior score per metric, gains range from 3.8–22.9 points on scene-level benchmarks and are 4.6/4.4/1.6 points at QMUL-ShoeV2 R@1/R@5/R@10. Those leaders rely on robust CLIP priors (Demić & Zajc, 2025), unlabelled-photo distillation (Sain et al., 2023b), or noisy-stroke filtering (Bhunia et al., 2022); ours couples semantic alignment with dense spatial memory using paired data alone.

4.3 Ablation and Representation Analysis

On QMUL-ShoeV2 we verify each design choice along the progressive path of Table 3.

Table 3: Ablation study on QMUL-ShoeV2.

Config	Design	R@1	R@5	R@10
E-1	Frozen DINOv3 embedding (zero-shot)	6.16	21.32	32.58
E0	Global representation with only $\mathcal{L}_{\text{clean}}$	38.74	72.07	81.23
E0_{1024}	Capacity-matched global representation	38.74	70.12	81.08
E1	+ $\mathcal{L}_{\text{pred}}$	38.44	73.57	83.18
E2	+ balanced slots ($\gamma = 0$)	43.09	75.38	83.93
E3	+ redundancy-aware mass ($\gamma = 1$)	44.14	76.73	85.74
E4	+ spatial memory $m(I)$ without $\mathcal{L}_{\text{sp}}$	46.40	78.83	85.74
E5	Full model with $\mathcal{L}_{\text{sp}}$	48.80	79.28	86.34

- **Paired alignment, not backbone strength, unlocks retrieval.** Frozen DINOv3 embeddings transfer poorly to sketch–photo retrieval, yielding only 6.16 R@1. After clean paired alignment, however, the same backbone reaches 38.74 R@1 (E0). The contrast indicates that the pretrained representation contains transferable semantics, but its cross-modal geometry must be reshaped before it supports direct sketch-to-photo matching. Descriptor width is not the limiting factor: E0_{1024} uses the full-model width yet leaves R@1 unchanged and slightly reduces R@5 and R@10. Thus, the decisive change is supervised pairing rather than backbone capacity.
- **Semantic compression delivers the largest single gain.** Balanced slot assignment ($\gamma = 0$, E2) lifts R@1 by 4.65, the largest step along the path, and redundancy-aware mass ($\gamma = 1$, E3) adds 1.05 on top of already balanced capacity. Adding $\mathcal{L}_{\text{pred}}$ alone (E1) leaves R@1 flat (38.44), with gains confined to R@5/R@10: the top-match bottleneck lies in how representational capacity is organized, not in the training objective itself. The capacity-matched global pooling baseline (E0_{1024}) confirms this interpretation: merely widening its descriptor to the full model's 1024 dimensions leaves R@1 unchanged and slightly lowers R@5/R@10. Figure 3 provides a concrete example: compression moves the ground-truth sneaker from rank 77 to rank 3.
- **The spatial branch adds comparable magnitude along an orthogonal axis.** Spatial memory plus $\mathcal{L}_{\text{sp}}$ jointly add 4.7 (E3→E5), matching the compression side (E0→E3, about 5.4); the spatial memory alone already brings 2.26 through cross-modal alignment (E4)—pooled, slot-compressed Stage 4 semantics lose explicit layout, whereas $m(I)$ keeps local structure at higher resolution (16×16) in fixed raster order. Geometry-conditioned region prediction ($\mathcal{L}_{\text{sp}}$, E5) then supervises inter-region relations, and its 2.40 concentrates almost entirely on R@1 (R@5 only +0.45): it resolves top-instance ambiguity rather than broadening recall. In Figure 3, it similarly promotes the ground-truth wedge boot from rank 9 to rank 3, while the E5 projection visualizes correctly paired sketches and photos within shared local neighborhoods. Comparable in magnitude and orthogonal in kind, the two paths jointly account for the full-model gain.

Low-rank interventions reveal asymmetric branch reliance. At 25% retained rank, truncating the global and slot projections lowers fused R@1 by only 0.75–2.25 points, whereas truncating the spatial readout costs 4.81 (Table 4). Because truncated SVD preserves the dominant singular directions, this contrast indicates that fusion absorbs semantic perturbations more readily than perturba-

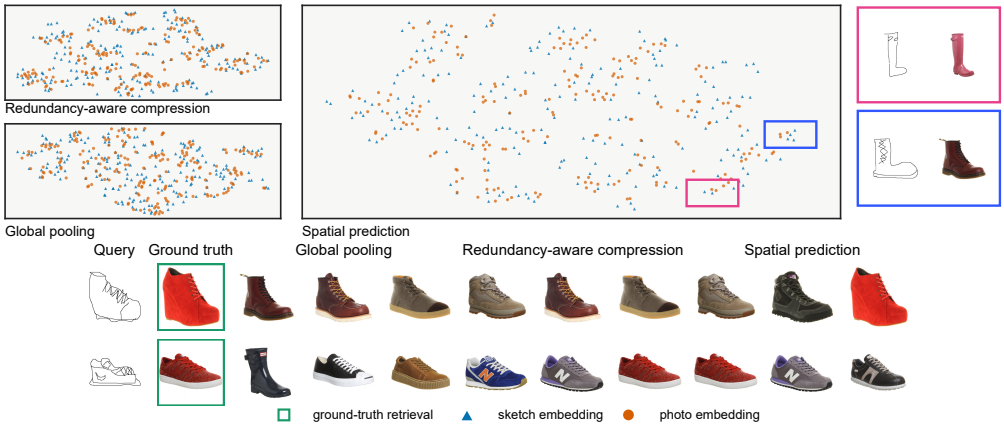

Figure 3: Qualitative analysis on QMUL-ShoeV2. Top: Top-3 retrievals from global pooling (E0) through redundancy-aware compression (E3) to spatial prediction (E5); green borders mark ground truth. Bottom: t-SNE projections of the same fixed sketch–photo pairs, with the enlarged E5 view linking two paired local neighborhoods to their examples.

tions to the ordered geometric evidence carried by spatial memory. Compression therefore identifies what distinguishes an instance, while spatial memory preserves where that evidence occurs.

Table 4: Effective-rank interventions on QMUL-ShoeV2.

Intervened projection	Rank (full $\rightarrow$ 25%)	ΔR@1
Global readout	$256 \rightarrow 64$	-0.75
Slot key projection	$64 \rightarrow 16$	-2.25
Slot content projection	$16 \rightarrow 4$	-1.20
Spatial readout	$256 \rightarrow 64$	$\mathbf{-4.81}$

Training-free transfer. Applying the FS-COCO model directly to SketchyCOCO-SL, without target-domain training, yields R@1/R@5/R@10=47.62/73.81/80.95. Its R@1 exceeds the reported scores of SceneDiff (34.3), Partially Does It (34.5), and SceneTrilogy (38.2) (Zuo et al., 2024; Chowdhury et al., 2022a; 2023), although it remains below target-trained leaders. The same model transfers poorly to fine-grained ShoeV2 (6.16/19.52/32.13). This contrast therefore tracks source–target data proximity: FS-COCO is much closer to SketchyCOCO-SL in scene composition and visual vocabulary than to the fine-grained shoe domain, so the stronger transfer reflects the smaller distribution shift between training and SketchyCOCO-SL rather than an inherent ability to match arbitrary object categories. More broadly, training the same architecture on larger-scale data may further improve its generalization.

5 Conclusion and Future Work

This paper addressed capacity allocation in compact cross-modal representations: repeated evidence should not claim shares by numerical advantage, and spatial layout should not vanish under aggregation and compression. Using deliberately constructed incomplete observations as supervision, our representation suppresses redundant encoding while preserving ordered structure across regions, freeing sketch–photo matching from the information-allocation limits of global pooling. The resulting unified representation leads on scene-level and fine-grained benchmarks while training only on paired sketch–photo data. Future work includes applying this capacity-allocation principle beyond global pooling, for example to other compact cross-modal architectures, and exploring retrieval and reconstruction from 2D sketches to 3D representations, as well as non-retrieval tasks such as classification and generation.

AI USE STATEMENT

In this work, the authors conceived and designed the core algorithm. Generative AI tools were used only to assist with implementing parts of the code, polishing the manuscript, and checking grammar and language. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide an anonymized reproduction archive containing the training and evaluation code, exact YAML configurations, dataset-split construction and audit utilities, resolved configurations, training logs, and metric files. Appendix A records the common settings and all dataset-specific deviations. Formal runs do not use an internal validation split or select checkpoints on the test set; each reported result is computed from the final checkpoint under the prescribed evaluation protocol.

REFERENCES

Shaden Naif Alshammari, John R. Hershey, Axel Feldmann, William T. Freeman, and Mark Hamilton. I-Con: A unifying framework for representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomáš Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3): 183–193, 1954.

Horace B. Barlow. Possible principles underlying the transformations of sensory messages. In Walter A. Rosenblith (ed.), *Sensory Communication*, pp. 217–234. MIT Press, 1961.

Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 999–1008, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9659–9670, 2020.

Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially does it: Towards scene-level FG-SBIR with partial input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2395–2405, 2022a.

Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. FS-COCO: Towards understanding of freehand sketches of common objects in context. In *Computer Vision – ECCV 2022*, pp. 253–270, 2022b.

Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. SceneTrilogy: On human scene-sketch and its complementarity with photo and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10972–10983, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Emil Demić and Luka Čehovin Zajc. Back to the drawing board: Rethinking scene-level sketch-based image retrieval. In *Proceedings of the British Machine Vision Conference*, 2025.

Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. SketchyCOCO: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5174–5183, 2020.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17658–17668, 2024.

Hervé Jégou and Andrew Zisserman. Triangulation embedding and democratic aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3310–3317, 2014.

Jianan Jiang, Hao Tang, Zhilin Jiang, Weiren Yu, and Di Wu. ARNet: Self-supervised FG-SBIR with unified sample feature alignment and multi-scale token recycling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3985–3993, 2025. doi: 10.1609/aaai.v39i4.32417.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to handle sketch-abstraction in sketch-based image retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16859–16869, 2024.

Hao Li, Yanhao Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. FreestyleRet: Retrieving images from style-diversified queries. In *Computer Vision – ECCV 2024*, pp. 258–274, 2024.

Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23349–23358, 2023.

Fang Liu, Changqing Zou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. SceneSketcher: Fine-grained image retrieval with scene sketches. In *Computer Vision – ECCV 2020*, pp. 718–734, 2020.

Fang Liu, Xiaoming Deng, Changqing Zou, Yu-Kun Lai, Keqi Chen, Ran Zuo, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. SceneSketcher-v2: Fine-grained scene-level sketch-based image retrieval using adaptive GCNs. *IEEE Transactions on Image Processing*, 31:3737–3751, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

Naila Murray and Florent Perronnin. Generalized max pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2473–2480, 2014.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. StyleMeUp: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8504–8513, 2021.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2765–2775, 2023a.

Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained SBIR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6873–6883, 2023b.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2132–2141, 2023.

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

Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11772–11781, 2021.

Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 799–807, 2016.

Ran Zuo, Haoxiang Hu, Xiaoming Deng, Cangjun Gao, Zhengming Zhang, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. SceneDiff: Generative scene-level image retrieval with text and sketch using diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1825–1833, 2024.

A Reproducibility Details

A.1 Software, Backbone, and Input Processing

The implementation is written in Python using PyTorch (version 2.3 or later), torchvision (0.18 or later), timm (1.0.20 or later), Pillow, PyYAML, NumPy, OpenCV, and tqdm. All reported experiments were run on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. We instantiate the pretrained DINOv3 ConvNeXt-B model distributed through timm as `convnext_base.dinov3_lvd1689m`. A single copy of its stem and four stages is shared by sketches and photos. Training uses automatic mixed precision, selecting bfloat16 when supported and float16 otherwise.

All inputs are converted to RGB, resized directly to 256×256 with bicubic interpolation and antialiasing, and normalized with the mean and standard deviation in the pretrained timm configuration. We resize the complete image rather than crop it so that no scene content is removed before feature extraction. At training time, paired sketches and photos receive the same weak affine transform with probability 0.5: rotation within $\pm 7^{\circ}$, isotropic scaling in $[0.97, 1.03]$, translation within 0.02 of the image extent, and an optional horizontal flip. Evaluation uses only the deterministic resize and normalization.

A.2 MAIN-REPRESENTATION TRAINING

The predictive view independently masks each modality on an 8×8 logical grid. We sample a left-, right-, top-, or bottom-connected macro region containing 25–50% of the grid cells, expand it by nearest-neighbor assignment to the native 32×32 Stage 2 feature map, and set the selected Stage 2 cells to zero. Thus the designated evidence is disrupted at Stage 2 rather than erased in input pixels. The clean branch uses the complete Stage 2 map. The cross-modal token predictor has two transformer blocks, eight attention heads, and an MLP ratio of four. Clean and predictive I-Con losses have equal weight; both use temperature $\tau = 0.07$ and debiasing $\alpha = 0.20$.

The global projection has 256 dimensions. The balanced compressor uses 32 ordered slots, a 64-dimensional assignment latent, a slot temperature of 0.10, and a 16-dimensional output per slot, producing a 512-dimensional complementary representation. Redundancy mass uses $\rho = 0.70$, $\beta = 0.08$, $\gamma = 1$, and eight Sinkhorn iterations. During main-model training, the normalized global and complementary blocks are fused with weights 0.5 and 0.5, respectively.

We optimize with AdamW and no learning-rate scheduler. Learning rates are 2×10^{-5} for the shared stem and Stage 1, 10^{-5} for Stages 2–4 and the backbone normalization, and 3×10^{-4} for the projection heads, compressor, and predictor. Weight decay is 10^{-4} and the gradient norm is clipped to 1.0. The logical batch is evaluated exactly through gradient caching: changing the physical microbatch after an out-of-memory event does not change the batch-level contrastive objective. Table 5 lists all dataset-specific settings.

Table 5: Dataset-specific settings for main-representation training. "Freeze" is the number of initial epochs for which the shared DINOv3 backbone is frozen.

Setting	Train/Test	Epochs	Freeze	Logical batch	Microbatch	Eval. batch
FS-COCO official unseen	7000/3000	20	2	256	32	768
FS-COCO official normal	7000/3000	10	0	256	32	768
FS-COCO user-disjoint unseen	7000/3000	10	0	256	32	768
SketchyCOCO full	11265/2816	20	2	60	15	768
SketchyCOCO-SL	1015/210	20	2	60	15	60
QMUL-ShoeV2	5982/666	20	2	60	15	60

All main-model runs use seed 20260816. During a frozen-backbone epoch, the nominal microbatch equals the logical batch (256 for FS-COCO and 60 otherwise); after unfreezing, the nominal microbatches are those shown in Table 5. Automatic fallback may reduce only the physical microbatch to fit memory. We save a resumable checkpoint after every epoch and an additional numbered checkpoint every five epochs. Formal configurations disable per-epoch test evaluation and evaluate `last.pt` only after training.

A.3 SPATIAL-MEMORY TRAINING

The completed main encoder and compressor are frozen before training the spatial branch. Its input is the Stage 3 map in $\mathbb{R}^{512\times16\times16}$. A 1×1 projection reduces it to 32 channels, followed by a depthwise 3×3 residual block, 2×2 average pooling, and fixed raster-order flattening of the resulting $32 \times 8 \times 8$ tensor. A two-layer head with hidden width 512 maps this ordered tensor to the normalized 256-dimensional spatial memory.

For position-conditioned semantic prediction, one Stage 3-aligned rectangle is sampled per pair with area ratio in $[0.25, 0.50]$ and aspect ratio in $[0.5, 2.0]$. The same rectangle is masked in the normalized sketch and photo inputs using the zero value, which corresponds to the pretrained mean color. A visible anchor cell is sampled outside the rectangle. The predictor combines the 256-dimensional masked spatial memory, a 128-dimensional anchor projection, and a 64-dimensional encoding of $[\Delta x, \Delta y, w, h]$; a 512-dimensional hidden layer predicts the 2048-dimensional target formed by the ordered 2×2 pooling of the complete Stage 3 target region. The target features are stop-gradient. Spatial I-Con uses the same $\tau = 0.07$ and $\alpha = 0.20$ as the main model, and the position-conditioned prediction loss has weight 0.2.

The extractor and predictor are optimized with AdamW at learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. The logical batch is 256 and Stage 3 features are extracted in microbatches of 128. The spatial branch uses the same number of epochs as its corresponding main run. Seeds are 20260820 for all three FS-COCO protocols and 20260816 for SketchyCOCO

full, SketchyCOCO-SL, and QMUL-ShoeV2. No test evaluation is performed while this branch is trained.

A.4 Retrieval and Evaluation Protocols

At inference, the normalized 256-dimensional global representation, 512-dimensional complementary representation, and 256-dimensional spatial memory are concatenated with equal weights $w_g = w_c = w_m = 1/3$ and normalized once more. Retrieval uses cosine similarity and introduces no query-specific optimization. Fusion evaluation uses a batch size of 256 for every dataset.

For FS-COCO, we follow the provided `val_normal.txt` and `val_unseen_user.txt` lists, each giving 7000 training and 3000 test pairs. We also report a separately constructed 7000/3000 user-disjoint split whose artist sets are verified to have empty intersection before training. SketchyCOCO full uses `trainInTrain` for 11265 training pairs and the union of `valInTrain` and `val` for 2816 test pairs. SketchyCOCO-SL uses the fixed published subset of 1015 training and 210 test pairs. QMUL-ShoeV2 follows the ARNet directory split `trainA/trainB` and `testA/testB`; each test sketch is ranked against the unique photos in `testB`, and its paired photo index defines the ground truth. For all settings we report sketch-to-photo Recall@K at $K \in \{1, 5, 10\}$.

A.5 Exact Ablation Definitions

All ablations in Table 3 use QMUL-ShoeV2, seed 20260816, 20 training epochs, and the same unique-photo evaluation gallery. E-1 directly applies the pretrained DINOv3 ConvNeXt-B representation with timm's native pooling and performs no training. E0 disables the balanced compressor and uses only $\mathcal{L}_{\text{clean}}$ (weight 1, $\mathcal{L}_{\text{pred}}$ weight 0). E0_{1024} is identical except that its global projection has 1024 dimensions, matching the fused width of E5. E1 retains the global-only representation but weights these two objectives equally. E2 enables the balanced compressor while setting $\gamma = 0$, which gives every token uniform Sinkhorn mass. E3 restores $\gamma = 1$ and is the completed main representation without the spatial branch. E4 adds $m(I)$ but sets its geometry-conditioned region-prediction weight $\mathcal{L}_{\text{sp}}$ to zero; its three retrieval blocks are still fused equally. E5 restores this weight to 0.2 and is the full model.

A.6 Reproduction Archive

The accompanying anonymous archive contains the six main and six spatial YAML configurations, five explicit ablation configurations, dataset preparation and split-audit scripts, one-command training and evaluation entry points, and associated run records, including resolved configurations, training logs, and final metric files. Dataset roots and checkpoint paths are intentionally excluded from YAML files and supplied at runtime. The archive also contains CPU-runnable unit and integration tests for feature-grid dimensions, masking, gradient caching, objective construction, optimizer sharing, spatial memory, and evaluation-protocol consistency.