INTERACTIVE SEMANTIC VERIFICATION FOR UNSUPERVISED DOMAIN ADAPTATION IN TEXT-TO-IMAGE PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation (UDA) has received increasing attention in text-to-image person re-identification (TIReID), as it enables models to adapt to an unlabeled target domain without costly annotations. However, under domain shift, existing feature-based adaptation methods face reliability challenges in both target domain pseudo correspondence construction and cross-domain representation learning. To this end, we propose an Interactive Semantic Verification framework (ISV) that leverages a multimodal large language model (MLLM). Specifically, to alleviate noisy pseudo image-text correspondences in the target domain caused by unreliable feature similarity, we develop Interactive Multi-evidence Learning (IML). It first performs balanced reciprocal matching to identify candidate correspondences and then employs an MLLM for semantic verification, leveraging external semantic knowledge to compensate for unreliable feature evidence, thereby removing mismatched pairs and recovering missed positive matches. Moreover, to improve cross-domain representation learning, we introduce Density-filtered Intermediate-domain Learning (DIL), which identifies and filters extreme feature outliers in low-density regions and explicitly models the difference between source and target features to construct more reliable intermediate representations. Extensive experiments on three TIReID benchmarks demonstrate that ISV consistently improves cross-domain retrieval performance and generalizes well across different source models.

1 INTRODUCTION

Text-to-image person re-identification (TIReID) retrieves pedestrian images using natural-language descriptions (Li et al., 2017; Zhang & Lu, 2018; Niu et al., 2020). In practice, however, a model trained on one dataset often faces a different target environment, where cameras, scenes, pedestrian appearances, and description styles may all change across datasets. Unsupervised domain adaptation (UDA) aims to transfer the source model without using target annotations (Long et al., 2015; Ganin et al., 2016; Fu et al., 2019; Liang et al., 2020). This setting is particularly challenging for TIReID because the source domain provides paired images and texts with identity labels, whereas the target domain provides neither identity labels nor the original image-text correspondences. Adaptation must therefore recover cross-modal alignment from unpaired target data. The reliability of the recovered correspondences is therefore crucial to subsequent adaptation. However, domain shift makes this recovery unreliable. As shown in Figure 1(a)(b), feature matching can select image-text pairs with clear semantic conflicts, while also missing valid but difficult pairs. Consequently, target pseudo supervision can be both incorrect and incomplete.

Many cross-domain retrieval methods (Jing et al., 2020; Hao et al., 2023; Luo et al., 2025; Niu et al., 2026; Gu et al., 2026) rely on learned features for adaptation. Specifically, some reduce the discrepancy between source and target feature distributions (Jing et al., 2020; Hao et al., 2023), while others mine target pseudo correspondences from feature similarities (Hao et al., 2023; Gu et al., 2026). Despite their different objectives, both strategies depend heavily on the current feature space. This dependence becomes particularly problematic under strong domain shift, because the same representation that has become unreliable is also used to construct the supervision for correcting itself (Ge et al., 2020a;b). As a result, feature errors can then affect adaptation in two ways: (1) they can distort

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.

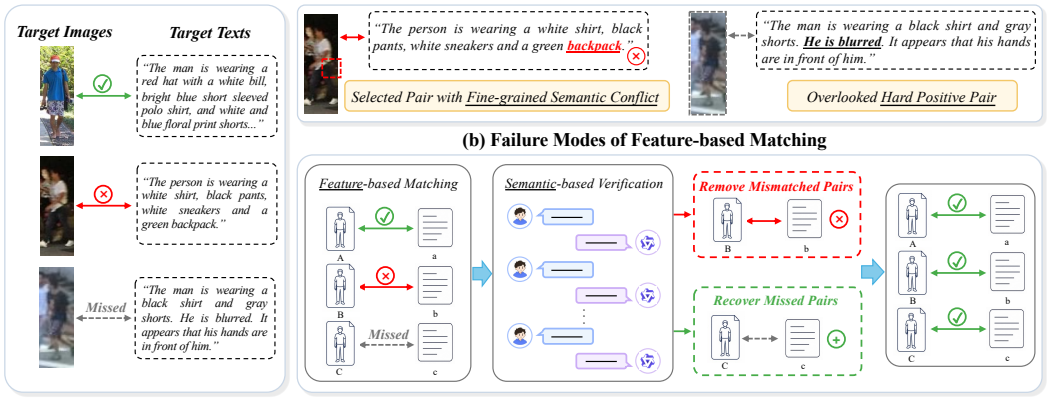


Figure 1: Motivation for complementary semantic verification in target correspondence learning. (a) Early feature-based matching identifies useful target image-text pairs, but can also produce incorrect or incomplete correspondences. (b) Two representative failure modes are shown: a selected pair with a fine-grained semantic conflict and an overlooked hard positive. (c) MLLM-assisted semantic verification provides complementary evidence to filter semantically conflicting pairs and recover plausible missed matches.

cross-domain representation modeling; and (2) they can introduce incorrect or incomplete pseudo pairs. These two effects may further reinforce each other during adaptation. More importantly, the problem is therefore deeper than ordinary pseudo-label noise: a shifted representation is effectively asked to certify its own adaptation supervision.

This observation motivates a reliability-oriented view of adaptation. On the one hand, feature similarity remains useful because it provides an efficient way to propose candidate correspondences (Hao et al., 2023). On the other hand, it should not be the only evidence used to judge whether a pair is reliable. In this context, additional semantic evidence can provide a complementary signal. A frozen multimodal large language model (MLLM) (Qin et al., 2025) offers one such source: it can inspect visible pedestrian attributes and compare them with textual descriptions, while its judgment is not produced by the target retrieval embedding itself. We do not assume that the MLLM is always correct. Instead, we use it as complementary evidence when feature-based matching is uncertain. Beyond correspondence learning, representation learning requires a similar treatment, since extreme and unreliable features should not dominate cross-domain modeling. Accordingly, reliable adaptation requires complementary reliability control at both the correspondence and representation levels.

Based on this principle, we propose Interactive Semantic Verification (ISV) for UDA TIReID. ISV improves reliability at both the correspondence and representation levels. Specifically, as illustrated in Figure 1(c), ISV introduces complementary semantic verification to filter conflicting pseudo pairs and recover plausible missed matches. For target correspondence learning, it first constructs candidates through balanced reciprocal matching, after which a frozen MLLM verifies selected candidates. Meanwhile, for cross-domain representation learning, ISV filters extreme low-density features before intermediate-domain modeling and explicitly models source-to-target feature differences during feature mixing.

We evaluate ISV on three public TIReID benchmarks and multiple cross-dataset transfer settings. The results show that ISV consistently improves cross-domain retrieval. Moreover, the gains remain when ISV is applied to different source models, including Implicit Relation Reasoning and Aligning (IRRA) (Jiang & Ye, 2023), Cross-Modal Adaptive Dual Association (CADA) (Lin et al., 2024), and Robust Dual Embedding (RDE) (Qin et al., 2024). This consistency suggests that the framework is not tied to a particular source architecture or initial representation. Furthermore, component studies show complementary gains from correspondence refinement, cross-domain representation modeling, and visual refinement.

Our contributions are threefold. First, we formulate unreliable feature self-supervision as an important challenge in UDA TIReID. Second, we develop a reliability-oriented framework that combines feature-based adaptation with complementary semantic verification. Third, we demonstrate consistent improvements across datasets, transfer directions, and source models.

2 Related Work

2.1 Text-to-Image Person Re-Identification

TIReID enables person retrieval through natural language descriptions, removing the need for image queries and making person search more flexible in practical scenarios. Existing methods can be broadly divided into three categories: conventional cross-modal alignment, vision-language pretraining, and approaches assisted by MLLMs. Conventional methods (Wang et al., 2020; Bai et al., 2023) learn discriminative image and text representations through fine-grained cross-modal alignment. LGUR (Shao et al., 2022) further addresses differences in feature granularity to reduce the modality gap. In contrast, vision-language pretraining methods (Jia et al., 2021; Li et al., 2022; Yao et al., 2022), such as Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), exploit pretrained image and text encoders to learn transferable and aligned cross-modal representations, while RDE improves TIReID learning by reducing the influence of noisy image-text correspondences (Qin et al., 2024). More recently, MLLM-assisted methods exploit richer semantic knowledge to generate diverse descriptions (Tan et al., 2024) or refine text queries through interaction (Qin et al., 2025). However, most existing methods focus on cross-modal alignment or direct transfer, while the lack of image–text correspondence in the target domain poses a major challenge to these methods. To address this challenge, ISV introduces external semantic knowledge from MLLM to complement unreliable feature evidence, thereby constructing more reliable pseudo correspondences in the unlabeled target domain.

2.2 Unsupervised Domain Adaptation for Text-to-Image Person Re-Identification

UDA reduces the need for costly target annotations by transferring knowledge from labeled source data to unlabeled target domains. In TIReID, however, domain shifts in visual appearance and textual descriptions, together with the absence of target image-text correspondence labels, make cross-modal adaptation particularly challenging (Niu et al., 2026). Early approaches such as the Moment Alignment Network (MAN) (Jing et al., 2020) alleviate domain and modality discrepancies through distribution alignment, but such alignment cannot guarantee reliable instance-level correspondence. To capture richer cross-domain relationships, Graph-Based Cross-Domain Knowledge Distillation (GCKD) (Luo et al., 2025) propagates information between source and target samples and transfers knowledge through contrastive distillation. Dynamic Mutual Alignment (DMA) (Gu et al., 2026) further coordinates visual and textual representations to reduce both domain and modality discrepancies. However, these methods primarily rely on learned features for domain adaptation, without sufficiently considering the reliability of feature-based target domain pseudo correspondences and cross-domain representation learning. In contrast to these methods, ISV improves cross-domain representations by filtering outlier features and explicitly modeling cross-domain differences. Moreover, it also complements unreliable target feature evidence with external MLLM semantics, thereby providing more reliable target pseudo correspondences.

3 Methodology

3.1 Notations and Problem Statement

For clarity, the labeled source domain is defined as $\mathcal{D}_s = \{(I_i^s, T_i^s, y_i^s)\}_{i=1}^{N_s}$, and the unlabeled target domain is defined as $\mathcal{D}_t = \{\{I_i^t\}_{i=1}^{N_t^I}, \{T_j^t\}_{j=1}^{N_t^T}\}$, where I, T, and y denote an image, text, and source identity. Source pairs and identities are labeled, whereas target identities and correspondences are unknown. Our goal is to adapt the source model to the target domain without target identity labels and image-text correspondences. Figure 2 shows ISV, which contains three modules: 1) Interactive Multi-evidence Learning (IML) checks feature-matched pairs with a frozen MLLM. 2) Density-filtered Intermediate-domain Learning (DIL) removes feature outliers and uses domain differences to build intermediate features. 3) Text-Guided Spatial Gating (TSG) uses source text to train an image-only gate that suppresses visual cues unrelated to identity.

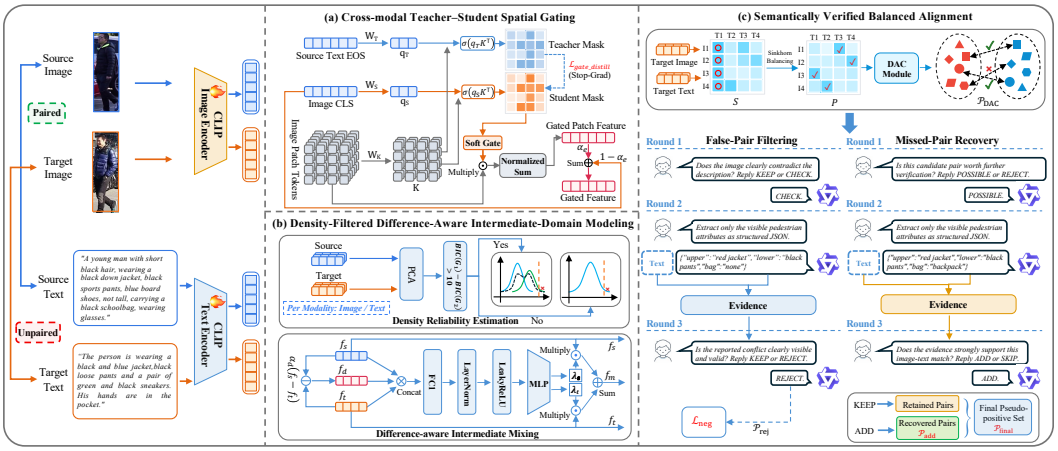


Figure 2: Overview of the proposed ISV framework for cross-dataset UDA TIReID. (a) TSG transfers source-text guidance to an image-only student through spatial gating, enhancing identity-relevant visual representations. (b) DIL filters low-density feature outliers and performs difference-aware intermediate-domain learning to improve cross-domain feature reliability. (c) IML combines Sinkhorn-balanced reciprocal matching with MLLM-based semantic verification to filter mismatched pairs and recover missed positive correspondences, improving target pseudo-supervision. Together, these three components enable more effective domain adaptation for TIReID.

3.2 Interactive Multi-evidence Learning

Domain shift can cause feature matching to select incorrect image-text pairs and miss difficult positive matches. To address both problems, IML selects candidate pairs using features and checks them with a frozen MLLM.

3.2.1 Balanced Reciprocal Matching

To avoid selecting too many pairs for only a few samples, we first balance image-text similarity scores. Let $F_t^I \in \mathbb{R}^{B_I \times D}$ and $F_t^T \in \mathbb{R}^{B_T \times D}$ denote normalized target image and text features, with cosine similarity $S = F_t^I (F_t^T)^\top$. Sinkhorn balancing (Cuturi, 2013) alternately normalizes the rows and columns of $\exp(S/\tau_{\text{sk}})$, with temperature parameter τ_{sk}, giving a detached matrix P used only for candidate selection.

Since balancing alone cannot ensure mutual matches, based on the reciprocal selection principle of Dual Alignment Consistency (DAC) (Hao et al., 2023), we retain only pairs that are selected in both retrieval directions and remove pairs with low matching scores:

$$\mathcal{P}_{\text{DAC}} = \left\{ (i,j) \;\middle|\; j \in \text{TopK}^k_{I\to T}(i),\ i \in \text{TopK}^k_{T\to I}(j),\ P_{ij} > \mathcal{Q}_\xi(P) \right\}, \tag{1}$$

where the Top-k sets give directional neighbors under P, and $\mathcal{Q}_\xi(P)$ is its ξ-quantile threshold.

3.2.2 False Pair Filtering

Due to unreliable feature similarities, reciprocal matches may contain conflicting image-text pairs, so we use the MLLM to check these pairs. To avoid excessive MLLM calls, we review only the lowest-scoring candidates in P.

To avoid removing correct pairs and improve the reliability of MLLM-based decisions, we adopt a multi-round interactive verification strategy from coarse to fine. Let $o_{ij}^{(r,\text{f})}$ be the Round r output, $\mathcal{H}_r$ the prompt used in Round r, and $\mathcal{R}^{\text{f}}$ the evidence extractor of Round 2 for false pair filtering. Round 1 performs a coarse-grained assessment of obvious visual conflicts:

$$o_{ij}^{(1,\text{f})} = \text{MLLM}(\mathcal{H}_1^{\text{f}}, I_i^t, T_j^t) \in \{\texttt{KEEP}, \texttt{CHECK}\}, \tag{2}$$

where `KEEP` retains the pair, while `CHECK` forwards uncertain candidates to Round 2 for fine-grained attribute extraction from the image, followed by evidence generation through comparison

with the textual description:

$$o_{ij}^{(2,\mathrm{f})} = \mathcal{R}^{\mathrm{f}}\big(\mathrm{MLLM}(\mathcal{H}_2, I_i^t), T_j^t\big) \in \{\texttt{STRONG}, \texttt{SUPPORT}, \varnothing\}, \tag{3}$$

where STRONG, SUPPORT, and $\varnothing$ denote strong, supporting, and absent conflict evidence. It should be noted that only candidate pairs classified as CHECK by the MLLM in Round 1 are sent to Round 2 for further semantic verification. Since blur or occlusion may make the generated evidence unreliable, Round 3 further verifies whether the evidence generated in Round 2 is visually reliable, clearly observable, and sufficient for the final decision:

$$o_{ij}^{(3,\mathrm{f})} = \mathrm{MLLM}\left(\mathcal{H}_3^{\mathrm{f}}, I_i^t, T_j^t, o_{ij}^{(2,\mathrm{f})}\right) \in \{\texttt{KEEP}, \texttt{REJECT}\}. \tag{4}$$

Only pairs marked REJECT enter $\mathcal{P}_{\mathrm{rej}}$. Unclear or invalid results default to KEEP to avoid removing possible true matches.

3.2.3 Missed Pair Recovery

However, unreliable feature similarities may also cause missed difficult positive pairs. To recover them without checking unrelated pairs, we use the MLLM to verify unmatched candidates whose counterpart ranks first in one direction and within Top-k in the other.

To avoid adding incorrect pairs and improve the reliability of MLLM-based decisions, recovery adopts a multi-round interactive verification strategy as well. Let $o_{ij}^{(r,\mathrm{a})}$ be each round's output and $\mathcal{R}^{\mathrm{a}}$ the evidence extractor for missed pair recovery:

$$o_{ij}^{(r,\mathrm{a})} = \begin{cases} \mathrm{MLLM}(\mathcal{H}_1^{\mathrm{a}}, I_i^t, T_j^t), & r = 1, \\ \mathcal{R}^{\mathrm{a}}\big(\mathrm{MLLM}(\mathcal{H}_2, I_i^t), T_j^t\big), & r = 2, \\ \mathrm{MLLM}(\mathcal{H}_3^{\mathrm{a}}, I_i^t, T_j^t, o_{ij}^{(2,\mathrm{a})}), & r = 3. \end{cases} \tag{5}$$

As in filtering, Round 1 coarsely identifies possible matches, and only candidate pairs classified as POSSIBLE by the MLLM in Round 1 are sent to the subsequent two rounds. Round 2 extracts fine-grained visual attributes and compares them with the text to generate evidence categorized as MATCH, STRONG, SUPPORT, or $\varnothing$. Round 3 verifies the evidence and returns ADD or SKIP, with unclear or invalid results defaulting to SKIP.

To keep matches one-to-one, we process $o_{ij}^{(3,\mathrm{a})} = \texttt{ADD}$ candidates in descending P_{ij} order. A pair enters $\mathcal{P}_{\mathrm{add}}$ only if both samples remain unused. The final set is

$$\mathcal{P}_{\mathrm{final}} = (\mathcal{P}_{\mathrm{DAC}} \setminus \mathcal{P}_{\mathrm{rej}}) \cup \mathcal{P}_{\mathrm{add}}. \tag{6}$$

3.2.4 Contrastive Learning with Rejection Penalty

To further improve target domain alignment using the refined pairs, we combine a weighted bidirectional contrastive loss (Radford et al., 2021) on the final positive pairs with a penalty that lowers the similarity of pairs rejected by the MLLM. For each positive pair, let $w_{ij} = \max(S_{ij}, \epsilon_w)$ and $Z_w = \sum_{(i,j)\in\mathcal{P}_{\mathrm{final}}} w_{ij} + \epsilon_Z$, where ϵ_w is a small positive constant that sets the minimum positive-pair weight, and ϵ_Z prevents division by zero. The directional probabilities $\pi_{i\to j}$ and $\pi_{j\to i}$ use the full similarity matrix S and temperature τ_d. The loss is

$$\mathcal{L}_{\mathrm{DAC}} = -\frac{1}{2Z_w} \sum_{(i,j)\in\mathcal{P}_{\mathrm{final}}} w_{ij} \left[\log \pi_{i\to j} + \log \pi_{j\to i}\right] + \beta_{\mathrm{neg}} \mathcal{L}_{\mathrm{neg}}. \tag{7}$$

The negative pair penalty is $\mathcal{L}_{\mathrm{neg}} = \sum_{(i,j)\in\mathcal{P}_{\mathrm{rej}}} \max(0, S_{ij} - m_h) / \max(1, |\mathcal{P}_{\mathrm{rej}}|)$, where β_{neg} weights the penalty, m_h is the similarity margin, and an empty rejected set contributes zero.

3.3 Density-filtered Intermediate-domain Learning

Although IML improves the reliability of target domain supervision, unreliable cross-domain intermediate representations may still reduce adaptation performance. Based on the feature mixing mechanism of the Intermediate Domain Module (IDM) (Dai et al., 2021), we introduce DIL to filter out extreme feature outliers before domain feature mixing and incorporate domain differences into the prediction of mixing weights.

3.3.1 DENSITY-BASED RELIABILITY ESTIMATION

To reduce the dimensionality of features, we first project source and target features onto a shared principal component analysis (PCA) (Jolliffe, 2002) basis for each modality. The distances to the global center and local neighbors are then used to construct a two-dimensional density descriptor. To identify unreliable samples, we then fit Gaussian mixture models (GMMs) (McLachlan & Peel, 2000) with one and two components on this two-dimensional descriptor and use the Bayesian Information Criterion (BIC) (Schwarz, 1978) for model selection. G_2 is selected when $\mathrm{BIC}(G_1) - \mathrm{BIC}(G_2) > 10$, and G_1 otherwise.

To find extreme outliers without removing normal samples, we normalize scores using the median and median absolute deviation (MAD) (Rousseeuw & Croux, 1993). Let ℓ_n be the negative log density of sample n, and $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_N]$ contain the N scores. The outlier score is

$$z_n = \kappa \frac{\ell_n - \mathrm{median}(\boldsymbol{\ell})}{\mathrm{MAD}(\boldsymbol{\ell}) + \epsilon_z}, \tag{8}$$

where κ is a scale factor, and ϵ_z prevents division by zero. We exclude only clear outliers from DIL mixing to keep useful samples.

3.3.2 INTERMEDIATE MIXING WITH FEATURE DIFFERENCES

To mix the filtered source and target features, we first truncate the larger group so that both groups have the same number of samples. Let f_s, f_t be their features and $\bar{f}_s, \bar{f}_t$ the normalized versions. To account for domain shift, the predictor uses both features and their difference $\bar{f}_s - \bar{f}_t$:

$$[\lambda_s, \lambda_t] = \mathrm{Softmax}\left(\frac{g_\phi([f_s; f_t; \alpha_d(\bar{f}_s - \bar{f}_t)])}{\tau_{\mathrm{mix}}} \right), \tag{9}$$

where λ_s, λ_t are the mixing weights, α_d scales the feature difference, g_ϕ is the predictor, and τ_{mix} is a learnable temperature. We use these mixing weights to form an intermediate feature:

$$f_m = \lambda_s f_s + \lambda_t f_t. \tag{10}$$

Because domain shifts differ by modality, images and text use separate predictors. Their bridge losses sum to $\mathcal{L}_{\mathrm{IDM}}$.

3.4 TEXT-GUIDED SPATIAL GATING

Although IML improves target pairs and DIL improves intermediate features, irrelevant image regions may still limit the domain adaptation ability of the model. To address this, TSG uses source text to teach an image-only student (Zagoruyko & Komodakis, 2017) to focus on identity-relevant regions. The student supplies features to DIL and IML.

To identify image regions that match a description, the teacher uses corresponding source text to query patches (Vaswani et al., 2017). The student learns to find these regions using only the global image feature, so it does not need text at test time. Let $V^s = [v_1^s, \ldots, v_{N_v}^s] \in \mathbb{R}^{D \times N_v}$ be the source image patch tokens (Dosovitskiy et al., 2021), where D is the feature dimension, and N_v is the number of image patch tokens. Define f_{eos}^s and f_{cls}^s as the source domain global text and image features, respectively. $q_{\mathrm{teacher}} = f_{\mathrm{eos}}^s$ and $q_{\mathrm{student}} = f_{\mathrm{cls}}^s$ are used to generate the source masks:

$$M_b^s = \sigma\left(\gamma_a (\widehat{W_b q_b})^\top \widehat{W_k V^s} \right), \tag{11}$$

where $b \in \{\mathrm{teacher}, \mathrm{student}\}$, W_b, W_k are query and key projections, hats denote ℓ_2 normalization, σ is the sigmoid function, and γ_a is a learnable scale. The detached teacher mask guides the student. The target domain mask is generated by the image-only student gate shared with the source domain.

To reduce background noise while keeping global information, we combine the global image feature with the gated patch feature f_p, obtained by a normalized weighted sum of image patch tokens:

$$f_{\mathrm{gate}} = (1 - \alpha_e) f_{\mathrm{cls}} + \alpha_e f_p, \tag{12}$$

where α_e sets the fusion strength at each epoch. Image DIL and target IML use student-gated features, while text DIL uses global features and Token Selection Embedding (TSE) (Qin et al., 2024) retains the original image tokens. Target masks are detached when forming gated features.

Table 1: Comparison with state-of-the-art methods across four transfer directions. R@1, R@5, and mAP (%) are reported. The best results are highlighted in bold, while the second-best results are underlined.

Methods	Venue	I→R			I→C			R→I			R→C		
		R@1	R@5	mAP	R@1	R@5	mAP	R@1	R@5	mAP	R@1	R@5	mAP
Source Only / Cross-domain													
CLIP	ICML'21	47.45	–	36.83	33.90	–	31.65	33.58	–	19.58	35.25	–	32.35
IRRA	CVPR'23	45.10	69.20	36.76	33.43	56.11	31.38	32.35	49.68	20.57	32.65	55.20	30.17
APTM	ACM MM'23	52.50	75.15	40.81	46.44	66.89	40.14	44.01	60.09	25.60	44.92	64.90	38.47
CFine	TIP'23	47.40	70.60	42.30	32.67	54.03	25.47	24.83	39.49	17.04	20.79	38.76	15.77
RDE	CVPR'24	49.25	72.10	38.46	38.11	59.24	34.16	42.17	58.32	26.37	36.94	58.22	33.65
ITSELF	WACV'26	51.30	73.30	41.07	41.05	63.30	37.79	43.32	59.17	27.57	40.32	61.70	36.65
UDA / Source + Target													
POUF	ICML'23	36.60	61.80	28.28	20.50	39.39	18.48	21.27	37.87	11.04	17.84	35.62	16.30
ReCLIP	WACV'24	50.35	73.25	42.75	33.48	55.25	29.19	27.18	42.51	17.83	21.04	40.29	18.92
GCKD	AAAI'25	59.95	79.05	**49.68**	52.70	72.76	45.97	46.40	61.17	26.06	51.01	70.29	43.71
Ours	–	**61.80**	**82.00**	47.35	**60.59**	**80.72**	**53.82**	**54.78**	**71.30**	**33.05**	**59.37**	**79.17**	**52.99**

3.5 Overall Training Procedure

To jointly optimize source supervision and cross-domain adaptation, the overall objective combines source Basic Global Embedding (BGE)/TSE supervision (Qin et al., 2024) with DIL, IML, and TSG losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BGE}} + \mathcal{L}_{\text{TSE}} + \lambda_{\text{IDM}}\mathcal{L}_{\text{IDM}} + \lambda_{\text{DAC}}\mathcal{L}_{\text{DAC}} + \mathcal{L}_{\text{att}}, \tag{13}$$

where λ_{IDM} and λ_{DAC} weight the two adaptation losses, and $\mathcal{L}_{\text{att}}$ trains the spatial gate.

4 Experiments

4.1 Experimental Settings

Datasets. To evaluate cross-dataset adaptation under diverse domain shifts, we conduct experiments on three widely used TIReID benchmarks: CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Shao et al., 2022; Jiang & Ye, 2023), and RSTPReid (Zhu et al., 2021), abbreviated as C, I, and R, respectively. Dataset statistics and official splits are provided in Appendix B.2.

Evaluation Metrics. To quantify retrieval quality from both top-ranked accuracy and overall ranking behavior, we use the standard TIReID metrics Recall at Rank-K (R@K), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). Higher values indicate better retrieval performance.

Implementation Protocol. We adopt RDE as the default source model and Qwen3-VL-8B-Instruct as the frozen MLLM. To evaluate the generalization ability of our proposed framework, we further integrate ISV with IRRA and CADA, while Qwen3-VL-32B-Instruct is used only for the MLLM scale analysis. The MLLM remains frozen throughout adaptation and participates only in semantic verification within IML. Additional implementation details are provided in Appendix B.3.

4.2 Comparison with State-of-the-Art Methods

To examine whether ISV remains effective when the source and target distributions change, we evaluate four transfers among the three benchmarks: I→R, I→C, R→I, and R→C. Table 1 compares ISV with methods that use only source-domain supervision and methods that also adapt using unlabeled target data. Details of the compared methods are provided in Appendix B.1. Based on the results, we can make the following observations: 1) ISV consistently outperforms the compared baselines across the four transfer settings, except for mAP on I→R against GCKD. On R→C, its R@1 exceeds GCKD by 8.36 percentage points. These results show that ISV improves top-ranked retrieval under domain shift by learning more reliable cross-domain representations and target domain correspondences. 2) Source-only methods generally perform less effectively when applied directly to different target datasets. Since they cannot utilize target-domain data during adaptation, their feature representations cannot adjust to changes in visual appearance and description patterns across

Table 2: Cross-domain retrieval with different source models. Gain (↑) reports the improvement over the corresponding source model.

Methods	Venue	C→I					I→C				
		R@1	R@5	R@10	mAP	mINP	R@1	R@5	R@10	mAP	mINP
IRRA	CVPR'23	42.55	62.19	69.68	21.84	1.97	33.43	56.11	66.23	31.38	19.01
+ISV	–	**55.27**	**72.69**	**79.40**	**29.85**	**4.02**	**59.16**	**79.21**	**86.58**	**53.01**	**37.03**
Gain (↑)	–	+12.72	+10.50	+9.72	+8.01	+2.05	+25.73	+23.10	+20.35	+21.63	+18.02
CADA	TMM'24	52.60	69.03	75.22	25.36	2.11	54.18	73.68	80.48	46.48	29.22
+ISV	–	**61.38**	**77.28**	**82.82**	**32.90**	**4.28**	**70.37**	**86.35**	**91.16**	**61.76**	**44.74**
Gain (↑)	–	+8.78	+8.25	+7.60	+7.54	+2.17	+16.19	+12.67	+10.68	+15.28	+15.52
RDE	CVPR'24	48.15	66.26	73.68	25.03	2.35	38.11	59.24	68.44	34.16	20.36
+ISV	–	**58.51**	**75.15**	**81.18**	**32.21**	**4.29**	**60.59**	**80.72**	**86.87**	**53.82**	**37.57**
Gain (↑)	–	+10.36	+8.89	+7.50	+7.18	+1.94	+22.48	+21.48	+18.43	+19.66	+17.21

Table 3: Effect of 8B and 32B MLLMs on retrieval performance on R→C and C→R.

Configuration	R→C					C→R				
	R@1	R@5	R@10	mAP	mINP	R@1	R@5	R@10	mAP	mINP
w/o IML	57.62	77.97	85.38	51.99	36.24	60.80	80.40	88.15	45.67	22.73
IML with 8B MLLM	59.37	79.17	86.16	52.99	36.98	62.10	81.50	88.60	46.92	23.83
IML with 32B MLLM	59.11	79.55	85.87	52.86	36.78	62.00	81.60	88.35	46.95	23.77

datasets, making target domain image-text similarity less reliable. 3) Existing UDA methods generally underperform ISV across the four transfer settings. On the one hand, these methods primarily rely on learned features, whose similarities may inaccurately reflect semantic correspondences under domain shift, leading to unreliable pseudo-correspondences. On the other hand, unreliable cross-domain representations may also degrade retrieval performance.

4.3 Adaptation with Different Source Models

To evaluate the generalization ability of ISV across different source models, we apply ISV to three representative source models: IRRA, CADA, and RDE. As shown in Table 2, ISV improves retrieval performance across all three source models in both C→I and I→C directions. For instance, when applied to IRRA, ISV achieves R@1 improvements of 12.72 and 25.73 percentage points in the two directions, respectively. Furthermore, the performance gains in I→C generally exceed those in C→I, possibly due to varying textual description granularities across datasets. Specifically, ICFG-PEDES tends to provide more fine-grained descriptions than CUHK-PEDES, leading to different levels of pseudo-correspondence noise during adaptation. By introducing external semantic evidence beyond feature similarity, ISV achieves greater improvements in the I→C direction.

4.4 Effect of IML and MLLM Scale

To evaluate the impact of MLLM scale on retrieval performance, we conduct experiments with 8B and 32B models on R→C and C→R, while keeping all other settings unchanged. As shown in Table 3, introducing IML improves retrieval performance in both directions by providing more reliable pseudo supervision. Moreover, the 32B model achieves performance comparable to the 8B model, possibly because the 8B model already provides sufficient semantic understanding for the current verification task. Considering the higher computational cost of the 32B model (see Appendix G), we adopt the 8B model in our final configuration.

4.5 Ablation Study

To evaluate the effectiveness of each proposed component, we conduct ablation studies on R→C and C→R. Specifically, we progressively add each component to the baseline, with the results shown in Table 4. From the results, adding the IML module improves retrieval performance, demonstrating

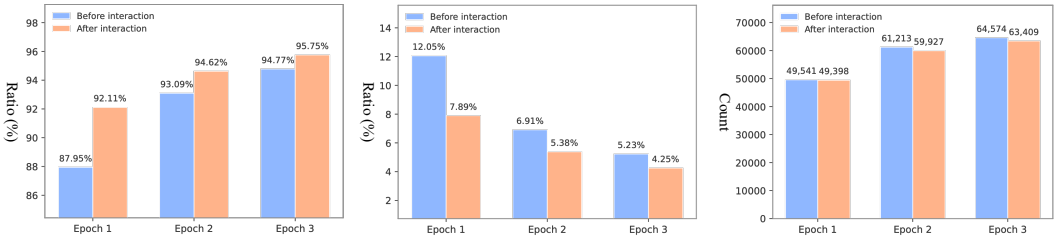
Table 4: Ablation studies on R→C and C→R.

Methods	R→C			C→R		
	R@1	R@10	mAP	R@1	R@10	mAP
Baseline	57.57	85.01	51.64	61.15	87.05	46.29
+ IML	58.95	86.06	52.80	60.95	87.50	46.53
+ IML + DIL	59.08	86.08	52.90	61.50	88.20	46.92
+ IML + DIL + TSG	**59.37**	**86.16**	**52.99**	**62.10**	**88.60**	**46.92**

the benefit of refining noisy pseudo correspondences. The proposed DIL module further improves performance, indicating its effectiveness in constructing more reliable intermediate-domain representations. By complementing these two modules with TSG to suppress identity-irrelevant visual responses, the model achieves further improvements in performance. Overall, the results demonstrate the effectiveness and importance of each proposed component in cross-dataset text-to-image person retrieval.

4.6 Pseudo-Pair Quality Analysis

To assess whether MLLM interaction improves the reliability of target pseudo supervision, we compare pseudo-pair quality before and after semantic interaction during the first three adaptation epochs on R→C. As shown in Figure 3, semantic interaction consistently increases the proportion of correct pairs and reduces that of incorrect pairs across the three early adaptation epochs, while keeping the total number of pseudo pairs largely unchanged. These results demonstrate that IML alleviates the impact of unreliable feature similarities in the target domain during early training stages. By jointly filtering incorrect pairs and recovering missed positive pairs, IML provides more reliable supervision for target-domain learning, thereby improving domain adaptation performance.



(a) Correct Pair Ratio (b) Incorrect Pair Ratio (c) Total Pair Count

Figure 3: Effect of MLLM interaction on target pseudo-pair quality during the first three adaptation epochs on R→C. (a) Proportion of correct pairs. (b) Proportion of incorrect pairs. (c) Total number of selected pairs per epoch.

4.7 Qualitative Results

To qualitatively assess whether ISV improves retrieval reliability under domain shift, we visualize the top-10 retrieval results of ISV and the baseline using the same queries and galleries on R→C. As shown in Figure 4, ISV retrieves more correct pedestrian images at higher ranks and reduces several errors made by the baseline. These results indicate that ISV establishes more reliable target domain pseudo correspondences and cross-domain intermediate representations, demonstrating the effectiveness of our method in improving cross-domain retrieval performance.

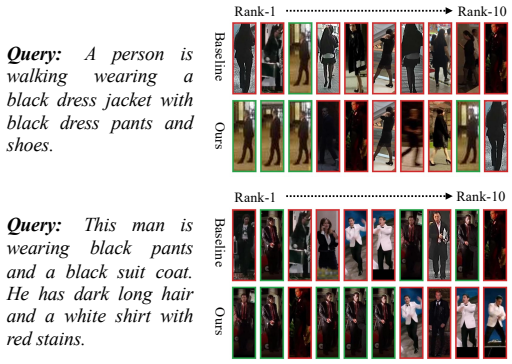


Figure 4: Top-10 retrieved results on R→C. Green and red borders indicate correctly and incorrectly retrieved images, respectively.

5 Conclusion

This paper introduces ISV for UDA TIReID. ISV addresses unreliable adaptation through three complementary components: 1) IML combines balanced reciprocal matching with semantic verification from a frozen MLLM to refine noisy target pseudo correspondences; 2) DIL filters extreme feature outliers and explicitly models source-to-target feature differences for more reliable cross-domain representation learning; 3) TSG transfers source textual guidance to an image-only student to suppress identity-irrelevant visual responses. Together, these components improve adaptation ability at the target correspondence, cross-domain representation, and visual feature levels. Extensive experiments on three TIReID benchmarks, multiple transfer directions, and different source models demonstrate the effectiveness and generalization of ISV. Future work could extend ISV to other cross-modal retrieval scenarios, such as video-text person retrieval, and explore more efficient or continual semantic verification for dynamically evolving target domains.

REFERENCES

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. RaSa: Relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 555–563, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. IDM: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11844–11854, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6112–6121, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020a.

Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11309–11321, 2020b.

Pengyu Gu, Qingchuan Tao, and Yanmei Yu. Dynamic mutual alignment for unsupervised domain adaptation text-to-image person re-identification. *IEEE Signal Processing Letters*, 33:1581–1585, 2026.

Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18962–18972, 2023.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, Cheng-Hao Kuo, and Ram Nevatia. ReCLIP: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2994–3003, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916, 2021.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2787–2797, 2023.

Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10678–10686, 2020.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2002.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12888–12900, 2022.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1970–1979, 2017.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6028–6039, 2020.

Dixuan Lin, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Cross-modal adaptive dual association for text-to-image person retrieval. *IEEE Transactions on Multimedia*, 26:6609–6620, 2024.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 97–105, 2015.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Bingjun Luo, Jinpeng Wang, Zewen Wang, Junjie Zhu, and Xibin Zhao. Graph-based cross-domain knowledge distillation for cross-dataset text-to-image person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 568–576, 2025.

Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000.

Tien-Huy Nguyen, Huu-Loc Tran, and Thanh Duc Ngo. ITSELF: Attention guided fine-grained alignment for vision-language retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1448–1458, 2026.

Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020.

Kai Niu, Qinzi Zhao, Jiahui Chen, and Yanning Zhang. Cache-aided cross-modal correlation correction for unsupervised cross-domain text-based person search. *Pattern Recognition*, 172:112521, 2026.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27197–27206, 2024.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via MLLMs for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14390–14399, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5566–5574, 2022.

Wentao Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of MLLMs for transferable text-to-image person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17127–17137, 2024.

Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. POUF: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 33816–33832, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. ViTAA: Visual-textual attributes alignment in person search by natural language. In *Proceedings of the European Conference on Computer Vision*, volume 12357, pp. 402–420, 2020.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. CLIP-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4492–4501, 2023.

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

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, volume 11205, pp. 707–723, 2018.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. DSSL: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 209–217, 2021.

APPENDIX

A Algorithm **14**

B Additional Implementation Details **14**

B.1 Baseline Details 14

B.2 Dataset Details 15

B.3 Implementation Details 15

B.4 Additional Method Details 16

B.5 Auxiliary Loss Definitions 17

C MLLM Prompt Details **18**

C.1 False Pair Filtering Prompts 19

C.2 Missed Pair Recovery Prompts 21

C.3 MLLM Inference Settings 22

D Stability Analysis **22**

E Additional Qualitative Results **23**

F Sensitivity to MLLM Activation Duration **24**

G Efficiency and Computational Cost **24**

H MLLM Failure Cases **25**

I The Use of Large Language Models **25**

A ALGORITHM

Algorithm 1 Training procedure of ISV.

Input: Labeled source data $\mathcal{D}_s$; unlabeled target data $\mathcal{D}_t$; encoders initialized on the source domain E_I, E_T; frozen MLLM $\mathcal{M}$; epochs E; MLLM cutoff E_m.

for $e = 1, 2, \ldots, E$ **do**
  if $e \geq 10$ **then**
    Update the dataset-level DIL reliability masks with Eq. (8).
  end if
  for each minibatch from the source and target domains **do**
    Encode source and target images/texts with E_I and E_T.
    Construct the teacher and student spatial masks with Eq. (11), obtain the student-gated image features with Eq. (12), and compute $\mathcal{L}_{\mathrm{att}}$.
    Apply the current reliability masks when available, align the retained source and target samples, construct the intermediate features with Eqs. (9)–(10), and compute $\mathcal{L}_{\mathrm{IDM}}$ with Eq. (30).
    Compute the target similarity matrix S and detached Sinkhorn assignment P, then construct the reciprocal candidate set $\mathcal{P}_{\mathrm{DAC}}$ with Eq. (1).
    if $e \leq E_m$ **then**
      Let $\mathcal{P}^e_{\mathrm{rev}} \subseteq \mathcal{P}_{\mathrm{DAC}}$ denote the low-scoring subset selected for MLLM review at epoch e. Verify $\mathcal{P}^e_{\mathrm{rev}}$ with Eqs. (2)–(4) to obtain rejected pairs $\mathcal{P}_{\mathrm{rej}}$.
      Check approximately reciprocal candidates with Eq. (5) to obtain recovered pairs $\mathcal{P}_{\mathrm{add}}$.
    else
      Set $\mathcal{P}_{\mathrm{rej}} = \varnothing$ and $\mathcal{P}_{\mathrm{add}} = \varnothing$.
    end if
    Form $\mathcal{P}_{\mathrm{final}} = (\mathcal{P}_{\mathrm{DAC}} \setminus \mathcal{P}_{\mathrm{rej}}) \cup \mathcal{P}_{\mathrm{add}}$ with Eq. (6).
    Compute $\mathcal{L}_{\mathrm{DAC}}$ with Eq. (7), then combine it with $\mathcal{L}_{\mathrm{BGE}}$, $\mathcal{L}_{\mathrm{TSE}}$, $\mathcal{L}_{\mathrm{IDM}}$, and $\mathcal{L}_{\mathrm{att}}$ to update the trainable parameters.
  end for
end for

Output: Adapted encoders $\{E_I, E_T\}$ and learned auxiliary parameters.

Algorithm 1 summarizes the complete training procedure. For each minibatch, spatial gating first produces the image representations used by the adaptation branches. Reliability-aware intermediate-domain learning then constructs the mixed source and target representations, while balanced reciprocal matching identifies candidate target correspondences. During the first E_m epochs, the frozen MLLM verifies uncertain candidates and searches for reliable missed matches. The refined target pseudo pairs are finally combined with source-domain supervision to update the retrieval model.

B ADDITIONAL IMPLEMENTATION DETAILS

B.1 BASELINE DETAILS

To provide a comprehensive evaluation of ISV against existing retrieval and domain adaptation approaches, the following representative baselines are selected for comparison:

- CLIP (Radford et al., 2021) learns transferable visual representations from a large collection of paired image and text data and serves as a general vision-language baseline.
- IRRA (Jiang & Ye, 2023) introduces cross-modal implicit relation reasoning and similarity distribution matching for text-to-image person retrieval.
- APTM (Yang et al., 2023) exploits large scale supervision from multiple attributes and language for text-based person retrieval.
- CFine (Yan et al., 2023) uses fine-grained alignment driven by CLIP to model detailed correspondences between text and images.
- RDE (Qin et al., 2024) learns robust dual embeddings to alleviate the influence of noisy correspondences between images and texts.
- ITSELF (Nguyen et al., 2026) uses fine-grained alignment guided by attention to improve cross-dataset generalization in vision-language person retrieval.

- POUF (Tanwisuth et al., 2023) uses unsupervised fine tuning with prompts to adapt large pretrained models to unlabeled target data.
- ReCLIP (Hu et al., 2024) refines CLIP representations through source-free domain adaptation.
- GCKD (Luo et al., 2025) transfers cross-domain knowledge through distillation with graph modeling for cross-dataset text-to-image person retrieval.
- CADA (Lin et al., 2024) models adaptive dual associations between visual and textual modalities for text-to-image person retrieval.

B.2 Dataset Details

To evaluate the generalization of ISV across different source and target domain distributions, three widely used TIReID benchmarks are adopted:

- CUHK-PEDES (Li et al., 2017) contains 40,206 images of 13,003 identities, with two textual descriptions for each image. The official split contains 11,003 identities for training, 1,000 for validation, and 1,000 for testing.
- ICFG-PEDES (Shao et al., 2022) contains 54,522 images of 4,102 identities, with one textual description for each image. We follow the standard split used in IRRA (Jiang & Ye, 2023), with 3,102 identities for training and 1,000 identities for testing.
- RSTPReid (Zhu et al., 2021) contains 20,505 images of 4,101 identities. The official split uses 3,701 identities for training, 200 for validation, and 200 for testing.

For each transfer, we initialize the retrieval model with the training data from the source domain. During adaptation, the target training data are treated as unlabeled and unpaired. Evaluation uses the official target test split.

B.3 Implementation Details

To ensure reproducibility and clarify the training configuration used throughout our experiments, this subsection summarizes the key implementation details. All experiments are conducted on a single NVIDIA L20X GPU.

The model is trained for 60 epochs using Adam (Kingma & Ba, 2015). The source and target batch sizes are both 32. The backbone learning rate is 2.5×10^{-6}, while the TSE embedding layers and TSG mask module use 10^{-3} and 10^{-4}, respectively. We use a linear warmup for 5 epochs followed by cosine decay (Loshchilov & Hutter, 2017). We set $\lambda_{\mathrm{IDM}} = 0.02$ and $\lambda_{\mathrm{DAC}} = 15$, both fixed throughout training. The mask regularization weight is 0.18 and the BGE weight for features gated by the teacher is 0.30.

TSG uses a query/key space with 256 dimensions, a soft gate lower bound of 0.82, and a warmup over 10 epochs. For DIL, we set the difference scale to 0.25. The learnable mixing temperature is initialized to 1.2 and constrained to $[1, 2]$. Density filtering starts from epoch 10 and retains at least 90% of the features. This delayed filtering avoids relying on unreliable density estimates during early adaptation. We set the robust scale to $\kappa = 0.6745$ and use 3.5 as the robust anomaly score threshold. PCA reduces the image and text features to 64 dimensions, and five nearest neighbors are used for density estimation.

For IML, the Sinkhorn temperature is 0.05 with three normalization iterations. Reciprocal candidate selection uses $k = 1$ and $\xi = 0.9$, and the DAC contrastive temperature is 0.1. We set the hard-negative margin $m_h = 0.30$, the rejected-pair weight $\beta_{\mathrm{neg}} = 0.15$, and the positive-pair weight floor $\epsilon_w = 10^{-6}$, with $\epsilon_Z = 10^{-8}$ used for normalization stability in the confidence-weighted DAC loss. During the first three epochs, the lowest 40%, 30%, and 20% of selected pairs by Sinkhorn similarity are sent to the MLLM for review, respectively. This progressively decreasing review ratio reduces verification costs as the target representations become more reliable during adaptation. Missed pair recovery uses forward top-1 and reverse top-3 neighbors from the unused samples. The remaining MLLM inference settings are provided in Appendix C.3, and Appendix D reports the stability analysis.

B.4 Additional Method Details

To provide a clearer view of the implementation logic and interactions among the proposed modules, this subsection presents additional methodological details.

TSG Gate Construction. To suppress irrelevant regions without discarding potentially useful identity cues, the mask M^x from Eq. (11) is converted into a bounded soft gate rather than being used as a hard patch selector. For branch $x \in \{\mathrm{tea}, \mathrm{stu}\}$, the gate is defined as

$$G^x_{\mathrm{soft}} = g_{\min} + (1 - g_{\min})M^x, \tag{14}$$

where tea and stu denote the teacher and student branches, respectively, and $g_{\min}$ is the lower bound of the soft gate. Because the spatial masks are less stable at the beginning of adaptation, the gate strength is gradually increased using the warmup factor

$$w_e = \min\left(1, \frac{e}{E_w}\right), \tag{15}$$

where e is the current epoch and E_w is the warmup duration. The resulting gate at epoch e is

$$G^x_e = 1 + w_e(G^x_{\mathrm{soft}} - 1)\,. \tag{16}$$

The gated patch feature is then obtained by a normalized weighted sum,

$$f^x_p = \frac{\sum_{n=1}^{N_v} G^x_{e,n} v_n}{\sum_{n=1}^{N_v} G^x_{e,n} + \epsilon_g}, \tag{17}$$

where $\epsilon_g = 10^{-6}$ is the numerical stability constant for the gated weighted sum; N_v and v_n retain the definitions from the main text. The fusion coefficient for each branch is scheduled as

$$\alpha^x_e = w_e \alpha_x, \tag{18}$$

where α_x is the branch-specific maximum fusion coefficient. Equation (12) is applied with the coefficient for each branch α^x_e. We use $g_{\min} = 0.82$, $E_w = 10$, $\alpha_{\mathrm{stu}} = 0.02$, and $\alpha_{\mathrm{tea}} = 0.45$. The teacher branch is used only on paired source samples, whereas the student branch is used for source images, target images, and test images; at test time $w_e = 1$. This asymmetric design allows reliable source text guidance to supervise the student during training while keeping target and test-time gating image-only. The teacher mask is detached only when it serves as the distillation target. The mask regularization and distillation objectives are given in Appendix B.5.

DIL Reliability and Mixing Details. Since DIL is intended to remove only extreme feature outliers rather than aggressively prune the data, reliability is estimated from both global deviation and local neighborhood density. For each modality, raw source and target CLS/EOS features share the same PCA projection, while image and text modalities use different PCA bases. After PCA and ℓ_2 normalization, each feature u_i is summarized by its distance to the global center and its average k_n-nearest-neighbor distance:

$$d^c_i = \|u_i - \mu\|_2. \tag{19}$$

$$d^n_i = \frac{1}{k_n} \sum_{j \in \mathcal{N}_{k_n}(i)} \|u_i - u_j\|_2, \tag{20}$$

where μ is the global center of the normalized features, k_n is the number of nearest neighbors used for density estimation, and $\mathcal{N}_{k_n}(i)$ denotes the corresponding neighbor set of u_i. The quantity d^c_i measures the global center deviation, while d^n_i characterizes the local density through the average nearest-neighbor distance. These two statistics form the density descriptor

$$q_i = [d^c_i, d^n_i]. \tag{21}$$

We use $k_n = 5$ whenever possible. GMMs with one and two components are fitted separately in each domain–modality set, and the model with two components is selected only when its BIC improves by more than 10. This model-selection rule avoids forcing a two-component density model when the feature distribution does not provide sufficient evidence for it. A sample is considered an outlier candidate only when its robust score in Eq. (8) exceeds 3.5 and the sum of the standardized center and neighbor distances is positive. Candidates are ranked by the robust score, but at least 90% of

each set is retained to prevent density filtering from removing a large portion of otherwise useful samples. These reliability masks affect only DIL: they are estimated from raw CLS/EOS features, while the image DIL branch mixes the image features gated by TSG.

After filtering, the longer source or target side is truncated within each modality so that mixed pairs have equal counts. This alignment is only for intermediate domain mixing and does not imply identity correspondence. The predictor g_ϕ in Eq. (9) contains two LayerNorm–LeakyReLU blocks followed by a linear layer with two outputs. Image and text use independent predictors. The learnable temperature is initialized to 1.2 and constrained to $[1, 2]$; the final linear layer uses a small initialization with standard deviation 10^{-3} and zero bias so that the initial source/target mixing weights are close to uniform, avoiding a premature bias toward either domain before meaningful difference cues are learned.

IML Candidate Control and Conservative Verification. To prevent reciprocal matching from concentrating on a small number of samples, candidate construction first balances the batch-level assignment before reciprocal selection. The Sinkhorn score is initialized as

$$P^{(0)} = \exp(S/\tau_{\mathrm{sk}}), \tag{22}$$

and each iteration alternates row and column normalization,

$$P_{ij} \leftarrow \frac{P_{ij}}{\sum_{j'} P_{ij'} + \epsilon_{\mathrm{sk}}}. \tag{23}$$

$$P_{ij} \leftarrow \frac{P_{ij}}{\sum_{i'} P_{i'j} + \epsilon_{\mathrm{sk}}}. \tag{24}$$

where $\epsilon_{\mathrm{sk}} = 10^{-8}$ is the numerical stability constant used in Sinkhorn row and column normalization. The first update normalizes each row and the second normalizes each column. We perform three iterations and detach P from gradient computation. After mutual Top-1 filtering and the 0.9 quantile threshold in Eq. (1), candidate construction uses the balanced score P, whereas pseudo positive weighting and target contrastive learning continue to use the original cosine similarity S.

To focus external semantic verification on the most uncertain pseudo pairs while limiting inference cost, only the selected candidates with the lowest scores are sent to the MLLM. In the first three epochs, the review ratios are 0.4, 0.3, and 0.2, with at least one pair reviewed for each nonempty batch. False pair filtering follows a conservative rule: Round 1 only flags possible contradictions, Round 2 extracts reusable image attributes and produces structured conflict evidence, and Round 3 rejects a pair only when the listed conflict is visually confirmed; ambiguous or invalid outputs default to `KEEP`. For missed pair recovery, endpoints already used by retained pseudo positives are excluded first. A remaining pair is considered only if one direction is Top-1 and the reverse direction is within Top-3. This near-reciprocal constraint restricts recovery to plausible missed matches rather than arbitrary unused pairs. Candidates accepted by the MLLM are processed in descending P_{ij} order and added only when both endpoints are still unused. Pairs already present in the original DAC set are not added back, so a rejected original pair cannot return unchanged. The exact prompt templates are provided in Appendix C.

B.5 Auxiliary Loss Definitions

To clarify the optimization objectives associated with TSG and DIL, we provide the effective losses used by the two modules. These objectives clarify how spatial guidance is regularized and how intermediate features are anchored to the two domains while complementing the main retrieval losses. We keep the definitions compact because the corresponding feature construction is already given in Secs. 3.4 and 3.3.

TSG Loss. To preserve retrieval-relevant information while transferring text-guided spatial attention from the teacher to the image-only student, TSG combines mask regularization, retrieval supervision, and teacher–student distillation in a unified attention objective. Let M^{tea} and M^{stu} denote the teacher and student masks computed on source samples, respectively. The effective attention loss used in training is

$$\begin{aligned}\mathcal{L}_{\mathrm{att}} = \eta_m \left[\mathcal{L}_{\mathrm{mask}}(M^{\mathrm{tea}}) + \eta_s \mathcal{L}_{\mathrm{mask}}(M^{\mathrm{stu}}) + \eta_c \mathcal{L}_{\mathrm{aux}}\right] \\ + \eta_r \mathcal{L}_{\mathrm{ret}}^{\mathrm{tea}} + \eta_d \, w_e \mathcal{L}_{\mathrm{dist}},\end{aligned} \tag{25}$$

where $\mathcal{L}_{\text{mask}}(\cdot)$ is the mask regularizer, $\mathcal{L}_{\text{aux}}$ is the auxiliary feature-consistency term, $\mathcal{L}_{\text{ret}}^{\text{tea}}$ is the teacher-gated retrieval loss, and $\mathcal{L}_{\text{dist}}$ is the teacher–student distillation loss. The coefficients η_m, η_s, η_c, η_r, and η_d weight the overall mask-regularization block, student-mask regularization, auxiliary consistency, teacher-gated retrieval, and distillation terms, respectively. The factor $w_e = \min(1, e/10)$ is the warmup factor over 10 epochs defined consistently with Eq. (15). The warmup prevents strong distillation from dominating before the student attention becomes sufficiently stable. All coefficients in this loss are kept fixed across the reported experiments. We use $(\eta_m, \eta_s, \eta_c, \eta_r, \eta_d) = (0.18, 0.1, 0.1, 0.30, 1.20)$. The mask regularizer is

$$\mathcal{L}_{\text{mask}}(M) = \zeta_a \mathcal{L}_{\text{area}}(M) + \zeta_s \mathcal{L}_{\text{smooth}}(M) + \zeta_b \mathcal{L}_{\text{binary}}(M), \tag{26}$$

where ζ_a, ζ_s, and ζ_b weight the area, smoothness, and weak binarization terms, respectively. We use $(\zeta_a, \zeta_s, \zeta_b) = (0.90, 0.006, 0.08)$. The three terms constrain the mean mask activation toward 0.36, encourage spatial smoothness, and weakly encourage binary responses, respectively, preventing the mask from becoming overly diffuse or collapsing to a trivial response. The auxiliary consistency and teacher–student distillation terms are

$$\mathcal{L}_{\text{aux}} = \frac{1}{BD} \sum_{i=1}^{B} \left\| f_{p,i}^{\text{stu}} - \text{sg}(f_{\text{cls},i}) \right\|_2^2. \tag{27}$$

$$\mathcal{L}_{\text{dist}} = \frac{1}{BN_v} \sum_{i=1}^{B} \sum_{n=1}^{N_v} \left(m_{i,n}^{\text{stu}} - \text{sg}(m_{i,n}^{\text{tea}}) \right)^2, \tag{28}$$

where B is the source-domain batch size, D is the feature dimension defined in the main text, $f_{p,i}^{\text{stu}}$ is the student-gated patch representation of sample i from Eq. (17), and $m_{i,n}^{\text{stu}}$ and $m_{i,n}^{\text{tea}}$ are the student and teacher mask values of patch n for sample i, respectively. The operator $\text{sg}(\cdot)$ denotes gradient stopping. The former loss aligns the local representation gated by the student with the global image feature, while the latter transfers the teacher mask to the student. $\mathcal{L}_{\text{ret}}^{\text{tea}}$ is the same bidirectional Triplet Alignment Loss (TAL) (Qin et al., 2024) retrieval objective used by the source BGE branch, but is evaluated on the image feature gated by the teacher and the source text feature. An auxiliary retrieval loss for the feature gated by the student is implemented but assigned zero weight in the default model, and therefore does not contribute to Eq. (25). All terms in $\mathcal{L}_{\text{att}}$ are computed on batches from the source domain, where reliable image–text pairs are available to provide teacher supervision without relying on noisy target pseudo correspondences.

DIL Bridge Loss. To keep the intermediate representation connected to both source and target domains rather than allowing it to drift toward either side, DIL uses a feature bridge loss that constrains each mixed feature by its corresponding source and target features. For modality $m \in \{I, T\}$ and a retained source–target pair indexed by b, let $f_{\text{mix},b}^{m}$ denote the mixed feature constructed by Eq. (10). Its feature bridge loss is

$$\ell_{\text{bridge},b}^{m} = \lambda_{s,b}^{m} \left\| f_{\text{mix},b}^{m} - f_{s,b}^{m} \right\|_2 + \lambda_{t,b}^{m} \left\| f_{\text{mix},b}^{m} - f_{t,b}^{m} \right\|_2, \tag{29}$$

where m indexes the modality, b indexes the retained pair, $f_{s,b}^{m}$ and $f_{t,b}^{m}$ are its source and target features, and $\lambda_{s,b}^{m}$ and $\lambda_{t,b}^{m}$ are the corresponding mixing weights predicted by Eq. (9). The image and text branches use independent DIL parameters because their domain shifts can differ, and the effective IDM objective is the sum over the retained samples in both modalities:

$$\mathcal{L}_{\text{IDM}} = \sum_{b=1}^{B_I'} \ell_{\text{bridge},b}^{I} + \sum_{b=1}^{B_T'} \ell_{\text{bridge},b}^{T}, \tag{30}$$

where B_I' and B_T' are the retained image and text counts after density filtering and count alignment between source and target samples within each modality. Thus, the current implementation uses the feature bridge term only; the prediction bridge and diversity terms from the original IDM formulation are not included in our training objective.

C MLLM Prompt Details

To ensure reproducibility of the semantic verification process, this section provides the exact prompt templates used by the frozen MLLM in IML. In the templates, {`caption`}, {`conflicts`}, and

{matches} denote the candidate description and the evidence passed from the previous step. False pair filtering and missed pair recovery use the same attribute prompt based only on the image in Round 2. We compare the extracted attributes with the candidate text using fixed rules, then pass the resulting conflict or match evidence to the final MLLM check. No MLLM is used during the evidence generation process. Table 5 summarizes the output of each round for the two branches.

Table 5: Outputs of the three verification rounds for false pair filtering and missed pair recovery.

Branch	Round 1	Round 2	Round 3
False pair filtering	KEEP/CHECK	attribute JSON → rule-based evidence	KEEP/REJECT
Missed pair recovery	POSSIBLE/REJECT	attribute JSON → rule-based evidence	ADD/SKIP

C.1 False Pair Filtering Prompts

Round 1: Coarse Contradiction Check.

You are an expert in text-based person re-identification.

Given one pedestrian image and one English text description, mark whether this image-text pair needs further checking.

Description: {caption}

Focus only on visible pedestrian attributes: upper clothing, inner clothing, lower clothing, shoes, bag, carried object, hair/hat, pose, viewpoint, and action.

Rules:

- KEEP when there is no clear visible contradiction.
- KEEP when the text is only partially verifiable but still plausible.
- KEEP when some attributes are invisible, blurry, occluded, cropped, or hard to verify.
- CHECK only when a clearly visible main pedestrian attribute directly contradicts the text.
- CHECK for obvious upper/lower clothing color conflicts, pants/shorts/skirt conflicts, clear bag/hat presence conflicts, or multiple clear main-attribute conflicts.
- Do not CHECK only because of shoe color, handheld color, subtle type differences, missing confirmation, blur, occlusion, or cropping.
- Do not guess invisible attributes.
- Do not use background as the main evidence.

Return strictly: DECISION: KEEP / CHECK

Round 2: Visible Attribute Extraction Shared by Both Branches.

You are an expert in pedestrian attribute recognition.

Given one pedestrian image, extract only visible attributes of the person.

Return strictly in JSON format:

```
{
  "upper_clothing": {"color": "", "type": ""},
  "inner_clothing": {"color": "", "type": ""},
  "lower_clothing": {"color": "", "type": ""},
  "shoes": {"color": "", "type": ""},
  "bag": {
    "presence": "yes/no/uncertain", "color": "", "type": ""
  },
  "handheld": {
    "presence": "yes/no/uncertain", "object": "", "color": ""
  },
  "head": {
    "hair": {
      "presence": "yes/no/uncertain", "color": "", "length": ""
    },
    "hat": {
      "presence": "yes/no/uncertain", "color": "", "type": ""
    }
  },
```

```
  "view_action": {"action": ""}
}
```

Rules:

- Only describe visible attributes.
- For "presence" fields, use "yes" only when the item or hair is visibly present; use "no" only when the relevant body region is clearly visible and the item or hair is clearly absent; use "uncertain" when the relevant body region is invisible, blurry, occluded, cropped out, or hard to judge.
- For any attribute without a "presence" field, including clothing/shoes color or type and `view_action.action`, use "uncertain" when it is invisible, blurry, occluded, cropped out, or hard to judge.
- If presence is "no" or "uncertain", set the associated color/type/object fields to "uncertain".
- If `hair.presence` is "no" or "uncertain", set hair color and length to "uncertain".
- Do not infer hidden details.
- Keep each value short and field-specific, such as "red" for color, "jacket" or "sneakers" for type, and "backpack" for object. Do not combine color and type in one value.
- If there are multiple colors, write only the main visible colors in the color field, such as "red" or "green and black".
- Do not describe background.

Round 3: Conflict Confirmation.

You are the final verifier for one image-text pair.

Pedestrian text description: {`caption`}

Suspected conflicts: {`conflicts`}

The conflicts were already generated by a previous checking stage. Your task is not to rescore the whole image-text pair. Your task is to confirm whether the listed conflicts are real visible contradictions.

Conflict labels:

- `[STRONG]` means high-confidence evidence from a main pedestrian attribute.
- `[SUPPORT]` means useful evidence that may be enough when it is a clear main-clothing or carried-object contradiction, or when several independent conflicts agree.

Rules:

- REJECT if any `[STRONG]` conflict is visible enough and directly contradicts the text.
- REJECT if one `[SUPPORT]` conflict is a clear `upper_clothing` or `lower_clothing` color/type contradiction on the main pedestrian.
- REJECT if one `[SUPPORT]` conflict is an explicit `handheld.presence` contradiction and the hand/object area is visible enough.
- REJECT if two or more `[SUPPORT]` conflicts from different body/object regions are visible enough, even when each one is moderate.
- For common colors such as black, white, gray, grey, navy, blue, or dark, REJECT when the main clothing region is clearly a different color. KEEP only when lighting, shadow, blur, or partial visibility can plausibly explain it.
- REJECT for clear type contradictions such as shirt vs coat/jacket, pants vs skirt, pants vs shorts, no bag vs visible bag, or text says carried object but no carried object is visible.
- KEEP only when every listed conflict is invalid, not visible enough, caused by occlusion/cropping/blur/lighting, a compatible synonym, or a minor unstable detail.
- KEEP for a single weak accessory/detail conflict only, such as shoe color, bag color, handheld color, hat color, or hair/action.
- Do not choose KEEP just because other non-listed attributes match.
- Do not reject because of background or hidden details.

Return strictly: `DECISION: KEEP / REJECT`

C.2 Missed Pair Recovery Prompts

Round 1: Coarse Candidate Check.

You are an expert in text-based person re-identification.
Pedestrian text description: {caption}
Your task is only a coarse entrance check: decide whether this candidate is worth sending to later evidence verification.
Rules:

- POSSIBLE when the main pedestrian is visible and upper/lower clothing can be roughly assessed, with no obvious hard contradiction to the text.
- POSSIBLE when the pair is only partially verifiable, generic, or incomplete, as long as there is no hard negative.
- POSSIBLE when accessories or objects are unclear; later rounds will decide whether they provide enough evidence.
- REJECT only for hard negatives that make this candidate clearly not worth checking.
- REJECT when the person is too blurry, too small, heavily cropped, or occluded so that upper/lower clothing cannot be assessed.
- REJECT when both upper and lower clothing are either impossible to assess or clearly incompatible with the text.
- REJECT for an obvious main-clothing mismatch, such as clearly light vs clearly dark, pants vs skirt/shorts, or outerwear vs short-sleeve shirt.
- REJECT when a prominent required object is clearly impossible, such as text requiring a backpack, handheld object, hat, or scarf while the relevant visible region clearly shows none.
- REJECT when the apparent support comes from the background or a different person, not the main pedestrian.
- Do not decide from background, pose, gender, hair, action, shoe color alone, or hidden details.

Return strictly: `DECISION: POSSIBLE / REJECT`

To ensure that recovery is supported by explicit visual evidence rather than coarse plausibility alone, Round 2 uses the same image-only visible attribute extraction prompt as false pair filtering. The extracted attributes are then compared with the candidate text using fixed rules to identify supporting and conflicting evidence, without an additional MLLM judgment. Only candidates with sufficient main-clothing support and no strong conflict continue to the final recovery check.

Round 3: Final Add/Skip Decision.

You are the final verifier for adding one candidate image-text pair.
Pedestrian text description: {caption}
Positive match evidence: {matches}
Possible conflict evidence: {conflicts}
Rules:

- ADD only when the positive evidence strongly supports this exact pair, not just a plausible person.
- ADD when `upper_clothing` and `lower_clothing` both support the text, and at least one main-clothing support is specific rather than generic black/dark/blue/gray clothing.
- ADD when a specific main-clothing cue is clearly supported, such as red, brown, orange, yellow, green, purple, beige, white outerwear, skirt, shorts, or vest, and the other main-clothing region also supports the text.
- ADD when one main-clothing region is specific, the other main-clothing region is compatible, and a strong stable detail also supports the text, such as carried-object type, hat color/type, scarf/inner clothing, non-common bag type, or a clearly distinctive non-black/non-white bag color.
- ADD when one main-clothing region has clear support and at least two independent stable details also support the text, such as carried-object type, hat color/type, scarf/inner clothing, non-common bag type, or a clearly distinctive non-black/non-white bag color.

- ADD when one specific main-clothing cue is clearly supported and one strong stable detail also clearly supports the text, but only if the possible conflict evidence is invalid or not visible enough.
- ADD only when the listed evidence is visibly about the main pedestrian, not the background.
- Treat listed conflict evidence as a risk signal. SKIP unless every listed conflict is clearly invalid, not visible enough, or caused by blur, occlusion, cropping, lighting, or a compatible synonym.
- SKIP if positive evidence does not include both `upper_clothing` and `lower_clothing` support, unless one main-clothing region has clear support and at least two independent stable details also support the text, or one specific main-clothing cue plus one strong stable detail both clearly support the text.
- SKIP if a main-clothing region is supported only by a generic type such as outerwear or pants, unless the other main-clothing region is specific and a strong stable detail also supports the text.
- SKIP if support is only generic black/dark/blue/gray outerwear or pants with no strong stable detail.
- SKIP if the only distinctive support comes from a common bag, handheld color, or shoe while main clothing remains generic.
- SKIP if the only auxiliary support is a common black/dark bag, backpack, shoulder bag, or shoe detail.
- SKIP if the only stable detail is simple bag/hat/handheld presence without a supported color, type, or object.
- SKIP if only one main-clothing region matches and there is neither two independent stable details nor one strong stable detail paired with a specific main-clothing cue.
- SKIP if the only strong auxiliary cue is shoe color or shoe type.
- SKIP if the match depends mainly on shoe color, handheld color, hair/action, background, or hidden details.
- SKIP when visibility, blur, occlusion, or cropping makes the positive evidence uncertain.
- SKIP when the pair is merely plausible but not strongly supported.

Return strictly: `DECISION: ADD / SKIP`

C.3 MLLM Inference Settings

Because IML directly uses MLLM decisions to refine pseudo correspondences during training, stochastic decoding would introduce unnecessary variation into the adaptation process. Accordingly, the MLLM temperature is set to 0, top-p to 1.0, and top-k to -1 to keep the verification process deterministic. Rounds 1 and 3 each generate at most 16 tokens because they only return short decisions, while Round 2 can generate up to 512 tokens for the structured JSON attributes. Because Round 2 depends only on the image, we extract the attributes of each target image only once and reuse them for both filtering and recovery. This avoids repeated attribute extraction and reduces unnecessary MLLM inference. The MLLM is frozen throughout adaptation and receives no gradients from the retrieval model.

D Stability Analysis

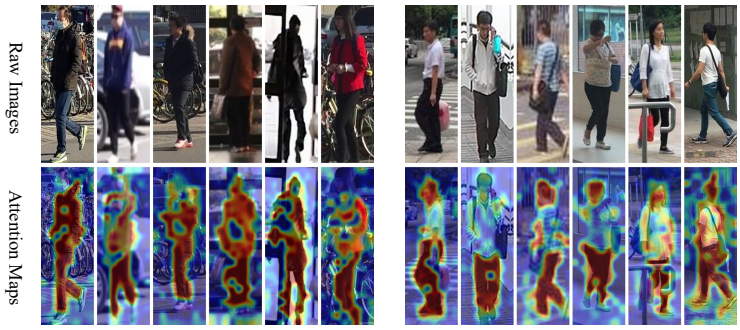
To evaluate whether the gains of ISV are robust to ordinary training randomness rather than a favorable initialization, we repeat the complete adaptation process with three random seeds under identical settings. As shown in Table 6, the variation remains small across all six transfer directions, with the standard deviations of the main retrieval metrics staying consistently limited. This stability across different source and target combinations indicates that the improvements of ISV are reproducible and are not driven by a particular random seed.

Table 6: Stability over six transfer directions. Mean±standard deviation is reported over three seeds.

Transfer	R@1	R@5	R@10	mAP	mINP
C→I	58.38 ± 0.14	75.14 ± 0.02	81.22 ± 0.07	32.22 ± 0.03	4.34 ± 0.06
C→R	62.07 ± 0.10	81.95 ± 0.40	88.67 ± 0.08	46.92 ± 0.25	23.89 ± 0.11
I→C	60.66 ± 0.32	80.68 ± 0.08	87.11 ± 0.40	53.89 ± 0.08	37.60 ± 0.04
I→R	61.88 ± 0.38	82.10 ± 0.26	88.45 ± 0.55	47.76 ± 0.38	24.90 ± 0.34
R→I	54.87 ± 0.09	71.50 ± 0.19	78.05 ± 0.17	33.06 ± 0.02	5.91 ± 0.03
R→C	59.24 ± 0.14	78.98 ± 0.20	85.90 ± 0.25	52.93 ± 0.08	36.79 ± 0.23

E Additional Qualitative Results

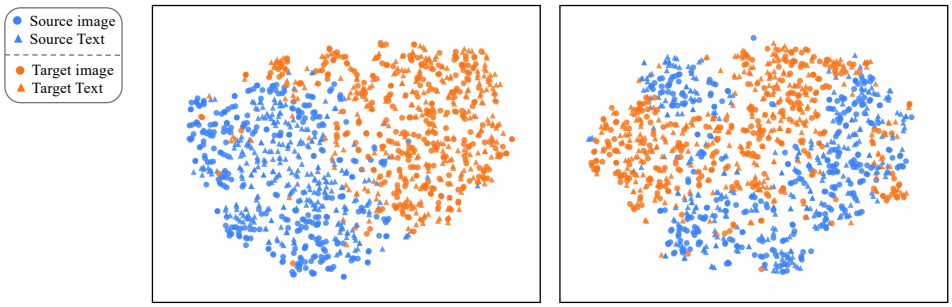
Student Attention Maps. To examine whether TSG can transfer source text guidance into spatial cues that remain useful without text at inference, we visualize the student attention in both domains. As shown in Figure 5, the responses are concentrated mainly on pedestrian regions rather than background areas in both the source and target domains. This behavior suggests that the student learns a transferable spatial prior from the teacher and can preserve pedestrian-focused attention after removing textual guidance.



(a)Source Domain (b)Target Domain

Figure 5: Student attention maps from TSG on R→C. Raw images are shown above the corresponding attention maps for (a) the source domain and (b) the target domain.

Feature Distribution. To examine whether ISV reduces domain discrepancy while maintaining image and text alignment, we visualize the learned representations with t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008). Figure 6 shows that the source and target samples exhibit stronger overlap under ISV than under the Baseline, while image and text features remain locally interleaved rather than forming separated modality clusters. This pattern indicates that ISV improves domain alignment without sacrificing cross-modal consistency, which is consistent with the retrieval gains observed in the quantitative experiments.



Baseline Ours

Figure 6: t-SNE visualization on R→C. Color denotes the domain and marker shape denotes the modality.

F Sensitivity to MLLM Activation Duration

To determine how long external semantic verification is beneficial during adaptation, we vary the number of early epochs in which the MLLM is activated on I→R while keeping all other settings fixed. It is worth noting that, to evaluate the effect of sustained semantic verification, we set the minimum activation duration to two epochs. For each candidate set, the lowest 30% of DAC candidates by similarity are sent for semantic verification to isolate the effect of activation duration. As shown in Figure 7, performance improves as the activation duration increases from two to three epochs and then remains relatively stable with further extension. This suggests that semantic evidence is most useful when target features and pseudo correspondences are still unreliable early in adaptation. We therefore use three MLLM active epochs in the final configuration, as this setting achieves a favorable balance between retrieval performance and verification cost.

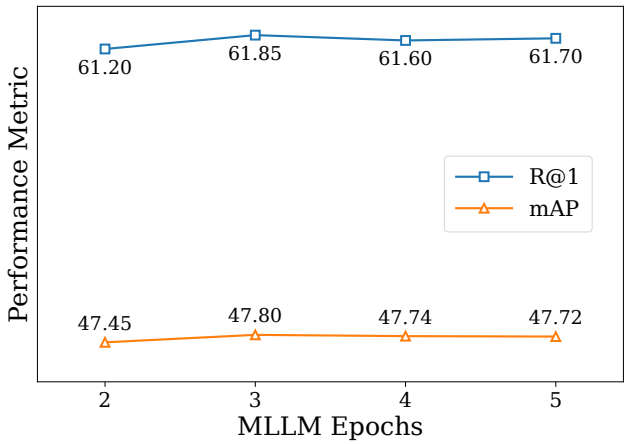


Figure 7: Sensitivity to MLLM activation duration on I→R. R@1 and mAP are reported for 2–5 MLLM epochs. The x-axis indicates the number of initial epochs during which the MLLM is activated.

G Efficiency and Computational Cost

To evaluate the additional cost introduced by semantic verification and determine whether a larger verifier is justified, we compare Qwen3-VL-8B and Qwen3-VL-32B on I→R under the same adaptation schedule and retrieval backbone. Both models remain frozen and are used only during the early adaptation stage. As shown in Table 7, the 32B verifier requires substantially more inference time, total training time, and GPU memory than the 8B model. Since the main scale experiment shows no consistent retrieval benefit from increasing the verifier size, the additional computation does not translate into a reliable performance gain. These results support Qwen3-VL-8B as the default verifier and also show that the MLLM cost is confined to training because no MLLM is required at test time.

Table 7: Training cost of Qwen3-VL-8B and Qwen3-VL-32B on I→R. GPU memory is measured during the first three epochs when the MLLM is active.

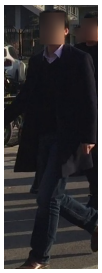
MLLM	MLLM Parameters	MLLM inference time (h)	Total training time (h)	GPU memory (GiB)	MLLM at test time
Qwen3-VL-8B-Instruct	8.77B	4.15	9.36	53.6	No
Qwen3-VL-32B-Instruct	33.36B	10.59	15.31	82.0	No

H MLLM FAILURE CASES

(a) False REJECT



Text: A man with a cream scarf, an orchid-purple coat, light tan trousers, black shoes, and a white-pink floral side bag.
Ground Truth: matched pair
MLLM Decision: `REJECT`
Evidence: purple coat vs. pink jacket.
Failure: a fine color difference is treated as a strong contradiction.

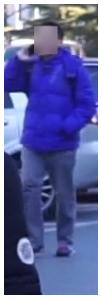


Text: A man in a black coat and black-white polo shirt, with gray jeans and dark-brown Oxford shoes.
Ground Truth: matched pair
MLLM Decision: `REJECT`
Evidence: gray vs. blue jeans; Oxford shoes vs. sneakers.
Failure: uncertain lower body color and shoe type lead to two false conflicts.

(b) False KEEP



Text: A young man in a navy-blue hooded down coat, carrying a black shoulder bag, with blue jeans and blue sneakers.
Ground Truth: mismatched pair
MLLM Decision: `KEEP`
Evidence: blue vs. gray upper clothing; blue vs. gray lower clothing.
Failure: two moderate color conflicts are detected, but the conservative verifier still keeps the pair.



Text: A young man in a blue puffer jacket over a white shirt, with black fitted pants and black running shoes.
Ground Truth: mismatched pair
MLLM Decision: `KEEP`
Evidence: black vs. gray lower clothing.
Failure: the matching blue upper clothing and a single moderate conflict are not sufficient to reject the pair.

Figure 8: Representative MLLM failure cases on ICFG-PEDES. Row (a) shows false rejections of matched pairs, while row (b) shows false keeps of mismatched pairs.

To understand where semantic verification still fails, we examine four image-text pairs incorrectly classified by the MLLM on ICFG-PEDES. Dataset identity annotations are used only for later analysis to determine whether each pair is matched or mismatched. As shown in Figure 8, the four examples include both false rejections of true pairs and false keeps of mismatched pairs. Figure 8(a) shows that false rejections arise when fine-grained visual attributes are ambiguous. Small color differences or uncertain shoe and clothing types can be interpreted as strong contradictions, causing a valid pair to be removed. In contrast, Figure 8(b) shows that false keeps are associated with the conservative rejection policy: moderate conflicts may be retained when the available evidence is not strong enough to justify rejection. Together, these cases reveal the tradeoff of conservative semantic verification. It protects difficult positives from aggressive filtering but can also preserve some mismatched pairs when visual evidence remains ambiguous or incomplete.

I THE USE OF LARGE LANGUAGE MODELS

We used large language models to polish the language and search for related work during manuscript preparation. This use is separate from Qwen3-VL in our experiments. Qwen3-VL remains frozen and is used only for semantic verification during adaptation. We take full responsibility for the paper.