CONSENSUS-ANCHORED RETRIEVAL WITH CROSS-MODAL ALIGNMENT AND SEMANTIC CALIBRATION FOR SKETCH-BASED PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sketch-based person re-identification is challenged by both the sketch–photo modality gap and strong sketch subjectivity, as sketches of the same identity may exhibit substantial style-dependent variations. Existing methods mainly fuse multiple sketches or suppress style discrepancies, but do not explicitly model the latent identity consensus shared across style-biased observations. We therefore propose a consensus-anchored retrieval framework for robust sketch-based person re-identification. Its core component, Style-Conditioned Prototype-Fusion Consensus (SPFC), jointly learns an identity-consensus prototype and a query-adaptive retrieval descriptor through two complementary style-conditioned estimators. Two auxiliary training objectives further support this representation: Global-to-Local Mutual Alignment (GLMA) uses the global descriptor of each modality to guide local aggregation in the other, while Semantic Affinity-Guided Contrastive Calibration (SACC) uses auxiliary identity-level semantics available only during training to calibrate ambiguous negatives. Inference remains entirely visual. Extensive experiments under single-query, multi-query, and cross-style protocols demonstrate the effectiveness and generalizability of the proposed framework. Upon acceptance, we will publicly release the source code and pre-trained checkpoints on GitHub.

1 INTRODUCTION

Person re-identification (Re-ID) (Gheissari et al., 2006; Leng et al., 2019; Gong et al., 2022; Qi et al., 2024; Ye et al., 2022) aims to retrieve a target person from a gallery captured by surveillance cameras. Most existing Re-ID systems assume that a photo query is available. In practical scenarios, however, a query image of the target may be unavailable when an event occurs. Sketch-based person Re-ID (Pang et al., 2018; Lin et al., 2023; Wang et al., 2019; Hu et al., 2025a; Wu et al., 2026) addresses this limitation by using hand-drawn sketches, typically provided by witnesses or artists, as queries to search a photo gallery. Compared with photo-based Re-ID, it must address both the sketch–photo modality gap and the subjectivity gap among sketches.

Recent studies have identified sketch subjectivity as a central challenge in sketch-based person Re-ID. Sketches of the same identity may vary substantially due to viewing perspective, memory uncertainty, drawing habit, and artistic style. Existing methods address this issue through multi-query fusion, subjectivity neutralization, or style-inclusive learning (Lin et al., 2023; Hu et al., 2025a; Wu et al., 2026). However, they primarily treat subjective sketches as samples to be fused, augmented, or style-normalized. As illustrated in Fig. 1, we instead regard them as style-biased observations of a latent identity and seek the identity consensus shared across these observations.

This perspective shifts the problem from simply combining multiple sketches to estimating a stable identity representation while preserving discriminative details. Naive averaging may suppress useful cues, whereas direct attention-based fusion may amplify unreliable sketches or style-specific patterns. Robust Re-ID further requires fine-grained sketch–photo alignment and discrimination among semantically similar identities. These observations motivate a framework that anchors retrieval in identity consensus and supports it with cross-modal alignment and semantic calibration.

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

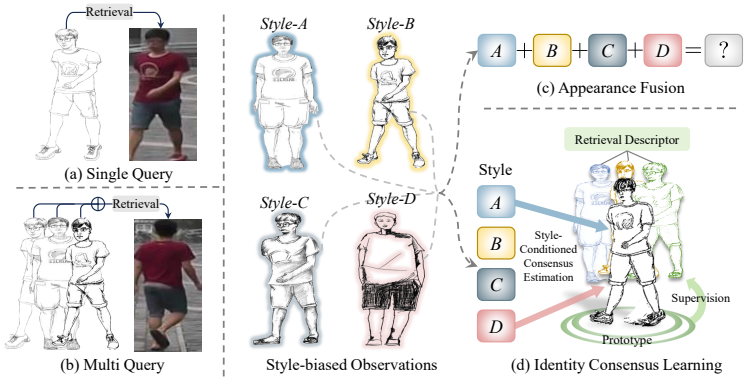

Figure 1: Task and motivation of our method. (a)–(b) Sketch-based person Re-ID retrieves a target identity from a photo gallery using one or multiple sketch queries. As multi-style sketches are style-biased observations of the same latent identity, direct appearance fusion may entangle identity cues with style-specific distortions (c). Our method instead estimates an identity-consensus prototype and uses it to regularize a discriminative retrieval descriptor (d).

To this end, we propose a consensus-anchored retrieval framework for sketch-based person Re-ID. Its core component, *Style-Conditioned Prototype-Fusion Consensus* (SPFC), jointly learns an identity-consensus prototype and a query-adaptive retrieval descriptor through two complementary style-conditioned estimators. The prototype aggregates same-identity sketches with style-conditioned weights and regularizes the retrieval descriptor. For multi-query (MQ) retrieval, the descriptor aggregates multiple sketch feature maps. For single-query (SQ) evaluation on multi-style benchmarks, it is extracted from one sketch and regularized by a prototype formed from same-identity sketches during training.

Two auxiliary training objectives further strengthen this representation. *Global-to-Local Mutual Alignment* (GLMA) uses each modality's global descriptor to guide local aggregation in the other, improving sketch–photo correspondence. *Semantic Affinity-Guided Contrastive Calibration* (SACC) uses identity-level semantics available only during training to calibrate ambiguous negatives. Retrieval at inference uses visual descriptors alone.

Our contributions are summarized as follows:

- We model sketches of the same identity as style-biased observations of a latent identity and formulate retrieval around their shared identity consensus.
- We develop SPFC to jointly learn a style-conditioned identity-consensus prototype and a query-adaptive retrieval descriptor for SQ and MQ retrieval.
- We introduce GLMA for cross-modal local alignment and SACC for calibrating semantically ambiguous negatives during training, while retaining visual-only inference.
- Controlled ablations verify the complementary roles of prototype consensus and style-aware fusion, while evaluations under SQ, MQ, and cross-style protocols demonstrate the effectiveness of the framework.

2 Related Work

2.1 Sketch Re-ID and Subjectivity

Sketch-based person Re-ID initially addressed the sketch–photo gap through adversarial learning, generation, and shared embeddings (Pang et al., 2018; Chen et al., 2021; Zhang et al., 2022a). Later work used attention, disentanglement, auxiliary modalities, and unified representations (Zhu et al., 2022; Chen et al., 2022; Liu et al., 2024; 2026); CLIP-based sketch retrieval, multimodal Re-ID, and few-shot generalization broadened the setting (Sain et al., 2023; Liu et al., 2025; Li et al., 2024a;

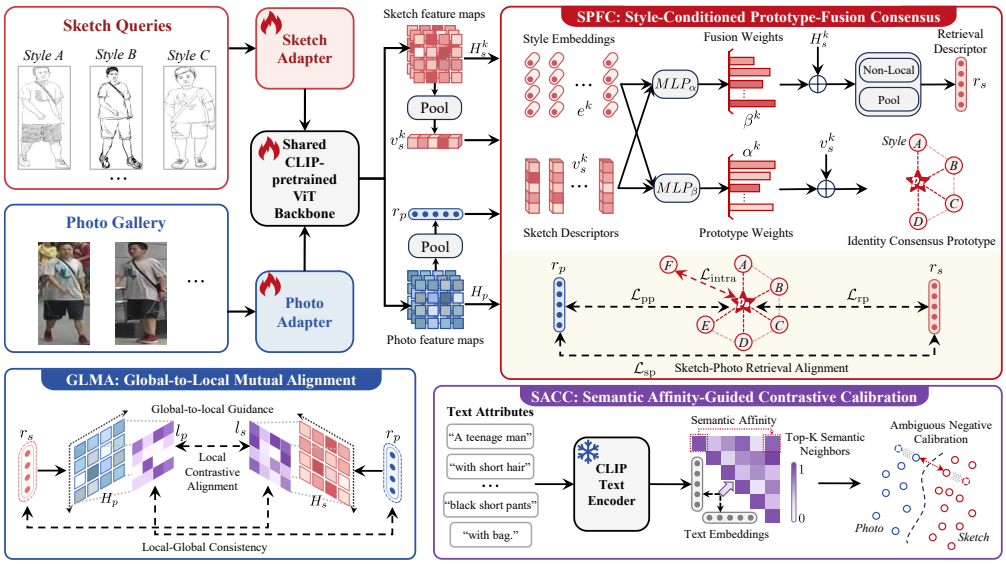

Figure 2: Overview of the proposed consensus-anchored retrieval framework. SPFC jointly learns an identity-consensus prototype and a retrieval-oriented sketch descriptor, while GLMA and SACC enhance fine-grained cross-modal alignment and semantic ambiguity discrimination, respectively.

Sun et al., 2025; Gong et al., 2026). These studies primarily advance cross-modal matching and generalization, whereas our focus is identity evidence shared across subjective sketches.

Market-Sketch-1K highlighted variation among sketches of the same identity (Lin et al., 2023). BDG combines sketch fusion with attribute guidance. CGC neutralizes subjectivity, FreeStyle promotes style-inclusive learning, CCS targets adaptation, and recent work separates style from content (Hu et al., 2025a; Wu et al., 2026; Hu et al., 2025b; Lu et al., 2025). Broader sketch retrieval studies user-dependent drawing styles, category–style generalization, and sketch abstraction (Sain et al., 2021; Hu et al., 2025c; Sain et al., 2020; Koley et al., 2024). Permutation-invariant set encoders and prototype-based metric learning provide general aggregation mechanisms (Zaheer et al., 2017; Lee et al., 2019; Snell et al., 2017), but do not distinguish a peer-sketch consensus target from retrieval-oriented fusion. SPFC makes this distinction through style-conditioned weighting, leave-one-out prototype supervision, and an independently weighted retrieval descriptor. GLMA uses global descriptors to select local evidence across modalities.

2.2 Semantic Guidance for Cross-Modal Retrieval

Semantic cues complement visual representations in cross-modal retrieval (Li et al., 2023). Attributes aid sketch–photo alignment (Lin et al., 2023), while prompt-based semantics and interactive text feedback support matching (Li et al., 2024b; Wu et al., 2025). SACC instead uses identity-level semantic affinity only during training to raise the logits of semantically similar negative identities in cross-modal contrastive learning, leaving inference entirely visual.

3 Methodology

Given a photo gallery and one or multiple sketch queries, sketch-based person retrieval aims to retrieve photos with the same identity as the sketch query. We denote a photo as x_p and a sketch as x_s. For multi-query retrieval, the sketch query is represented as $\mathcal{S} = \{(x_s^k, \sigma^k)\}_{k=1}^{K}$, where σ^k is the sketch style and K is the number of available sketches.

Our framework adopts a dual-branch CLIP-ViT encoder. Photos and sketches are first processed by modality-specific adapters and then encoded by a shared ViT backbone, producing deep feature maps and global retrieval features. The training consists of two stages. We first conduct a lightweight visual contrastive warm-up to initialize the sketch–photo embedding space. Then, the main stage

optimizes SPFC as the shared retrieval path, with GLMA and SACC serving as auxiliary objectives for cross-modal local alignment and semantic ambiguity calibration. The overall pipeline is shown in Fig. 2.

3.1 VISUAL CONTRASTIVE WARM-UP

Before the main identity consensus learning stage, we perform a lightweight visual contrastive warm-up to initialize the shared sketch–photo embedding space. Although the visual encoder is initialized from CLIP, the gap between sparse sketches and natural photos remains substantial, which may destabilize the subsequent consensus learning. Given a mini-batch of photo-sketch pairs, we obtain the global features $z_p = \mathrm{Pool}(F_p(x_p))$ and $z_s = \mathrm{Pool}(F_s(x_s))$. For multi-query samples, each sketch is encoded independently and the warm-up sketch feature is obtained by masked averaging over valid sketches. This stage does not involve style-conditioned aggregation, prototypes, text features, or local alignment.

We optimize a bidirectional supervised contrastive loss between normalized photo and sketch features:

$$\mathcal{L}_w = \frac{1}{2}\big(\mathcal{L}^{con}_{p\to s} + \mathcal{L}^{con}_{s\to p}\big), \tag{1}$$

where positives are samples sharing the same identity across modalities. This warm-up provides a coarse but stable cross-modal visual initialization for the following SPFC-based identity consensus learning.

3.2 STYLE-CONDITIONED PROTOTYPE-FUSION CONSENSUS

A sketch contains both identity evidence and style-dependent bias. Consequently, treating different sketches equally may propagate unreliable style patterns, whereas aggressively suppressing style may also remove useful identity details. Our key insight is that style should not merely be discarded as a nuisance; it can serve as contextual information for estimating how each sketch contributes to identity consensus. Based on this observation, we propose Style-Conditioned Prototype-Fusion Consensus (SPFC), which jointly learns a stable consensus anchor and a discriminative retrieval representation.

Given K same-identity sketches, the sketch encoder independently extracts their deep feature maps and global descriptors:

$$H^k_s = F_s(x^k_s), \qquad v^k_s = \mathrm{Pool}(H^k_s). \tag{2}$$

A learnable embedding $e^k = E(\sigma^k)$ encodes the corresponding style label σ^k. We concatenate visual and style information as $u^k_s = [v^k_s; e^k]$. Two independent scoring heads then estimate prototype and fusion weights:

$$\omega^b_k = \frac{\exp(h_b(u^k_s))}{\sum_{j\in\mathcal{V}} \exp(h_b(u^j_s))}, \quad b \in \{p, f\}, \tag{3}$$

where $\mathcal{V}$ denotes the valid sketch set. We use $\alpha_k = \omega^p_k$ to construct an identity-consensus prototype

$$p_s = \sum_{k\in\mathcal{V}} \alpha_k v^k_s, \tag{4}$$

which summarizes identity cues consistently supported across different styles.

The fusion weights $\beta_k = \omega^f_k$ are learned by a separate scoring head, allowing prototype construction and retrieval aggregation to adapt to their respective objectives. In the multi-query setting, we apply style-aware fusion (SAF) to the deep feature maps:

$$\widetilde{H}_s = \sum_{k\in\mathcal{V}} \beta_k H^k_s, \qquad r_s = \mathrm{Pool}\left(\mathrm{NL}(\widetilde{H}_s)\right). \tag{5}$$

For SQ, the descriptor is $r_s = \mathrm{Pool}(\mathrm{NL}(H_s))$; no cross-sketch aggregation is required. The photo branch produces $H_p = F_p(x_p)$ and $r_p = \mathrm{Pool}(H_p)$. Here, r_s and r_p denote the final retrieval-oriented sketch and photo descriptors, respectively. They are used for cross-modal learning and directly compared at inference. During SQ training on multi-style data, p_s is constructed from

same-identity sketches with different styles, while r_s is extracted from the individual query sketch. Thus, SQ and MQ share the same consensus-learning principle but differ in query-time aggregation. For single-sketch data, SPFC reduces to $K = 1$, $\alpha_1 = 1$, and $p_s = v_s$.

SPFC couples the prototype and retrieval branches through four objectives. Let $\mathcal{R}_s = \{r_s^i\}_i$, $\mathcal{R}_p = \{r_p^j\}_j$, and $\mathcal{P}_s = \{p_s^i\}_i$ denote the mini-batch collections of sketch descriptors, photo descriptors, and sample-aligned prototypes, respectively. Let $\mathcal{L}_{\text{mpc}}$ denote bidirectional multi-positive contrastive learning and $d_{\cos}(u, v) = 1 - \cos(u, v)$. We define

$$\mathcal{L}_{\text{sp}} = \mathcal{L}_{\text{mpc}}(\mathcal{R}_s, \mathcal{R}_p), \quad \mathcal{L}_{\text{rp}} = \tfrac{1}{M} \sum_i d_{\cos}(r_s^i, p_s^i), \quad \mathcal{L}_{\text{pp}} = \mathcal{L}_{\text{mpc}}(\mathcal{P}_s, \mathcal{R}_p), \tag{6}$$

where M is the number of sketch-side samples.

For $\mathcal{L}_{\text{intra}}$, each identity c with at least two sketches provides a randomly selected anchor $v_c^{a_c}$. We recompute and renormalize the prototype weights after excluding the anchor to obtain $p_c^{-a_c}$. With $q_{cd} = \hat{v}_c^{a_c \top} \hat{p}_d^{-a_d} / \tau_p$, we use

$$\mathcal{L}_{\text{intra}} = -\frac{1}{|\mathcal{C}_+|} \sum_{c \in \mathcal{C}_+} \log \frac{\exp(q_{cc})}{\sum_{d \in \mathcal{C}_+} \exp(q_{cd})}, \tag{7}$$

where $\mathcal{C}_+$ contains the valid multi-sketch identities; the loss is set to zero if $\mathcal{C}_+ = \varnothing$. The complete objective is

$$\mathcal{L}_{\text{SPFC}} = \lambda_{\text{sp}} \mathcal{L}_{\text{sp}} + \lambda_{\text{proto}} \left(\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{rp}} + \eta \mathcal{L}_{\text{pp}} \right). \tag{8}$$

On multi-style data, unlike a learnable class center or a single set pooler, SPFC constructs p_s from same-identity sketches through learned attention and trains the prototype estimator with leave-one-out contrastive supervision, while r_s is learned through an independent map-level estimator. Detailed batchwise definitions are provided in the supplementary material.

3.3 Global-to-Local Mutual Alignment

SPFC produces the final sketch retrieval descriptor r_s, while the photo branch provides its counterpart r_p. Although these global descriptors are optimized for identity discrimination, global pooling may discard fine-grained cues such as clothing patterns and carried accessories. Direct patch-to-patch alignment is also unreliable because sketch and photo regions are not spatially registered. We therefore propose Global-to-Local Mutual Alignment (GLMA), which uses the global semantics of one modality to select identity-relevant local evidence from the other.

During training, each sketch is paired with one same-identity photo at the corresponding batch index. Let $X_s = \{x_{s,n}\}_{n=1}^N$ and $X_p = \{x_{p,n}\}_{n=1}^N$ denote their final retrieval-path tokens. In MQ, X_s is obtained from the fused sketch map, whereas SQ uses the individual sketch map. Bidirectional global guidance is defined as

$$\begin{aligned} a_{s,n} &= \text{softmax}_n \left(\cos(x_{s,n}, r_p) / \tau_a \right), \quad \ell_s = \sum_{n=1}^N a_{s,n} x_{s,n}, \\ a_{p,n} &= \text{softmax}_n \left(\cos(x_{p,n}, r_s) / \tau_a \right), \quad \ell_p = \sum_{n=1}^N a_{p,n} x_{p,n}. \end{aligned} \tag{9}$$

Unlike dense token-to-token cross-attention, GLMA uses each global descriptor as a single semantic query over the other modality, avoiding explicit spatial correspondence while producing a compact local descriptor.

We optimize GLMA by

$$\mathcal{L}_{\text{GLMA}} = \mathcal{L}_{\text{mpc}}(\mathbf{L}_s, \mathbf{L}_p) + \frac{1}{2} \left[d_{\cos}(\ell_s, r_s) + d_{\cos}(\ell_p, r_p) \right], \tag{10}$$

where $\mathbf{L}_s$ and $\mathbf{L}_p$ collect the cross-guided local descriptors in the mini-batch, and the consistency terms are averaged over the batch. Guidance is restricted to the sampled positive pair and remains fully differentiable, while other batch samples participate through the identity-based contrastive objective. GLMA is used only during training, and inference independently compares r_s and r_p.

3.4 Semantic Affinity-Guided Contrastive Calibration

Standard cross-modal contrastive learning treats different-identity samples as negatives but does not explicitly distinguish semantically unrelated negatives from ambiguous identities with highly similar appearance descriptions. The latter provide more informative supervision and should compete more strongly with the positive identity. We therefore propose Semantic Affinity-Guided Contrastive Calibration (SACC), which uses language semantics to calibrate negative difficulty rather than directly participating in retrieval.

We encode sample attribute descriptions using the CLIP text encoder and average same-identity text features to obtain an identity-level semantic representation t_c. The semantic affinity between identities c and d is $A_{cd} = \cos(t_c, t_d)$. For each identity, we retain its top-K positively correlated identities. The calibration term is $\Delta_{cd} = \mathbf{1}[c \neq d]\, M_{cd}\, [A_{cd}]_+$, where M_{cd} is the top-K affinity mask and $[x]_+ = \max(x, 0)$.

Given the original sketch–photo contrastive logit, SACC calibrates it as

$$l_{ij} = \frac{\cos(r_s^i, r_p^j)}{\tau_c}, \qquad \widetilde{l}_{ij} = l_{ij} + \gamma \Delta_{y_i, y_j}, \tag{11}$$

where γ controls the calibration strength. Thus, only semantically similar negative identities receive larger logits and become stronger competitors in contrastive learning.

Let $\widetilde{\mathbf{L}} = [\widetilde{l}_{ij}]$ denote the calibrated logit matrix within a mini-batch. The SACC objective is

$$\mathcal{L}_{\mathrm{SACC}} = \frac{1}{2}\left[\mathcal{L}_{s\to p}^{\mathrm{mpc}}(\widetilde{\mathbf{L}}) + \mathcal{L}_{p\to s}^{\mathrm{mpc}}(\widetilde{\mathbf{L}}^\top)\right]. \tag{12}$$

Unlike hard-negative selection, SACC retains the full negative set and calibrates only the logits of identity-level semantic neighbors within the visual multi-positive objective. Text is discarded at inference.

3.5 Overall Objective and Inference

After the visual contrastive warm-up, the main stage jointly optimizes identity discrimination, the SPFC representation, and its two auxiliary objectives:

$$\mathcal{L} = \mathcal{L}_{\mathrm{id}} + \mathcal{L}_{\mathrm{tri}} + \mathcal{L}_{\mathrm{SPFC}} + \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{GLMA}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{SACC}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{id}}$ and $\mathcal{L}_{\mathrm{tri}}$ denote the identity classification and triplet losses, respectively, while λ_{g} and λ_{s} balance GLMA and SACC. GLMA and SACC operate on the SPFC retrieval descriptors, making consensus learning the shared representation path, while $\mathcal{L}_w$ is optimized only in the preceding warm-up stage.

At inference, each gallery photo is encoded as r_p. A single sketch directly produces r_s, whereas multiple sketches are aggregated by the style-aware fusion branch to obtain a unified query descriptor. After ℓ_2 normalization, we compute retrieval similarity as $s(x_s, x_p) = \hat{r}_s^\top \hat{r}_p$ and rank gallery images in descending order. Only visual retrieval descriptors are used for matching; textual semantics and training objectives introduce no additional retrieval-time computation.

4 Experiments

4.1 Datasets and Evaluation Protocols

Market-Sketch-1K (Lin et al., 2023) contains 996 identities and sketches drawn in six styles. Following the official zero-shot split, 498 identities are used for training and the remaining 498 for testing. We evaluate single-query (SQ), multi-query (MQ), and cross-style protocols.

PKU-Sketch (Pang et al., 2018) contains 200 identities, each associated with one sketch and two photos. We follow the standard protocol with 150 training identities and 50 disjoint testing identities, and report the average results over ten trials. Unlike Market-Sketch-1K, PKU-Sketch provides one sketch per identity. Here, SPFC reduces to its single-sketch form, allowing us to evaluate the framework without cross-sketch consensus. For SACC, both datasets are supplemented with identity-level

Table 1: Comparison with state-of-the-art methods on Market-Sketch-1K under single-query (SQ) and multi-query (MQ) protocols. All results are reported in percentages. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Venue	Single Query (SQ)			Multi Query (MQ)		
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
DCLNet (Sun et al., 2022)	MM'22	13.45	12.24	29.20	–	–	–
DSCNet (Zhang et al., 2022b)	TIFS'22	14.73	13.84	30.55	–	–	–
DEEN (Zhang & Wang, 2023)	CVPR'23	12.62	12.11	25.44	–	–	–
BDG (Lin et al., 2023)	MM'23	19.61	18.10	38.95	24.45	24.70	50.40
SketchTrans$^+$ (Chen et al., 2023)	TPAMI'24	28.45	19.42	56.45	–	–	–
OLTM (Li et al., 2024b)	NeurIPS'24	38.35	<u>36.75</u>	63.88	<u>62.55</u>	<u>69.48</u>	<u>90.36</u>
CGC (Hu et al., 2025a)	TIFS'25	34.17	31.73	57.68	54.18	57.23	81.93
KTCAA (Gong et al., 2026)	AAAI'26	22.70	21.30	–	–	–	–
FreeStyle (Wu et al., 2026)	TIP'26	<u>41.16</u>	35.02	**67.09**	–	–	–
Ours	–	**41.34**	**40.21**	<u>65.02</u>	**69.84**	**74.50**	**93.78**

attribute annotations or textual descriptions provided by prior works (Zhai et al., 2022; Lin et al., 2019). All semantic representations are computed using training identities only, and no textual information from test identities is involved.

Evaluation metrics. We report mean Average Precision (mAP) and cumulative matching characteristic accuracy at Rank-1 and Rank-5 on both datasets.

4.2 Implementation Details

All images are resized to 288×144 and augmented with random cropping and horizontal flipping. The visual encoder uses CLIP ViT-B/16 (Radford et al., 2021; Dosovitskiy et al., 2020), while the text encoder uses its paired CLIP Transformer and remains frozen. We train with AdamW, using 5 epochs of visual warm-up followed by 50 epochs of main training. Learning rates are 1×10^{-5} for pretrained encoders, 1×10^{-4} for adapters and SPFC, and 5×10^{-4} for prediction heads. We balance identities during sampling. Market-Sketch-1K SQ draws sketches of the same identity across styles, MQ groups available sketches in different styles into each query set, and PKU-Sketch uses one sketch per identity. The temperatures for cross-modal and masked prototype contrast are 0.07 and 0.1. We set $\lambda_{\mathrm{g}} = 0.2$, retain the two strongest semantic neighbors for SACC, and use $\gamma = 0.2$ and $\lambda_{\mathrm{s}} = 0.05$. All experiments use one NVIDIA GeForce RTX 5090 GPU. Exact batch compositions, complete loss weights, and training schedules are provided in the supplementary material.

4.3 Comparison with State-of-the-Art Methods

We compare our method with recent state-of-the-art approaches on Market-Sketch-1K and PKU-Sketch. Table 1 reports the comparison on Market-Sketch-1K. Our method achieves the highest mAP and Rank-1 under the SQ protocol, exceeding FreeStyle and OLTM by 0.18 and 3.46 percentage points, respectively, while FreeStyle remains stronger at Rank-5. Under the MQ protocol, our method outperforms OLTM by 7.29 and 5.02 percentage points in mAP and Rank-1, respectively. The MQ result is consistent with the intended role of style-conditioned fusion in aggregating identity evidence from multiple subjective sketches.

As shown in Table 2, our method achieves 94.21% mAP, 94.40% Rank-1, and 99.80% Rank-5 on PKU-Sketch, obtaining the best performance across all three reported metrics. It surpasses the previous highest mAP reported by KTCAA by 1.91 percentage points, the previous highest Rank-1 reported by CGC by 0.40 points, and the previous highest Rank-5 reported by FreeStyle by 1.14 points. These results demonstrate the effectiveness of the proposed framework under the conventional single-sketch setting.

Table 2: Comparison with state-of-the-art methods on PKU-Sketch. All results are reported in percentages. The best and second-best results are highlighted in bold and underlined, respectively.

Method	mAP	Rank-1	Rank-5
BDG (Lin et al., 2023)	66.37	70.00	–
SketchTrans$^+$ (Chen et al., 2023)	–	85.80	96.00
DALNet (Liu et al., 2024)	86.20	90.00	98.60
CGC (Hu et al., 2025a)	89.60	<u>94.00</u>	98.00
FreeStyle (Wu et al., 2026)	83.78	87.33	<u>98.66</u>
KTCAA (Gong et al., 2026)	<u>92.30</u>	92.70	–
Ours	**94.21**	**94.40**	**99.80**

4.4 Ablation Study

Component Analysis. Table 3 reports MQ ablations on Market-Sketch-1K. All variants use the same adapters and CLIP ViT-B/16 encoder, while the baseline excludes warm-up, SPFC, GLMA, and SACC. Warm-up raises mAP from 62.68% to 64.05%. On this baseline, SPFC gives the largest individual gain of 2.84 mAP and 2.01 Rank-1 points; GLMA and SACC also improve both metrics. Every pair beats its individual variants, and SACC further improves SPFC with GLMA. SPFC with GLMA reaches 69.19% mAP and 73.69% Rank-1, while the full model reaches 69.84% and 74.50%. Relative to the initial baseline, the full gains are 7.16 and 10.64 points.

Table 3: Ablation study of different components under the multi-query protocol on Market-Sketch-1K.

Setting	Warm-up	SPFC	GLMA	SACC	mAP(%)	Rank-1(%)
(a)	–	–	–	–	62.68	63.86
(b)	✓	–	–	–	64.05	68.07
(c)	✓	✓	–	–	66.89	70.08
(d)	✓	–	✓	–	65.92	69.08
(e)	✓	–	–	✓	65.67	68.88
(f)	✓	✓	✓	–	<u>69.19</u>	<u>73.69</u>
(g)	✓	✓	–	✓	67.41	71.29
(h)	✓	–	✓	✓	66.35	69.48
(i)	✓	✓	✓	✓	**69.84**	**74.50**

Table 4: Ablation of SPFC design choices on Market-Sketch-1K under the MQ protocol.

Variant	Prototype	Uses style embedding	Fusion weights	mAP	Rank-1
Mean Fusion	–	–	Uniform	66.35	69.48
Attention Fusion	–	–	Learned	67.05	69.32
Style Fusion	–	✓	Learned	67.70	70.88
Prototype Only	✓	✓	Uniform	68.43	72.09
Shared Weights	✓	✓	Shared	<u>68.96</u>	<u>73.29</u>
Full SPFC	✓	✓	Learned	**69.84**	**74.50**

Effect of SPFC Design Choices. With warm-up, GLMA, and SACC fixed, Table 4 isolates SPFC design choices. Attention without style embeddings raises mAP but slightly lowers Rank-1 relative to uniform fusion. Style conditioning improves both metrics. Prototype Only outperforms Style Fusion despite using uniform fusion. Sharing prototype and fusion weights adds 0.53 mAP and 1.20 Rank-1 points over Prototype Only; learning independent weights adds another 0.88 and 1.21 points. These results support learning distinct weights for prototype construction and retrieval fusion.

Cross-style Generalization. We train and test on disjoint Market-Sketch-1K style sets. Table 5 shows that our method leads FreeStyle in 8 of 15 valid SQ splits. Its largest lead occurs for $AB \rightarrow F$,

Table 5: Cross-style SQ comparison on Market-Sketch-1K in mAP (%). "–" denotes overlapping training and testing styles.

Train \ Test	FreeStyle (Wu et al., 2026)					Ours				
	F	EF	DEF	$C\cdots F$	$B\cdots F$	F	EF	DEF	$C\cdots F$	$B\cdots F$
A	25.05	**25.87**	20.10	**25.92**	23.72	**27.10**	24.06	**23.15**	23.50	**23.75**
AB	26.17	26.22	**27.91**	27.99	–	**30.16**	**26.96**	25.78	**28.22**	–
ABC	**32.90**	**31.43**	**32.54**	–	–	32.58	30.45	31.14	–	–
$A\cdots D$	36.18	**35.38**	–	–	–	**36.48**	33.84	–	–	–
$A\cdots E$	36.93	–	–	–	–	**37.10**	–	–	–	–

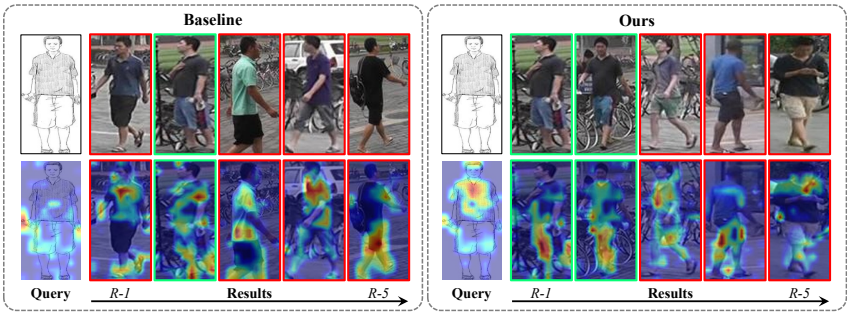

Figure 3: Visualization of retrieval results and similarity-based Grad-CAM maps. Green and red boxes denote correct and incorrect retrieval results, respectively.

at 30.16% versus 26.17% mAP. FreeStyle leads in the other 7 splits, including all three trained on ABC, so the gains depend on the style split.

Computational Cost. Training uses 150.09M parameters, while inference uses the visual branch with 86.66M. SQ costs 28.04 GFLOPs and 0.81 ms per query. MQ with six sketches costs 168.24 GFLOPs and 6.03 ms. Timings exclude preprocessing, data transfer, and similarity search.

4.5 Visualization Analysis

Figure 3 compares retrieval rankings and similarity-based Grad-CAM (Selvaraju et al., 2017; Hu et al., 2021) maps for a sketch query with distracting objects, including bicycles. The baseline ranks a correct match second, whereas our method retrieves it first. In the query and matched photo, baseline responses are dispersed and sometimes extend beyond the pedestrian. Our method shows stronger activation over the person, including upper-body appearance and lower-body clothing, while reducing responses to nearby clutter. Although some background activation remains, the ranking and maps together suggest that the learned descriptor relies more on person-related evidence for this cross-modal match.

5 Conclusion

In this paper, we present a consensus-anchored framework for sketch-based person Re-ID that models sketches of the same identity as style-biased observations of a latent identity. At its core, SPFC jointly learns a style-conditioned identity-consensus prototype and a query-adaptive retrieval descriptor, using the prototype to regularize retrieval in both SQ and MQ settings. GLMA uses each modality's global descriptor to guide local aggregation in the other, improving sketch-photo correspondence; SACC uses training-only identity-level semantics to calibrate ambiguous negatives. Together, these objectives strengthen a visual-only inference model. Experiments on Market-Sketch-1K and PKU-Sketch show strong SQ and MQ retrieval performance and competitive cross-style results, supporting the combined treatment of sketch subjectivity, cross-modal alignment, and semantic ambiguity. The dependence on provided style labels for multi-style weighting and robustness to severely degraded sketches remain directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, and language polishing. Generative AI was not used to design the research method or experiments, implement experimental code, generate or process datasets, or create scientific figures. The authors reviewed the AI-assisted material and take full responsibility for the final content, citations, claims, and results.

REFERENCES

Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis. In *proceedings of the 30th ACM international conference on multimedia*, pp. 4012–4020, 2022.

Cuiqun Chen, Mang Ye, Meibin Qi, and Bo Du. Sketchtrans: Disentangled prototype learning with transformer for sketch-photo recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2950–2964, 2023.

Qingshan Chen, Zhenzhen Quan, Kun Zhao, Yifan Zheng, Zhi Liu, and Yujun Li. A cross-modality sketch person re-identification model based on cross-spectrum image generation. In *International Forum on Digital TV and Wireless Multimedia Communications*, pp. 312–324. Springer, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 1528–1535, 2006.

Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4313–4322, 2022.

Yunpeng Gong, Yongjie Hou, Jiangming Shi, Kim Long Diep, and Min Jiang. A theory-inspired framework for few-shot cross-modal sketch person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4284–4292, 2026.

Shaojun Gui, Yu Zhu, Xiangxiang Qin, and Xiaofeng Ling. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing*, 403:294–303, 2020.

Hanjiang Hu, Hesheng Wang, Zhe Liu, and Weidong Chen. Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization. *IEEE/CAA Journal of Automatica Sinica*, 9(2):313–328, 2021.

Zechao Hu, Zhengwei Yang, Hao Li, and Zheng Wang. Contrastive-generative-contrastive: Neutralize subjectivity in sketch re-identification. *IEEE Transactions on Information Forensics and Security*, 2025a.

Zechao Hu, Zhengwei Yang, Hao Li, Zheng Wang, and Yixiong Zou. Cross-category subjectivity generalization for style-adaptive sketch re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22644–22653, 2025b.

Zechao Hu, Zhengwei Yang, Hao Li, Yixiong Zou, Fengbin Zhu, and Zheng Wang. Unified category and style generalization for instance-level sketch retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 864–873, 2025c.

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.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019.

He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17459–17469, 2024a.

Rui Li, Tingting Ren, Jie Wen, and Jinxing Li. Optimal transport-based labor-free text prompt modeling for sketch re-identification. *Advances in Neural Information Processing Systems*, 37: 119198–119226, 2024b.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023.

Kejun Lin, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, and Shin'ichi Satoh. Beyond domain gap: Exploiting subjectivity in sketch-based person retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 2078–2089, 2023.

Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019.

Manman Liu, Xu Cheng, Yan Jiang, and Baowei Wang. Colorsketchnet: Unifying color, sketch and texture for modality-agnostic multi-modal person re-identification. *Neural Networks*, pp. 108374, 2025.

Xingyu Liu, Xu Cheng, Haoyu Chen, Hao Yu, and Guoying Zhao. Differentiable auxiliary learning for sketch re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3747–3755, 2024.

Xingyu Liu, Yan Jiang, Xu Cheng, Hao Yu, Haoyu Chen, and Guoying Zhao. Crombo: Cross-modality bootstrapping for unified sketch–photo representation learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(1):1–18, 2026.

Lingyi Lu, Xin Xu, and Xiao Wang. Style separation and content recovery for generalizable sketch re-identification and a new benchmark. In *International Conference on Multimedia Modeling*, pp. 114–127. Springer, 2025.

Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 609–617, 2018.

Lei Qi, Ziang Liu, Yinghuan Shi, and Xin Geng. Generalizable metric network for cross-domain person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34 (10):9039–9052, 2024.

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*, pp. 8748–8763. PMLR, 2021.

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. *arXiv preprint arXiv:2007.15103*, 2020.

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8500–8509, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.00840. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00840.

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, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 5333–5341, 2022.

Zhen Sun, Lei Tan, Yunhang Shen, Chengmao Cai, Xing Sun, Pingyang Dai, Liujuan Cao, and Rongrong Ji. Flexireid: Adaptive mixture of expert for multi-modal person re-identification. *arXiv preprint arXiv:2510.15595*, 2025.

Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019.

Xinyi Wu, Cuiqun Chen, Hui Zeng, Zhiping Cai, Bo Du, and Mang Ye. Interactive sketch-based person re-identification with text feedback. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Xinyi Wu, Cuiqun Chen, Hui Zeng, Zhiping Cai, and Mang Ye. Freestyle: Towards style-inclusive sketch-based person retrieval. *IEEE Transactions on Image Processing*, 2026.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2872–2893, 06 2022. doi: 10.1109/TPAMI.2021.3054775.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 63–71, 2022.

Yafei Zhang, Yongzeng Wang, Huafeng Li, and Shuang Li. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3347–3355, 2022a.

Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18:1554–1565, 2022b.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2153–2162, 2023.

Fengyao Zhu, Yu Zhu, Xiaoben Jiang, and Jiongyao Ye. Cross-domain attention and center loss for sketch re-identification. *IEEE Transactions on Information Forensics and Security*, 17:3421–3432, 2022.

APPENDIX

A ADDITIONAL METHODOLOGICAL DETAILS

This section provides additional definitions and protocol-specific details omitted from the main paper for conciseness.

A.1 BATCHWISE MULTI-POSITIVE CONTRASTIVE OBJECTIVE

Let $\mathbf{U} = \{u_i\}_{i=1}^{M}$ and $\mathbf{V} = \{v_j\}_{j=1}^{N}$ denote two feature collections with identity labels y_i^U and y_j^V, respectively. Their pairwise contrastive logits are

$$l_{ij} = \frac{\hat{u}_i^\top \hat{v}_j}{\tau}, \tag{A.1}$$

where $\hat{u}_i$ and $\hat{v}_j$ are ℓ_2-normalized features and τ is the contrastive temperature. For each anchor u_i, the positive set in $\mathbf{V}$ is

$$\mathcal{P}_{U\to V}(i) = \{j \mid y_j^V = y_i^U\}. \tag{A.2}$$

The directional multi-positive contrastive loss is

$$\mathcal{L}_{U\to V} = -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathcal{P}_{U\to V}(i)|}\sum_{j\in\mathcal{P}_{U\to V}(i)}\log\frac{\exp(l_{ij})}{\sum_{t=1}^{N}\exp(l_{it})}. \tag{A.3}$$

The bidirectional objective is

$$\mathcal{L}_{\mathrm{mpc}}(\mathbf{U},\mathbf{V}) = \frac{1}{2}\left(\mathcal{L}_{U\to V} + \mathcal{L}_{V\to U}\right). \tag{A.4}$$

Thus, all samples in the opposite feature collection that share the anchor identity are treated as positives, while the remaining samples participate as negatives. This formulation is used by $\mathcal{L}_{\mathrm{sp}}$, $\mathcal{L}_{\mathrm{pp}}$, and the local contrastive term in GLMA. In our identity-balanced batches, each anchor has at least one positive in the opposite collection by construction. For SACC, the same positive-set construction is applied directly to the calibrated logit matrix, while the identity labels and positive sets remain unchanged.

A.2 LEAVE-ONE-OUT PROTOTYPE CONSTRUCTION

The prototype p_s used by $\mathcal{L}_{\mathrm{rp}}$ and $\mathcal{L}_{\mathrm{pp}}$ is constructed from all valid same-identity sketches. In contrast, $\mathcal{L}_{\mathrm{intra}}$ adopts a leave-one-out construction to prevent direct anchor self-inclusion.

For each valid multi-sketch identity c, we randomly select an anchor index a_c and exclude it from the prototype support set. The remaining prototype weights are recomputed and renormalized as

$$\alpha_{c,k}^{-a_c} = \frac{\exp(h_p(u_{s,c}^k))}{\sum_{j\in\mathcal{V}_c\setminus\{a_c\}}\exp(h_p(u_{s,c}^j))}, \tag{A.5}$$

where $k \in \mathcal{V}_c \setminus \{a_c\}$ and $\mathcal{V}_c$ is the valid sketch set of identity c. The leave-one-out prototype is then

$$p_c^{-a_c} = \sum_{k\in\mathcal{V}_c\setminus\{a_c\}}\alpha_{c,k}^{-a_c} v_{s,c}^k. \tag{A.6}$$

The anchor descriptor is required to identify its own leave-one-out prototype against those of other valid identities, as defined by $\mathcal{L}_{\mathrm{intra}}$ in the main paper. This design encourages the prototype estimator to capture identity evidence supported by peer sketches rather than relying on the anchor itself. When $|\mathcal{C}_+| = 1$, the softmax contains only the positive prototype and the resulting loss is zero. When $\mathcal{C}_+ = \varnothing$, the term is explicitly set to zero. Therefore, $\mathcal{L}_{\mathrm{intra}}$ provides effective contrastive supervision only when at least two valid multi-sketch identities are available.

A.3 PROTOCOL-SPECIFIC CONSENSUS LEARNING

The sketch support set used during training should be distinguished from the query available at inference. This distinction enables SQ and MQ to share the same prototype-based consensus supervision while adopting different query-time aggregation strategies.

For SQ training on multi-style data, the retrieval descriptor r_s is extracted from an individual sketch, whereas its sample-aligned prototype p_s is constructed from all valid same-identity sketches sampled across styles in the current mini-batch. This full prototype is used by $\mathcal{L}_{\mathrm{rp}}$ and $\mathcal{L}_{\mathrm{pp}}$; leave-one-out construction is applied only to $\mathcal{L}_{\mathrm{intra}}$. The consistency loss $\mathcal{L}_{\mathrm{rp}}$ transfers the shared identity evidence represented by p_s to the individual retrieval descriptor. At inference, only the single query sketch is required.

For MQ training, each sketch-side sample contains a valid multi-style query set. The prototype estimator applies $\{\alpha_k\}$ to the pooled sketch descriptors to construct p_s, whereas the fusion estimator applies the independently learned weights $\{\beta_k\}$ to the corresponding deep feature maps to construct r_s. The two estimators therefore operate on the same valid sketch set but serve different purposes: the former provides a stable consensus target, whereas the latter preserves retrieval-oriented information through query-adaptive fusion.

On single-sketch data, the prototype estimator naturally degenerates to

$$K = 1, \qquad \alpha_1 = 1, \qquad p_s = v_s. \tag{A.7}$$

Because no peer sketch remains after excluding the only observation, $\mathcal{L}_{\mathrm{intra}}$ is set to zero. The other prototype-related objectives retain their original definitions. Therefore, PKU-Sketch evaluates the applicability of the shared framework under a single-observation setting, rather than the complete multi-style consensus-learning capability.

A.4 GLMA PAIR CONSTRUCTION AND BATCHWISE OBJECTIVE

During training, the identity-balanced sampler produces index-aligned sketch-side and photo samples. Each sketch-side sample, which may be either an individual sketch or an MQ sketch set, is paired with one same-identity photo at the corresponding batch index.

GLMA constructs one cross-guided local descriptor pair (ℓ_s^i, ℓ_p^i) for each index-aligned positive pair. The global photo descriptor r_p^i guides local aggregation over the sketch tokens, while r_s^i reciprocally guides aggregation over the photo tokens. We do not enumerate all possible combinations among same-identity sketch and photo samples.

Let

$$\mathbf{L}_s = \{\ell_s^i\}_{i=1}^M, \qquad \mathbf{L}_p = \{\ell_p^i\}_{i=1}^M. \tag{A.8}$$

The local–global consistency term in the main paper is implemented as the batch average

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{2M} \sum_{i=1}^{M} \left[d_{\cos}(\ell_s^i, r_s^i) + d_{\cos}(\ell_p^i, r_p^i) \right]. \tag{A.9}$$

Accordingly,

$$\mathcal{L}_{\mathrm{GLMA}} = \mathcal{L}_{\mathrm{mpc}}(\mathbf{L}_s, \mathbf{L}_p) + \mathcal{L}_{\mathrm{cons}}. \tag{A.10}$$

Samples at other batch indices do not participate in the paired guidance operation, but they contribute positives or negatives to the subsequent multi-positive contrastive objective according to their identity labels. This paired conditioning is used only during training and is discarded at inference.

A.5 BATCHWISE SEMANTIC NEIGHBOR SELECTION

SACC dynamically determines semantic neighbors from the identities appearing in each mini-batch. Let $\mathcal{I}_c$ denote the samples of identity c in the current batch. Their text features are averaged and normalized to obtain

$$t_c = \mathrm{Norm}\left(\frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} t_i \right). \tag{A.11}$$

Let $\mathcal{C}_B$ denote the set of identities present in the current mini-batch. The identity-level affinity $A_{cd} = \cos(t_c, t_d)$ is computed for $c, d \in \mathcal{C}_B$. For each identity c, we define its semantic neighbor set as

$$\mathcal{N}_c = \text{Top-K}_{d \in \mathcal{C}_B \setminus \{c\},\, A_{cd} > 0}\, A_{cd}, \tag{A.12}$$

and construct the mask

$$M_{cd} = \mathbf{1}[d \in \mathcal{N}_c]. \tag{A.13}$$

We use $K = 2$ in all experiments; therefore, $|\mathcal{N}_c| \leq 2$ when fewer than two positively correlated identities are available.

The neighbor relation is recomputed for every mini-batch and depends on its identity composition. Identities that are not selected as semantic neighbors remain in the contrastive denominator with their original logits. The calibrated sketch-to-photo logit matrix is transposed for the reverse photo-to-sketch direction, following the bidirectional objective defined in the main paper. Hence, SACC does not remove conventional negatives; it only strengthens the competition from semantically ambiguous identities during training. No semantic information is used at inference.

B Additional Implementation Details

This section specifies the mini-batch construction, objective coefficients, optimization schedule, and execution environment used by the fixed reference configuration. Unless explicitly stated otherwise, the same model and objective settings are used across protocols; only the organization of the sketch observations and the number of identities per mini-batch differ.

B.1 Protocol-Specific Mini-Batch Construction

We use identity-balanced sampling in both modalities. The *identity batch size* is the number of identities sampled per iteration, rather than the number of image tensors passed through one modality. Table B.1 summarizes the resulting composition. The photo and sketch-side entries are index-aligned by identity, as required by GLMA.

Table B.1: Exact training mini-batch composition. "Views" in MQ denotes independently augmented occurrences of the same identity-level query set; each set itself contains six tensor slots, of which only the available styles are valid.

Protocol	IDs/batch	Sketch-side samples/ID	Photos/ID	Sketch-side batch	Photo batch
Market-Sketch-1K (SQ)	6	4 individual sketches	4	24 sketches	24 photos
Market-Sketch-1K (MQ)	6	4 views of one multi-style set	4	24 sets	24 photos
PKU-Sketch (SQ)	16	1 individual sketch	1	16 sketches	16 photos

Market-Sketch-1K(SQ). Each mini-batch contains six identities. For every identity, the sampler draws four photos and four sketches, giving 24 samples per modality. Sketch sampling first shuffles the available style labels and selects one sketch from each distinct style until four slots have been filled. If fewer than four styles are available, the remaining slots are sampled from all sketches of that identity; sampling with replacement is used only when the number of available sketches is smaller than the number of remaining slots. Thus, style diversity is preferred but cannot be guaranteed for an identity with insufficient observations. Photo slots are sampled with replacement and therefore need not correspond to four distinct photo files.

The four sampled sketch slots form the same-identity support set used to construct the ordinary prototype p_s. For $\mathcal{L}_{\text{intra}}$, one slot is randomly selected as the anchor and excluded, and the prototype weights are recomputed over the other three slots. Consequently, every leave-one-out target has three sampled peer slots, although fewer than three *distinct* images may remain when sampling had to duplicate an observation. The retrieval descriptor r_s is always extracted from the individual sketch serving as the current SQ sample, not from the four-sketch aggregate.

Market-Sketch-1K(MQ). Each mini-batch likewise contains six identities and four photos per identity. The sketch record for an identity is an identity-level query set with six tensor slots corresponding to styles A–F. The set contains one sketch for every available style, rather than requiring all six styles to be present. Missing styles are represented by zero-valued image tensors and the padding style index -1. A valid-sketch mask, derived from the number of available styles, removes these padded positions from prototype construction, style-aware weighting, and feature-map fusion.

For the ABCDEF training split, the 498 identities contain between three and six valid sketches: 1 identity has three, 250 have four, 154 have five, and 93 have all six styles. Since preprocessing stores one MQ record per identity, the identity sampler produces four independently augmented occurrences of that set and pairs them with four sampled photos. Prototype and fusion weights are computed within each valid multi-style set; padded positions never contribute.

PKU-Sketch. PKU-Sketch uses 16 identities per mini-batch. Each identity contributes its only sketch and one randomly sampled photo, so both modality batches contain 16 images and $K = 1$. The singleton prototype and fusion weights satisfy $\alpha_1 = \beta_1 = 1$, and the prototype reduces to $p_s = v_s$. No valid peer remains for leave-one-out construction; therefore, $\mathcal{L}_{\text{intra}} = 0$. The prototype–retrieval and prototype–photo objectives retain their definitions, while style-aware fusion becomes an identity-preserving singleton mapping.

B.2 Complete Objective and Loss Coefficients

The main-stage implementation is equivalently written as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}} + \lambda_{\text{g}}\mathcal{L}_{\text{GLMA}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} \\ + \lambda_{\text{proto}}\left(\mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{rp}} + \eta\mathcal{L}_{\text{pp}}\right) + \lambda_{\text{s}}\mathcal{L}_{\text{SACC}}.\end{aligned} \tag{B.1}$$

Table B.2 gives the fixed reference values and their command-line counterparts.

Table B.2: Loss coefficients in the fixed reference configuration.

Quantity	Value	Implementation coefficient
λ_{sp}	0.05	fused sketch–photo
λ_{proto}	0.05	intra and retrieval–prototype
η	0.20	prototype–photo: $0.01/0.05$
λ_{g}	0.20	GLMA
λ_{s}	0.05	SACC
γ	0.20	SACC logit calibration
SACC neighbors	2	dynamic batchwise top-K

Both $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{tri}}$ have unit weight. Identity classification uses ordinary cross-entropy without label smoothing. The triplet term is the weighted-regularized triplet objective: positive and negative distances are softly aggregated and optimized with a soft-margin loss. It therefore has no fixed hinge margin. Although the training interface retains a legacy `margin=0.3` argument, this value is not consumed by the current triplet implementation.

The contrastive temperature is 0.07 for visual warm-up, direct sketch–photo alignment, GLMA, and SACC. The masked leave-one-out prototype contrast uses temperature 0.1. The coefficients in Table B.2 are shared by SQ and MQ. They are also retained in the single-sketch PKU formulation, where the $\lambda_{\text{proto}}\mathcal{L}_{\text{intra}}$ contribution is identically zero because $K = 1$.

B.3 Optimization and Training Schedule

All images are resized to 288×144. Training augmentation applies 10-pixel padding, random cropping back to 288×144, and random horizontal flipping. Evaluation uses only resizing and CLIP normalization. We use four data-loading workers.

Training has two consecutive stages. First, five epochs of visual contrastive warm-up optimize only $\mathcal{L}_w$ with temperature 0.07. The warm-up optimizer updates the shared CLIP visual backbone at 1×10^{-5} and the two modality adapters at 1×10^{-4}. These learning rates remain constant

throughout the five warm-up epochs. No classification, triplet, SPFC, GLMA, or SACC loss is used in this stage.

The same in-memory network is then passed directly to 50 main-stage epochs numbered 0–49; no intermediate checkpoint is serialized and reloaded. A separate main-stage AdamW is constructed before warm-up but receives no updates during that stage. Its moment estimates are consequently empty when main training begins, while the visually warmed model weights are retained. The main-stage learning rates are 1×10^{-5} for the CLIP visual backbone, 1×10^{-4} for the modality adapters, NonLocal block, and SPFC parameters, and 5×10^{-4} for the batch-normalization and classification heads. Thus, parameter groups deliberately use different learning rates in both stages.

The visual warm-up and main-stage optimizers use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 10^{-2}. The CLIP text encoder is frozen and used only to provide fixed semantic features for SACC; it has no optimizer parameter group. During the main stage, a piecewise schedule linearly scales every group from 0.1 to 1.0 of its base learning rate over epochs 0–9, keeps the base rate for epochs 10–19, and uses 0.1 of the base rate for epochs 20–49. This learning-rate warm-up is distinct from the preceding visual contrastive warm-up; no cosine scheduler is used.

GLMA is active from main-stage epoch 0. The explicit SPFC losses $\mathcal{L}_{\mathrm{sp}}$, $\mathcal{L}_{\mathrm{intra}}$, $\mathcal{L}_{\mathrm{rp}}$, and $\mathcal{L}_{\mathrm{pp}}$, together with SACC, are activated at main-stage epoch 10. The SPFC fusion path itself is already part of the MQ retrieval descriptor before epoch 10 and can receive gradients from the identity, triplet, and GLMA objectives. We do not use gradient clipping, automatic mixed precision, or a gradient scaler.

B.4 Reproducibility and Computing Infrastructure

All reported runs use deterministic execution. Python, NumPy, CPU-side data-loader workers, and all CUDA devices are seeded. We disable cuDNN benchmarking and TF32, enable deterministic cuDNN and PyTorch algorithms, set float32 matrix-multiplication precision to its highest setting, and use `CUBLAS_WORKSPACE_CONFIG=:4096:8`. The identity sampler is re-seeded by the sum of the experiment seed and current epoch, and each data-loader worker receives a deterministic derived seed.

Experiments are run on a single NVIDIA GeForce RTX 5090 GPU with 32 GB memory. The host contains two Intel Xeon Silver 4410Y processors (24 physical cores and 48 logical CPUs in total) and 256 GB system memory, and runs Ubuntu 22.04.5 LTS with Linux kernel 6.8.0. The software environment comprises Python 3.9.25, PyTorch 2.8.0, torchvision 0.23.0, CUDA 12.8, and cuDNN 9.10.2. Principal supporting packages are NumPy 2.0.2, SciPy 1.13.1, Pillow 11.3.0, OpenCV 4.13.0, scikit-learn 1.6.1, and tensorboardX 2.6.5. We use the repository-local OpenAI CLIP implementation with the ViT-B/16 visual and text checkpoints.

C Additional Experimental Results

C.1 Complete Comparison on PKU-Sketch

Table C.1 presents the complete comparison on PKU-Sketch. Our method achieves 94.21% mAP, 94.40% Rank-1, 99.80% Rank-5, 100.00% Rank-10, and 91.97% mINP. It surpasses the previous best results by 1.91 percentage points in mAP, 0.40 points in Rank-1, and 1.14 points in Rank-5, while matching the best Rank-10 performance. Among the methods reporting mINP, our method further improves over FreeStyle by 16.66 points. These results demonstrate that the proposed framework remains effective on PKU-Sketch, where only one sketch is available per identity and SPFC naturally reduces to a single-observation identity anchor.

Table C.1: Comparison with state-of-the-art methods on PKU-Sketch. R-1, R-5, and R-10 denote Rank-1, Rank-5, and Rank-10. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Venue	mAP(%)	R-1(%)	R-5(%)	R-10(%)	mINP(%)
CDAFL (Pang et al., 2018)	MM'18	–	34.00	56.30	72.50	–
LMDI (Gui et al., 2020)	Neuro.'20	–	49.00	70.40	80.20	–
CDAC (Zhu et al., 2022)	TIFS'22	–	60.80	80.60	88.80	–
BDG (Lin et al., 2023)	MM'23	66.37	70.00	–	–	–
SketchTrans^{+} (Chen et al., 2023)	TPAMI'24	–	85.80	96.00	99.00	–
DALNet (Liu et al., 2024)	AAAI'24	86.20	90.00	98.60	**100.00**	–
CGC (Hu et al., 2025a)	TIFS'25	89.60	<u>94.00</u>	98.00	**100.00**	–
FreeStyle (Wu et al., 2026)	TIP'26	83.78	87.33	<u>98.66</u>	<u>99.35</u>	<u>75.31</u>
KTCAA (Gong et al., 2026)	AAAI'26	<u>92.30</u>	92.70	–	–	–
Ours	–	**94.21**	**94.40**	**99.80**	**100.00**	**91.97**

C.2 Complete Cross-Style Generalization Results

We provide the complete cross-style comparisons under both single-query (SQ) and multi-query (MQ) protocols on Market-Sketch-1K. Following the protocol used in the main paper, the model is trained on a subset of sketch styles and directly evaluated on disjoint unseen styles. Concatenated letters denote style sets; for example, AB represents training on styles A and B, while DEF denotes testing on styles D, E, and F.

Table C.2: Complete cross-style generalization comparison on Market-Sketch-1K under the single-query (SQ) protocol. Results are reported in mAP (%). The best and second-best results across all four methods for each train–test configuration are highlighted in bold and underlined, respectively.

Train	CGC (Hu et al., 2025a)					OLTM (Li et al., 2024b)				
	F	EF	DEF	$C\cdots F$	$B\cdots F$	F	EF	DEF	$C\cdots F$	$B\cdots F$
A	9.96	7.93	7.17	7.36	7.25	<u>25.83</u>	22.41	<u>22.71</u>	23.47	22.96
AB	10.76	9.35	9.46	9.93	–	27.10	22.72	23.18	26.47	–
ABC	17.65	15.96	15.57	–	–	25.75	26.60	26.83	–	–
$A\ldots D$	23.35	19.48	–	–	–	32.60	30.07	–	–	–
$A\ldots E$	24.12	–	–	–	–	33.87	–	–	–	–

Train	FreeStyle (Wu et al., 2026)					Ours				
	F	EF	DEF	$C\cdots F$	$B\cdots F$	F	EF	DEF	$C\cdots F$	$B\cdots F$
A	25.05	**25.87**	20.10	**25.92**	<u>23.72</u>	**27.10**	<u>24.06</u>	**23.15**	<u>23.50</u>	**23.75**
AB	26.17	<u>26.22</u>	**27.91**	<u>27.99</u>	–	**30.16**	**26.96**	<u>25.78</u>	**28.22**	–
ABC	**32.90**	**31.43**	**32.54**	–	–	<u>32.58</u>	<u>30.45</u>	<u>31.14</u>	–	–
$A\ldots D$	<u>36.18</u>	**35.38**	–	–	–	**36.48**	<u>33.84</u>	–	–	–
$A\ldots E$	<u>36.93</u>	–	–	–	–	**37.10**	–	–	–	–

SQ cross-style comparison. Table C.2 reports the complete unseen-style generalization results under the SQ protocol. Our method achieves the best performance in 8 out of the 15 valid train–test configurations and ranks second in all remaining ones, indicating stable generalization across diverse unseen-style settings. The advantage is especially clear when only a small number of training styles are available. For example, our method improves $A \to F$ and $A \to DEF$ to 27.10% and 23.15% mAP, respectively, surpassing the corresponding second-best results of 25.83% and 22.71%. With two observed styles, it further improves $AB \to F$ from 27.10% to 30.16% and $AB \to EF$ from 26.22% to 26.96%.

FreeStyle performs better on several configurations with broader training style coverage, particularly $ABC \to F$, $ABC \to EF$, and $ABC \to DEF$, where our method consistently ranks second.

When the training-style coverage expands further, our method again attains the best performance on $A\cdots D\rightarrow F$ and $A\cdots E\rightarrow F$, reaching 36.48% and 37.10% mAP, respectively. Overall, these results show that the proposed framework does not rely on a specific style split, but maintains competitive generalization across different numbers and compositions of observed styles.

Table C.3: Complete cross-style generalization comparison on Market-Sketch-1K under the multi-query (MQ) protocol. Results are reported in mAP (%). Style ranges such as D–F include all intermediate styles. The best and second-best results for each train–test configuration are highlighted in bold and underlined, respectively.

Train	BDG (Lin et al., 2023)				CGC (Hu et al., 2025a)				Ours			
	EF	D–F	C–F	B–F	EF	D–F	C–F	B–F	EF	D–F	C–F	B–F
A	4.93	5.17	4.21	3.93	<u>7.77</u>	<u>8.06</u>	<u>9.10</u>	<u>9.08</u>	**25.26**	**25.92**	**30.42**	**29.15**
AB	5.62	5.63	4.27	–	<u>10.21</u>	<u>11.29</u>	<u>14.22</u>	–	**33.11**	**32.71**	**38.93**	–
ABC	6.84	6.09	–	–	<u>18.89</u>	<u>19.97</u>	–	–	**40.21**	**41.04**	–	–
$A\ldots D$	8.91	–	–	–	<u>24.94</u>	–	–	–	**43.44**	–	–	–

MQ cross-style comparison. Table C.3 presents the complete unseen-style generalization results under the MQ protocol. Our method achieves the best performance in all 10 valid train–test configurations, with CGC consistently ranking second. Compared with CGC, the improvements range from 17.49 to 24.71 percentage points. For example, our method improves $A\rightarrow C\cdots F$ from 9.10% to 30.42% mAP and $AB\rightarrow C\cdots F$ from 14.22% to 38.93% mAP.

Moreover, the performance increases steadily as more training styles are observed. On the unseen EF test set, our method improves from 25.26% with training style A alone to 33.11%, 40.21%, and 43.44% as the training-style set expands to AB, ABC, and $A\cdots D$, respectively. This trend suggests that the proposed style-conditioned prototype-fusion mechanism can effectively exploit increasingly diverse subjective observations to learn transferable identity evidence. The much larger margins under MQ further support the benefit of style-conditioned aggregation when multiple unseen-style sketches are jointly available as the query.

C.3 Backbone Analysis

We further analyze the influence of different visual backbones on Market-Sketch-1K, as shown in Table C.4. Compared with CNN-based CLIP backbones, transformer-based CLIP-ViT models achieve consistently better retrieval performance, indicating that the global self-attention mechanism is more effective for capturing identity-related patterns under large style and modality variations. Moreover, increasing the backbone capacity from CLIP-RN50 to CLIP-RN101 improves the retrieval accuracy, especially under the multi-query setting, where stronger representations facilitate the aggregation of complementary sketch cues. Based on these observations, we adopt CLIP-ViT-B/16 as the final backbone, which provides a stronger visual representation for subsequent identity consensus learning and cross-modal alignment.

Table C.4: Performance comparison of different visual backbones on Market-Sketch-1K. Results are reported under both single-query (SQ) and multi-query (MQ) protocols.

Backbone	SQ		MQ	
	mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)
CLIP-RN50	27.49	25.40	43.22	43.78
CLIP-RN101	31.38	29.96	50.61	54.02
CLIP-ViT-B/32	33.02	30.80	52.31	53.01
CLIP-ViT-B/16	**41.34**	**40.21**	**69.84**	**74.50**

C.4 HYPERPARAMETER SENSITIVITY

Table C.5: One-factor hyperparameter sensitivity on Market-Sketch-1K under the MQ protocol. One parameter is varied at a time while the others are fixed to the selected configuration. Per-metric optima and selected parameter values are highlighted in bold.

Parameter	Value	Rank-1(%)	Rank-5(%)	Rank-10(%)	mAP(%)	mINP(%)
λ_{sp}	0.025	73.90	93.37	**97.99**	68.70	56.07
	0.05	**74.50**	93.78	96.99	**69.84**	**56.96**
	0.10	73.69	**93.97**	97.59	69.02	56.72
	0.20	73.90	**93.97**	96.59	69.29	56.46
λ_{proto}	0.025	73.49	**93.78**	97.19	69.55	56.74
	0.05	**74.50**	**93.78**	96.99	69.84	**56.96**
	0.10	74.29	93.77	**97.59**	**69.88**	56.91
	0.20	72.09	92.37	**97.59**	68.18	55.54
λ_{g}	0.10	73.90	93.78	96.59	69.40	56.37
	0.20	**74.50**	93.78	96.99	**69.84**	**56.96**
	0.30	74.29	93.57	96.99	68.56	56.08
	0.50	72.49	**94.18**	**97.19**	69.03	56.53
λ_{s}	0.01	72.69	93.57	96.99	69.37	**57.03**
	0.05	**74.50**	93.78	96.99	**69.84**	56.96
	0.10	73.69	92.37	**97.99**	68.80	55.88
	0.20	73.69	**94.38**	97.39	69.37	56.53
SACC K	1	74.09	93.78	96.79	69.68	56.19
	2	**74.50**	93.78	96.99	**69.84**	**56.96**
	3	73.90	**93.98**	**97.39**	69.62	56.61
	4	73.09	93.57	96.99	69.34	56.66
Warm-up epochs	0	68.88	90.96	96.39	66.07	53.34
	5	**74.50**	**93.78**	96.99	**69.84**	**56.96**
	10	74.09	93.37	**97.39**	69.20	56.36

Table C.5 reports a one-factor sensitivity analysis under the MQ protocol on Market-Sketch-1K. Each parameter is varied independently while the remaining parameters are fixed to the selected configuration. Overall, the model remains relatively stable around moderate loss weights, whereas excessively weak or strong regularization generally reduces the retrieval performance.

For SPFC, $\lambda_{\text{sp}} = 0.05$ achieves the best Rank-1, mAP, and mINP, while larger or smaller values provide only isolated gains at higher ranks. For λ_{proto}, the setting of 0.10 produces a marginally higher mAP than 0.05 by 0.04 percentage points, whereas 0.05 yields the best Rank-1 and mINP and therefore provides a better overall balance. Increasing λ_{proto} to 0.20 causes a clear degradation, indicating that overly strong prototype regularization may restrict retrieval-oriented discrimination.

For the auxiliary objectives, $\lambda_{\text{g}} = 0.20$ and $\lambda_{\text{s}} = 0.05$ achieve the best mAP and Rank-1. Although larger weights occasionally improve Rank-5 or Rank-10, they reduce the more discriminative Rank-1 and mAP metrics. SACC performs best overall with $K = 2$ semantic neighbors. Increasing K to 3 improves the higher-rank accuracy but slightly decreases mAP and Rank-1, suggesting that including additional, less relevant semantic neighbors may weaken the calibration of genuinely ambiguous identities.

The visual warm-up is also important. Compared with training without warm-up, five warm-up epochs improve mAP from 66.07% to 69.84% and Rank-1 from 68.88% to 74.50%. Extending the warm-up to ten epochs provides no further improvement in the principal metrics. These results support the selected configuration of $\lambda_{\text{sp}} = 0.05$, $\lambda_{\text{proto}} = 0.05$, $\lambda_{\text{g}} = 0.20$, $\lambda_{\text{s}} = 0.05$, $K = 2$, and five warm-up epochs.

D Additional Visualization Results

D.1 Intra-Class and Inter-Class Distance Distributions

Figure D.1 compares the cosine-distance distributions of same-identity and different-identity pairs under the SQ and MQ protocols. We quantify their separation using the mean gap $d = \mu_{\text{inter}} - \mu_{\text{intra}}$, where a larger value indicates clearer separation between different identities relative to within-identity variation.

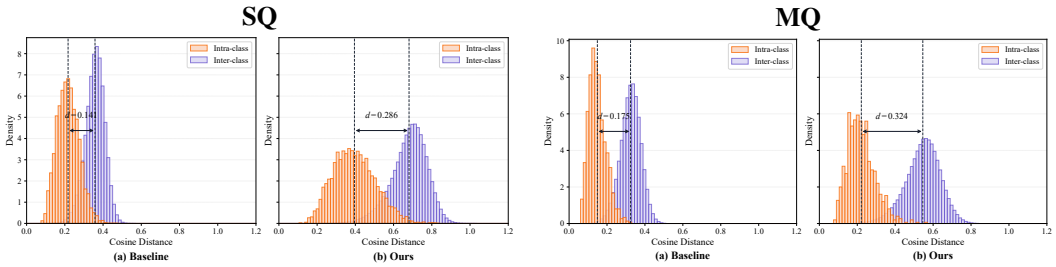

Figure D.1: Intra-class and inter-class cosine-distance distributions under SQ (left pair) and MQ (right pair). Each pair shows the baseline and our method. Dashed lines indicate the corresponding mean distances, and $d = \mu_{\text{inter}} - \mu_{\text{intra}}$ denotes their separation.

Under the SQ protocol, the baseline obtains mean intra-class and inter-class distances of 0.217 and 0.358, respectively, resulting in a mean gap of 0.141. The full model shifts these means to 0.395 and 0.681 and enlarges the gap to 0.286. The separation therefore increases by 0.145 and is more than twice that of the baseline. This result indicates that the proposed training objectives substantially strengthen identity discrimination even when only one sketch is available at inference.

A similar trend is observed under the MQ protocol. The baseline produces mean intra-class and inter-class distances of 0.149 and 0.324, yielding a gap of 0.175. The full model shifts these means to 0.221 and 0.546, respectively, expanding the gap to 0.324. The resulting improvement of 0.149 shows that style-conditioned consensus learning and multi-sketch fusion produce a more discriminative embedding geometry by separating different identities more strongly relative to within-identity variation.

Overall, the enlarged mean gaps under both protocols demonstrate that the proposed framework consistently improves the relative separation between same-identity and different-identity pairs. This more discriminative embedding structure is consistent with the retrieval gains observed under both SQ and MQ settings.