SEMANTICS WITHIN, RELATIONS BETWEEN: STRUCTURING VISION-LANGUAGE REPRESENTATIONS FOR PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent CLIP-based person re-identification (ReID) methods increasingly exploit language-aligned semantics through textual prompts, descriptions, or image-derived pseudo-words. However, image-derived semantics are typically constructed only after visual encoding, limiting their interaction with visual representation formation, while conventional sample-level objectives do not explicitly organize the relative ordering of hard competing identities. These limitations can leave fine-grained visual evidence underexploited and ambiguous identity neighborhoods insufficiently structured for retrieval. We present **SWR**, a framework that addresses these two limitations through *Semantics Within* and *Relations Between*. Its In-Encoder Visual Semantic Tokenization (IVST) introduces learnable semantic tokens into the visual Transformer, allowing them to evolve jointly with image patches, and collects their intermediate states across visual layers to construct image-conditioned pseudo-word representations. Its Identity-Relational Ranking (IRR) aggregates observations into identity prototypes and applies list-wise supervision over hard identity neighborhoods in the fused pre-BN representation. On Market-1501 and MSMT17, SWR improves the PromptSG baseline by 2.0 and 1.7 mAP points, reaching 96.6% and 88.9% mAP, respectively. It further achieves 72.7% mAP on Occluded-Duke and 86.0% on CUHK03-D. Compared with IADT under matched training settings, SWR improves mAP by 0.4 and 0.5 points on Market-1501 and MSMT17 while using 29.1% fewer parameters and reducing training time by 29.6–33.6%. Controlled analyzes further validate multi-layer state selection, relational supervision location, hard-identity mining, and retrieval-boundary behavior.

1 INTRODUCTION

Person re-identification (ReID) aims to retrieve the same person across non-overlapping camera views. Large intra-identity variations and subtle inter-identity differences make successful retrieval depend not only on holistic identity discrimination but also on fine-grained appearance cues such as clothing, carried objects, body regions, and local visual patterns (Zheng et al., 2015; Hermans et al., 2017; Luo et al., 2019; He et al., 2021b). Vision–language pre-training provides a natural source of such semantic information. CLIP (Radford et al., 2021) learns a shared image–text embedding space, and recent ReID methods exploit this knowledge through prompt learning, textual descriptions, or visual-to-text inversion (Li et al., 2022; Zhai et al., 2023; Han et al., 2024; Yang et al., 2024; Shang et al., 2025). These advances demonstrate the value of language-aligned representations for person retrieval.

Despite this progress, two aspects remain underexplored in image-driven vision-language ReID. First, existing pseudo-word based approaches generally derive language-side semantics from representations that have already passed through the visual encoder. Although such modules can influence visual features through end-to-end training, they do not introduce semantic tokens that interact with patch tokens during the visual forward pass or expose how image-conditioned representations evolve across visual layers. Second, standard identity classification and metric-learning objectives primarily optimize individual discrimination or pairwise separation, whereas ReID is ultimately evaluated

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.

by the relative ordering of multiple competing identities. These observations motivate us to investigate both *within-instance representation formation* and *between-identity relational organization*.

To model representation formation within an instance, we introduce **In-Encoder Visual Semantic Tokenization (IVST)**. Instead of constructing all pseudo-words from post-encoder visual features, IVST inserts a small number of Visual Semantic Tokens (VSTs) directly into the visual Transformer. The VSTs participate in self-attention with image patches throughout the encoder, and their intermediate states across visual layers form an image-conditioned candidate pool. Selected states are projected into the CLIP language space and mapped to pseudo-word embeddings, which are combined with a global identity pseudo-word and processed by the frozen CLIP text encoder. IVST therefore allows language-side representations to be constructed from states that are formed jointly with the visual hierarchy, rather than exclusively from already encoded visual outputs.

To organize relations between identities, we introduce **Identity-Relational Ranking (IRR)**. Rather than applying another sample-level metric constraint, IRR aggregates multiple observations of each identity into retrieval and textual prototypes and constructs directional relations among competing identities. It then applies listwise ranking supervision to hard identity neighborhoods in the fused pre-BN representation, where visual and textual information have already interacted. For datasets with large identity spaces, a detached prototype memory enlarges the candidate set beyond the current mini-batch, allowing IRR to discover harder competing identities. In this way, IRR regularizes how ambiguous identities are relatively organized instead of forcing heterogeneous visual and textual representations to match feature values directly.

Together, IVST and IRR form **SWR—Semantics Within, Relations Between**. IVST determines how image-conditioned pseudo-word representations are formed from states evolving inside the visual encoder, while IRR uses the resulting textual representation as a semantic reference for organizing hard identity relations in the fused retrieval branch. Figure 1 illustrates the complete framework. Our main contributions are summarized as follows:

- We present **SWR** to address two limitations of existing vision–language ReID: post-encoder semantic construction and insufficient modeling of relations among competing identities. SWR jointly structures within-instance representations and between-identity relations.
- We introduce **IVST**, which lets VSTs evolve with image patches throughout the visual encoder and uses their multi-layer states to construct image-conditioned pseudo-words. Full-hierarchy state collection outperforms restricted layer ranges without requiring external captions or attribute annotations.
- We introduce **IRR**, which applies listwise supervision to hard identity neighborhoods in the fused pre-BN space and enlarges the candidate set with cross-batch prototypes. ListMLE consistently outperforms MSE, KL, and RankNet under the same relational setting.
- SWR achieves 96.6% and 88.9% mAP on Market-1501 and MSMT17, improving PromptSG by 2.0 and 1.7 points, respectively. Compared with IADT, it also uses 29.1% fewer parameters and 29.6–33.6% less training time while achieving higher accuracy.

2 Related Work

Visual representation learning for person ReID. Person ReID has traditionally been formulated as discriminative visual representation learning. Identity classification and metric-learning objectives provide strong supervision for separating identities (Zheng et al., 2015; Hermans et al., 2017; Luo et al., 2019), while subsequent work improves discrimination through part-based representations, semantic regions, attention mechanisms, and multi-scale features (Sun et al., 2018; Wang et al., 2018; Zhou et al., 2019; Wang et al., 2020a; Zhu et al., 2020; Zhang et al., 2020). Transformer-based approaches further model long-range interactions among visual tokens and strengthen global-local representation learning (He et al., 2021a; Zhu et al., 2022; 2024; Li et al., 2023). These methods establish strong identity-discriminative visual representations. SWR builds on this foundation but introduces language-aligned representations during visual encoding and explicitly regularizes relations among competing identities.

Vision-language learning for person ReID. CLIP (Radford et al., 2021) enables ReID models to exploit knowledge acquired from large-scale image–text pre-training. CLIP-ReID (Li et al., 2022) learns identity-specific textual prompts, while MP-ReID (Zhai et al., 2023) and CLIP-SCGI (Han et al., 2024) incorporate richer pedestrian descriptions through multiple prompts, attribute text, or synthesized captions. Such explicit textual supervision can provide rich cross-modal information, but requires additional annotations or caption-generation pipelines and may introduce descriptions that are not reliable for every image. A complementary direction derives language-side representations directly from visual inputs. PromptSG (Yang et al., 2024) maps a global pedestrian representation to an instance-specific pseudo-word, while IADT (Shang et al., 2025) further separates identity and attribute-related representations using encoded visual features.

These ReID approaches are closely related to broader prompt-learning and visual-to-text inversion methods. CoOp (Zhou et al., 2022b) learns continuous textual contexts, CoCoOp (Zhou et al., 2022a) conditions prompts on individual images, and MaPLe (Khattak et al., 2023) introduces coupled visual and textual prompts across multiple stages. Textual inversion and image-to-word mapping instead represent visual concepts directly in the language embedding space (Gal et al., 2022; Saito et al., 2023; Baldrati et al., 2023). Visual Prompt Tuning (VPT) (Jia et al., 2022) inserts learnable tokens into a visual Transformer primarily for parameter-efficient adaptation.

IVST is related to these approaches but differs in how the additional visual tokens are used. Rather than constructing pseudo-words only from visual features after encoding, IVST inserts VSTs at the visual input and tracks their evolving states throughout the encoder. Selected intermediate states are then projected into the CLIP language space and mapped to pseudo-word embeddings. Unlike conventional visual prompting, the VST states therefore serve not only as learnable visual tokens but also as inputs to the subsequent language-side representation construction.

Relational and ranking-based representation learning. Metric learning commonly optimizes pairwise or tuple-based relations between samples, with triplet loss being particularly influential in ReID (Hermans et al., 2017). Supervised contrastive learning extends positive–negative comparisons over a mini-batch (Khosla et al., 2021), while ranked-list objectives directly optimize relative ordering among examples (Wang et al., 2019). Listwise learning-to-rank methods further model the ordering of an entire candidate set (Xia et al., 2008).

IRR adopts the listwise perspective but applies it to identity-level relations in ReID. Multiple observations are first aggregated into identity prototypes, and relational supervision is constructed over hard competing identities rather than individual image pairs. Retrieval and textual prototypes induce within- and cross-representation relation lists, whose relative orderings are used for supervision. For large identity spaces, cross-batch prototype memory follows the general principle of enlarging the effective comparison set beyond the current mini-batch (Wang et al., 2020b); in IRR, it is used specifically to broaden the candidate identity set available for hard-neighbor mining.

3 Methodology

3.1 Overview

SWR consists of two components that operate at different stages of the vision–language ReID pipeline. In-Encoder Visual Semantic Tokenization (IVST) inserts learnable VSTs into the CLIP visual encoder and collects their intermediate states across visual layers. Selected states are projected into the language space and mapped to pseudo-word embeddings, which are combined with a global identity pseudo-word and encoded by the frozen CLIP text encoder.

The resulting textual representation is then combined with the visual token sequence through the fixed cross-modal interaction module. Identity-Relational Ranking (IRR) operates on the fused representation immediately after this interaction and before the BN neck. It aggregates multiple observations into identity prototypes and applies listwise relational supervision over hard competing identities. The complete framework is illustrated in Fig. 1.

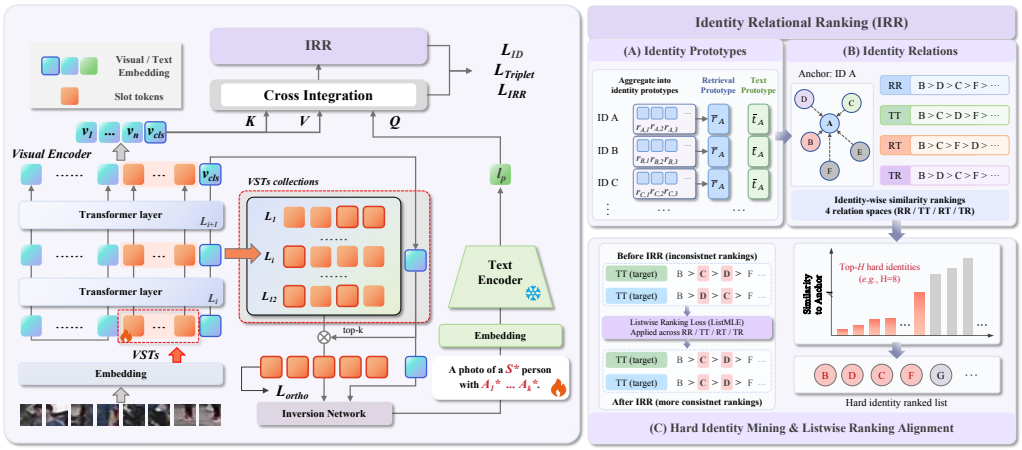

Figure 1: Overview of the proposed **SWR** framework. IVST constructs instance-conditioned semantics within the visual encoder by collecting and selecting visual semantic token states across visual depths, while IRR organizes relations between identities through identity prototypes, hard identity mining, and listwise ranking alignment.

3.2 Semantics Within: In-Encoder Visual Semantic Tokenization

Existing pseudo-word based ReID methods typically construct language-side representations from visual features after encoding. IVST instead inserts additional learnable tokens into the visual encoding process itself. Given an input image, M Visual Semantic Tokens (VSTs) are appended to the class token and N patch tokens before entering the CLIP visual Transformer:

$$\mathbf{Z}^0 = [\mathbf{z}^0_{\text{cls}}, \mathbf{z}^0_1, \ldots, \mathbf{z}^0_N, \mathbf{s}^0_1, \ldots, \mathbf{s}^0_M].$$

The VSTs participate in self-attention together with the class and patch tokens throughout the encoder. We collect their intermediate states from all L visual Transformer layers, producing $M \times L$ candidate states rather than using only the final-layer outputs. Each candidate is projected into the shared CLIP embedding space and scored by its dot-product similarity with the projected global visual representation. The top K states are retained for pseudo-word construction.

The global visual representation is mapped to an identity pseudo-word S^*, while the selected VST states are mapped to VST-derived pseudo-words $\{A^*_i\}_{i=1}^K$. These embeddings are inserted into the continuous prompt *"A photo of a S^* person with $A^*_1, \ldots, A^*_K$."* and processed by the frozen CLIP text encoder to obtain an image-conditioned textual representation.

The two types of pseudo-words are constructed from complementary visual sources: S^* is derived from the global visual representation, whereas $\{A^*_i\}$ are derived from selected VST states collected across visual layers. We do not impose predefined attribute labels on the VST-derived pseudo-words; their role is learned jointly through the downstream ReID objectives.

To reduce redundancy among the selected VST states, we additionally apply a diversity regularizer during training. Its formulation, together with the identity classification, triplet, and vision–language contrastive objectives, is provided in Appendix A.1. The textual representation is subsequently combined with the visual token sequence through the fixed cross-modal interaction module, described in Appendix A.2. .

3.3 Relations Between: Identity-Relational Ranking

ReID is evaluated by ranking gallery identities relative to a query, whereas standard classification and metric objectives do not explicitly supervise the ordering among multiple competing identities. IRR introduces an additional identity-level relational objective for this purpose.

Let $\mathbf{r}_i$ denote the fused representation of image i immediately after cross-modal interaction and before the BN neck, and let $\mathbf{t}_i$ denote its textual representation. For identity c with N_c observations, we aggregate the two representations into identity prototypes:

$$\bar{\mathbf{r}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{r}_i, \qquad \bar{\mathbf{t}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{t}_i.$$

After normalization, retrieval and textual prototypes define four directional relation scores:

$$R^{ab}_{c,j} = \mathrm{sim}(\bar{\mathbf{a}}_c, \bar{\mathbf{b}}_j), \qquad a, b \in \{r, t\}, \quad j \neq c,$$

yielding R^{rr}, R^{tt}, R^{rt}, and R^{tr}. Here, RR and TT measure within-representation identity relations, while RT and TR measure directional cross-representation relations.

For each anchor identity, self-relations are excluded and the H highest-ranked competing identities under a teacher relation are retained. Given teacher relation R^t, let

$$\pi^t = \mathrm{argsort}_{\downarrow}(R^t)$$

denote the descending ordering of the selected candidate identities. Given a student relation R^s, we optimize the normalized ListMLE objective and aggregate the directional objectives as

$$\mathcal{L}_{\mathrm{rank}}(R^s, \pi^t) = -\frac{1}{H} \sum_{j=1}^{H} \log \frac{\exp\left(R^s_{\pi^t_j}/\tau\right)}{\sum_{k=j}^{H} \exp\left(R^s_{\pi^t_k}/\tau\right)}, \qquad \mathcal{L}_{\mathrm{IRR}} = \frac{1}{|\mathcal{P}|} \sum_{(s,t)\in\mathcal{P}} \mathcal{L}_{\mathrm{rank}}(R^s, \pi^t).$$

Here, τ controls ranking sharpness and $\mathcal{P}$ denotes the directional teacher–student relation pairs used by IRR. The same candidate identities selected by the teacher relation are evaluated in the corresponding student relation, so the objective supervises their relative ordering rather than requiring different relation spaces to share identical similarity values.

IRR is applied to the fused pre-BN representation immediately after cross-modal interaction. At this point, the textual representation has already interacted with the visual token sequence, while the retrieval-specific BN transformation has not yet been applied. The choice of supervision location is examined experimentally in Sec. 4.4.

For datasets with a large identity space, we maintain a detached FIFO memory of identity prototypes to enlarge the candidate set available for hard-neighbor mining. Historical prototypes do not receive gradients and serve only as additional candidate identities. The complete IRR training procedure, including prototype aggregation, hard-neighbor selection, stop-gradient operations, and memory updates, is provided in Appendix A.3.

3.4 Training Objective

SWR retains the identity classification, triplet, and supervised vision–language objectives used by the baseline. IVST additionally introduces a diversity regularizer over the selected VST states, while IRR contributes the relational ranking objective defined above. Complete loss definitions, weights, and optimization schedules are provided in Appendix A.1 and Appendix A.4.

4 Experiments

We evaluate SWR on standard and challenging ReID benchmarks and analyze the key design choices of IVST and IRR. Additional diagnostics are provided in the appendix. Unless otherwise stated, results are reported without re-ranking or test-time adaptation.

4.1 Experimental Setup

We evaluate SWR on Market-1501 (Zheng et al., 2015), MSMT17 (Wei et al., 2018), Occluded-Duke (Miao et al., 2019; 2021), and CUHK03-D (Li et al., 2014) under the detected new protocol (Zhong et al., 2017). We report mAP and Rank-1 accuracy. SWR uses CLIP ViT-B/16 with a frozen text encoder. Unless otherwise stated, we use $M = 2$, $K = 16$, $H = 8$, and $\tau = 0.07$. MSMT17 employs a detached prototype memory of size 256, whereas Market-1501 uses the candidate identities available in the current training context. Complete optimization, augmentation, loss-weight, and memory settings are provided in Appendix A.4.

Table 1: Comparison with representative state-of-the-art methods on Market-1501, MSMT17, Occluded-Duke, and CUHK03-D. CUHK03-D denotes the detected setting under the new protocol. "–" indicates that a directly comparable result was not reported.

Backbone	Method	Reference	Market-1501		MSMT17		Occluded-Duke		CUHK03-D	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
CNN	OSNet	ICCV'19	84.9	94.8	52.9	78.7	–	–	67.8	72.3
	HOReID	CVPR'20	84.9	94.2	–	–	43.8	55.1	–	–
	RGA-SC	CVPR'20	88.4	96.1	57.5	80.3	–	–	74.5	79.6
	ISP	ECCV'20	84.9	94.2	–	–	52.3	62.8	71.4	75.2
	CDNet	CVPR'21	86.0	95.1	54.7	78.9	–	–	–	–
	CAL	ICCV'21	87.0	94.5	56.2	79.5	–	–	–	–
	ALDER	TIP'21	88.9	95.6	59.1	82.5	–	–	–	–
	LTReID	TMM'22	86.9	94.7	58.6	81.0	–	–	–	–
ViT	TransReID	ICCV'21	88.9	95.2	67.4	85.3	59.2	66.4	–	–
	DCAL	CVPR'22	87.5	94.7	64.0	83.1	–	–	–	–
	AAFormer	TNNLS'23	88.0	95.4	65.6	84.4	58.2	67.1	77.2	78.1
	PHA	CVPR'23	90.2	96.1	68.9	86.1	–	–	80.3	83.2
CLIP	CLIP Baseline	–	86.4	93.3	66.1	84.4	53.5	60.8	–	–
	CLIP-ReID	AAAI'23	89.6	95.5	73.4	88.7	59.5	67.1	78.6	80.5
	PromptSG	CVPR'24	94.6	97.0	87.2	92.6	–	–	–	–
	CLIP-SCGI	arXiv'24	96.0	97.6	88.2	92.9	59.2	67.5	–	–
	MP-ReID	AAAI'24	95.5	97.7	–	–	–	–	–	–
	FLaN-Net	IJCAI'25	90.5	95.7	–	–	65.5	75.2	84.8	**87.1**
	IADT*	MM'25	96.2	97.5	88.4	92.9	70.9	74.6	–	–
	SWR	–	**96.6**	**98.3**	**88.9**	**93.3**	**72.7**	**75.4**	**86.0**	87.0

* denotes our re-implementation.

4.2 Overall Performance

Table 1 compares SWR with representative CNN-based methods, including OSNet (Zhou et al., 2019), HOReID (Wang et al., 2020a), RGA-SC (Zhang et al., 2020), ISP (Zhu et al., 2020), CDNet (Li et al., 2021), CAL (Rao et al., 2021), ALDER (Zhang et al., 2022), and LTReID (Wang et al., 2023); Transformer-based methods, including TransReID (He et al., 2021a), DCAL (Zhu et al., 2022), AAFormer (Zhu et al., 2024), and PHA (Zhang et al., 2023); and recent vision–language approaches, including CLIP-ReID (Li et al., 2022), PromptSG (Yang et al., 2024), CLIP-SCGI (Han et al., 2024), MP-ReID (Zhai et al., 2023), FLaN-Net (Chen et al., 2025), and IADT (Shang et al., 2025). Across all four benchmarks, SWR achieves the strongest mAP among the directly comparable methods: 96.6% on Market-1501, 88.9% on MSMT17, 72.7% on Occluded-Duke, and 86.0% on CUHK03-D. These results span standard, large-scale, detector-noisy, and heavily occluded retrieval conditions. Importantly, SWR achieves this performance without manually annotated pedestrian attributes or externally generated captions.

4.3 What Forms the Semantics Within?

We next examine how the design of IVST affects the representations used for pseudo-word construction. We study three aspects: how many VST states are needed, from which visual layers the selected states arise, and whether the resulting textual representation changes the subsequent cross-modal interaction.

Semantic capacity. We first study the interaction between the number of inserted VSTs M and the number of retained states K under the full SWR setting. As shown in Fig. 2(a), the two-dimensional M–K search exhibits a broad high-performing region rather than a monotonic improvement as either quantity increases. In particular, simply introducing more VSTs does not consistently improve retrieval, indicating that semantic capacity is governed by the balance between the number of semantic carriers and the number of retained states. We therefore adopt $M = 2$ and $K = 16$, which achieves 96.6% mAP and 98.3% Rank-1 under the full SWR setting.

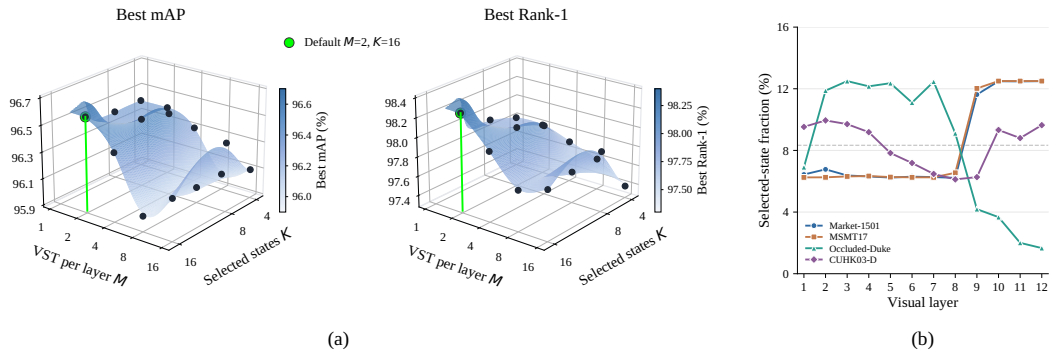

Figure 2: **Semantic capacity and layer-wise state selection.** (a) M–K sensitivity of SWR on Market-1501. (b) Layer-wise selection distributions across four benchmarks.

Layer-wise state selection. Figure 2(b) shows that the selected VST states are not restricted to the final visual layer, and that their depth distribution differs substantially across datasets. Market-1501 and MSMT17 favor states from later visual layers, while CUHK03-D exhibits a broader mixture of intermediate and deep states. Occluded-Duke presents a markedly different pattern, with substantially greater use of shallow and intermediate layers and reduced selection from the deepest layers.

These observations show that the states retained by IVST arise from different parts of the visual hierarchy under different retrieval conditions. The shift on Occluded-Duke is consistent with a greater reliance on local or intermediate visual evidence when complete holistic appearance information is less available; however, the layer statistics themselves do not assign a specific human-interpretable semantic role to individual VST states.

Because IVST uses a raw dot product for candidate scoring, we additionally check whether the observed layer preferences could be explained by layer-dependent feature norms. Recomputing the selection with cosine similarity preserves the major cross-layer patterns, whereas selecting candidates according to feature norm alone does not reproduce the observed distributions. The layer preferences therefore cannot be explained solely by differences in feature magnitude. Detailed diagnostics are provided in Appendix B.4.

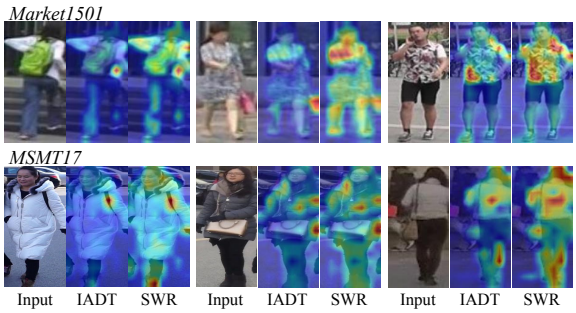

Figure 3: Effect of IVST on cross-modal interaction.

Effect on cross-modal interaction. Finally, we examine how IVST-derived semantics propagate into cross-modal fusion. In the fixed interaction module, the textual representation serves as the query while visual tokens provide keys and values, making cross-attention a direct view of the visual evidence involved in multimodal interaction. Figure 3 shows that replacing the post-hoc semantic construction of IADT with IVST consistently changes the spatial distribution of attended visual evidence across Market-1501 and MSMT17. IVST therefore affects not only the language-side representation, but also the subsequent visual evidence used during cross-modal integration.

4.4 How Are Relations Organized Between Identities?

We next analyze the design of IRR from three perspectives: where the relational objective should be applied, which competing identities provide useful supervision, and how the candidate identity context affects hard-neighbor discovery. We additionally examine in the appendix whether the resulting relational regularization propagates to the final positive–negative retrieval boundary.

Where should relational supervision act? We first compare three candidate supervision locations along the retrieval pipeline. Applying IRR to the projected visual representation yields 96.3% mAP and 97.9% Rank-1 on Market-1501, while supervising the post-BN retrieval representation obtains 96.3% / 98.0%. Applying IRR to the fused pre-BN representation immediately after cross-modal interaction reaches 96.6% / 98.3%, compared with 96.4% / 98.2% for IVST without IRR.

We further examine representation geometry using centered kernel alignment (CKA) (Kornblith et al., 2019). As shown in Fig. 4, the fused pre-BN representation has substantially stronger alignment with the textual representation than the other candidate retrieval spaces. On Market-1501, the corresponding CKA is 0.880, compared with 0.581 for the projected visual CLS and 0.257 for the post-BN representation. Occluded-Duke shows the same pattern, with values of 0.894, 0.474, and 0.238, respectively. Thus, cross-modal interaction produces a fused representation whose geometry is strongly aligned with the textual representation, while the subsequent BN transformation substantially changes that geometry.

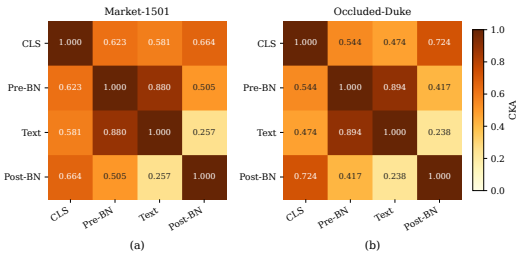

Figure 4: **Representation alignment measured by CKA.** Pairwise CKA among CLS, fused pre-BN, text, and post-BN representations on (a) Market-1501 and (b) Occluded-Duke.

Setting	Representation	Top-8	ρ	τ
No IRR	Projected visual	0.665	0.846	0.665
No IRR	Fused pre-BN	0.392	0.333	0.227
IRR	Projected visual	0.565	0.611	0.452
IRR	**Fused pre-BN**	**0.418**	**0.780**	**0.607**

Table 2: Relation-space localization on Market-1501. ρ and τ denote Spearman and Kendall correlations, respectively.

A complementary rank-based diagnostic measures whether identities are ordered similarly across representation spaces. Spearman's ρ measures agreement between complete rankings, whereas Kendall's τ measures pairwise ordering agreement (Spearman, 1904; Kendall, 1938). As shown in Table 2, applying IRR to the fused pre-BN representation increases its Spearman correlation with the textual identity ordering from 0.333 to 0.780 and its Kendall correlation from 0.227 to 0.607. Together with the supervision-location ablation and the CKA analysis, these results support the fused pre-BN representation as the most effective supervision location under the current IRR design.

Which identities provide useful relational supervision? ReID errors are not distributed uniformly over all negative identities. For a given query, the highest-ranked incorrect identities constitute the competitors most likely to interfere with retrieval. IRR therefore focuses its relational objective on the local hard-identity neighborhood rather than treating all negative identities equally.

For each anchor identity, we compare the Top-8 highest-similarity competing identities with randomly sampled and easiest negatives. Across all four relation spaces, their mean similarities follow *hard* > *random* > *easy*: 0.852 / 0.633 / 0.263 for RR, 0.997 / 0.972 / 0.916 for TT, 0.042 / −0.198 / −0.303 for RT, and 0.536 / −0.200 / −0.598 for TR. The absolute scales differ substantially across relation spaces, especially between TT and the cross-representation relations. This motivates organizing relative identity orderings rather than directly requiring heterogeneous relation scores to take the same numerical values.

(a) Market-1501: hard identities		
Scope	mAP	Rank-1
All identities	96.4	98.0
Hard Top-8	**96.6**	**98.3**
(b) MSMT17: candidate context		
Context	mAP	Rank-1
No relational ranking	87.2	92.6
Mini-batch	88.1	92.7
Memory-256	**88.9**	**93.3**

Table 3: Controlled analysis of hard identities and candidate context.

As shown in Table 3(a), restricting relational supervision to the hard Top-8 identities reaches 96.6% mAP and 98.3% Rank-1, compared with 96.4% / 98.0% when all candidate identities are used. Under the evaluated setting, focusing on the hard neighborhood therefore provides a more useful relational constraint than uniformly supervising the full candidate set.

Although IRR directly regularizes relations among competing identities rather than explicit positive–negative pairs, we further examine whether this effect propagates to the final retrieval boundary. The

analysis in Appendix C.3 shows that the downstream effect is concentrated on ambiguous queries, where IRR improves mINP, hardest-positive rank, and the positive–hard-negative retrieval margin.

Why is broader identity context necessary? MSMT17 contains 1,041 training identities, while the $P16K4$ sampler exposes only 16 identities in a mini-batch, corresponding to 1.54% of the training identity set. A 256-entry prototype memory increases the observed candidate coverage to 23.56% on average.

More importantly, only 5.66–9.76% of the Top-8 hard identities identified from the broader memory-based candidate set are present in the current mini-batch across the four relation spaces. Thus, more than 90% of the hard competitors identified from this broader candidate context would be unavailable to a batch-only objective. The memory-based candidate set also yields harder Top-8 neighborhoods while leaving the median similarity of the overall negative population nearly unchanged; detailed statistics are reported in Appendix C.2.

The retrieval results in Table 3(b) follow the same trend. Under the relational-context control setting, the reference model without relational ranking obtains 87.2% mAP and 92.6% Rank-1 on MSMT17. Mini-batch IRR improves this result to 88.1% / 92.7%, while enlarging the candidate context with Memory-256 reaches 88.9% / 93.3%. These results support the usefulness of a broader candidate identity context for hard-neighbor discovery under the evaluated IRR configuration.

Taken together, the current analyses show that IRR benefits from supervision at the fused pre-BN representation, from concentrating on hard competing identities, and from exposing the relational objective to a broader candidate set. We separately compare ListMLE with similarity regression, distribution matching, and pairwise ranking under the same hard-identity setting in Appendix C.1.

4.5 Additional Analysis

Training efficiency. SWR also provides a favorable accuracy–efficiency trade-off relative to IADT under matched training settings. SWR contains 95M parameters compared with 134M for IADT, a reduction of 29.1%. Training time decreases from 3563s to 2510s on Market-1501 and from 9547s to 6341s on MSMT17, corresponding to reductions of 29.6% and 33.6%, respectively. At the same time, mAP increases from 96.2% to 96.6% on Market-1501 and from 88.4% to 88.9% on MSMT17. Complete hardware, batch-size, timing, and model-size statistics are reported in Appendix D.

Effect under occlusion. We additionally analyze relational supervision under severe occlusion. Adding IRR to the baseline without IVST improves Occluded-Duke from 72.3% / 76.0% to 73.5% / 77.1% without introducing pose estimation, visibility prediction, part parsing, or an occlusion-specific test-time matching procedure. Detailed controlled results and comparisons on Occluded-Duke and Occluded-REID are provided in Appendix E.

5 Conclusion

We presented **SWR**, a framework for structuring vision-language representations in person ReID through *Semantics Within* and *Relations Between*. IVST constructs image-conditioned pseudo-word representations from VST states evolving across the visual hierarchy, while IRR organizes hard competing identities through listwise relational ranking and broader identity context. Experiments across standard and challenging ReID benchmarks demonstrate strong retrieval performance, with controlled analyses validating multi-layer state selection and hard-identity relational supervision. SWR further achieves higher accuracy than IADT with substantially fewer parameters and lower training cost.

Reproducibility Statement

We provide the main architectural and optimization details required to reproduce SWR in Sec. 4.1 and the supplementary material. Appendix A specifies the complete training objectives, cross-modal interaction module, IRR training procedure, optimization settings, and prototype-memory configuration. Additional controlled analyses of IVST and IRR are provided in Appendices B and C. Training hardware, batch size, and computational cost are reported in Appendix D.

Ethics Statement

Person ReID can support legitimate retrieval and research applications, but it may also facilitate privacy-invasive surveillance when deployed without appropriate safeguards. Practical deployment should therefore require a lawful and clearly defined purpose, restricted system access, appropriate dataset governance, and evaluation of demographic and environmental biases. Our method does not introduce demographic labels or infer sensitive identity attributes. Nevertheless, improved retrieval capability does not eliminate the broader privacy and misuse risks associated with person ReID systems.

AI use statement

Generative AI tools were used primarily to assist with language editing and polishing of the manuscript, including improving grammar, clarity, readability, and presentation. They were also used to assist in organizing and refining the presentation of experimental analyses based on results obtained by the authors. The proposed methodology, implementation, experiments, and reported results were developed, conducted, and verified by the authors. All AI-assisted content was critically reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

References

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Minghui Chen, Dayan Wu, Chenxu Yang, Qinghang Su, and Zheng Lin. Categorical attention: Fine-grained language-guided noise filtering network for occluded person re-identification. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 801–809. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/90. URL https://doi.org/10.24963/ijcai.2025/90. Main Track.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Qianru Han, Xinwei He, Zhi Liu, Sannyuya Liu, Ying Zhang, and Jinhai Xiang. Clip-scgi: Synthesized caption-guided inversion for person re-identification, 2024. URL https://arxiv.org/abs/2410.09382.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15013–15022, October 2021a.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *IEEE International Conference on Computer Vision*, pp. 15013–15022, 2021b.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017. URL https://arxiv.org/abs/1703.07737.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727, 2022.

Mengxi Jia, Yifan Sun, Yunpeng Zhai, Xinhua Cheng, Yi Yang, and Ying Li. Semi-attention partition for occluded person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 998–1006, 2023. doi: 10.1609/aaai.v37i1.25180.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning, 2023. URL https://arxiv.org/abs/2210.03117.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. URL https://arxiv.org/abs/2004.11362.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Dengjie Li, Siyu Chen, Yujie Zhong, and Lin Ma. Dip: Learning discriminative implicit parts for person re-identification, 2023. URL https://arxiv.org/abs/2212.13906.

Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6729–6738, 2021.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. *arXiv preprint arXiv:2211.13977*, 2022.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019.

Jiaxu Miao, Yu Wu, and Yi Yang. Identifying visible parts via pose estimation for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification, 2021. URL https://arxiv.org/abs/2108.08728.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval, 2023. URL https://arxiv.org/abs/2302.03084.

Rui Shang, Min Liu, Xueping Wang, Yuan Bian, and Yaonan Wang. Decoupled identity and attribute tokenization for person re-identification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 6213–6222, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755580. URL https://doi.org/10.1145/3746027.3755580.

Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *Computer Vision – ECCV 2024*, volume 15137 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024. doi: 10.1007/978-3-031-72986-7_13.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), 2018. URL https://arxiv.org/abs/1711.09349.

Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5070–5078, 2024. doi: 10.1609/aaai.v38i5.28312.

Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6449–6458, 2020a.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, MM '18, pp. 274–282. ACM, October 2018. doi: 10.1145/3240508.3240552. URL http://dx.doi.org/10.1145/3240508.3240552.

Pingyu Wang, Zhicheng Zhao, Fei Su, and Hongying Meng. Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification. *IEEE Transactions on Multimedia*, 25:4610–4622, 2023. doi: 10.1109/TMM.2022.3179902.

Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M. Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5207–5216, 2019.

Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6388–6397, 2020b.

Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4754–4763, 2022.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1192–1199, 2008.

Jiaer Xia, Lei Tan, Pingyang Dai, Mingbo Zhao, Yongjian Wu, and Liujuan Cao. Attention disturbance and dual-path constraint network for occluded person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6198–6206, 2024. doi: 10.1609/aaai.v38i6.28437.

Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re- identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17343–17353, 2024. doi: 10.1109/CVPR52733.2024.01642.

Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification, 2023. URL https://arxiv.org/abs/2312.16797.

Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14133–14142, June 2023.

Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Transactions on Image Processing*, 31:352–365, 2022. doi: 10.1109/TIP.2021.3128330.

Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3186–3195, 2020.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding, 2017. URL https://arxiv.org/abs/1701.08398.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *IEEE International Conference on Computer Vision*, pp. 3702–3712, 2019.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022a. URL https://arxiv.org/abs/2203.05557.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022b.

Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification, 2022. URL https://arxiv.org/abs/2205.02151.

Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pp. 346–363, 2020.

Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification, 2024. URL https://arxiv.org/abs/2104.00921.

A Additional Method and Implementation Details

A.1 Training Objectives and Regularization

For completeness, we provide the training objectives omitted from the main paper. SWR preserves the identity-discriminative and vision–language supervision of the baseline, while IVST introduces an additional diversity regularizer over the selected VST states and IRR contributes the relational ranking objective described in the main paper.

Identity classification loss. Given a visual representation $\mathbf{v}_i$ with identity label y_i, the standard identity classification objective is

$$\mathcal{L}_{\text{id}} = -\frac{1}{B}\sum_{i=1}^{B} \log p(y_i \mid \mathbf{v}_i),$$

where B denotes the mini-batch size and $p(y_i \mid \mathbf{v}_i)$ is the predicted probability of the ground-truth identity.

Triplet loss. For an anchor $\mathbf{v}_i$, its hardest positive $\mathbf{v}_i^+$ and hardest negative $\mathbf{v}_i^-$ within the mini-batch are used to define

$$\mathcal{L}_{\text{tri}} = \frac{1}{B}\sum_{i=1}^{B} \left[m + d(\mathbf{v}_i, \mathbf{v}_i^+) - d(\mathbf{v}_i, \mathbf{v}_i^-)\right]_+ ,$$

where m is the triplet margin, $d(\cdot,\cdot)$ denotes the feature distance, and $[\cdot]_+ = \max(0, \cdot)$.

Vision–language contrastive loss. Following the baseline, the visual and textual representations are jointly optimized using a supervised vision–language contrastive objective. Let $\mathbf{v}_i$ and $\mathbf{t}_j$ denote normalized visual and textual representations, respectively, and let

$$s_{ij} = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\tau_{\text{vl}}}$$

denote their temperature-scaled similarity. The image-to-text and text-to-image objectives are written as

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j:y_j=y_i} \exp(s_{ij})}{\sum_{j=1}^{B} \exp(s_{ij})},$$

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{j:y_j=y_i} \exp(s_{ji})}{\sum_{j=1}^{B} \exp(s_{ji})},$$

and the symmetric vision–language objective is

$$\mathcal{L}_{\text{sup}} = \frac{1}{2}\left(\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}}\right).$$

VST-state diversity regularization. IVST selects K VST states from the multi-layer candidate pool. To reduce redundancy among the retained states, we apply an orthogonality-based diversity regularizer. Let $\tilde{\mathbf{A}} \in \mathbb{R}^{K\times d}$ denote the row-normalized selected VST representations. We define

$$\mathcal{L}_{\text{orth}} = \left\| \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top - \mathbf{I} \right\|_F^2 .$$

This objective discourages highly correlated selected states without requiring predefined pedestrian attribute labels.

Overall objective. The baseline objective is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{id}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}}.$$

The complete SWR objective further incorporates the VST-state diversity regularizer and Identity-Relational Ranking:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{IRR}}\mathcal{L}_{\text{IRR}}.$$

IRR is disabled during the first five training epochs, linearly warmed up during the following five epochs, and subsequently optimized with its target weight $\lambda_{\text{IRR}} = 0.1$. The remaining optimization settings are reported in Appendix A.4.

A.2 Cross-Modal Interaction Module

The cross-modal interaction module is inherited from the baseline architecture and kept unchanged across the controlled comparisons. It is therefore not a contribution of SWR and is described here only for completeness.

Let $\mathbf{V} \in \mathbb{R}^{B\times 129\times 512}$ denote the projected visual token sequence, containing one class token and 128 patch tokens, and let $\mathbf{t} \in \mathbb{R}^{B\times 1\times 512}$ denote the textual representation. Cross-modal interaction is implemented using multi-head attention with the textual representation as query and the visual sequence as keys and values:

$$\mathbf{Q} = \mathbf{t}, \qquad \mathbf{K} = \mathbf{V}, \qquad \mathbf{V}_{\text{attn}} = \mathbf{V}.$$

We use eight attention heads. The resulting attention tensor has shape $B \times 8 \times 1 \times 129$. The resulting fused representation is subsequently passed through the same downstream normalization and retrieval pipeline for all compared methods.

IRR is applied to the fused representation immediately after cross-modal interaction and before the subsequent BN-neck transformation. Thus, the performance changes attributed to IRR do not arise from modifying the interaction architecture itself.

Cross-attention visualization. For Fig. 3, we extract the cross-attention weights under the same visualization protocol for IADT and SWR. The class-token component is removed before spatial visualization, and the remaining 128 patch weights are reshaped according to the 16×8 patch grid. Attention heads are aggregated consistently for both models.

A.3 IRR Training Procedure

Algorithm 1 summarizes the training procedure of IRR. IRR operates on identity-level prototypes rather than individual observations. For each mini-batch, fused pre-BN and textual representations belonging to the same identity are first aggregated into prototypes. Hard competing identities are then selected from the available candidate set, and listwise relational supervision is applied over the resulting identity neighborhood. On MSMT17, this candidate set is augmented with detached prototypes stored from preceding mini-batches.

Directional supervision. The set $\mathcal{P}$ contains the directional teacher–student relation pairs used by IRR. For each pair, the teacher relation determines the candidate identities and their target ordering, while the student relation is evaluated on exactly the same identities. IRR therefore transfers relative identity ordering rather than regressing absolute similarity values across heterogeneous representation spaces.

Gradient flow. The hard-identity indices and target ordering obtained from the teacher relation are treated as non-differentiable. Gradients propagate through the student relation scores and current-batch representations. Historical prototypes stored in the cross-batch memory are detached and therefore serve only as additional candidates.

Cross-batch candidate set. Market-1501 uses candidate identities available in the current training context. MSMT17 additionally uses a detached FIFO prototype memory of size 256. The memory affects only the candidate set available for hard-neighbor discovery and introduces no additional inference-time ranking module.

A.4 Implementation Details

Training configuration. All experiments use a pretrained CLIP ViT-B/16 visual encoder with the CLIP text encoder kept frozen. Input images are resized to 256×128 and augmented with random horizontal flipping, padding, random cropping, and random erasing. Models are optimized for 60 epochs using Adam. Each mini-batch contains 64 images sampled from 16 identities with four observations per identity. The learning rate is 5×10^{-6} for the pretrained visual encoder and 5×10^{-5} for randomly initialized modules, and is decayed by a factor of 0.1 at epochs 20 and 40.

Algorithm 1 Identity-Relational Ranking (IRR)

Require: Mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}$; hard-neighborhood size H; temperature τ; directional relation pairs $\mathcal{P}$; optional prototype memory $\mathcal{M}$
Ensure: IRR loss $\mathcal{L}_{\mathrm{IRR}}$

Obtain fused pre-BN representations $\mathbf{r}_i$ and textual representations $\mathbf{t}_i$
for each identity c appearing in $\mathcal{B}$ **do**
    $\bar{\mathbf{r}}_c \leftarrow \mathrm{Norm}\Big(\frac{1}{N_c}\sum_{i:y_i=c}\mathbf{r}_i\Big)$
    $\bar{\mathbf{t}}_c \leftarrow \mathrm{Norm}\Big(\frac{1}{N_c}\sum_{i:y_i=c}\mathbf{t}_i\Big)$
end for
$\mathcal{C} \leftarrow$ BUILDCANDIDATEPOOL$(\{(\bar{\mathbf{r}}_c, \bar{\mathbf{t}}_c, c)\}, \mathcal{M})$
$\mathcal{L}_{\mathrm{IRR}} \leftarrow 0$
for each anchor identity c in the current mini-batch **do**
    Remove self-identity candidates from $\mathcal{C}$
    for each directional pair $(s, t) \in \mathcal{P}$ **do**
        Compute teacher relation scores $R^t_{c,j}$ for all candidate identities j
        $\mathcal{H}^t_c \leftarrow \mathrm{TopH}(R^t_{c,:}, H)$
        $\pi^t \leftarrow \mathrm{StopGrad}\left(\mathrm{argsort}_{\downarrow}(R^t_{c,\mathcal{H}^t_c})\right)$
        Compute student relation scores $R^s_{c,\mathcal{H}^t_c}$ on the same candidate identities
        $\mathcal{L}_{\mathrm{IRR}} \leftarrow \mathcal{L}_{\mathrm{IRR}} + \mathcal{L}_{\mathrm{rank}}(R^s_{c,\mathcal{H}^t_c}, \pi^t; \tau)$
    end for
end for
Normalize $\mathcal{L}_{\mathrm{IRR}}$ by the number of anchors and directional relation pairs
if prototype memory is enabled **then**
    Detach current identity prototypes
    $\mathcal{M} \leftarrow$ UPDATEFIFO$(\mathcal{M}, \{(\bar{\mathbf{r}}_c, \bar{\mathbf{t}}_c, c)\})$
end if
return $\mathcal{L}_{\mathrm{IRR}}$

Table 4: Ablation study on the two pseudo-word components of IVST.

VSTs	ID	Market-1501		MSMT17	
		mAP	Rank-1	mAP	Rank-1
–	–	92.0	96.3	85.3	91.6
✓	–	96.1	97.5	87.3	92.6
–	✓	94.6	97.0	87.2	92.6
✓	✓	**96.4**	**98.2**	**88.9**	**93.3**

IVST configuration. The default configuration inserts $M = 2$ VSTs into the input token sequence and collects their intermediate states from all 12 visual Transformer layers, yielding 24 candidate VST states. Candidates are projected into the CLIP embedding space and ranked by raw dot-product similarity with the projected global visual representation. The top $K = 16$ states are retained for pseudo-word construction.

IRR configuration. Identity observations are aggregated into normalized prototypes before relational ranking. We use $H = 8$ hard competing identities, $\tau = 0.07$, and a target IRR weight of $\lambda_{\mathrm{IRR}} = 0.1$. IRR is disabled during the first five epochs, linearly warmed up during the following five epochs, and kept at its target weight thereafter.

Prototype memory. For MSMT17, we maintain a detached FIFO prototype memory with capacity 256. Historical prototypes are stop-gradient representations and are used only to expand the candidate identity context for IRR. IRR is a training-time objective and introduces no additional inference-time ranking module.

B Additional Analysis of IVST

B.1 Component Ablation

Table 4 isolates the contributions of the VST-derived pseudo-word branch and the global identity pseudo-word. Both components improve over the visual-only configuration, while their combination gives the strongest retrieval performance on both Market-1501 and MSMT17.

B.2 Layer-Range Ablation

The layer-selection statistics in the main paper show that Market-1501 selects VST states from later visual layers more frequently. This raises a natural question: are states from the remaining layers still necessary? We therefore restrict the candidate pool to several predefined layer ranges while keeping the remaining IVST configuration unchanged. All layer-range variants, including the full-layer reference, are evaluated within the same controlled ablation setting. The full-layer result therefore serves as the matched reference for this experiment.

As shown in Table 5, all restricted variants perform worse than collecting candidate states from the complete visual hierarchy. In particular, using only the deep layers does not match the full-layer configuration despite their higher average selection frequency on Market-1501. Removing the intermediate layers also causes a measurable performance drop. These results indicate that selection frequency does not directly reflect layer necessity: less frequently selected states can still provide useful candidates for particular instances. Collecting VST states across all visual layers therefore provides a more complete candidate set than restricting semantic construction to a predefined depth range.

Layer range	mAP	Rank-1
Shallow only (L1–4)	96.2	97.4
Deep only (L9–12)	96.2	97.6
Shallow + deep (L1–4, 9–12)	96.3	97.7
All layers (L1–12)	**96.4**	**98.0**

Table 5: Effect of restricting the visual layers used for VST-state collection on Market-1501.

B.3 Complete Semantic-Capacity Grid

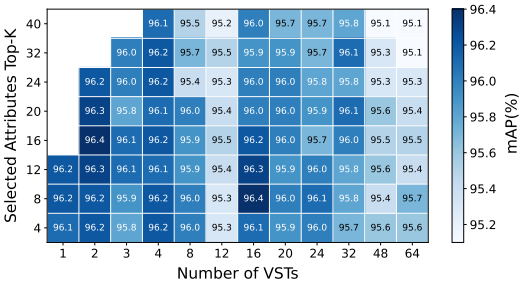

Figure 5: Complete M–K sensitivity analysis for IVST without IRR. Blank entries denote invalid configurations.

Figure 5 provides the complete $M \times K$ sweep for IVST without IRR. This experiment is separated from the full-SWR capacity study in the main paper and isolates the effect of VST capacity itself. Performance remains high over a broad region but does not improve monotonically with either the number of inserted VSTs or the number of retained states. Increasing token capacity alone is therefore insufficient; retrieval performance depends on the joint choice of M and K.

B.4 Controlling for Feature Norms in VST Selection

IVST ranks projected VST states using their raw dot-product similarity with the projected global visual representation. Since

$$\mathbf{a}^\top \mathbf{g} = \|\mathbf{a}\| \, \|\mathbf{g}\| \cos\theta,$$

the score can in principle be affected by layer-dependent feature magnitude. We therefore examine whether the layer-selection distributions reported in the main paper can be explained primarily by differences in VST feature norms.

For each visual layer, we record the mean ℓ_2 norm of projected VST candidates and recompute Top-K selection using three criteria: the original raw dot product, cosine similarity, and candidate norm alone. The cosine criterion removes the candidate-magnitude term, while the norm-only criterion provides a direct control for a purely magnitude-driven selection rule.

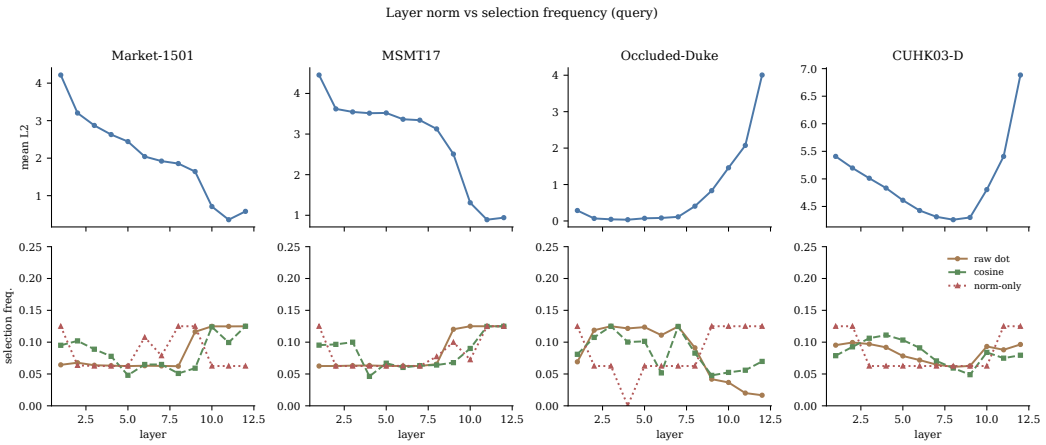

Figure 6: Feature-norm control for layer-wise VST selection.

As shown in Fig. 6, feature magnitude varies substantially across visual layers, but these variations do not directly explain the observed selection distributions. On Market-1501 and MSMT17, candidate norms generally decrease toward later layers, whereas the original raw-dot-product selection increasingly favors later states. Occluded-Duke provides a complementary case: the deepest layers exhibit rapidly increasing feature norms, while their raw-dot-product selection frequency decreases.

Removing the magnitude term through cosine similarity preserves the major cross-layer tendencies, whereas ranking candidates by feature norm alone produces substantially different distributions. Feature magnitude can therefore affect individual raw dot-product scores, but the layer-selection patterns observed in the main paper cannot be reduced to a simple preference for high-norm candidates.

C Additional Analysis of IRR

C.1 Comparison of Relational Objectives

To distinguish the effect of listwise ordering from that of generic relational supervision, we compare IRR with alternative objectives under the same hard-identity setting. The supervision location, candidate identities, Top-H neighborhood, and training configuration are kept fixed; only the relational objective is changed. We compare direct similarity regression (MSE), distribution matching (KL), pairwise ranking (RankNet), and the listwise ListMLE objective used by IRR.

As shown in Table 6, ListMLE provides the strongest overall performance on both datasets. On Market-1501, the differences among relational objectives are relatively small, although ListMLE still achieves the best result of 96.6% mAP and 98.3% Rank-1. The distinction becomes clearer on the larger MSMT17 identity space: ListMLE reaches 88.9%/93.3%, compared with 88.3%/92.8% for MSE, 87.7%/92.6% for KL, and 87.5%/92.4% for RankNet.

Objective	Market-1501		MSMT17	
	mAP	Rank-1	mAP	Rank-1
MSE	96.5	97.8	88.3	92.8
KL	96.4	97.9	87.7	92.6
RankNet	96.4	98.0	87.5	92.4
ListMLE	**96.6**	**98.3**	**88.9**	**93.3**

Table 6: Comparison of relational objectives under the same hard-identity setting.

These controlled comparisons indicate that the benefit of IRR cannot be attributed solely to introducing relational supervision over hard identities. Under the same candidate setting, supervising the ordering of the complete hard-identity list is more effective than directly regressing relation scores, matching similarity distributions, or decomposing the neighborhood into pairwise preferences. The larger gap on MSMT17 further suggests that listwise organization becomes particularly useful when the identity space and competing neighborhoods are more complex.

C.2 Cross-Batch Hard-Identity Coverage

We further quantify how prototype memory changes the candidate identity context available to IRR on MSMT17. The analysis is performed using a frozen checkpoint and a deterministic replay of 200 training batches under the $P16K4$ sampler. Each mini-batch contains 16 identities, whereas the memory contains 245.25 unique identities on average.

Relation	Top-8 visible	Batch Top-8	Memory Top-8
RR	9.76%	0.646	**0.745**
TT	5.70%	0.989	**0.996**
RT	5.77%	0.002	**0.022**
TR	5.66%	0.013	**0.053**

Table 7: Effect of cross-batch relational context on hard-identity discovery. "Visible" denotes the fraction of the broader Top-8 neighborhood already present in the current mini-batch.

As shown in Table 7, only 5.66–9.76% of the Top-8 identities identified from the broader memory-based candidate set are already present in the current mini-batch. The broader candidate context also increases the mean similarity of the selected Top-8 competitors in all four relation spaces.

Importantly, the median similarity over the complete negative population remains nearly unchanged between batch-only and memory-based candidate sets. The prototype memory therefore primarily expands the hard tail available to IRR rather than uniformly shifting the overall negative distribution.

C.3 Does Relational Organization Improve the Retrieval Boundary?

IRR directly regularizes the relative organization of competing identities rather than explicitly optimizing the positions of positive gallery samples. We therefore examine whether this relational regularization propagates to the actual positive–negative retrieval boundary.

For a query q, let $\mathcal{P}_q$ and $\mathcal{N}_q$ denote its valid positive and negative gallery samples under the standard ReID evaluation protocol. We define the positive–hard-negative margin as

$$m(q) = \min_{p \in \mathcal{P}_q} s(q,p) - \max_{n \in \mathcal{N}_q} s(q,n),$$

where $s(\cdot,\cdot)$ denotes cosine similarity. A negative margin indicates that at least one negative sample is ranked ahead of the weakest positive. We additionally examine mINP and the rank of the hardest positive, i.e., the last valid positive in the ranked gallery.

For Market-1501, this comparison isolates the effect of IRR under the same retrieval architecture. The MSMT17 IRR configuration additionally uses the Memory-256 candidate context and should therefore be interpreted as evidence for the complete relational training configuration rather than an IRR-only causal comparison.

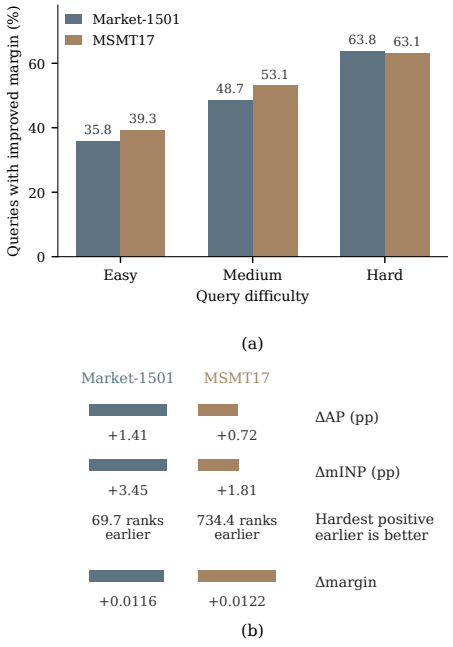

Figure 7: **Effect of IRR on difficult retrieval boundaries.** (a) Queries with harder baseline margins are more likely to obtain an improved positive–hard-negative margin. (b) On queries with baseline margin $m(q) < 0$, relational supervision improves AP, mINP, hardest-positive rank, and retrieval margin.

As shown in Fig. 7(a), the downstream effect of relational supervision is concentrated on more difficult queries. On Market-1501, the fraction of queries whose positive–hard-negative margin improves rises from 35.8% for easy queries to 63.8% for hard queries. MSMT17 shows the same trend, increasing from 39.3% to 63.1%.

We further focus on queries whose baseline margin is already negative. Figure 7(b) shows that the effect is not confined to internal reordering among negative identities. On Market-1501, AP and mINP increase by 1.41 and 3.45 percentage points, respectively, while the hardest positive moves

Table 8: Comparison of training efficiency on Market-1501 and MSMT17.

Method	Params	Training Time	Time↓	mAP
Market-1501				
CLIP-ReID	89M	4689s	1.00×	89.6
PromptSG	94M	2417s	0.52×	94.6
CLIP-SCGI*	113M	4451s	0.95×	96.0
IADT	134M	3563s	0.76×	96.2
SWR	95M	2510s	0.54×	**96.6**
MSMT17				
CLIP-ReID	89M	12904s	1.00×	73.4
PromptSG	94M	6108s	0.47×	87.2
CLIP-SCGI†	113M	13534s	1.05×	88.2
IADT	134M	9547s	0.74×	88.4
SWR	95M	6341s	0.49×	**88.9**

† = our re-implementation

forward by 69.7 ranks on average and the retrieval margin increases by 0.0116. On MSMT17, the corresponding improvements are 0.72 and 1.81 percentage points, the hardest positive moves forward by 734.4 ranks, and the margin increases by 0.0122.

For Market-1501, 58 queries move from a negative to a positive margin after IRR. Among Rank-1 errors, 14 previously incorrect queries are corrected, whereas only three previously correct queries become incorrect. These results show that the effect of relational regularization can propagate beyond the relative ordering of negative identities to the placement of valid positives near difficult retrieval boundaries.

The effect is substantially smaller on already well-separated queries. IRR should therefore be interpreted as a relational regularizer whose downstream benefit is concentrated around ambiguous identity neighborhoods rather than as a direct surrogate for the final retrieval metric.

D Training Efficiency

All efficiency experiments are conducted on a single NVIDIA RTX 3090 GPU with a batch size of 64 under the same 60-epoch training protocol. Training time is measured over the complete training procedure. Table 8 compares parameter count, training time, and retrieval accuracy.

Compared with IADT, SWR reduces the parameter count from 134M to 95M (29.1% fewer parameters), while reducing training time by 29.6% on Market-1501 and 33.6% on MSMT17. At the same time, mAP increases from 96.2% to 96.6% on Market-1501 and from 88.4% to 88.9% on MSMT17.

SWR also remains within approximately 4% of PromptSG's training time while improving mAP by 2.0 percentage points on Market-1501 and 1.7 percentage points on MSMT17.

IRR is used only during training and introduces no additional inference-time ranking module or IRR-specific parameter. Its contribution therefore incurs no additional IRR-specific inference stage.

E Analysis under Occlusion

E.1 Effect of IRR on Occluded-Duke

Occlusion provides a challenging setting for examining relational supervision because incomplete observations increase ambiguity among visually similar identities. We therefore isolate IRR on Occluded-Duke by comparing the same baseline model with and without relational supervision. Here, **Baseline** denotes the model without IVST or IRR, while **Baseline + IRR** adds only the IRR objective.

Adding IRR improves mAP from 72.3% to 73.5% and Rank-1 from 76.0% to 77.1%, corresponding to gains of 1.2 and 1.1 percentage points. IRR introduces no pose estimation, visibility prediction,

Table 9: Comparison with representative occluded-ReID methods in terms of mAP.

Method	Reference	Occluded-Duke	Occluded-REID
PGFA (Miao et al., 2019)	ICCV'19	37.3	–
HOReID (Wang et al., 2020a)	CVPR'20	43.8	70.2
FED (Wang et al., 2022)	CVPR'22	56.4	79.3
SAP (Jia et al., 2023)	AAAI'23	62.2	76.8
ADP (Xia et al., 2024)	AAAI'24	63.8	85.1
DPM-SPT (Tan et al., 2024)	AAAI'24	63.0	81.1
KPR (Somers et al., 2024)	ECCV'24	67.1	79.1
SWR	–	**72.7**	**88.0**

body-part parsing, or occlusion-specific test-time matching. The result therefore shows that identity-level relational supervision can improve retrieval under severe occlusion without introducing an occlusion-specific architecture.

E.2 Comparison with Occlusion-Specific Methods

We further compare the complete SWR model with representative methods designed for occluded person ReID.

As shown in Table 9, SWR achieves 72.7% mAP on Occluded-Duke and 88.0% on Occluded-REID. Unlike many occlusion-specific approaches, SWR does not require explicit body-part parsing, pose estimation, visibility prediction, or specialized test-time matching. Together with the controlled IRR result in Table 10, these results show that the proposed framework remains competitive under severe occlusion without introducing a dedicated occlusion-handling pipeline.

Method	mAP	Rank-1
Baseline	72.3	76.0
Baseline + IRR	**73.5**	**77.1**

Table 10: Effect of IRR on Occluded-Duke.