CORRESPONDENCE-REPRESENTATION CO-EVOLUTION FOR WEAKLY SUPERVISED VISIBLE-INFRARED PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised visible-infrared person re-identification learns from independently labeled visible and infrared images of the same training identities, without access to their cross-modal correspondences. When provisional matches supervise representation learning, early matching errors can distort the representation and be reinforced in subsequent correspondence inference. In this paper, we propose ReCUE, a correspondence-representation co-evolution framework that uses agreement among competing global assignments to guide supervision while keeping every correspondence revisable. Specifically, Global Correspondence Hypothesis Inference (GCHI) enumerates the K lowest-cost feasible one-to-one assignments and measures each candidate match's support by its frequency across these assignments. Structural Uncertainty-Aware Learning (SUAL) translates this global support into differentiated supervision: highly supported matches receive hard targets, ambiguous matches retain soft targets over competing counterparts, and weakly supported matches receive no correspondence-derived positive supervision. As the representation evolves, Correspondence-Representation Co-Evolution (CRCE) periodically reconstructs the complete correspondence state and associated supervision, allowing previously accepted matches to be reassigned or downgraded. Experiments on SYSU-MM01 and LLCM validate the effectiveness of ReCUE, which achieves 73.0%/69.0% Rank-1/mAP on SYSU-MM01 All Search and 57.0%/60.5% on LLCM VIS→IR. The code will be available.

1 INTRODUCTION

Visible-infrared person re-identification (VI-ReID) aims to match pedestrian identities across visible and infrared modalities (Li et al., 2026b; Shi et al., 2026). Conventional supervised methods require consistent identity annotations across both modalities, which involves manually associating identities across substantial cross-spectrum appearance differences, making the process time-consuming and labor-intensive. Weakly supervised VI-ReID reduces this annotation burden by requiring identity labels only within each modality, without manually annotated cross-modal identity correspondences (Zhang et al., 2025). This preserves intra-modal supervision while removing the need for manual identity association across modalities. The resulting challenge is to infer latent correspondences while learning a representation that generalizes to unseen identities.

Previous works explore this problem through complementary strategies. Pairwise and memory-based methods improve local matching by comparing cross-modal features or using historical identity information (Yang et al., 2022; 2023a; Shi et al., 2024; Ye et al., 2025). Reliability-aware methods adjust supervision using estimated confidence, while ranked assignment reveals alternatives through feasible bijections (Zhang et al., 2025; Yang et al., 2023b; Zhao et al., 2024; Dehler et al., 2024). Despite these advances, two limitations remain. First, correspondence evidence is often unreliable early in training, when cross-modal representations are still immature. Local pairwise scores are therefore noisy, while retaining only the minimum-cost global assignment collapses several near optimal mappings into a single decision, obscuring whether a selected edge is genuinely stable. Second, reliability-aware methods mainly reduce the influence of uncertain pairs through scalar confidence estimates, without characterizing their competition under the full assignment constraint. More importantly, correspondence inference and representation learning form a feedback

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.

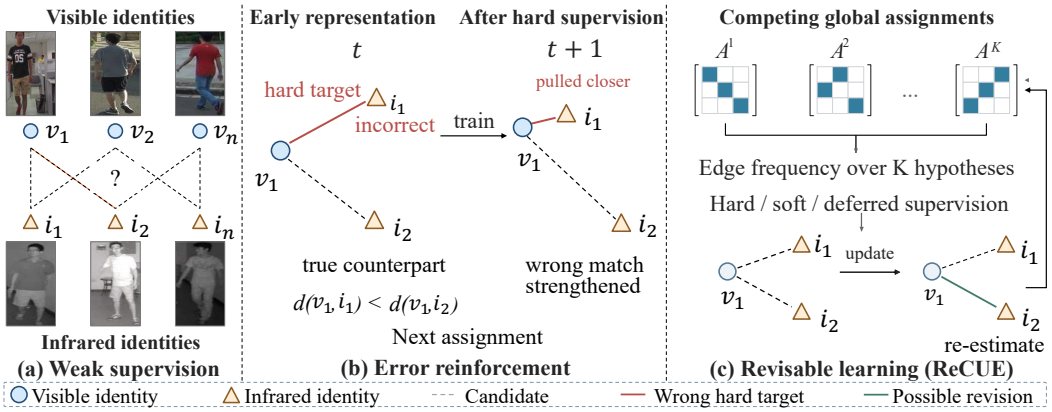


Figure 1: **Motivation for revisable cross-modal correspondence learning.** (a) WS-VIReID provides independent identity labels within each modality, while cross-modal correspondences remain unknown. (b) With an immature representation, an incorrect counterpart may appear closer than the true match. Using this provisional match as a hard target pulls the wrong pair closer, reinforcing the error in subsequent assignments. (c) ReCUE measures correspondence support by edge frequency across the lowest-cost feasible one-to-one assignments. This support determines whether to assign hard targets, retain soft targets over competing counterparts, or defer correspondence-derived positive supervision. As the representation evolves, ReCUE periodically reconstructs the complete correspondence state, allowing earlier matches and supervision decisions to be revised.

loop: provisional correspondences supervise the representation, which in turn determines later correspondences. As a result, early matching errors may be reused and progressively reinforced.

In this paper, we propose ReCUE, a **R**evisable **C**orrespondence and **U**ncertainty-guided Representation Co-**E**volution method for weakly supervised VI-ReID that jointly evolves latent cross-modal correspondences and identity representations. Specifically, ReCUE first uses Global Correspondence Hypothesis Inference (GCHI) to combine cached identity-prototype geometry with bidirectional identity-predictor evidence and enumerate the K lowest-cost feasible one-to-one assignments. Their edge-wise agreement forms an unweighted truncated marginal that measures each candidate correspondence's support across competing global hypotheses. Structural Uncertainty-Aware Learning (SUAL) then converts this support into differentiated supervision: highly supported relations receive hard targets, ambiguous relations retain consensus-based soft targets, and unreliable relations receive no correspondence-derived positive supervision. Because these supervision decisions depend on the evolving representation, Correspondence-Representation Co-Evolution (CRCE) periodically reconstructs the complete correspondence state from updated prototypes and matching costs, allowing previously selected relations to be promoted, demoted, or replaced. Together, the three components form a revisable correspondence-representation learning loop rather than committing to a fixed set of early pseudo-labels. The main contributions are summarized as follows:

- We propose ReCUE, a framework that jointly learns a latent one-to-one cross-modal correspondence and an identity representation without access to cross-modal identity links.
- We propose GCHI to estimate correspondence reliability from the unweighted truncated edge marginals of the K-best feasible bijections, thereby capturing global competition among candidate matches.
- We propose SUAL and CRCE to route uncertainty-aware supervision and periodically reconstruct correspondences, enabling continuous revision.
- Experimental results on the SYSU-MM01 and LLCM benchmarks consistently demonstrate the effectiveness and robustness of ReCUE.

2 Related Work

VI-ReID with incomplete cross-modal supervision. Fully supervised methods assume a shared identity space across visible and infrared modalities (Zhang & Wang, 2023; Yu et al., 2025), whereas

recent unsupervised approaches combine intra-modal clustering, memory banks, and cross-modal association (Ye et al., 2025; Yang et al., 2025; Pang et al., 2025; Wang et al., 2025; Cheng et al., 2026). The weakly supervised protocol considered here retains identity labels within each modality while withholding the unknown cross-modal bijection (Zhang et al., 2025). It is distinct from unpaired and sparsely paired settings, where cross-modal supervision is either absent or only partially available (Yao et al., 2025; Lu et al., 2026). Among methods most closely related to our optimization framework, PGM alternates graph matching with representation learning, whereas MIGOT and prototype-transport methods repeatedly solve global OT-based association problems (Wu & Ye, 2023; Zhang et al., 2024; 2026b). These approaches motivate joint correspondence learning, but repeatedly selecting a single refreshed optimum does not reveal the stability of individual edges across competing global mappings.

Assignment uncertainty. Ranked assignment enumerates feasible solutions in increasing order of cost, and m-best marginals have been used to characterize ambiguity across multiple global assignments (Paul et al., 2025; Dehler et al., 2024). However, these techniques are primarily used to quantify association ambiguity rather than to control how uncertain correspondences supervise representation learning. ReCUE uses ranked global support directly as a supervision signal.

Reliability-aware learning and iterative correction. Methods for noisy correspondence learning typically soften, reweight, or filter supervision using local, bidirectional, or learned reliability estimates (Yang et al., 2023b; Zhao et al., 2024; Zha et al., 2024; Li et al., 2026a). Iterative correction further allows correspondences to be updated as representations improve (Ding et al., 2025; Zhou et al., 2026; Zhang et al., 2026a; Lv et al., 2026; Tian et al., 2026). However, reliability is usually estimated for individual relations, while iterative updates often operate on a single current correspondence state; neither explicitly represents how an edge competes across multiple globally feasible assignments while keeping earlier decisions fully revisable. ReCUE combines ranked global support with non-absorbing correspondence reconstruction to address these two limitations.

3 Methodology

3.1 Overview and Formulation

As shown in Fig. 2, ReCUE consists of three coupled components. At the beginning of each training stage, Global Correspondence Hypothesis Inference (GCHI) constructs the K lowest-cost one-to-one assignments and measures the support of each candidate match across them. Structural Uncertainty-Aware Learning (SUAL) uses this support to select hard, soft, or deferred cross-modal supervision. Correspondence-Representation Co-Evolution (CRCE) rebuilds these assignments and supervision decisions after the representation has been updated.

Let $\mathcal{D}^v = \{(x_p^v, y_p^v)\}_{p=1}^{N_v}$ and $\mathcal{D}^i = \{(x_q^i, y_q^i)\}_{q=1}^{N_i}$ denote the visible and infrared training sets, respectively. Identity labels are available within each modality, but their cross-modal correspondence is unknown. Since the two label spaces describe the same N training identities, their unknown cross-modal identity mapping is a permutation. Accordingly, the feasible assignment space Ω is:

$$\Omega = \left\{ A \in \{0,1\}^{N\times N} \;\middle|\; A\mathbf{1}_N = \mathbf{1}_N,\; A^\top \mathbf{1}_N = \mathbf{1}_N \right\}. \tag{1}$$

For partially overlapping label spaces, this formulation would require partial assignments with an explicit unmatched state. Let $A^\star \in \Omega$ denote the unobserved training-identity permutation. Neither $A^\star$ nor any of its edges is available for constructing assignments, routing supervision, tuning hyperparameters, or selecting a model. The learner must infer training correspondences from the two intra-modal label spaces and learn a representation that transfers to disjoint test identities. The withheld $A^\star$ is accessed only after training for correspondence-recovery analysis, while retrieval on the standard test split remains the primary objective. Let $F^{(t)}$ denote the feature representation at training stage t, and let θ collect the trainable encoder and identity-predictor parameters. At stage t, $F^{(t)}$ generates global hypotheses $A_{1:K}^{(t)}$ and ranked-assignment support $\rho^{(t)}$, which guide the next representation update while keeping correspondence decisions revisable.

3.2 Global Correspondence Hypothesis Inference

At refresh stage t, identity c in modality $\mu \in \{v, i\}$ is summarized by the cached prototype:

$$P_c^{\mu,(t)} = \mathrm{sg}\left[\mathrm{Norm}\left(|\mathcal{I}_c^\mu|^{-1} \sum_{p\in\mathcal{I}_c^\mu} E_{\theta^{(t)}}^\mu (x_p^\mu) \right) \right]. \tag{2}$$

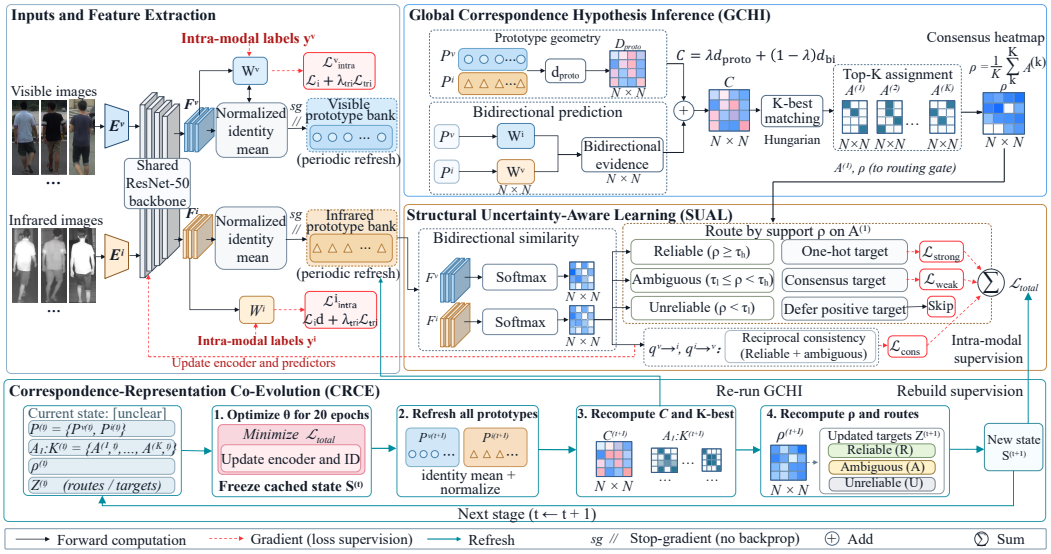


Figure 2: Overview of ReCUE. GCHI combines prototype geometry and bidirectional predictor evidence to construct the assignment cost, enumerates the K lowest-cost feasible one-to-one assignments, and computes edge-wise support ρ. SUAL converts this support into hard, soft, or deferred cross-modal supervision. CRCE alternates representation updates with full prototype and correspondence reconstruction, allowing previously selected matches and supervision decisions to remain revisable.

The prototype bank remains fixed until the next refresh, and stage superscripts are omitted below. For an identity pair (a, b), the prototype distance is $d_{\text{proto}}(a, b) = [1 - (P_a^v)^\top P_b^i]/2$.

Bidirectional evidence and joint cost. We complement prototype geometry with directional evidence from the two modality-specific identity predictors. Let $\bar{w}_c^\mu = w_c^\mu / \|w_c^\mu\|_2$, with the normalized weights collected in $\bar{W}^\mu$. For $(\mu, \nu) \in \{(v, i), (i, v)\}$, define $q_{c\to d}^{\mu\to\nu}$ as the d-th entry of $\text{Softmax}((\bar{W}^\nu)^\top P_c^\mu / \tau_c)$. Then, the bidirectional discrepancy and assignment cost are:

$$d_{\text{bi}}(a, b) = 1 - \sqrt{q_{a\to b}^{v\to i} q_{b\to a}^{i\to v}}, \qquad C_{ab} = \lambda d_{\text{proto}}(a, b) + (1 - \lambda) d_{\text{bi}}(a, b). \tag{3}$$

A low d_{bi} therefore requires consistent evidence in both directions, while $\lambda \in [0, 1]$ balances prototype geometry and predictor evidence. Rather than retaining only the minimum-cost assignment, we extract the K lowest-cost globally feasible solutions:

$$A_{1:K} = \text{KBest}_{A\in\Omega} \langle C, A \rangle_F = \left\{ A^{(1)}, A^{(2)}, \ldots, A^{(K)} \right\}, \tag{4}$$

where the assignments are ordered by nondecreasing cost. The first assignment is the preferred correspondence, while the remaining assignments expose competitive alternatives. We enumerate the K lowest-cost feasible assignments exactly using Murty's algorithm (Murty, 1968), with a Hungarian solver for its constrained assignment subproblems. Appendix A.1 formalizes the resulting edge stability, and Appendix B.2 provides complexity details.

3.3 Ranked-Assignment Support and Uncertainty-Aware Learning

We measure the structural support of a candidate edge across the K lowest-cost globally feasible assignments. Unlike an isolated pair score, this quantity reflects whether a relation survives competition under the complete one-to-one matching constraint. We define ρ_{ab} as an unweighted truncated edge marginal over the enumerated bijections. Related m-best marginals have been used to characterize ambiguity in probabilistic matching and data association (Rezatofighi et al., 2016; Paul et al., 2025). For an identity pair (a, b), its ranked-assignment support is:

$$\rho_{ab}^{(t)} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\left[A_{ab}^{(k,t)} = 1 \right], \qquad u_{ab}^{(t)} = 1 - \rho_{ab}^{(t)}. \tag{5}$$

where $\rho_{ab}^{(t)}$ is the ranked-assignment support of edge (a, b) and $u_{ab}^{(t)} = 1 - \rho_{ab}^{(t)}$ is its structural uncertainty. A high value of $\rho_{ab}^{(t)}$ means that the edge persists across the enumerated competitive bijections, whereas a low value means that it is replaced by alternative globally feasible relations.

Assignment-marginal structure. Because every $A^{(k,t)}$ is a permutation matrix, their average satisfies $\rho^{(t)}\mathbf{1} = \mathbf{1}$ and $\rho^{(t)\top}\mathbf{1} = \mathbf{1}$. Thus, $\rho^{(t)}$ is doubly stochastic and can serve as a valid soft target in both retrieval directions. The marginal also admits a global stability interpretation. Let J_K denote the cost of the Kth assignment, and let J_e^+ and J_e^- denote the minimum feasible costs when edge e is forced to be present and absent, respectively. Then the following cost-gap conditions identify edges that are guaranteed to appear in or be excluded from all K-best assignments:

$$J_e^- > J_K \Rightarrow \rho_e = 1, \qquad J_e^+ > J_K \Rightarrow \rho_e = 0. \tag{6}$$

These conditions show that extreme marginal values certify agreement within the enumerated global alternatives. Accordingly, ρ is a truncated structural support measure rather than a calibrated posterior probability over correspondences. The proof is provided in Appendix A.1. For each correspondence selected by the current best hypothesis $A^{(1,t)}$, we classify it as reliable if $\rho_{ab}^{(t)} \geq \tau_{\text{high}}$, ambiguous if $\tau_{\text{low}} \leq \rho_{ab}^{(t)} < \tau_{\text{high}}$, and unreliable otherwise. We denote the three sets by $\mathcal{R}^{(t)}$, $\mathcal{A}^{(t)}$, and $\mathcal{U}^{(t)}$, respectively. Routing during training is determined directly by ρ, while $u = 1 - \rho$ is used only for post-hoc error detection.

Uncertainty-controlled supervision. Let $f_s^\mu = \text{Norm}(E_\theta^\mu(x_s^\mu))$ denote an online sample embedding. During each refresh stage, the cached opposite-modality prototypes serve as fixed cross-modal targets, while gradients are propagated through the online embeddings. Let $p_s^{v\to i}(\tau)$ denote the softmax distribution induced by the similarities between a visible sample and all cached infrared prototypes, with the reverse direction defined symmetrically.

The ranked-assignment support $\rho^{(t)}$ determines not only the strength but also the form of cross-modal supervision. For a correspondence selected by $A^{(1,t)}$, reliable relations use the selected partner as a hard target, ambiguous relations use $\rho_{a:}^{(t)}$ for visible anchors and $\rho_{:b}^{(t)}$ for infrared anchors as soft targets, and unreliable relations receive no correspondence-derived positive target. Since $\rho^{(t)}$ is doubly stochastic, the same consensus matrix defines valid targets in both retrieval directions without additional normalization.

Let $\mathcal{L}_{\text{strong}}^{(t)}$ denote the bidirectional hard-target cross-entropy over reliable relations, and let $\mathcal{L}_{\text{weak}}^{(t)}$ denote the bidirectional consensus-target cross-entropy over ambiguous relations. The complete stage objective is:

$$\mathcal{L}^{(t)} = \mathcal{L}_{\text{id}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_s\mathcal{L}_{\text{strong}}^{(t)} + \lambda_w\mathcal{L}_{\text{weak}}^{(t)} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}^{(t)}, \tag{7}$$

where $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{tri}}$ are the intra-modal identity-classification and batch-hard triplet losses, respectively. For $\mathcal{S}^{(t)} = \mathcal{R}^{(t)} \cup \mathcal{A}^{(t)}$, $\mathcal{L}_{\text{cons}}^{(t)}$ enforces reciprocal agreement between the two modality-specific predictors:

$$\mathcal{L}_{\text{cons}}^{(t)} = \frac{\sum_{(a,b)\in\mathcal{S}^{(t)}} \rho_{ab}^{(t)} \left(q_{a\to b}^{v\to i} - q_{b\to a}^{i\to v}\right)^2}{\sum_{(a,b)\in\mathcal{S}^{(t)}} \rho_{ab}^{(t)}}. \tag{8}$$

and is set to zero when $\mathcal{S}^{(t)}$ is empty. This routing changes the learning signal qualitatively rather than merely rescaling a pseudo-label: reliable relations receive a one-hot cross-modal target, ambiguous relations preserve competing globally feasible alternatives through $\rho_{a:}^{(t)}$, and unreliable relations contribute no correspondence-derived target gradient during the current stage. Cached prototypes and routing variables are fixed within a stage, whereas the encoder and identity predictors remain trainable. Appendix A gives directional loss definitions and mini-batch normalization details.

3.4 Correspondence-Representation Co-Evolution

The co-evolution procedure closes the feedback loop between correspondence inference and representation learning. At each refresh stage, the current representation induces ranked assignments and routed supervision, while the updated representation reconstructs the complete correspondence state for the next stage. Unlike methods that carry selected pseudo-labels forward, ReCUE allows all identity relations to re-enter the feasible assignment problem after each refresh. This design combines alternating correspondence learning (Wu & Ye, 2023; Zhang et al., 2024; Zhou et al., 2026;

Table 1: Comparison with state-of-the-art VI-ReID methods on SYSU-MM01 and LLCM (%).

Type	Method	SYSU-MM01						LLCM			
		All Search			Indoor Search			VIS→IR		IR→VIS	
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	R1	mAP
S-VIReID	AGW (Ye et al., 2022)	47.5	47.7	35.3	54.2	63.0	59.2	51.5	55.3	43.6	51.8
	LbA (Park et al., 2021)	55.4	54.1	-	58.5	66.3	-	50.8	55.6	43.8	53.1
	CAJ (Ye et al., 2021)	69.9	66.9	53.6	76.3	80.4	76.8	56.5	59.8	48.8	56.6
	DEEN (Zhang & Wang, 2023)	74.7	71.8	-	80.3	83.3	-	62.5	65.8	54.9	62.9
	MUN (Yu et al., 2023)	76.2	73.8	-	79.4	82.1	-	-	-	-	-
	CSDN (Yu et al., 2025)	76.7	73.0	-	84.5	86.8	-	63.7	66.5	55.8	63.5
SS-VIReID	MAUM-50 (Liu et al., 2022)	28.8	36.1	-	-	-	-	-	-	-	-
	MAUM-100 (Liu et al., 2022)	38.5	39.2	-	-	-	-	-	-	-	-
	OTLA (Wang et al., 2022)	48.2	43.9	-	47.4	56.8	-	44.2	48.2	36.2	42.2
	DPIS (Shi et al., 2023)	58.4	55.6	-	63.0	70.0	-	-	-	-	-
US-VIReID	PGM (Wu & Ye, 2023)	57.3	51.8	35.0	56.2	62.7	58.1	44.9	49.0	39.4	45.3
	PCAL (Yang et al., 2025)	57.9	52.9	36.9	60.1	66.7	62.1	-	-	-	-
	MMM (Shi et al., 2024)	61.6	57.9	-	64.4	70.4	-	-	-	-	-
	DLM (Ye et al., 2025)	62.2	58.4	43.7	67.3	72.7	68.9	-	-	-	-
	DMDL (Li et al., 2026b)	65.9	61.9	47.5	70.7	75.5	71.7	51.8	55.4	45.3	50.9
	HIL (Shi et al., 2026)	66.3	65.0	52.6	71.8	77.5	74.3	-	-	-	-
	CLNS (Zhao et al., 2026)	69.5	65.6	51.2	74.7	79.0	75.3	54.6	57.8	46.9	52.6
WS-VIReID	HECCL (Zhang et al., 2025)	70.4	66.6	52.7	76.5	80.2	76.6	55.3	58.7	47.3	53.3
	Ours	**73.0**	**69.0**	**55.7**	**78.7**	**82.5**	**79.5**	**57.0**	**60.5**	**49.2**	**55.6**

Zhang et al., 2026b) with assignment-level supervision routing and non-absorbing reconstruction. After the intra-modal warm-up, stage $t = 0$ begins at epoch 20. As illustrated in Fig. 2, each stage alternates four steps: we first optimize the encoder and identity predictors for 20 epochs while keeping the cached prototypes, correspondence hypotheses, and routing decisions fixed; we then refresh all identity prototypes, recompute the assignment cost and K-best hypotheses, and finally rebuild the support matrix and routed supervision for the next stage. The complete correspondence state is then recomputed as follows:

$$\begin{aligned} \theta^{(t+1)} &= \mathrm{Update}_{20}\Big(\theta^{(t)}, \nabla_\theta \mathcal{L}(\theta; \mathcal{R}^{(t)}, \mathcal{A}^{(t)}, \mathcal{U}^{(t)})\Big), \\ A_{1:K}^{(t+1)} &= \mathrm{KBest}_{A\in\Omega}\langle C^{(t+1)}, A\rangle_F, \quad \rho^{(t+1)} = \Phi(A_{1:K}^{(t+1)}), \end{aligned} \tag{9}$$

where Update_{20} denotes 20 epochs of optimizer updates rather than an exact minimization, and Φ is the marginal computation defined in Eq. 5. If $z_{ab}^{(t)}$ denotes the routing decision for edge (a, b), each refresh recomputes $z_{ab}^{(t+1)} = \Gamma(\rho_{ab}(C(\theta^{(t+1)})))$ without imposing the absorbing constraint $z_{ab}^{(t)} = \mathcal{R} \Rightarrow z_{ab}^{(t+1)} = \mathcal{R}$. Previous supervision continues to influence θ, but no edge is forced to persist. Stages $t = 0, \ldots, 4$ are used for model training, and the sixth reconstruction at epoch 120 is used for final analysis. Appendix A.1 proves that complete reconstruction preserves access to every feasible bijection, whereas locking an incorrect edge may exclude the true correspondence. The complete procedure is given in Appendix B.1.

3.5 Experimental Setup

Datasets and evaluation protocols. We evaluate ReCUE on SYSU-MM01 (Wu et al., 2017), LLCM (Zhang & Wang, 2023). SYSU-MM01 contains 395 training and 96 test identities. We follow the standard single-shot All-Search and Indoor-Search protocols and average over 10 random gallery samplings. LLCM contains 713 training and 351 test identities and is evaluated in both VIS→IR and IR→VIS directions. We report Rank-1, mAP, and mINP, using mAP as the primary retrieval metric. The ground-truth cross-modal mapping is never used for training, hyperparameter selection, or model selection, and correspondence statistics are used only for post-hoc analysis. Error detection is evaluated by AUROC and AUPRC, treating incorrect correspondences as positives.

3.6 Comparison with State-of-the-Art Methods

We compare ReCUE with fully supervised (S-VIReID), semi-supervised (SS-VIReID), unsupervised (US-VIReID), and weakly supervised VI-ReID (WS-VIReID) methods under their published protocols. Tab. 1 summarizes the results on SYSU-MM01 and LLCM.

Table 2: Ablation studies on SYSU-MM01. (a) Component ablation, where Correct is the number of recovered cross-modal correspondences out of 395. (b) Matched-budget comparison of structural uncertainty representations; AUROC/AUPRC treat incorrect correspondences as positives. (c) Representation-and-update controls comparing locked versus fully revisable updating, with full-plan soft OT included as a routing-free control. (d) Routing ablation comparing hard supervision, continuous confidence weighting, soft targets, and deferral.

(a) Component Ablation

Setting	Correct	R1	mAP
Baseline	286	70.7	66.3
+ Global	294	71.0	67.1
+ Stability	307	72.3	68.0
+ Full rebuild	**316**	**73.0**	**69.0**

(b) Structural Uncertainty Comparison

Source	AUROC↑	AUPRC↑	R1↑	mAP↑
Local	0.688	0.407	72.32	68.04
Sinkhorn	0.771	0.493	72.54	68.34
Weighted K	0.790	0.489	**73.00**	68.84
Uniform K	**0.803**	**0.520**	72.96	**68.98**

(c) Representation and Update Controls

Representation	Update	R1	mAP
Full-plan OT	Soft target	72.38	68.10
Local	Locked	72.34	68.10
Local	Full	72.32	68.04
Sinkhorn	Locked	72.40	68.32
Sinkhorn	Full	72.54	68.34
Uniform K	Locked	72.50	68.52
Uniform K	Full	**72.96**	**68.98**

(d) Uncertainty-Aware Routing

Routing	Corr.	R1	mAP
All hard	77.0	71.8	67.3
Confidence weighted	78.2	72.2	67.8
Hard + soft, no defer	79.0	72.5	68.3
Hard + defer, no soft	78.7	72.4	68.5
Hard + soft + defer	**80.0**	**73.0**	**69.0**

On SYSU-MM01, ReCUE achieves 73.0% Rank-1, 69.0% mAP, and 55.7% mINP under All Search, and 78.7%, 82.5%, and 79.5% under Indoor Search, respectively, exceeding the reported HECCL results by 2.2-3.0% across the two settings. The consistent gains in Rank-1, mAP, and mINP indicate that ReCUE improves not only the top retrieved match but also the broader ordering of same-identity samples. This is consistent with its revisable correspondence supervision, which reduces the influence of unstable early cross-modal matches during representation learning. On LLCM, ReCUE achieves 57.0% Rank-1 and 60.5% mAP for VIS→IR, and 49.2% Rank-1 and 55.6% mAP for IR→VIS, outperforming HECCL by 1.7-2.3 percentage points. The improvement in both retrieval directions suggests that the benefit is not specific to one query modality. Together with the SYSU-MM01 results, this supports the effectiveness of revisable correspondence learning across different cross-modal retrieval conditions.

3.7 Analysis of the Proposed Method

Component ablation. Tab. 2(a) reveals a clear progression in how correspondence information is converted into retrieval performance. Introducing a globally constrained Top-1 assignment improves the correspondence state, but the resulting retrieval gain remains modest, indicating that enforcing one-to-one consistency alone is not sufficient. A larger improvement appears when Top-K support is used to route supervision, showing that the competing globally feasible assignments provide useful information beyond the single lowest-cost solution. In particular, their agreement determines whether a selected relation should be trusted, softened, or deferred during representation learning. Full-state reconstruction then provides an additional gain by allowing these supervision decisions

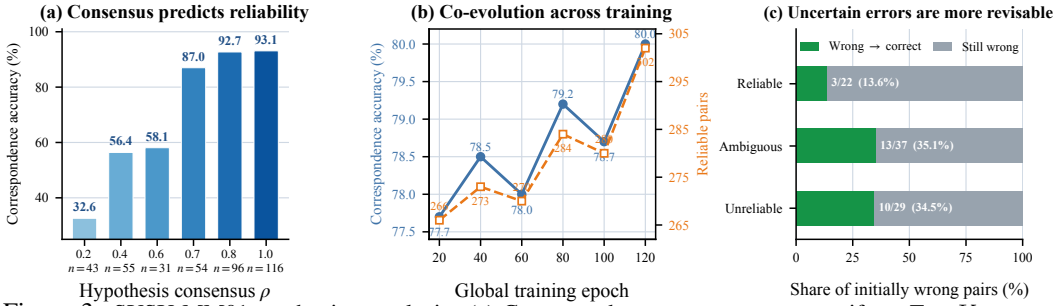


Figure 3: SYSU-MM01 mechanism analysis. (a) Correspondence accuracy across uniform Top-K consensus levels, with n denoting the number of selected edges. (b) Co-evolution from epochs 20 to 120, showing correspondence accuracy and reliable-set size. (c) Recovery of epoch-20 errors by their initial routing state, evaluated at epoch 120. Green denotes wrong-to-correct transitions and gray denotes errors that remain wrong.

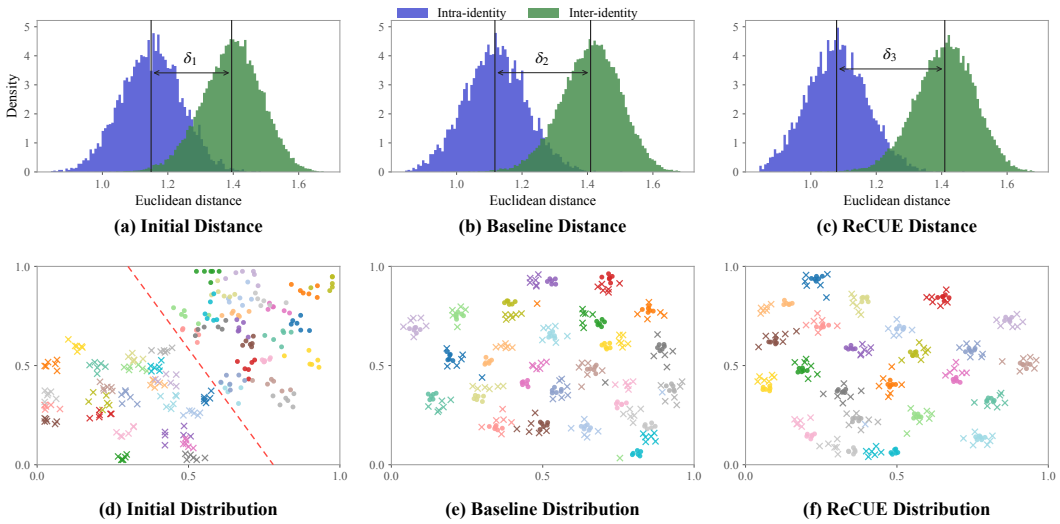


Figure 4: (a-c) show the intra-identity and inter-identity distances of cross-modality features for the initial representation, baseline, and ReCUE, respectively. The intra-identity and inter-identity distances are indicated in blue and green, respectively. (d-f) show the distributions of feature embeddings in the 2D feature space. A total of 20 identities are randomly selected from the test set, with the same identities and samples used across all three representations. Samples with the same color belong to the same identity. The dot and cross markers denote images from the VIS and IR modalities, respectively.

to be revised as the representation changes, instead of turning early correspondence choices into persistent constraints.

Structural uncertainty from hypothesis consensus. Tab. 2(b) compares Top-K support with local and Sinkhorn alternatives. The Top-K family is strongest overall: weighted consensus leads Rank-1, while uniform consensus leads AUROC, AUPRC, and mAP. This indicates that agreement across competing feasible assignments provides a stronger reliability signal than isolated local or transport-based scores.

Representation and update controls. Tab. 2(c) crosses local, Sinkhorn, and uniform Top-K representations with locked and revisable updating. Revision changes little under local or Sinkhorn uncertainty but improves both metrics under Top-K. Full-plan OT remains competitive, indicating that the strongest result comes from coupling structured reliability with reconstruction rather than from alternation alone.

Hard, soft, and deferred supervision. Tab. 2(d) isolates routing. Making every selected edge a hard positive performs worst. Continuous weighting helps but does not retain competing targets or defer low-support relations. Removing either soft targets or deferral hurts, whereas their combination is strongest, distinguishing the interface from confidence rescaling alone.

Mechanism analysis. Fig. 3(a) shows that higher Top-K consensus is associated with cleaner correspondences at the first checkpoint, supporting its use as a correspondence-reliability signal. Extreme marginal values also admit a global certificate against the enumerated bijections rather than an isolated pair score. Appendix C provides additional reliability analyses. Fig. 3(b) shows the reference Full trajectory from epochs 20 to 120. Correspondence accuracy and reliable-set coverage increase overall but exhibit intermediate reversals, illustrating that reconstruction is non-absorbing rather than monotonic. Fig. 3(c) further shows that initially wrong correspondences routed as ambiguous or unreliable are corrected much more often than those already deemed reliable. This indicates that the routing score not only separates correspondence reliability but also identifies errors with greater potential for later revision.

Feature geometry and retrieval behavior. Fig. 4 shows that ReCUE does not simply contract the embedding space: it improves cross-modal alignment within identities while largely preserving inter-identity separation. This behavior reflects uncertainty-aware routing, with strong, soft, or deferred supervision according to correspondence support. Consequently, ReCUE suppresses modality-driven offsets without sacrificing the identity structure required for discrimination. The

Table 3: SYSU-MM01 correspondence revision under clean and perturbed initialization. (a) Correctness transitions from epoch 20 to 120 under different update policies, where W→C and C→W denote wrong-to-correct and correct-to-wrong changes. (b) Recovery of initially wrong correspondences under Full, grouped by their epoch-20 routing states. (c) Net correction under different initial perturbation strengths α. (d) Full-minus-Locked gains in Rank-1 and mAP across the same 10 random gallery samplings; brackets report 95% t-confidence intervals computed from the paired sampling-wise differences.

(a) Correctness flow by update policy

Policy	W→C	C→W	Net	Final
Fixed	0	0	0	307
Naive	18	15	+3	310
Locked	15	9	+6	313
Full	**26**	17	+9	**316**

(b) Full recovery by initial route

Route	Wrong	W→C	Recovery	C→W
Reliable	22	3	13.6	5
Ambiguous	37	13	35.1	8
Unreliable	29	10	34.5	4
Total	88	26	29.5	17

(c) Net correction under perturbation

α	Initial	Naive	Locked	Full
0	307	+3	+6	+9
0.15	298	+8	+11	+15
0.30	284	+12	+16	+25
0.50	261	+19	+24	+37

(d) Paired Full–Locked gain (95% CI)

α	Δ Rank-1	Δ mAP
0	+0.46 [0.27, 0.65]	+0.46 [0.27, 0.65]
0.15	+0.36 [0.04, 0.68]	+0.50 [0.25, 0.75]
0.30	+0.38 [0.04, 0.72]	+0.58 [0.42, 0.74]
0.50	+0.44 [0.30, 0.58]	+0.72 [0.52, 0.92]

corresponding CMC behavior provides a retrieval-level view of the same mechanism. The advantage is most visible near the head of the ranked list and then narrows at larger ranks, indicating that ReCUE primarily resolves the ordering of visually similar cross-modal candidates rather than merely making the correct identity appear somewhere in a broad candidate set. Competing global assignments are most useful in this regime: consensus stabilizes reliable neighbors, while reconstruction can replace initially plausible but incorrect ones as the representation improves. Together, these results support the intended correspondence-representation feedback loop rather than mere feature compactness.

Revisable correspondence dynamics. Tab. 3 further examines how correspondence states change under different update strategies. Under Full reconstruction, wrong-to-correct transitions outnumber correct-to-wrong transitions, yielding the largest positive net correction among the compared policies. In addition, initially ambiguous and unreliable correspondences are substantially more recoverable than those already routed as reliable, with recovery rates of 35.1% and 34.5% compared with 13.6%. This indicates that uncertain relations remain open to revision rather than being prematurely fixed. The perturbation results show the same pattern: as the initial correspondence quality degrades, Full reconstruction achieves increasingly larger net corrections, indicating that its benefit becomes more pronounced when early assignments are less reliable. Together, these transitions show that the method does not simply accumulate previously accepted relations; it can promote, demote, and replace correspondences as the representation evolves.

Tab. 3 separates correction from churn. Naive Top-1 refresh both repairs and introduces errors, while locking reduces churn but prevents an accepted relation from being reconsidered. Full reconstruction keeps every relation eligible for reassignment and yields the largest net correction in the clean setting. Under shared ground-truth-independent perturbation, Full recovers more errors as the initial correspondence degrades and, for $\alpha \geq 0.30$, also introduces fewer new errors than naive refresh.

4 Conclusion

In this paper, we proposed ReCUE, a revisable correspondence-representation co-evolution method for weakly supervised VI-ReID. ReCUE addresses the challenge of learning cross-modal representations without access to the training identity mapping and mitigates the self-reinforcing errors caused by prematurely fixed correspondences. Specifically, Global Correspondence Hypothesis Inference (GCHI) constructs globally competitive assignments and estimates their edge-level stability, Structural Uncertainty-Aware Learning (SUAL) converts this stability into hard, soft, or deferred supervision, and Correspondence-Representation Co-Evolution (CRCE) reconstructs the complete correspondence state as the representation evolves. Together, these modules form a revisable learning loop rather than relying on fixed pseudo-labels. Experiments on SYSU-MM01 and LLCM demonstrate the effectiveness of ReCUE. Complete reconstruction provides the clearest benefit under Top-K assignment uncertainty on SYSU-MM01, while LLCM shows consistent improvements in both retrieval directions.

ETHICS STATEMENT

This work uses established public benchmarks, introduces no new identity data, and evaluates representation learning rather than a deployed identification system. Operational use would require appropriate consent and governance, human oversight, and auditing across populations, sensors, and environments. Benchmark performance alone does not establish fitness for deployment.

AI USE STATEMENT

Generative AI tools were used solely for language editing, proofreading, and improving the readability of the manuscript. The authors independently conceptualized the research, designed and executed the experiments, and analyzed the results. The authors reviewed all AI-assisted text and take full responsibility for the complete content of this work.

REPRODUCIBILITY STATEMENT

Section 3.5 specifies evaluation and Appendix B gives complete training and implementation details. Appendices A to C provide expanded losses, comparator definitions, controlled variants and additional diagnostics.

REFERENCES

Zhuxuan Cheng, Minmin Yang, Zhijia Zhang, Huijie Fan, and Wu Ding. Unsupervised visible-infrared person re-identification with pseudo-label refinement and modality-unified memory. *Digital Signal Processing*, 168:105658, 2026. doi: 10.1016/j.dsp.2025.105658.

Robin Dehler, Martin Herrmann, Jan Strohbeck, and Michael Buchholz. A graph neural network approach for solving the ranked assignment problem in multi-object tracking. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2646–2652, 2024. doi: 10.1109/IV55156.2024.10588674. Preprint available as arXiv:2604.01696.

Qijie Ding, Jie Yin, Daokun Zhang, and Junbin Gao. Combating confirmation bias: A unified pseudo-labeling framework for entity alignment. *Data Mining and Knowledge Discovery*, 39(5): 51, 2025. doi: 10.1007/s10618-025-01128-0.

Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, and Xi Peng. Learning with dual-level noisy correspondence for multi-modal entity alignment. In *International Conference on Learning Representations*, 2026a.

Jiaze Li, Yan Lu, Bin Liu, Guojun Yin, and Mang Ye. Dual-level modality debiasing learning for unsupervised visible-infrared person re-identification. *Pattern Recognition*, 176:113257, 2026b. doi: 10.1016/j.patcog.2026.113257.

Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19366–19375, 2022. doi: 10.1109/CVPR52688.2022.01876.

Zefeng Lu, Yunfei He, Meijun Chen, Yue Yuan, Hui Lu, Yuyu He, and Zhihong Tian. Weakly supervised learning meets vis-nir re-identification: Progressive cross-modality alignment for representation learning. *IEEE Transactions on Information Forensics and Security*, 21:6004–6019, 2026. doi: 10.1109/TIFS.2026.3707451.

Yuan Lv, Jun Zhang, Guo Li, and Jingtian Zhu. Structural consensus and collaborative correction for unsupervised visible-infrared person re-identification. *Infrared Physics & Technology*, 156: 106621, 2026. doi: 10.1016/j.infrared.2026.106621.

Katta G. Murty. Letter to the editor—an algorithm for ranking all the assignments in order of increasing cost. *Operations Research*, 16(3):682–687, 1968. doi: 10.1287/opre.16.3.682.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. doi: 10.3390/s17030605.

Zhiqi Pang, Chunyu Wang, Lingling Zhao, and Junjie Wang. Augmented and softened matching for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11100–11109, 2025. doi: 10.1109/ICCV51701.2025.01033.

Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12046–12055, 2021. doi: 10.1109/ICCV48922.2021.01183.

Richard D. Paul, Johannes Seiffarth, David Rügamer, Katharina Nöh, and Hanno Scharr. How to make your cell tracker say "i dunno!". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6914–6923, 2025. doi: 10.1109/ICCV51701.2025.00650.

Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic matching using m-best solutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, 2016. doi: 10.1109/CVPR.2016.22.

Haonan Shi, Yubin Wang, De Cheng, Lingfeng He, Nannan Wang, and Xinbo Gao. Hierarchical identity learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 35:8413–8428, 2026. doi: 10.1109/TIP.2026.3713464.

Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11218–11228, 2023. doi: 10.1109/ICCV51070.2023.01030.

Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *Computer Vision – ECCV 2024*, pp. 456–474, 2024. doi: 10.1007/978-3-031-72649-1_26.

Moyao Tian, Shijia Liu, Yan Yang, Xin Yuan, Minshi Chen, Wei Wang, and Xiao Wang. Structural-semantic reciprocal learning for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2607.15220*, 2026.

Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *Computer Vision – ECCV 2022*, pp. 93–109, 2022. doi: 10.1007/978-3-031-20053-3_6.

Xiao Wang, Lekai Liu, Bin Yang, Mang Ye, Zheng Wang, and Xin Xu. TokenMatcher: Diverse tokens matching for unsupervised visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7934–7942, 2025. doi: 10.1609/aaai.v39i8.32855.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5380–5389, 2017. doi: 10.1109/ICCV.2017.575.

Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9548–9558, 2023. doi: 10.1109/CVPR52729.2023.00921.

Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2843–2851, 2022. doi: 10.1145/3503161.3548198.

Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11069–11079, 2023a. doi: 10.1109/ICCV51070.2023.01016.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19883–19892, 2023b. doi: 10.1109/CVPR52729.2023.01904.

Yiming Yang, Weipeng Hu, and Haifeng Hu. Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1290–1304, 2025. doi: 10.1109/TIFS.2025.3527356.

Haoyu Yao, Bin Yang, Wenke Huang, Bo Du, and Mang Ye. Unsupervised visible-infrared person re-identification under unpaired settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11916–11926, 2025. doi: 10.1109/ICCV51701.2025.01108.

Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13567–13576, 2021. doi: 10.1109/ICCV48922.2021.01331.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

Mang Ye, Zesen Wu, and Bo Du. Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3815–3829, 2025. doi: 10.1109/TPAMI.2025.3541053.

Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11185–11195, 2023. doi: 10.1109/ICCV51070.2023.01027.

Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. CLIP-driven semantic discovery network for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 27:4137–4150, 2025. doi: 10.1109/TMM.2025.3535353.

Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 852–861. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657806.

Yafei Zhang, Lingqi Kong, Huafeng Li, and Jie Wen. Weakly supervised visible-infrared person re-identification via heterogeneous expert collaborative consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12659–12669, 2025. doi: 10.1109/ICCV51701.2025.01176.

Yifeng Zhang, Canlong Zhang, Haifei Ma, Zhixin Li, Zhiwen Wang, and Chunrong Wei. Adaptive confidence-driven learning and cross-modal hard sample mining for unsupervised visible-infrared person re-identification. *Information Processing & Management*, 63(1):104346, 2026a. doi: 10.1016/j.ipm.2025.104346.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2153–2162, 2023. doi: 10.1109/CVPR52729.2023.00214.

Zhizhong Zhang, Jiangming Wang, Xin Tan, Yanyun Qu, Junping Wang, Yong Xie, and Yuan Xie. Mutual information guided optimal transport for unsupervised visible-infrared person re-identification. *arXiv preprint arXiv:2407.12758*, 2024.

Zhizhong Zhang, Jiaqi Chen, Xin Tan, Yuan Xie, and Yan Gao. Transporting the cross-modal prototypes for unsupervised visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 35:3751–3764, 2026b. doi: 10.1109/TIP.2026.3680065.

Sicheng Zhao, Wei Lu, Shihao Li, Sibao Chen, Chris H. Q. Ding, Futian Wang, Jin Tang, and Bin Luo. CLNS: Camera-aware label noise suppression for unsupervised visible-infrared person re-identification. *Pattern Recognition*, 179:113873, 2026. doi: 10.1016/j.patcog.2026.113873.

Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, and Yanfeng Wang. Mitigating noisy correspondence by geometrical structure consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27381–27390, 2024. doi: 10.1109/CVPR52733.2024.02585.

Xiying Zheng, Yukang Zhang, Yang Lu, and Hanzi Wang. Semi-supervised visible-infrared person re-identification via modality unification and confidence guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5761–5770, 2024. doi: 10.1145/3664647.3680735.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *International Conference on Learning Representations (ICLR)*, 2026.

A Expanded Optimization Objective

This section expands the compact notation in Eq. 7. For $m \in \{v, i\}$, let $p(a)$ and $n(a)$ denote the batch-hardest positive and negative for anchor a. The intra-modal loss is

$$\begin{aligned}\mathcal{L}_{\text{id}} &= -\sum_{m} \frac{1}{|\mathcal{B}^m|} \sum_{(x,y)\in\mathcal{B}^m} \log p^m(y \mid x),\\ \mathcal{L}_{\text{tri}} &= \sum_{m} \frac{1}{|\mathcal{B}^m|} \sum_{a} [\gamma + d(f_a^m, f_{p(a)}^m) - d(f_a^m, f_{n(a)}^m)]_+.\end{aligned} \tag{10}$$

We use $d(f_s, f_{s'}) = \|f_s - f_{s'}\|_2$ on normalized embeddings. For online samples and cached opposite-modality prototypes, define the two directional prediction distributions

$$\begin{aligned}p_{s\to b}^{v\to i}(\tau) &= \frac{\exp(f_s^{v\top}\,\mathrm{sg}(P_b^i)/\tau)}{\sum_\ell \exp(f_s^{v\top}\,\mathrm{sg}(P_\ell^i)/\tau)},\\ p_{s\to a}^{i\to v}(\tau) &= \frac{\exp(f_s^{i\top}\,\mathrm{sg}(P_a^v)/\tau)}{\sum_\ell \exp(f_s^{i\top}\,\mathrm{sg}(P_\ell^v)/\tau)}, \qquad \pi_{ab}^{(t)} = \rho_{ab}^{(t)}.\end{aligned} \tag{11}$$

Let $b^{(t)}(a)$ and $a^{(t)}(b)$ be the partners induced by $A^{(1,t)}$. For a routed edge set $\mathcal{G} \in \{\mathcal{R}^{(t)}, \mathcal{A}^{(t)}\}$, the eligible sample subsets are

$$\begin{aligned}\mathcal{B}_{\mathcal{G}}^v &= \{s \in \mathcal{B}^v : (y_s^v, b^{(t)}(y_s^v)) \in \mathcal{G}\},\\ \mathcal{B}_{\mathcal{G}}^i &= \{s \in \mathcal{B}^i : (a^{(t)}(y_s^i), y_s^i) \in \mathcal{G}\}.\end{aligned} \tag{12}$$

The VIS and IR identity samples are independent and consequently these four sets can have different cardinalities. Define $\eta_{\mathcal{G}}^m = \mathbb{I}[|\mathcal{B}_{\mathcal{G}}^m| > 0]$ and $\mathrm{Avg}_{s\in\mathcal{B}}[g_s] = |\mathcal{B}|^{-1}\sum_{s\in\mathcal{B}} g_s$ for nonempty $\mathcal{B}$, and zero otherwise. The two hard-target directions and their nonempty-direction average are

$$\begin{aligned}\ell_{\text{strong}}^{v\to i} &= -\,\mathrm{Avg}_{s\in\mathcal{B}_{\mathcal{R}}^v}\left[\log p_{s\to b^{(t)}(y_s^v)}^{v\to i}(\tau_s)\right],\\ \ell_{\text{strong}}^{i\to v} &= -\,\mathrm{Avg}_{s\in\mathcal{B}_{\mathcal{R}}^i}\left[\log p_{s\to a^{(t)}(y_s^i)}^{i\to v}(\tau_s)\right],\\ \mathcal{L}_{\text{strong}} &= \frac{\eta_{\mathcal{R}}^v \ell_{\text{strong}}^{v\to i} + \eta_{\mathcal{R}}^i \ell_{\text{strong}}^{i\to v}}{\eta_{\mathcal{R}}^v + \eta_{\mathcal{R}}^i}.\end{aligned} \tag{13}$$

For ambiguous anchors, consensus retains all alternatives selected by the ranked hypotheses. Using $\pi^{(t)}$ from Eq. 11,

$$\begin{aligned}
\ell_{\text{weak}}^{v\to i} &= -\operatorname{Avg}_{s\in\mathcal{B}_{\mathcal{A}}^{v}}\left[\sum_{b}\pi_{y_s^v b}^{(t)}\log p_{s\to b}^{v\to i}(\tau_w)\right],\\
\ell_{\text{weak}}^{i\to v} &= -\operatorname{Avg}_{s\in\mathcal{B}_{\mathcal{A}}^{i}}\left[\sum_{a}\pi_{a y_s^i}^{(t)}\log p_{s\to a}^{i\to v}(\tau_w)\right],\\
\mathcal{L}_{\text{weak}} &= \frac{\eta_{\mathcal{A}}^{v}\ell_{\text{weak}}^{v\to i}+\eta_{\mathcal{A}}^{i}\ell_{\text{weak}}^{i\to v}}{\eta_{\mathcal{A}}^{v}+\eta_{\mathcal{A}}^{i}}.
\end{aligned} \tag{14}$$

Each displayed ratio is defined as zero when both indicators in its denominator are zero. Thus each modality is normalized by its own routed sample count, an empty direction is omitted rather than downweighted by a fixed factor of $1/2$, and an identity absent from the independently sampled mini-batch creates no correspondence-loss term. For $\mathcal{S}^{(t)}=\mathcal{R}^{(t)}\cup\mathcal{A}^{(t)}$, reciprocal predictor agreement is

$$\mathcal{L}_{\text{cons}}^{(t)}=\frac{\sum_{(a,b)\in\mathcal{S}^{(t)}}\rho_{ab}^{(t)}(q_{a\to b}^{v\to i}-q_{b\to a}^{i\to v})^2}{\sum_{(a,b)\in\mathcal{S}^{(t)}}\rho_{ab}^{(t)}}, \tag{15}$$

and is set to zero when $\mathcal{S}^{(t)}$ is empty. Cached prototypes and routing variables remain fixed within a stage, so the strong and weak losses update the encoder through online sample embeddings. The consistency loss updates the trainable normalized predictor weights.

A.1 Structural Properties and Non-Reducibility of the Learning Interface

We now separate the algorithmic content of the proposed interface from its implementation choices. At a fixed refresh stage, let $J(A)=\langle C,A\rangle_F$, let $J_K=J(A^{(K)})$, and use $A_e\in\{0,1\}$ for edge e. Ties in ranked assignment are resolved by a fixed rule independent of correspondence ground truth. The statements below establish three properties used by the method: the retained alternatives obey global one-to-one constraints, soft routing changes the direction rather than only the scale of the learning signal and reconstruction preserves assignment reachability. They do not treat the truncated marginal as a calibrated posterior and do not claim that m-best assignment marginalization itself is new.

Proposition 1 (truncated assignment marginal). The matrix $\rho=K^{-1}\sum_{k=1}^{K}A^{(k)}$ is doubly stochastic. For

$$J_e^{+}=\min_{A\in\Omega:A_e=1}J(A),\qquad J_e^{-}=\min_{A\in\Omega:A_e=0}J(A), \tag{16}$$

$J_e^{-}>J_K$ implies $\rho_e=1$, whereas $J_e^{+}>J_K$ implies $\rho_e=0$. Thus the extreme consensus values certify stability relative to the enumerated global alternatives, not merely a large local similarity.

Proof. Each $A^{(k)}$ is a permutation matrix, and their uniform convex combination therefore has unit row and column sums. If $J_e^{-}>J_K$, no assignment among the first K can exclude e, hence every indicator in the average is one. The second statement follows symmetrically because no assignment among the first K can contain e when $J_e^{+}>J_K$. □

The same property explains why the soft target in Eq. 11 is simply $\pi=\rho$. For a fixed row, cross-entropy satisfies

$$H(\rho_{a:},p_{a:})=H(\rho_{a:})+\mathrm{KL}(\rho_{a:}\|p_{a:}), \tag{17}$$

so ambiguous supervision is minimized by preserving the truncated alternative distribution rather than collapsing it to the single edge in $A^{(1)}$.

Proposition 2 (reachability under reconstruction). Let $L^{(t)}$ be the edges made absorbing by a locked-update control and assume its constrained assignment set is nonempty:

$$\Omega_L^{(t)}=\{A\in\Omega:A_e=1\text{ for all }e\in L^{(t)}\}\subseteq\Omega. \tag{18}$$

At any later cost matrix $C^{(t+1)}$,

$$\min_{A\in\Omega}\langle C^{(t+1)},A\rangle_F\le\min_{A\in\Omega_L^{(t)}}\langle C^{(t+1)},A\rangle_F. \tag{19}$$

Moreover, if a locked edge is absent from the hidden bijection $A^\star$, then $A^\star \notin \Omega_L^{(t)}$. Full reconstruction continues to optimize over Ω and therefore preserves the reachability of $A^\star$; it does not, by itself, guarantee that the current representation ranks $A^\star$ first.

Proof. The cost inequality follows from set inclusion. A single constraint $A_e = 1$ that contradicts $A_e^\star = 0$ excludes $A^\star$ from the constrained set, while $A^\star \in \Omega$ by the problem definition. □

Corollary 2.1 (a sufficient recovery condition). Suppose the representation at a later refresh produces a cost matrix $C^{(t+1)}$ for which the hidden bijection is the unique optimum,

$$\langle C^{(t+1)}, A^\star \rangle_F < \langle C^{(t+1)}, A \rangle_F \quad \text{for every } A \in \Omega \setminus \{A^\star\}. \tag{20}$$

Full reconstruction returns $A^\star$ at that refresh. A locked method cannot do so whenever a previously locked edge contradicts $A^\star$.

Proof. The strict inequality makes $A^\star$ the unique minimizer over the full set Ω. Proposition 2 shows that a contradictory lock removes it from the restricted set $\Omega_L^{(t)}$. □

This corollary makes the role of reconstruction precise. The method does not claim that optimization must create the gap in Eq. 20; rather, if representation learning does create it, the update rule is capable of acting on that evidence. An absorbing wrong decision can prevent the same recovery even under an otherwise perfect later cost ordering.

Proposition 3 (local evidence does not determine assignment stability). An edge can have the same local cost and belong to the same unique optimum in two problems, yet have different support among the two best bijections.

Construction. For $M > \epsilon > 0$, consider

$$C^{(a)} = \begin{bmatrix} 0 & \epsilon & M \\ \epsilon & 0 & M \\ M & M & 0 \end{bmatrix}, \qquad C^{(b)} = \begin{bmatrix} 0 & M & M \\ M & 0 & \epsilon \\ M & \epsilon & 0 \end{bmatrix}. \tag{21}$$

Both matrices have the identity permutation as their unique optimum and the selected edge $e = (1, 1)$ has cost zero in both. In $C^{(a)}$, however, the second-best permutation swaps identities 1 and 2, giving $\rho_e^{(2)} = 1/2$; in $C^{(b)}$ it swaps identities 2 and 3, giving $\rho_e^{(2)} = 1$. Hence an isolated pair score together with its membership in $A^{(1)}$ cannot identify stability under globally competing bijections. This does not claim that ranked-assignment marginals are new; it formalizes why they contain supervision-routing information absent from the selected edge alone. □

Proposition 4 (a sufficient perturbation certificate). Let $\widetilde{C} = C + \Delta C$ with $\|\Delta C\|_\infty \le \varepsilon$, and write $J_K(C)$ for the Kth assignment cost under C. Then

$$J_e^-(C) - J_K(C) > 2N\varepsilon \implies \widetilde{\rho}_e = 1, \qquad J_e^+(C) - J_K(C) > 2N\varepsilon \implies \widetilde{\rho}_e = 0. \tag{22}$$

Proof. Every feasible assignment contains N edges, so its cost changes by at most $N\varepsilon$. The perturbed Kth order statistic is at most $J_K(C) + N\varepsilon$, because the original top-K assignments remain feasible. Under the first inequality, every assignment excluding e has perturbed cost at least $J_e^-(C) - N\varepsilon > J_K(C) + N\varepsilon$ and therefore cannot enter the perturbed top K. The second implication is symmetric for assignments containing e. □

The certificate is deliberately sufficient rather than necessary: it explains when an extreme route is invariant to a bounded cost change, without treating ρ as a calibrated posterior or assuming that all intermediate scores are stable.

Proposition 5 (soft targets vary continuously with the represented alternatives). For logits s, target $q \in \Delta^{N-1}$, temperature $\tau > 0$, and

$$\ell(q, s) = -\sum_j q_j \log \mathrm{softmax}(s/\tau)_j, \tag{23}$$

the logit gradient is $\nabla_s \ell = (p - q)/\tau$, where $p = \mathrm{softmax}(s/\tau)$. Consequently, for any q, q',

$$\|\nabla_s \ell(q, s) - \nabla_s \ell(q', s)\|_1 = \frac{\|q - q'\|_1}{\tau}. \tag{24}$$

A switch between two hard one-hot targets can therefore change the gradient by $2/\tau$, whereas revising a consensus target changes it in direct proportion to the mass reassigned among alternatives. This identity motivates soft supervision for ambiguous relations; it is an optimization property, not a claim that soft targets are necessarily correct. □

Corollary 5.1 (deviation from an oracle target). For analysis only, let $e_{j^\star}$ denote the one-hot target of the hidden correct partner. Proposition 5 gives the exact identity

$$\|\nabla_s \ell(q,s) - \nabla_s \ell(e_{j^\star}, s)\|_1 = \frac{2(1-q_{j^\star})}{\tau}. \tag{25}$$

Thus a wrong hard target has the maximum deviation $2/\tau$, while a soft target reduces that deviation exactly to the extent that it retains mass on the correct alternative. The identity is explanatory rather than operational: $j^\star$ is never observed or used by the routing rule.

Proposition 6 (one-to-one capacity couples edge support). For any visible-index subset I and infrared-index subset J, the truncated marginal satisfies

$$\sum_{a\in I}\sum_{b\in J} \rho_{ab} \le \min\{|I|, |J|\}. \tag{26}$$

In particular, support assigned to one candidate consumes row and column capacity that cannot simultaneously be assigned to a conflicting edge.

Proof. By Proposition 1, each row and column of ρ sums to one and all entries are nonnegative. Therefore

$$\sum_{a\in I, b\in J} \rho_{ab} \le \sum_{a\in I, b} \rho_{ab} = |I|, \qquad \sum_{a\in I, b\in J} \rho_{ab} \le \sum_{a, b\in J} \rho_{ab} = |J|. \tag{27}$$

Taking the smaller bound proves the claim. □

This conservation law is the precise sense in which the reliability signal is assignment-level. Independently scored edges need not obey both directional capacities; averaging feasible bijections does so automatically, including for every finite truncation K.

Proposition 7 (consensus supervision is expected hypothesis supervision). Let $b_a^{(k)}$ be the unique infrared partner of visible identity a in $A^{(k)}$, and let $p_{a:}$ be any strictly positive predicted distribution. Then

$$\frac{1}{K}\sum_{k=1}^{K}\big[-\log p_{a,b_a^{(k)}}\big] = -\sum_b \rho_{ab}\log p_{ab} = H(\rho_{a:}, p_{a:}). \tag{28}$$

The analogous identity holds in the reverse direction. Moreover, if s_a are the logits of $p_{a:} = \mathrm{softmax}(s_a/\tau)$, then

$$\nabla_{s_a} H(\rho_{a:}, p_{a:}) = \frac{1}{K}\sum_{k=1}^{K} \nabla_{s_a}\big[-\log p_{a,b_a^{(k)}}\big]. \tag{29}$$

Proof. Substituting $\rho_{ab} = K^{-1}\sum_k \mathbb{I}[b_a^{(k)} = b]$ into the right-hand side of Eq. 28 and exchanging the finite sums gives the first identity. Differentiation commutes with this finite average, which yields Eq. 29. The reverse direction follows because every $A^{(k)}$ is also one-to-one by columns. □

Thus the ambiguous loss is not an arbitrary interpolation between pairwise targets: it is exactly the average hard-target loss, and exactly the average hard-target gradient, over the retained globally feasible explanations. A cost-weighted Top-K marginal satisfies the same identities after replacing $1/K$ by its normalized hypothesis weights.

Proposition 8 (the routed objective is not scalar confidence weighting). Let a be the partner selected by $A^{(1)}$ and suppose an ambiguous target q assigns positive mass to some $b \neq a$. There are no fixed constants $w \ge 0$ and c such that

$$H(q,p) = w[-\log p_a] + c \quad \text{for every } p \in \operatorname{int}\Delta^{N-1}. \tag{30}$$

Moreover, for an unreliable identity c, the routed cross-modal target gradient is zero, whereas a full-plan soft method with target $Q_{c:}$ has gradient $(p_{c:} - Q_{c:})/\tau$, which is nonzero unless the current prediction already equals the transport row.

Proof. If Eq. 30 held for all logits, its gradients would satisfy $p - q = w(p - e_a)$ for all interior p. Evaluating two distinct values of p forces $w = 1$, after which equality requires $q = e_a$, contradicting $q_b > 0$. The second statement follows directly from the routed objective: unreliable identities occur in neither $\mathcal{L}_{\text{strong}}$, $\mathcal{L}_{\text{weak}}$, nor $\mathcal{L}_{\text{cons}}$, while differentiation of the full-plan cross-entropy gives $(p_{c:} - Q_{c:})/\tau$. □

This proposition separates the proposed interface from two common reductions. Ambiguity changes the direction of the target gradient rather than merely its magnitude, and deferral removes correspondence-derived influence rather than assigning every identity a dense soft target. Together with Proposition 2, it also shows why periodic reconstruction is a distinct operation: the target representation controls how a stage learns, whereas reconstruction controls which bijections remain reachable at the next stage.

Proposition 9 (finite-truncation update bound). Let $\rho^{(K)} = K^{-1}\sum_{k=1}^{K} A^{(k)}$ and let $L > K$. For every row a,

$$\left\|\rho_{a:}^{(L)} - \rho_{a:}^{(K)}\right\|_1 \le \frac{2(L-K)}{L}, \qquad \left\|\rho_{a:}^{(K+1)} - \rho_{a:}^{(K)}\right\|_1 \le \frac{2}{K+1}. \tag{31}$$

The same bounds hold for every column.

Proof. Write the larger truncation as

$$\rho^{(L)} = \frac{K}{L}\rho^{(K)} + \frac{L-K}{L}\bar{A}, \qquad \bar{A} = \frac{1}{L-K}\sum_{k=K+1}^{L} A^{(k)}. \tag{32}$$

Both $\rho_{a:}^{(K)}$ and $\bar{A}_{a:}$ are probability vectors, whose ℓ_1 distance is at most two. Therefore $\|\rho_{a:}^{(L)} - \rho_{a:}^{(K)}\|_1 = (L-K)L^{-1}\|\bar{A}_{a:} - \rho_{a:}^{(K)}\|_1$ gives the first bound; setting $L = K+1$ gives the second. Column-wise normalization follows identically from the bijection constraint. □

The bound is deterministic and requires no probabilistic interpretation of the ranked list. It controls how much the represented target distribution can change when more hypotheses are appended; it does not guarantee that a hard route remains unchanged when a score lies exactly at a threshold. Appendix Tab. 14 therefore reports route agreement separately from distributional deviation.

Corollary 9.1 (one-step route stability). For every edge,

$$\left|\rho_e^{(K+1)} - \rho_e^{(K)}\right| = \frac{\left|A_e^{(K+1)} - \rho_e^{(K)}\right|}{K+1} \le \frac{1}{K+1}. \tag{33}$$

Hence appending one hypothesis cannot change an edge's route when its current support is farther than $1/(K+1)$ from both routing thresholds. At $K = 10$, the support count $c_e = \sum_{k=1}^{10} A_e^{(k)}$ gives the exact discrete rule $c_e \le 3$ (unreliable), $4 \le c_e \le 6$ (ambiguous), and $c_e \ge 7$ (reliable). The corollary characterizes list-truncation sensitivity; it does not imply stability for edges adjacent to the $3/4$ or $6/7$ count boundaries.

Lemma 9.2 (cost-scale invariance). For $a > 0$ and any scalar b, replacing the cost by $C' = aC + b\mathbf{1}\mathbf{1}^\top$ leaves the ordering of all feasible bijections, and therefore uniform $\rho^{(K)}$, unchanged. The cost-weighted Top-K representation is also unchanged whenever its numerical floor is inactive, because all assignment-cost gaps and their median temperature T_g scale by the same factor a.

Proof. Every feasible assignment has exactly N selected edges, so $\langle C', A\rangle_F = a\langle C, A\rangle_F + bN$. Positive affine transformation preserves the ranked list. For the weighted variant, $(J'_k - J'_1)/T'_g = a(J_k - J_1)/(aT_g)$ in the nondegenerate case. □

Finite-list boundary. If the hidden permutation $A^\star$ is absent from the first K solutions, the truncated marginal provides no guarantee that its edges receive positive mass. If $A^\star$ occurs in the list, each of its edges has support at least $1/K$, but this alone need not cross τ_{low}; with $K = 10$, one appearance gives 0.1 and is still deferred. Thus the method requires the ranked list to expose correct alternatives repeatedly enough for routing; the larger-K study measures approximation to a longer truncation, not posterior calibration or inclusion of $A^\star$.

Theorem 10 (operator non-equivalence to three update classes). Consider a stage at which (i) an ambiguous identity has a non-one-hot target q, (ii) an unreliable identity is deferred, and (iii) a later reconstructed optimum excludes an edge that a locked update would retain. Over all possible model logits and this cost sequence, the resulting learning-and-refresh operator cannot be reproduced by any one of the following classes: (i) scalar weighting of the current top-1 hard target; (ii) dense full-plan soft supervision applied to every identity; or (iii) refinement constrained to preserve the locked edge.

Proof. For the ambiguous identity, Proposition 8 shows that no scalar multiple of a top-1 hard-target gradient equals the gradient induced by q for every logit vector. For the unreliable identity, the proposed correspondence-derived gradient is zero, whereas dense full-plan supervision gives $(p - Q)/\tau$ and is nonzero for some logits. Finally, Proposition 2 shows that the locked feasible set excludes every later assignment that removes the retained edge, while full reconstruction searches all of Ω. Each simpler class therefore disagrees with at least one of the three operations; none realizes the complete operator. □

The theorem is a non-reducibility statement, not a universal performance guarantee. It identifies the technical object evaluated by the factorial and routing controls: a structured target representation, a three-way gradient interface, and a non-absorbing reconstruction operator. Backbone choice, optimizer, and refresh frequency instantiate this operator but do not define it. Conversely, ranked enumeration alone is insufficient: without routing it does not specify a representation update, and without reconstruction it does not keep earlier decisions revisable.

Proposition 11 (bidirectional coherence of the structured target). For the forward target $\pi_a^{v\to i} = \rho_{a:}$ and reverse target $\pi_b^{i\to v} = \rho_{:b}^\top$,

$$\sum_b \pi_{ab}^{v\to i} = 1, \qquad \sum_a \pi_{ba}^{i\to v} = 1, \qquad \pi_{ab}^{v\to i} = \pi_{ba}^{i\to v} = \rho_{ab}. \tag{34}$$

Consequently, the two directional losses use the same edge mass and cannot independently assign incompatible total mass to a visible or infrared identity.

Proof. The first two equalities are the row and column sums from Proposition 1. The third follows because both directional targets read the same entry of ρ; the reverse target transposes the column but does not renormalize its edge mass. □

This property is shared by any doubly stochastic plan, including Sinkhorn, but not by two independently normalized local conditionals in general. Ranked consensus contributes a different piece of information within this structured class: each mass value is the observed inclusion frequency of an edge across discrete feasible bijections.

Proposition 12 (route-dependent target-gradient exposure). Consider one directional cross-modal target term before the encoder Jacobian, and omit mini-batch normalization. Its logit gradient is

$$g_a^{\text{corr}} = \begin{cases} \lambda_s(p_{a:} - e_{b_a})/\tau_s, & a \in \mathcal{R}, \\ \lambda_w(p_{a:} - \rho_{a:})/\tau_w, & a \in \mathcal{A}, \\ 0, & a \in \mathcal{U}, \end{cases} \tag{35}$$

where b_a is the top-1 assigned partner. Hence

$$\|g_a^{\text{corr}}\|_1 \le \begin{cases} 2\lambda_s/\tau_s, & a \in \mathcal{R}, \\ 2\lambda_w/\tau_w, & a \in \mathcal{A}, \\ 0, & a \in \mathcal{U}. \end{cases} \tag{36}$$

For analysis only, if $e_{b^\star}$ is the hidden correct target, the ambiguous target-gradient discrepancy is exactly

$$\frac{\lambda_w}{\tau_w}\,\|\rho_{a:} - e_{b^\star}\|_1 = \frac{2\lambda_w(1-\rho_{ab^\star})}{\tau_w}. \tag{37}$$

Proof. Eq. 35 follows from the softmax cross-entropy gradient and the routing definition. Any two probability vectors have ℓ_1 distance at most two, proving Eq. 36. Since $\|\rho_{a:} - e_{b^\star}\|_1 = (1-\rho_{ab^\star}) + \sum_{b\neq b^\star}\rho_{ab} = 2(1-\rho_{ab^\star})$, the final equality follows. □

The bound concerns the correspondence-target term, not the full encoder gradient or a convergence guarantee. It nevertheless makes the supervision interface explicit: reliable relations may exert the full hard-target signal, ambiguous relations receive both a distributed direction and the smaller coefficient λ_w, and unreliable relations contribute no correspondence-derived target gradient during that stage.

Together, these properties isolate the method's three interfaces. Ranked hypotheses reveal globally coupled ambiguity that a selected edge alone cannot encode; routing converts that structure into controlled gradients; and full reconstruction avoids turning an early routed edge into a permanent feasibility constraint. The empirical analyses test whether these sufficient structural properties translate into correspondence recovery and retrieval improvement.

B Protocol and Controlled-Variant Details

Implementation details. We implement the framework in PyTorch with an ImageNet-pretrained ResNet-50 and resize images to 288×144. Training uses style augmentation, random cropping, horizontal flipping, and random erasing. Each mini-batch independently samples eight visible and eight infrared identities with four images per identity (64 images); the two identity sets are never aligned through the hidden cross-modal mapping. Adam uses weight decay 5×10^{-4}, encoder/predictor learning rates $3\times10^{-4}/6\times10^{-4}$, a 10-epoch learning-rate warm-up, and 0.1 decays at epochs 30 and 70. A 20-epoch intra-modal stage is followed by 100 co-evolution epochs (120 total).

We set $\lambda = 0.5$ in Eq. 3, $\tau_c = \tau_s = \tau_w = 0.07$, and normalize prototypes, predictor identity weights, and sample embeddings. The batch-hard triplet loss uses Euclidean distance and margin $\gamma = 0.3$. We use $K = 10$ and infer correspondences at epochs 20, 40, 60, 80, 100, and 120. Thresholds $(\tau_{\text{low}}, \tau_{\text{high}}) = (0.4, 0.7)$ map 0-3, 4-6, and 7-10 supporting hypotheses to unreliable, ambiguous, and reliable routing. Loss weights are $\lambda_{\text{tri}} = 0.25$, $\lambda_s = 1$, $\lambda_w = 0.25$, and $\lambda_{\text{cons}} = 1$.

All hyperparameters are fixed across SYSU-MM01 and LLCM without access to ground-truth correspondence and were specified before the sensitivity analyses reported in the main text and Appendix. Ground truth is used only for post-hoc correspondence diagnostics, never for training, assignment construction, routing, checkpoint selection, or model selection.

B.1 Complete Co-Evolution Procedure

Algorithm 1 summarizes the complete training procedure. After the intra-modal warm-up, correspondence inference and representation learning alternate with a fixed refresh interval $\Delta = 20$. Each refresh reconstructs the complete correspondence state rather than accumulating previous assignments.

Algorithm 1 Revisable correspondence-representation co-evolution

Require: $\mathcal{D}^v, \mathcal{D}^i, K, \tau_{\text{low}}, \tau_{\text{high}}; \Delta = 20$
1: $\theta^{(0)} \leftarrow \text{WarmUp}(\mathcal{D}^v, \mathcal{D}^i, 20)$
2: **for** $t = 0, \ldots, 4$ **do**
3: $\quad P^{(t)} \leftarrow \text{sg}(\text{Proto}(E_{\theta^{(t)}}))$; $C^{(t)} \leftarrow \text{Cost}(P^{(t)})$
4: $\quad A^{(t)}_{1:K} \leftarrow \text{KBest}_{A\in\Omega}\langle C^{(t)}, A\rangle_F$; $\rho^{(t)} \leftarrow K^{-1}\sum_k A^{(k,t)}$
5: $\quad z^{(t)} \leftarrow \text{Route}(A^{(1,t)}, \rho^{(t)}; \tau_{\text{low}}, \tau_{\text{high}})$
6: $\quad \theta^{(t+1)} \leftarrow \text{Train}_\Delta(\theta^{(t)}; P^{(t)}, A^{(1,t)}, \rho^{(t)}, z^{(t)})$
7: **end for**
8: Reconstruct $P^{(5)}, C^{(5)}, A^{(5)}_{1:K}$ at epoch 120
Ensure: Retrieval encoder $E_{\theta^{(5)}}$

Baseline reimplementation. The state-of-the-art tables quote HECCL values from the cited source; the baseline results used in the controlled comparisons come from our separately trained reimplementation. It uses pairwise reliability-aware weak supervision with ResNet-50, 288×144 inputs, 120 training epochs, encoder and predictor learning rates of 3×10^{-4} and 6×10^{-4}, a 10-epoch learning-rate warm-up, decays at epochs 30 and 70, prototype coefficient 0.8, and $\lambda_1 = \lambda_2 = 0.25$. The baseline and all proposed variants use the same 20-epoch first stage and 100-epoch second stage. Correspondence accuracy is measured at the same model checkpoint as retrieval.

B.2 Comparator and Update-Control Definitions

Partial-overlap extension. Eq. 1 matches the evaluated equal-cardinality benchmark protocol. If only $M \le \min(N_v, N_i)$ training identities overlap, the inference layer can instead use

$$\Omega_M = \left\{A \in \{0,1\}^{N_v\times N_i} : A\mathbf{1} \le \mathbf{1}, A^\top\mathbf{1} \le \mathbf{1}, \langle A, \mathbf{1}\mathbf{1}^\top\rangle_F = M\right\}. \tag{38}$$

Equivalently, dummy unmatched nodes with a fixed rejection cost turn the problem into a square ranked assignment. Averaging the K best partial solutions yields a sub-stochastic real-edge matrix; adding $\rho_{a\emptyset} = 1 - \sum_b \rho_{ab}$ (and the analogous reverse mass) restores a directional probability target. A best-assignment edge to $\emptyset$ receives no cross-modal positive supervision, while real edges retain the same hard/soft/deferred routing and full reconstruction. If the assumed overlap size or rejection cost is misspecified, the solver may force spurious real matches or overuse the unmatched state. The reported benchmark claims use the exact-bijection set Ω; Eq. 38 specifies the inference change required when training identities are missing from one modality.

Ranked-assignment complexity. For N identities, the cost matrix requires $O(N^2)$ storage and each Hungarian subproblem costs $O(N^3)$. Murty enumeration solves additional constrained subproblems as K grows. We use exact enumeration with the fixed $K = 10$ at the benchmark identity scales; larger settings can use approximate ranking, while partial overlap uses Eq. 38 or an unbalanced matching solver.

Diagnostic reliability scores. All diagnostics rank the same N edges selected by $A^{(1)}$. Local cost uses $s^{\text{cost}}_{ab} = -C_{ab}$. The single-optimum score is the edge-exclusion margin $s^{\text{margin}}_{ab} = J^{(-ab)} - J^\star$, where $J^\star$ is the optimal assignment cost and $J^{(-ab)}$ is the optimum after forbidding the selected edge. These are scalar confidence scores; in particular, local cost is distinct from the local-softmax distribution defined below. The Sinkhorn and ranked-assignment diagnostics use their selected-edge marginals.

Complete uncertainty representations. The matched-routing-budget replacement study requires both a scalar routing score and a directional soft target. For the local alternative, we form two conditional distributions from the common cost:

$$\begin{aligned} Q^{\text{local},v\to i}_{ab} &= \frac{\exp(-C_{ab}/\tau_{\text{local}})}{\sum_\ell \exp(-C_{a\ell}/\tau_{\text{local}})}, \\ Q^{\text{local},i\to v}_{ba} &= \frac{\exp(-C_{ab}/\tau_{\text{local}})}{\sum_\ell \exp(-C_{\ell b}/\tau_{\text{local}})}, \qquad \tau_{\text{local}} = 0.07. \end{aligned} \tag{39}$$

Its routing reliability is the bidirectional geometric mean $r_{ab}^{\text{local}} = \sqrt{Q_{ab}^{\text{local},v\to i} Q_{ba}^{\text{local},i\to v}}$. The two ambiguous targets are the corresponding VIS→IR row and IR→VIS row distributions.

For optimal transport, we apply Sinkhorn scaling to $\exp(-C/\varepsilon)$ with uniform marginals $a = b = \frac{1}{N}\mathbf{1}$, $\varepsilon = 0.05$, and at most 100 iterations. Iterations stop when the maximum absolute row or column marginal residual is below 10^{-6}. If P is the resulting plan, we define $Q^{\text{OT}} = NP$, so every row and column sums to one.

For the cost-weighted Top-K representation, let $J_k = \langle C, A^{(k)} \rangle_F$ and

$$w_k = \frac{\exp[-(J_k - J_1)/T_g]}{\sum_{\ell=1}^{K} \exp[-(J_\ell - J_1)/T_g]}, \quad T_g = \max\big(10^{-6}, \text{median}_{k=2,\dots,K}(J_k - J_1)\big), \tag{40}$$

and set $Q_{ab}^{\text{Gibbs}} = \sum_k w_k A_{ab}^{(k)}$. Uniform Top-$K$ instead uses $Q_{ab}^{\text{uniform}} = \rho_{ab} = K^{-1} \sum_k A_{ab}^{(k)}$. Both use $K = 10$. Because each is a convex combination of permutation matrices, its rows and columns sum to one. Sinkhorn, weighted Top-K, and uniform Top-K route by the selected-edge entry of their Q and use $Q_{a:}$ and $Q_{:b}^{\top}$ as the two directional ambiguous targets.

Representation	Routing reliability on $A^{(1)}$	Ambiguous targets ($v\to i,\ i\to v$)	Fixed parameters
Local-softmax	$\sqrt{Q_{ab}^{\text{local},v\to i} Q_{ba}^{\text{local},i\to v}}$	$Q_{a:}^{\text{local},v\to i},\ Q_{b:}^{\text{local},i\to v}$	$\tau_{\text{local}} = 0.07$
Sinkhorn / OT	$Q_{ab}^{\text{OT}} = NP_{ab}$	$Q_{a:}^{\text{OT}},\ Q_{:b}^{\text{OT}\top}$	$\varepsilon = 0.05$; ≤ 100 iterations; tol. 10^{-6}
Cost-weighted Top-K	$Q_{ab}^{\text{Gibbs}} = \sum_k w_k A_{ab}^{(k)}$	$Q_{a:}^{\text{Gibbs}},\ Q_{:b}^{\text{Gibbs}\top}$	$K = 10$; T_g = median assignment-cost gap
Uniform Top-K	$\rho_{ab} = K^{-1} \sum_k A_{ab}^{(k)}$	$\rho_{a:},\ \rho_{:b}^{\top}$	$K = 10$

Table 4: Uncertainty representations in the matched-routing-budget replacement study. Every alternative supplies both routing reliability and directional ambiguous targets.

Full-plan Sinkhorn soft co-evolution. This control is distinct from the routed Sinkhorn representation above. At each refresh it forms

$$P^{(t)} = \text{Sinkhorn}\Big(\exp[-C^{(t)}/\varepsilon]\Big), \qquad Q^{(t)} = \text{sg}\Big(NP^{(t)}\Big), \tag{41}$$

with the same uniform marginals, $\varepsilon = 0.05$, stopping rule, and refresh schedule. Thus every row and column of $Q^{(t)}$ sums to one, and the plan is cached until the next refresh. The prediction distributions use the same $\tau_w = 0.07$. All identities receive the two directional soft targets:

$$\begin{aligned} \mathcal{L}_{\text{OT}}^{v\to i} &= -\frac{1}{|\mathcal{B}^v|} \sum_{s\in\mathcal{B}^v} \sum_b Q_{y_s^v b}^{(t)} \log p_{s\to b}^{v\to i}, \\ \mathcal{L}_{\text{OT}}^{i\to v} &= -\frac{1}{|\mathcal{B}^i|} \sum_{s\in\mathcal{B}^i} \sum_a Q_{a y_s^i}^{(t)} \log p_{s\to a}^{i\to v}, \\ \mathcal{L}_{\text{full-plan}} &= \mathcal{L}_{\text{id}} + 0.25\mathcal{L}_{\text{tri}} + \tfrac{1}{2}(\mathcal{L}_{\text{OT}}^{v\to i} + \mathcal{L}_{\text{OT}}^{i\to v}). \end{aligned} \tag{42}$$

It uses neither Top-K hypotheses nor ρ, $\mathcal{R}/\mathcal{A}/\mathcal{U}$, $\mathcal{L}_{\text{strong}}$, $\mathcal{L}_{\text{weak}}$, or $\mathcal{L}_{\text{cons}}$. For comparable post-hoc correspondence accuracy, the final representation is evaluated through the same one-to-one optimum $A_{\text{eval}} = \arg\min_{A\in\Omega} \langle C^{(T)}, A \rangle_F$, rather than row-wise $\arg\max Q$.

Matched routing budgets. The replacement comparison controls route cardinality rather than applying incommensurate numerical thresholds to different scores. For every estimator and refresh stage, we use the current $C^{(t)}$ to enumerate uniform Top-K hypotheses. On the N edges in its current $A^{(1,t)}$, the reference thresholds determine

$$\begin{aligned} n_R^{(t)} &= \#\{(a,b) \in A^{(1,t)} : \rho_{ab}^{(t)} \ge 0.7\}, \\ n_A^{(t)} &= \#\{(a,b) \in A^{(1,t)} : 0.4 \le \rho_{ab}^{(t)} < 0.7\}, \\ n_U^{(t)} &= N - n_R^{(t)} - n_A^{(t)}. \end{aligned} \tag{43}$$

The alternative representation ranks those same N selected edges by its own reliability: the top $n_R^{(t)}$ enter $\mathcal{R}$, the next $n_A^{(t)}$ enter $\mathcal{A}$, and the rest enter $\mathcal{U}$. It also supplies its own ambiguous target.

Hence all variants have identical routing budgets at each refresh, but may route different pairs and train with different soft targets. Budget computation uses the current cost state, never a separate uniform-model trajectory. A tie at a budget boundary is resolved by a fixed deterministic hash $h(a, b) = \mathrm{SHA256}(s_0 \| a \| b)$ of the modality-specific identity indices. The public constant s_0 and canonical encoding are shared across estimators; the rule is independent of ground truth, model scores, and evaluation metrics.

All comparator parameters are prespecified from the shared cost representation or inherited from the training temperatures. No correspondence ground truth, correspondence accuracy, reliable-set precision, or test retrieval metric is used to set τ_{local}, ε, the Sinkhorn stopping rule, or T_g. These comparator definitions are used in the matched-budget uncertainty comparison reported in Tab. 2(b).

Correspondence-update controls. Fixed Top-K infers ranked hypotheses and routing once. Naive best-A refresh recomputes only $A^{(1)}$ every 20 epochs and applies a hard target to each selected edge. Reliable-edge locking uses hard/soft/deferred routing but preserves a pair once it enters $\mathcal{R}$; this is the accumulative update control. Full co-evolution reconstructs the complete assignment over all identities at every refresh.

C Additional Experimental Analysis

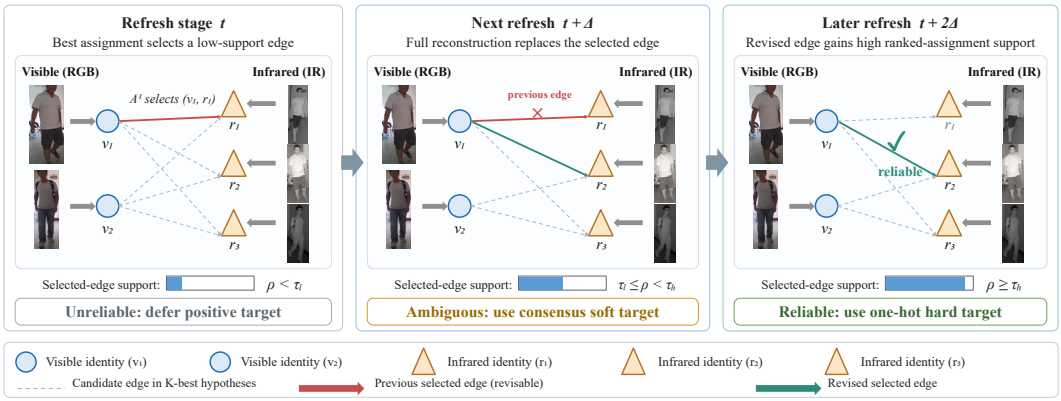


Figure 5: Revisable correspondence during co-evolution: an uncertain match is deferred, reassigned after reconstruction, and confirmed by later evidence.

Qualitative retrieval results. Fig. 6 complements the aggregate CMC and mAP results with query-level behavior. Successful cases show that ReCUE retrieves cross-modal positives despite substantial illumination and viewpoint changes. The difficult cases contain visually similar negative identities near the top of the ranking, but recover the correct identity within the Top-10 list. The remaining failures are primarily associated with weak appearance evidence, severe occlusion, or visually ambiguous clothing patterns.

Additional benchmark results on RegDB. We additionally evaluate ReCUE on RegDB (Nguyen et al., 2017) under the standard VIS→IR and IR→VIS retrieval protocols. Tab. 5 compares ReCUE with representative methods under different supervision settings.

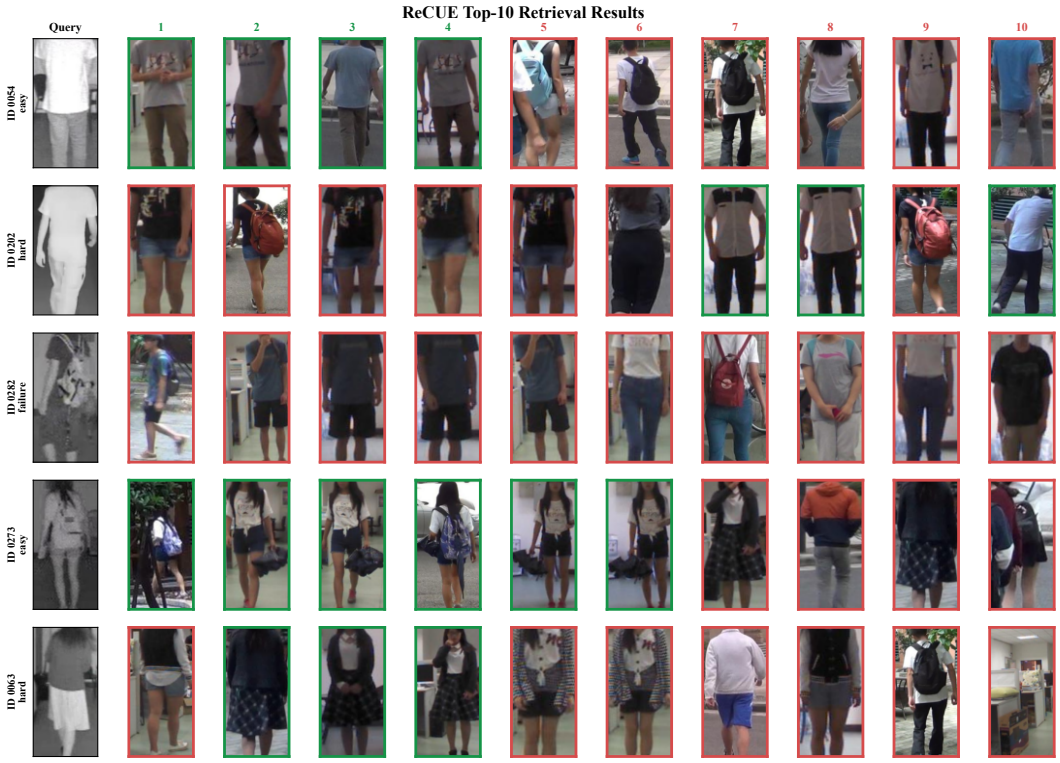


Figure 6: Top-10 cross-modal retrieval examples on SYSU-MM01. Query images are shown in the first column. Green and red borders denote correct and incorrect gallery matches, respectively. The examples include straightforward queries, difficult queries whose first correct match appears at a later rank, and representative failure cases.

Type	Method	VIS→IR		IR→VIS	
		R1	mAP	R1	mAP
S-VIReID	AGW (Ye et al., 2022)	70.1	66.4	70.5	65.9
	CAJ (Ye et al., 2021)	85.0	79.1	84.8	77.8
SS-VIReID	DPIS (Shi et al., 2023)	62.3	53.2	61.5	52.7
	MUCG (Zheng et al., 2024)	86.9	76.7	83.7	74.1
US-VIReID	GUR (Yang et al., 2023a)	73.9	70.2	75.0	69.9
WS-VIReID	HECCL (Zhang et al., 2025)	86.9	80.1	86.4	80.6
	Ours	**88.16**	**81.52**	**87.68**	**81.84**

Table 5: Comparison with previous VI-ReID methods on RegDB (%).

ReCUE achieves 88.16% Rank-1 and 81.52% mAP for VIS→IR, and 87.68% Rank-1 and 81.84% mAP for IR→VIS. It exceeds the reported HECCL results in both retrieval directions, providing additional evidence that the improvements observed on SYSU-MM01 and LLCM are also maintained on RegDB.

Complete co-evolution trajectory. Tab. 6 provides the integer counts underlying Fig. 3(b). The non-monotonic changes at epochs 60 and 100 are consistent with a refresh that reconstructs the global assignment rather than accumulating accepted pairs. Across the full trajectory, the reliable set expands while retaining above 91% precision.

Early-error correction and routing-state revision. Tab. 7 compares correspondence transitions from the shared epoch-20 checkpoint to the final assignment. Fixed routing cannot change an edge, while naive refresh both corrects and introduces errors. Locking the initially reliable route prevents its transitions by construction, not by discovering that every such edge is correct. Full reconstruction

Epoch	Correct / 395	Corr. Acc.	Reliable	Rel. correct	Precision	Coverage
20	307	77.7	266	244	91.7	67.3
40	310	78.5	273	250	91.6	69.1
60	308	78.0	270	246	91.1	68.4
80	313	79.2	284	259	91.2	71.9
100	311	78.7	280	255	91.1	70.9
120	**316**	**80.0**	**302**	**276**	**91.4**	**76.5**

Table 6: Complete SYSU-MM01 co-evolution statistics. Accuracy, precision, and coverage are percentages.

has the largest wrong-to-correct flow and net gain. Its recovery is concentrated among initially ambiguous and unreliable mistakes, consistent with soft supervision and deferral leaving more room for later revision.

(a) Overall correctness flow

Strategy	W→C	C→W	Net	Final correct	Acc.
Fixed Top-K	0	0	0	307	77.7
Naive best-A refresh	18	15	+3	310	78.5
Reliable edges locked	15	9	+6	313	79.2
Full revisable	**26**	17	+9	**316**	**80.0**

(b) Flow by initial route

Strategy	Route	Initial wrong	W→C	Recovery	C→W
Naive	R	22	2	9.1	6
	A	37	10	27.0	6
	U	29	6	20.7	3
Locked	R	22	0	0.0	0
	A	37	9	24.3	6
	U	29	6	20.7	3
Full	R	22	3	13.6	5
	A	37	13	35.1	8
	U	29	10	34.5	4

Table 7: Correspondence transitions under controlled update strategies on SYSU-MM01. Route labels are assigned at the shared epoch-20 checkpoint; recovery and accuracy are percentages.

For Full, 26 of 88 initial errors are corrected and 17 of 307 initially correct relations become wrong. Tab. 8 gives the underlying correctness and routing-state transitions. Routing also moves in both directions: 20 initially reliable pairs are later downgraded, while 56 initially non-reliable pairs become reliable.

(a) Correctness transition

Transition	Pairs
Correct → Correct	290
Correct → Wrong	17
Wrong → Correct	26
Wrong → Wrong	62

(b) Recovery by initial route

Route	Wrong	W→C	Recovery
Reliable	22	3	13.6
Ambiguous	37	13	35.1
Unreliable	29	10	34.5

(c) Routing-state transition

Initial \ Final	R	A	U	Total
R	246	16	4	266
A	48	30	8	86
U	8	12	23	43
Total	302	58	35	395

Table 8: Revision dynamics from global epoch 20 to 120.

C.1 Initial-Cost Perturbation Stress Test

Intervention protocol. To intervene on the quality of the initial correspondence without consulting ground truth, we perturb only the cost passed to the assignment solver at global epoch 20:

$$C^{(20)}_{\text{pert}} = C^{(20)} + \alpha s_C Z, \qquad Z_{ab} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{44}$$

where

$$s_C = 1.4826 \operatorname*{median}_{a,b} \left| C^{(20)}_{ab} - \operatorname*{median}_{c,d} C^{(20)}_{cd} \right| \tag{45}$$

is the median-absolute-deviation scale of the current cost. We use $\alpha \in \{0, 0.15, 0.30, 0.50\}$. The perturbation matrix is generated without ground-truth correspondence and is shared by Fixed, Naive, Locked, and Full for each perturbation strength. Prototypes and the unperturbed $C^{(20)}$ are identical across policies; only the cost supplied to the first assignment inference is changed. From epoch 40 onward, every policy that refreshes correspondence uses its normal, unperturbed current cost. Thus Naive and Full receive the same opportunity to use new evidence, and their difference isolates how that evidence is routed and whether earlier decisions remain revisable. The noise strengths were fixed without inspecting correspondence accuracy or retrieval performance.

Correspondence flow. Tab. 9 shows that stronger cost perturbation degrades the common initial assignment. Full consistently recovers the most initial errors and produces the largest net correction. It need not introduce fewer new errors at every strength: in the clean run its C→W count is 17, compared with 15 for Naive, and the two are equal at $\alpha = 0.15$. At $\alpha \geq 0.30$, Full both corrects more initial errors and introduces fewer new ones than Naive.

α	Strategy	Initial correct	W→C	C→W	Net	Final correct	Corr. Acc.	Rank-1	mAP
0	Fixed	307	0	0	0	307	77.7	72.3	68.0
	Naive	307	18	15	+3	310	78.5	72.1	67.8
	Locked	307	15	9	+6	313	79.2	72.6	68.5
	Full	307	**26**	17	+9	**316**	**80.0**	**73.0**	**69.0**
0.15	Fixed	298	0	0	0	298	75.4	71.8	67.4
	Naive	298	27	19	+8	306	77.5	72.0	67.4
	Locked	298	20	9	+11	309	78.2	72.3	68.1
	Full	298	**34**	19	+15	**313**	**79.2**	**72.8**	**68.7**
0.30	Fixed	284	0	0	0	284	71.9	70.9	66.5
	Naive	284	34	22	+12	296	74.9	71.5	67.0
	Locked	284	27	11	+16	300	75.9	71.9	67.5
	Full	284	**45**	20	+25	**309**	**78.2**	**72.4**	**68.2**
0.50	Fixed	261	0	0	0	261	66.1	69.7	65.1
	Naive	261	43	24	+19	280	70.9	70.4	65.9
	Locked	261	36	12	+24	285	72.2	71.1	66.6
	Full	261	**59**	22	+37	**298**	**75.4**	**71.5**	**67.3**

Table 9: Correspondence recovery after perturbing the shared initial cost. Fixed cannot change by definition. Bold within each strength marks the best final result; the low C→W count for Locked partly follows from preserving reliable edges by construction.

(a) High-consensus initial errors

α	Reliable	Wrong	Precision
0	266	22	91.7
0.15	267	26	90.3
0.30	271	35	87.1
0.50	275	49	82.2

(b) Recovery of initially wrong reliable edges

α	Wrong R	Naive	Locked	Full
0	22	2 (9.1)	0 (0.0)	**3 (13.6)**
0.15	26	7 (26.9)	0 (0.0)	**8 (30.8)**
0.30	35	10 (28.6)	0 (0.0)	**13 (37.1)**
0.50	49	14 (28.6)	0 (0.0)	**20 (40.8)**

Table 10: Effect of the intervention. Panel (a) shows that noise can create high-consensus wrong correspondences by breaking competition among nearby assignments. Panel (b) reports recovered counts and recovery rates in parentheses. Locked has zero reliable-edge recovery by definition.

Routing controls. Continuous confidence weighting applies a one-hot hard target to every edge in $A^{(1)}$, weighted by ρ_{ab}:

$$\mathcal{L}_{\text{cw}} = \frac{\sum_{(a,b)\in A^{(1)}} \rho_{ab} \, \text{CE}(p_{a:}, \delta_b)}{\sum_{(a,b)\in A^{(1)}} \rho_{ab}}. \tag{46}$$

It uses neither soft targets nor deferral. In "Hard + Soft, no defer," reliable pairs are hard and both ambiguous and unreliable pairs receive the hypothesis-distribution target. In "Hard + Defer,

no soft," only reliable pairs receive correspondence-derived positive supervision; ambiguous and unreliable pairs are both deferred. Main Tab. 2(d) reports the corresponding controlled comparison. Its small reversal between the two partial designs also shows that correspondence accuracy and retrieval quality need not move in lockstep.

Initial consensus levels. The six support levels in Fig. 3(a) partition all 395 selected edges and recover the initial 307/395 correspondence accuracy (Tab. 11). At the training threshold $\rho \geq 0.7$, 244 of 266 edges are correct, giving 91.7% precision at 67.3% coverage.

ρ	0.2	0.4	0.6	0.7	0.8	1.0	Total
Pairs	43	55	31	54	96	116	395
Correct	14	31	18	47	89	108	307
Incorrect	29	24	13	7	7	8	88
Accuracy	32.6	56.4	58.1	87.0	92.7	93.1	77.7

Table 11: Observed uniform Top-K consensus levels at the initial SYSU-MM01 checkpoint. Accuracy is a percentage.

Reliability score	AUROC↑	AUPRC↑	Empirical AURC↓	P@67.3↑
Local cost $(-C_{ab})$	0.688	0.407	0.1231	84.2
Single-optimum margin	0.742	0.458	0.1096	87.2
Sinkhorn / OT marginal	0.771	0.493	0.1012	89.5
Cost-weighted Top-K	0.790	0.489	**0.0946**	90.2
Uniform Top-K	**0.803**	**0.520**	0.0974	**91.7**

Table 12: Reliability-score diagnostics on the common 395-edge initial checkpoint. Incorrect correspondence is positive for AUROC/AUPRC; P@67.3 is reliable-set precision (%). AURC is exact tie-aware empirical area.

Risk-coverage evaluation. For AUROC and AUPRC, incorrect correspondence is positive and scores are oriented toward unreliability; both metrics use their standard grouped treatment of tied scores. Risk-coverage instead selects edges in descending reliability. At a displayed coverage c, we use $k = \mathrm{round}(cN)$. If the cutoff falls within a tied score group g containing n_g edges and c_g correct edges, expected correctness under all within-group orderings is

$$\bar{C}(k) = C_{<g} + (k - N_{<g})\frac{c_g}{n_g}, \qquad R(k) = 1 - \frac{\bar{C}(k)}{k}. \tag{47}$$

The exact tie-aware empirical area is $\mathrm{AURC} = N^{-1}\sum_{k=1}^{N} R(k)$. Ranking and tie handling are independent of ground-truth correspondence; ground truth is used only after ranking to evaluate correctness and risk. Tab. 13 shows that uniform consensus is strongest at medium-to-high coverage, while its coarse support levels limit fine ordering at low coverage and give the cost-weighted marginal a slightly lower overall AURC.

Reliability score	P@20	P@40	P@55	P@67.3	P@80	P@90	P@100
Local cost $(-C_{ab})$	93.7	91.8	87.6	84.2	80.1	78.7	77.7
Single-optimum margin	96.2	93.7	90.3	87.2	82.9	79.8	77.7
Sinkhorn / OT marginal	**97.5**	**95.6**	92.2	89.5	84.5	80.6	77.7
Cost-weighted Top-K	96.2	94.9	**93.1**	90.2	85.4	80.9	77.7
Uniform Top-K	93.1	93.0	92.8	**91.7**	**86.3**	**82.7**	77.7

Table 13: Tie-aware precision (%) across coverage on the 395 $A^{(1)}$ edges at the initial SYSU-MM01 checkpoint. Displayed coverages use $k = \mathrm{round}(cN)$; for uniform Top-K, cutoffs inside a consensus tie use expected precision over all within-tie orderings.

Truncation stability. At the initial checkpoint, C is fixed and $A^{(1)}$ is the same for every K. We compare the 395 selected-edge values $\{\rho_e^{(K)} : e \in A^{(1)}\}$ with a larger-K reference at $K = 100$. Spearman correlation uses average ranks for ties, and route agreement is the fraction assigned to the same R/A/U state as the reference. The convergence of MAE, rank correlation, and routing shows that $K = 10$ approximates the larger truncation without being a sharp retrieval optimum. The $K = 100$ row is a reference, not ground truth or an exact assignment marginal.

K	Reliable	Correct	Precision	Coverage	MAE to K=100↓	Spearman↑	Route agreement↑
5	261	240	**92.0**	66.1	0.074	0.914	348/395 = 88.1
10	266	244	91.7	67.3	0.043	0.958	370/395 = 93.7
20	268	246	91.8	67.8	0.031	0.974	376/395 = 95.2
50	270	247	91.5	68.4	0.011	0.993	389/395 = 98.5
100	271	248	91.5	68.6	0 (ref.)	1.000 (ref.)	- (ref.)

Table 14: Initial-checkpoint uniform Top-K truncation stability on the fixed 395 edges selected by $A^{(1)}$. Accuracy-derived quantities are percentages.

Hyperparameter sensitivity. Complementing the truncation analysis above, Tab. 15 examines retrieval sensitivity to the number of hypotheses, routing thresholds, and correspondence refresh interval. Performance varies only slightly across the tested values of K, while the default $(\tau_l, \tau_h) = (0.4, 0.7)$ gives the strongest retrieval among the tested thresholds. Periodic refresh clearly improves over keeping correspondence fixed, whereas increasing the refresh frequency from 20 to 10 epochs provides no additional benefit. Overall, the default configuration lies in a stable operating region rather than at a sharp numerical optimum.

Hypotheses K

K	Corr.	R1	mAP
5	79.2	72.9	68.8
10	**80.0**	73.0	**69.0**
20	79.5	**73.1**	68.9

Routing thresholds

(τ_l, τ_h)	R1	mAP
$(0.3, 0.7)$	72.8	68.7
$(0.4, 0.7)$	**73.0**	**69.0**
$(0.4, 0.8)$	72.7	68.8

Refresh interval

Interval	Corr.	R1	mAP
None	77.7	72.3	68.0
40	79.0	72.5	68.4
20	**80.0**	**73.0**	**69.0**
10	79.5	72.8	68.7

Table 15: Sensitivity to the number of hypotheses K, routing thresholds, and correspondence refresh interval on SYSU-MM01 All Search. Corr. denotes correspondence accuracy; bold indicates the best value within each sensitivity block.

Assignment-cost balance and consistency regularization. Tab. 16 varies the balance between prototype geometry and bidirectional identity evidence in Eq. 3, and the consistency weight in Eq. 7. Moderate settings form a broad performance plateau rather than a sharp optimum. Using either evidence source alone is weaker overall, while correspondence accuracy and retrieval do not always move in lockstep: $\lambda = 0.75$ slightly improves Rank-1 over the default despite lower correspondence accuracy and mAP, and $\lambda_{\text{cons}} = 0.5$ gives the highest correspondence accuracy without the best retrieval.

The two assignment cues have complementary failure modes: prototype distance summarizes identity-level feature geometry, whereas bidirectional predictor evidence tests whether modality-specific classifiers support the same cross-modal identity relation. Their balanced combination therefore improves the ordering of competing assignments even though neither cue is sufficient on its own.

(a) Assignment-cost balance

λ	Corr. Acc.	Rank-1	mAP
0.00	78.0	72.3	68.2
0.25	79.0	72.8	68.6
0.50	**80.0**	73.0	**69.0**
0.75	79.5	**73.1**	68.8
1.00	77.7	72.2	68.0

(b) Consistency weight

λ_{cons}	Corr. Acc.	Rank-1	mAP
0.0	79.2	72.6	68.5
0.5	**80.3**	72.9	68.8
1.0	80.0	**73.0**	**69.0**
2.0	79.5	72.8	68.7

Table 16: Sensitivity to the assignment-cost balance and bidirectional consistency weight on SYSU-MM01 All Search. Moderate settings form a broad plateau; no single setting dominates every metric.