Decoupled Prompting and Topology-Aware Optimal Transport for CLIP-Based Unsupervised Cross-Modal Re-Identification

Xiaohao Xie, Wenhua Jiao, and Wei Meng

Abstract. Unsupervised Visible-Infrared Person Re-identification (USL-VI-ReID) is critical for cross-modal intelligent surveillance. While vision-language models (e.g., CLIP) present powerful representational capabilities, directly fine-tuning them for USL-VI-ReID often causes catastrophic *feature collapse* and *prompt degradation* due to massive domain gaps and noisy pseudo-labels. Furthermore, traditional discrete matching and heuristic denoising strategies suffer from severe cross-modal *information starvation* and *numerical bias* against hard positives. To address these challenges, we propose a robust, CLIP-based unsupervised cross-modal fine-tuning framework. First, we design an *implicit adapter* fine-tuning strategy coupled with *decoupled multi-dimensional semantic prompting* to isolate domain biases without destroying pre-trained priors. Second, a Cluster-Aware Cross-Modal Semantic Alignment (CCSA) mechanism maps dynamic visual centers to modality-shared textual proxies via visual-conditioned prompting, facilitating an implicit soft alignment decoupled from hard clustering noise. Third, we frame cross-modal association as a Topology-Aware Optimal Transport (TOTO) problem. Utilizing Fused Gromov-Wasserstein (FGW) constraints and Argmax assignments, TOTO injects potent hard regularization to overcome optimization inertia on difficult samples. Finally, our Pure Relative Confidence Ratio and Dual Adaptive Denoising (RCR-DAD) module eliminates numerical bias, formulating a robust self-paced learning trajectory. Extensive experiments on SYSU-MM01 and RegDB demonstrate our framework achieves state-of-the-art performance. The code will be released.

Keywords: Unsupervised VI-ReID · Vision-Language Models · Optimal Transport · Cross-Modal Alignment.

1 Introduction

Visible-Infrared Person Re-identification (VI-ReID) [22] is critical for 24-hour intelligent surveillance. To avoid prohibitive cross-modal annotation costs, Unsupervised VI-ReID (UVI-ReID) [11] has garnered significant attention, predominantly relying on Cluster Contrastive Learning [6, 12] using pseudo-labels from visual clustering. While pre-trained vision-language models like CLIP [16, 9] provide powerful generic representations, directly fine-tuning them for UVI-ReID suffers from massive domain gaps. The combination of inaccurate pseudo-labels

and full-parameter fine-tuning easily destroys the pre-trained feature space, causing *feature collapse*. To adapt to modality discrepancies, visual prompting [39] has been introduced. However, simply stacking generic prompts lacks explicit constraints, allowing them to take optimization shortcuts by learning modality-specific noise, which inevitably leads to *prompt degradation*.

Furthermore, mainstream cluster-driven cross-modal paradigms [26, 2] essentially lack direct cross-modal identity supervision. Relying solely on the visual space makes clustering highly susceptible to background clutter, leading to rapid accumulation of pseudo-label noise. This instability is exacerbated by the dynamic conflicts inherent in cross-modal discrete matching. Translating continuous distances into discrete labels forces a dilemma: multi-round strict bipartite matching (e.g., PGM [24]) enforces global one-to-one constraints, which inevitably forces mismatches on unbalanced clusters, injecting lethal hard noise into the momentum dictionary. Conversely, confidence-based heuristic matching [36, 34] typically evades challenging samples, leading to severe *information starvation* and depriving the network of cross-domain gradients.

Finally, instance-level pseudo-label denoising presents a critical hurdle between noise removal and hard sample optimization. Existing mechanisms typically rely on absolute prediction confidence [28], introducing a fatal *numerical bias*: highly valuable hard positives naturally exhibit lower confidence due to their ambiguous features. Traditional methods mistakenly suppress these critical samples, causing severe optimization inertia.

To bridge the modality gap without destroying the pre-trained knowledge, we propose a robust, CLIP-based unsupervised cross-modal fine-tuning framework. Our main contributions are summarized as follows:

1. **Robust CLIP Fine-tuning Framework:** We design a phased strategy featuring a *frozen warm-up* and *implicit adapters*. This achieves an optimal balance between preserving pre-trained knowledge and adapting to target domains, effectively avoiding the catastrophic forgetting associated with full fine-tuning.
2. **Decoupled Multi-dimensional Semantic Prompting:** We propose a composite input stream integrating modality, camera, and common prompts. Modality and camera prompts explicitly absorb domain biases, while auxiliary supervision strictly guides the common prompt to extract pure identity features, preventing representation degradation.
3. **Cluster-Aware Cross-Modal Semantic Alignment (CCSA):** We map dynamic visual cluster centers to domain-invariant semantic text anchors. This injects explicit semantic regularization into the unsupervised visual clustering, enhancing heterogeneous matching robustness.
4. **Topology-Aware Optimal Transport (TOTO):** We frame cross-modal association via Fused Gromov-Wasserstein (FGW) topological constraints and Argmax local hard assignments. This replaces traditional multi-round graph matching, injecting potent "hard regularization" gradients to overcome optimization inertia towards challenging samples.
5. **Pure Relative Confidence Ratio and Dual Adaptive Denoising (RCR-DAD):** We propose a scale-invariant confidence evaluator. Through unidi-

rectional truncation, it inherently protects hard samples and achieves optimization equality, driving a robust self-paced learning trajectory coupled with adaptive temperature and soft-masking mechanisms.

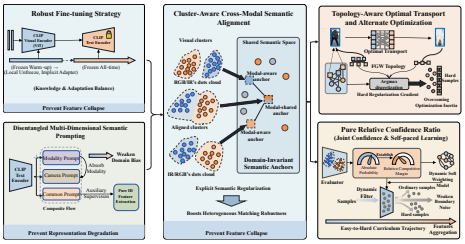

Fig. 1. The proposed unsupervised VI-ReID framework. The framework consists of five core modules organized across three stages. Left (Input Prompting): (1) Robust Fine-Tuning Strategy (Top): A phased approach combining *frozen warm-up* and *implicit adapters* to prevent feature collapse. (2) Decoupled Multi-Dimensional Semantic Prompting (Bottom): Explicit isolation of domain biases via specific prompts and pure identity feature extraction via auxiliary supervision. Center (Semantic Alignment): (3) Cluster-Aware Cross-Modal Semantic Alignment (CCSA): Explicit semantic regularization pulling dynamic visual clusters toward domain-invariant semantic anchors. Right (Matching Denoising): (4) Topology-Aware Optimal Transport (Top): Injecting strong "hard regularization" gradients through Fused Gromov-Wasserstein (FGW) constraints and Argmax assignment. (5) Pure Relative Confidence Ratio (Bottom): A joint evaluator driving an adaptive, easy-to-hard self-paced learning trajectory by filtering noise while protecting hard samples.

2 Related Works

2.1 Unsupervised Cross-Modal Re-Identification

Unsupervised ReID relies heavily on clustering [12] and contrastive learning paradigms [6]. To mitigate pseudo-label noise in single-modality scenarios, various refinement strategies have been proposed, including probability distillation [3], stochastic training [13], pseudo-pair similarities [23], transformer architectures [14], and part-based refinements [5]. However, these methods struggle against the extreme heterogeneous modality discrepancy inherent in Visible-Infrared ReID (VI-ReID) [22].

To eliminate expensive cross-modal annotations, Unsupervised VI-ReID (USL-VI-ReID) has rapidly evolved from homogeneous-to-heterogeneous transitions [11] to deep cross-modal association mining, utilizing dual-contrastive aggregation [26] and progressive graph matching [24]. Recently, handling high-noise environments and hard sample matching has become the core focus, driving innovations in adaptive confidence-driven learning [36], self-paced curriculum association [25], and dual-level outlier filtering [34]. These milestones directly inspire the noise-resistant alignment mechanisms designed in this paper.

2.2 Vision-Language Models and Prompt Learning

Large-scale vision-language models (VLMs) like CLIP [16] have revolutionized visual representation learning. To adapt these foundation models to downstream

tasks efficiently, continuous prompt learning [39] has replaced manual templates, with extensions to instance-conditional dynamic tokens [38] and dense predictions [17]. Consequently, CLIP-ReID [9] successfully introduced CLIP into supervised ReID via explicit text prompts.

However, directly applying conventional prompt learning to USL-VI-ReID faces severe adaptability bottlenecks. Lacking explicit cross-modal supervision, generic prompts easily entangle with modality-specific background noise and camera biases, leading to "Prompt Degradation". To resolve this feature entanglement dilemma in unsupervised environments, this paper proposes a decoupled multi-dimensional semantic prompting and cluster alignment mechanism to explicitly isolate domain noise and extract pure pedestrian identity features.

3 Methodology

As illustrated in Fig.1, our framework alternates between clustering-driven cross-modal alignment and representation optimization. We establish a robust fine-tuning strategy to prevent feature collapse (Sec. 3.1) and extract pure identity features via decoupled semantic prompting (Sec. 3.2). For pseudo-label refinement, we align cross-modal semantics (Sec. 3.3) and apply topology-aware optimal transport (Sec. 3.4). Finally, an adaptive denoising mechanism drives robust self-paced learning (Sec. 3.5).

3.1 Overall Architecture and Robust Fine-Tuning Strategy

Our method builds upon the ClusterContrast [6] paradigm. During the **E-step**, the network extracts features to conduct independent clustering. We introduce text-anchor-based semantic alignment and Fused Gromov-Wasserstein (FGW) optimal transport to generate pure cross-modal pseudo-labels. In the **M-step**, the model employs a momentum-updated ClusterNCE Loss combined with hard regularization and dynamic soft-weighting to aggregate cross-modal features robustly.

Implicit Adapter Fine-Tuning. Replacing CNNs with a massive CLIP Vision Transformer (ViT) and applying full fine-tuning easily destroys pre-trained priors, causing *feature collapse*. To balance knowledge retention and domain adaptation efficiently, we adopt an asymmetric unfreezing strategy. We selectively unfreeze only the deepest Transformer Block and all internal projection mapping components (i.e., attention projections and MLP down-projection layers) across the backbone. Crucially, the massive MLP expansion layers and layer normalizations are strictly frozen. This transforms the unfrozen projection components into an "Implicit Adapter." Such an asymmetric design ensures that the core non-linear representational priors of the foundation model remain intact, forcing the network to fit the Re-ID task using existing high-order semantics rather than destroying its generalized receptive field.

3.2 Decoupled Multi-Dimensional Semantic Prompting

Directly applying a single text prompt easily entangles with the massive VI gap and camera biases. We propose an explicitly decoupled prompting architecture. Given an image, its token sequence fed into the ViT is constructed as

$X = [E_{cls}, P_{common}, P_{modal}, P_{cam}, E_{patch}]$, where E_{cls} and E_{patch} are class and patch embeddings. The modality prompt $P_{modal} \in \{P_v, P_{ir}\}$ absorbs spectral domain biases, while the camera prompt $P_{cam} \in \{P_{c1}, \ldots, P_{cK}\}$ absorbs viewpoint noises. With underlying variances explicitly isolated, the common identity prompt P_{common} is enforced to capture pure domain-invariant pedestrian features.

Auxiliary Supervision. To prevent P_{common} from degrading into useless context, we introduce a token-level auxiliary supervision. Specifically, we extract the output tokens associated with P_{common} from the visual encoder and perform mean pooling to derive the common identity feature Z_{common}. Following independent clustering, we establish intra-modal memory banks M_m for modality $m \in \{v, ir\}$. Given a baseline feature f_i (i.e., Z_{cls}) with pseudo-label y_i, the standard intra-modal contrastive loss is

$$\mathcal{L}_{intra}(f) = \sum_{m} -\frac{1}{|B|} \sum_{i} \log \frac{\exp(f_i \cdot M_m[y_i]/\tau)}{\sum_k \exp(f_i \cdot M_m[k]/\tau)}. \tag{1}$$

We explicitly extend this supervision to the common prompt output Z_{common} as an auxiliary loss. Furthermore, to explicitly preserve instance-specific discriminability against cluster homogenization, an instance-level contrastive loss $\mathcal{L}_{inst}$ is introduced to maximize the similarity between two augmented views. Mathematically, it shares the same InfoNCE formulation as Eq.1. The joint decoupled prompting loss is: $\mathcal{L}_{DMSP} = \mathcal{L}_{intra}(Z_{cls}) + \lambda_{aux}\mathcal{L}_{intra}(Z_{common}) + \lambda_{inst}\mathcal{L}_{inst}$.

3.3 Cluster-Aware Cross-Modal Semantic Alignment (CCSA)

Directly clustering unaligned VI features often causes cluster centers to oscillate. We leverage text priors to construct a modality-shared semantic space. For K pseudo-label identities, instead of static text, we project visual centroids into context tokens v^k via a lightweight Meta-Network (Linear Layer). By injecting v^k into a base template and passing it through CLIP's frozen text encoder, we generate epoch-stable semantic anchors f_t^k, forming a beacon set $\mathcal{A} = \{f_t^1, \ldots, f_t^K\}$.

Instead of orthogonal penalties, we compel individual visual instances (V_i^M) to map toward their corresponding semantic anchors via an instance-to-anchor Vision-to-Text (V2T) alignment loss:

$$\mathcal{L}_{v2t}^{M} = -\frac{1}{|B_M|} \sum_{i=1}^{|B_M|} \log \frac{\exp(\text{sim}(V_i^M, f_t^{\hat{y}_i})/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(V_i^M, f_t^j)/\tau)} \quad (M \in \{v, ir\}) \tag{2}$$

where $\mathcal{L}_{v2t} = \mathcal{L}_{v2t}^{v} + \mathcal{L}_{v2t}^{ir}$. This mechanism suppresses cluster shifts caused by noise and implicitly pulls heterogeneous features of the same identity closer, replacing rigid geometric constraints with robust semantic regularization.

3.4 Topology-Aware Optimal Transport (TOTO)

To overcome severe misassignments caused by greedy visual matching, we frame pseudo-label generation as an Optimal Transport problem utilizing Fused Gromov-Wasserstein (FGW) distance. Given visible and infrared cluster centers $\mathbf{F}^v$ and

$\mathbf{F}^{ir}$, along with their intra-modality topological affinity matrices $\mathbf{D}^{v}$ and $\mathbf{D}^{ir}$, the joint cost matrix $\mathbf{C}_{fused}^{(t)}$ simultaneously aligns feature similarity and geometric manifold structure:

$$\mathbf{C}_{fused}^{(t)} = (1-\beta)\mathbf{C}_{feat} + \beta(\mathbf{C}_{feat} \odot (\mathbf{D}^{v}\mathbf{P}^{(t-1)}(\mathbf{D}^{ir})^{T})) \tag{3}$$

Utilizing the Sinkhorn-Knopp algorithm, we solve the entropy-regularized optimal transport objective to obtain a globally coordinated soft assignment matrix $\mathbf{P}^{*}$. However, continuously fluctuating soft weights deprive hard positives of optimization gradients. Therefore, we apply strict Argmax local discretization: $\hat{y}_{v\to ir}^{(i)} = \arg\max_j \mathbf{P}_{i,j}^{*}$ (and vice versa). This translates global topological coordination into binary *Hard Regularization*, providing drastic and sustained cross-domain pulling gradients. Finally, we employ Alternate Cross-Modal Contrastive Learning (ACCL) [24]:

$$\mathcal{L}_{cross} = \begin{cases} \mathcal{L}_{InfoNCE}\left(f_{ir}, \mathcal{M}_{v}, \hat{y}_{ir\to v}\right), & \text{if epoch\%2=1} \\ \mathcal{L}_{InfoNCE}\left(f_{v}, \mathcal{M}_{ir}, \hat{y}_{v\to ir}\right), & \text{if epoch\%2=0} \end{cases} \tag{4}$$

This physically disentangles optimization gradients, forcing one modality to converge toward the frozen momentum dictionary of the other alternately.

3.5 Pure Relative Confidence Ratio and Dual Adaptive Denoising

Bounded by hard assignments, traditional absolute thresholding blindly discards valuable hard positives, causing optimization inertia. We propose the Pure Relative Confidence Ratio (RCR), which evaluates reliability based on relative significance rather than absolute magnitude. Given target probability $p_{target}^{(i)}$ and the strongest competitor probability $p_{comp}^{(i)}$, the truncated score is:

$$R_i = \max\left(w_{min}, \min\left(1.0, \left(\frac{p_{target}^{(i)}}{p_{comp}^{(i)} + \epsilon}\right)^{\gamma}\right)\right) \tag{5}$$

This achieves *Optimization Equality*: as long as $p_{target}^{(i)} > p_{comp}^{(i)}$, even extreme hard samples gain a full weight of 1.0. If mismatch occurs, it triggers an exponential decay via γ. Guided by RCR, we introduce the decoupled Dual Adaptive Denoising (DAD):

(1) **Intra-Modal Denoising (IMD):** This module acts on the intra-modal contrastive loss $\mathcal{L}_{intra}(Z_{cls})$. It dynamically adjusts the temperature parameter $\tau_i = \tau_{base} + \lambda(1-R_i)$ based on the reliability score. For suspected noise (low R_i), the increased τ_i flattens the output probability distribution, effectively smoothing the abnormal gradients during backpropagation. This soft-tolerance strategy prevents label noise from destroying the stability of intra-modal clustering manifolds.

(2) **Cross-Modal Denoising (CMD):** To prevent ambiguous or mismatched visual samples from polluting the pure semantic text space, we reconstruct the

Table 1. Comparison with state-of-the-art methods on SYSU-MM01 and RegDB datasets. * denotes pre-training on extra labeled data; † denotes results without camera information. Rank-1 (R1), mAP, and mINP metrics are reported in (%). **Bold** indicates the best result among unsupervised methods, and underline denotes the global best across all methods.

	Method	Venue	SYSU-MM01 (All Search)			SYSU-MM01 (Indoor)			RegDB (Vis → Ther)			RegDB (Ther → Vis)		
			R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
Supervised	AGW [33]	TPAMI'21	47.50	47.65	35.30	54.17	62.97	59.23	70.05	66.37	50.19	70.49	65.90	51.24
	DDAG [31]	ECCV'20	54.75	53.02	–	61.02	67.98	–	69.34	63.46	–	68.06	61.80	–
	CAJ [32]	ICCV'21	69.88	66.89	–	76.26	80.37	–	85.03	79.14	–	84.75	77.82	–
	PartMix [8]	CVPR'23	77.78	74.62	–	81.52	84.38	–	85.66	82.27	–	84.93	82.52	–
	DEEN [37]	CVPR'23	74.70	71.80	–	80.30	83.30	–	91.10	85.10	–	89.50	83.40	–
	ProtoHPE [35]	MM'23	71.90	70.60	–	77.80	81.30	–	88.70	83.70	–	88.70	82.00	–
	SAAI [7]	ICCV'23	75.90	77.03	–	83.20	88.01	–	91.07	91.45	–	92.09	92.01	–
	PMWGCN [19]	TIFS'24	66.80	64.90	–	72.60	76.20	–	90.60	85.10	–	91.10	84.80	–
Unsupervised	OTLA* [21]	ECCV'22	29.90	27.10	–	29.80	38.80	–	32.90	29.70	–	32.10	28.60	–
	ADCA [26]	MM'22	45.51	42.73	28.29	50.60	59.11	55.17	67.20	64.05	52.67	68.48	63.81	49.62
	PGM [24]	CVPR'23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	–	69.85	65.17	–
	DOTLA* [4]	MM'23	50.36	47.36	32.40	53.47	61.73	57.35	85.63	76.71	61.58	82.91	74.97	58.60
	MBCCM [2]	MM'23	53.14	48.16	32.41	55.21	61.98	57.13	83.79	77.87	65.04	82.82	76.74	61.73
	CCLNet [1]	MM'23	54.03	50.19	–	56.68	65.12	–	69.94	65.53	–	70.17	66.66	–
	GUR† [27]	ICCV'23	60.95	56.99	41.85	64.22	69.49	64.81	73.91	70.23	8.88	75.00	69.94	56.21
	CHCR [15]	TCSVT'23	47.70	45.30	–	–	–	–	68.20	63.80	–	70.00	65.90	–
	SCA-RCP [10]	TKDE'24	51.41	48.52	33.56	56.77	64.19	59.25	85.59	79.12	–	82.41	75.73	–
	MMM [18]	ECCV'24	61.60	57.90	–	64.40	70.40	–	89.70	80.50	–	85.80	77.00	–
	N-ULC [20]	AAAI'25	61.81	58.92	45.01	67.04	73.08	69.42	88.75	82.14	68.75	88.17	81.11	66.05
	PCAL [29]	TIFS'25	54.39	51.95	38.09	59.69	66.72	62.44	86.43	82.51	72.33	86.21	81.23	68.71
	MCL [30]	ICCV'25	62.95	62.71	**50.63**	67.81	74.19	70.82	**89.83**	**83.12**	**72.86**	**88.64**	**82.04**	**69.12**
	Ours	–	**63.61**	**62.99**	45.85	**68.28**	**75.51**	**72.64**	81.50	76.67	64.21	81.26	75.96	63.52

vision-text alignment loss $\mathcal{L}_{v2t}$ utilizing a 1D dynamic soft-mask R_i:

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B} R_i \log \frac{\exp(\text{sim}(V_i, f_t^{\hat{y}_i})/\tau)}{\sum_{k=1}^{K}\exp(\text{sim}(V_i, f_t^{k})/\tau)} \tag{6}$$

During optimization, this mask functions as a flow control gate: only when a visual instance itself exhibits high reliability is its alignment gradient allowed to pass. This strictly confines pseudo-label pollution outside the semantic space.

Ultimately, the overall joint optimization loss is defined as $\mathcal{L}_{all} = \mathcal{L}_{DMSP} + \lambda_{v2t}\mathcal{L}_{v2t} + \mathcal{L}_{cross}$.

4 Experiments

4.1 Experimental Setup

Datasets: We evaluate our method on two publicly available datasets: SYSU-MM01 (6 cameras, dual modalities) and RegDB (1 camera, dual modalities). **Implementation Details:** We adopt CLIP ViT-B/16 as our backbone. During training, we employ a phased unfreezing strategy where only the prompts and BNNeck are trained in the first epoch, followed by unfreezing the 11th Transformer layer. Unsupervised clustering is performed via DBSCAN. **Detailed hyperparameter settings (e.g., learning rates, optimizer, batch size, and DBSCAN eps) are provided in Appendix A.**

Table 2. Comprehensive ablation study of each component in our proposed framework on the SYSU-MM01 dataset. MP: Multi-Dimensional Prompt; AL: Auxiliary Loss; IL: Instance Loss; CC: CCSA (V2T Loss); TO: TOTO; RC: RCR. **Bold** indicates the best performance.

Index	Proposed Components						All Search (%)		Indoor Search (%)	
	MP	AL	IL	CC	TO	RC	mAP	R1	mAP	R1
1							46.23	46.07	62.53	53.61
2	✓						49.39	50.32	63.04	54.97
3	✓	✓					49.88	51.33	66.99	58.86
4	✓		✓				50.34	51.89	65.54	57.02
5	✓	✓	✓				52.66	53.37	67.11	59.34
6			✓				49.44	50.57	64.21	56.35
7				✓			49.97	51.52	67.23	58.77
8					✓		47.45	52.10	70.55	62.85
9				✓		✓	52.04	53.02	66.02	58.04
10	✓	✓	✓	✓			53.74	54.51	67.96	60.43
11	✓	✓	✓		✓		57.34	59.65	73.00	65.95
12	✓	✓	✓			✓	53.65	53.97	67.94	60.10
13	✓	✓	✓	✓	✓		58.23	60.12	73.78	66.71
14	✓	✓	✓		✓	✓	60.34	62.43	73.42	66.81
15	✓	✓	✓	✓		✓	54.02	54.90	69.12	61.34
16	✓	✓	✓	✓	✓	✓	**62.99**	**63.61**	**75.51**	**68.28**

Table 3. Ablation study on the backbone unfreezing strategy and implicit adapters on SYSU-MM01. *Internal Proj.* refers to unfreezing all projection layers within the Transformer blocks (acting as Implicit Adapter). N denotes the number of unfrozen Transformer blocks.

Variant	Fine-Tuning Strategy Summary	Internal Proj.	Unfrozen Blocks (N)	All Search Rank-1 / mAP	Indoor Search Rank-1 / mAP
(a)	Base	×	0	4.67 / 7.26	6.64 / 14.55
(b)	Local Blocks Only	×	1	16.00 / 16.52	22.45 / 32.26
(c)	Pure Implicit Adapter	✓	0	54.21 / 48.08	61.89 / 69.90
(d)	**Hybrid Fine-Tuning (Ours)**	✓	**1**	52.10 / 47.45	**62.85 / 70.55**
(e)	Hybrid Fine-Tuning (Deeper)	✓	2	59.55 / 54.76	62.68 / 70.35
(f)	Hybrid Fine-Tuning (Over-unfrozen)	✓	3	48.84 / 45.72	56.44 / 64.85
(g)	Full Fine-Tuning (FFT)	–	12 (All)	**59.91 / 60.25**	60.25 / 65.69

4.2 Comparison with the State-of-the-Arts

To comprehensively evaluate the effectiveness of our proposed framework, we compare it with 21 state-of-the-art (SOTA) VI-ReID methods, including 8 supervised methods (e.g., PMWGCN [19], SAAI [7]) and 13 unsupervised methods (e.g., MCL [30], N-ULC [20], MMM [18]). The detailed quantitative results on the SYSU-MM01 and RegDB datasets are summarized in Tab.1.

Performance on SYSU-MM01. SYSU-MM01 is currently the most challenging large-scale benchmark due to its complex indoor and outdoor scenarios and severe cross-camera view variations. Under the All-Search setting, our method achieves 63.61% in Rank-1 and 62.99% in mAP, outperforming the latest ICCV'25 method MCL (62.95% Rank-1) and AAAI'25 method N-ULC (61.81% Rank-1). More remarkably, under the Indoor-Search setting, our method reaches 68.28% in Rank-1 and 75.51% in mAP. It is highly noteworthy that our mean Inverse Negative Penalty (mINP) reaches a staggering **72.64%**, which not only

Table 4. Sensitivity analysis of the common prompt sequence length.

Prompt Length	mAP/R1 (all)	mAP/R1 (indoor)
1	47.9/48.06	62.6/54.75
3	48.89/49.05	63.89/55.88
5	49.38/49.55	64.54/56.44
7	48.98/49.15	64.02/55.99
9	48.39/48.56	63.25/55.31
11	47.65/47.82	62.28/54.46

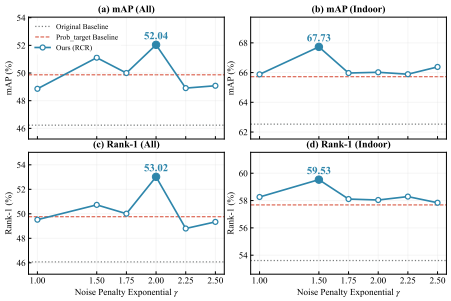

Fig. 2. Sensitivity of the noise penalty γ.

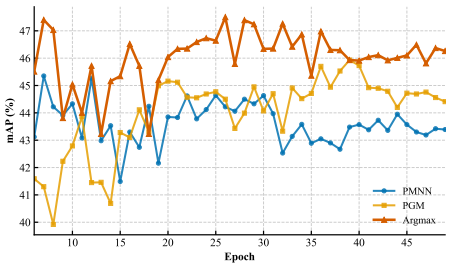

Fig. 3. Dynamics of matching strategies.

sets a new record among all unsupervised methods but also **surpasses all reported fully supervised methods**. Since mINP is a critical metric reflecting the model's ability to retrieve the hardest matching samples, this global superiority strongly demonstrates that our topology-aware optimal transport (TOTO) and adaptive denoising (RCR-DAD) mechanisms effectively protect hard samples and excavate pure identity semantics from severe background clutter.

Performance on RegDB. On the RegDB dataset, our framework also exhibits highly competitive performance, achieving 81.50% / 76.67% (Vis $\rightarrow$ Ther) and 81.26% / 75.96% (Ther $\rightarrow$ Vis) in Rank-1 / mAP, respectively. While our results on RegDB do not surpass the absolute highest scores (e.g., MCL), this phenomenon is theoretically consistent with our architectural design. RegDB is collected using only a single overlapping camera view, lacking the cross-camera topological variance and view biases present in real-world deployments. Consequently, our decoupled camera-specific prompts (P_{cam}) and the high-order

Table 5. Decoupled ablation study of the dual adaptive denoising mechanism (All Search). Baseline+V2T denotes the baseline with Visual-Text Alignment Loss without RCR. IMD (Intra-Modal Denoising) represents the intra-modal adaptive temperature smoothing, and CMD (Cross-Modal Denoising) represents the cross-modal dynamic soft-mask blocking.

Baseline+V2T	IMD	CMD	mAP	Rank-1
✓			49.97	51.52
✓	✓		50.85	51.94
✓		✓	51.1	52.07
✓	✓	✓	52.04	53.02

structural alignments cannot fully exert their intended advantages. Nevertheless, our method still significantly outperforms classical unsupervised paradigms like GUR [27] and MBCCM [2], proving its robust generalization capabilities.

Comparison with Supervised Methods. Although a predictable performance gap remains in Rank-1 accuracy compared to top-tier fully supervised architectures (e.g., PartMix [8], DEEN [37]) that utilize massive manually annotated cross-modal pairs, our unsupervised paradigm remarkably narrows this gap. By leveraging explicit semantic prompting and cluster-aware alignment, our method circumvents the expensive annotation costs while achieving matching robustness that rivals early supervised baselines like DDAG [31].

4.3 Ablation Study and Analysis

To deeply verify the effectiveness of each core component in the proposed framework, we conduct comprehensive ablation studies primarily on the highly challenging SYSU-MM01 dataset. As systematically summarized in Tab.2, the integration of individual modules into the baseline (Index 1) yields steady and progressive improvements, with the full framework (Index 16) achieving the optimal performance across all evaluation settings. To provide a rigorous explanation of these phenomena, the discussion in this section will strictly correspond to the innovative modules of our methodology, conducting an in-depth analysis from five dimensions: Robust Fine-Tuning Strategy, Decoupled Multi-Dimensional Semantic Prompting, CCSA, TOTO, and RCR-DAD.

Effectiveness of Robust Fine-Tuning Strategy. To evaluate the proposed Implicit Adapter mechanism and explicitly determine the optimal balance between pre-trained knowledge retention and domain adaptation, we conduct a comprehensive unified ablation study on the unfreezing strategies. Note that across all configurations, the final layer normalization and the subsequent visual projection layer (Base) are consistently unfrozen to project the pre-trained generalized features into the task-specific metric space. As shown in Tab.3, we progressively toggle the unfreezing switch for all internal projection layers across the backbone (Internal Proj.) and adjust the depth of unfrozen transformer blocks (N), leading to the following objective observations:

- **Limited Capacity of Base or Local Blocks.** Relying solely on Base (Row a) or unfreezing local blocks while keeping all internal projections frozen (Row b) yields highly suboptimal performance. This indicates that without appropriate feature-space adaptation, the network lacks sufficient representational capacity to bridge the massive cross-modal gap.
- **Effectiveness of Implicit Adapters and FFT.** Activating all internal projections across the backbone as pure implicit adapters (Row c) significantly boosts performance, confirming its effectiveness in cross-modal alignment. On the other extreme, Full Fine-Tuning (FFT, Row g) demonstrates highly competitive overall performance, particularly excelling in the All Search setting. However, FFT necessitates updating the entire massive parameter space of the backbone, which not only incurs prohibitive computational costs but also performs slightly worse in the complex Indoor Search scenario, hinting at a potential risk of overfitting to specific domain noises.

– **Parameter-Performance Trade-off.** Our hybrid strategies (Rows d and e) effectively balance capacity and efficiency. Variant (e) with $N = 2$ yields superior performance in the All Search setting, whereas Variant (d) with $N = 1$ achieves the global best in the Indoor Search setting. Although increasing the unfrozen depth ($N = 2$) or adopting FFT provides further notable gains in All Search retrieval, we ultimately select $N = 1$ (Variant d) as our default framework configuration. This decision is strictly driven by the principle of parameter efficiency: $N = 1$ secures highly competitive state-of-the-art accuracy while maintaining a significantly smaller footprint of learnable parameters. It gracefully avoids excessive reliance on large-scale parameter tuning and preserves the generalized priors of the foundation model. Notably, unfreezing further ($N = 3$, Row f) breaks this balance and leads to performance degradation.

Effectiveness of Decoupled Multi-Dimensional Semantic Prompting. Integrating explicit semantic prompting and auxiliary supervisions significantly enhances feature decoupling capabilities. Specifically, the combination of Modality, Camera, and Common prompts consistently outperforms using single prompts by absorbing distinct domain biases. **Extensive hyperparameter analyses—including prompt sequence lengths, presence of specific prompts, and weight sensitivities of auxiliary and instance losses—are comprehensively detailed in Appendix B.**

Effectiveness of CCSA. To verify the necessity of mapping visual instances to semantic anchors, we investigate the contribution of the V2T alignment loss ($\mathcal{L}_{v2t}$). Integrating CCSA explicitly guides heterogeneous visual manifolds and prevents the collapse of the pre-trained space, significantly boosting baseline performance. Empirically, the model exhibits an "inverted-U" sensitivity regarding the constraint weight, achieving a precise sweet spot at $\lambda_{v2t} = 0.01$. **The detailed fine-grained weight analysis curve and corresponding discussions are deferred to Appendix B.**

Dynamic Exploration of TOTO. Within the topology-aware optimal transport framework, the discretization of the soft probability matrix directly determines the model's optimization dynamics. Since practical cross-modal clustering inevitably encounters unbalanced cluster numbers, this subsection compares our local greedy hard assignment (Argmax) with two complex, intuitive strategies to explore their profound impacts under this unbalanced setting. The compared strategies are defined as follows: 1) Progressive Graph Matching (PGM) [24]: Emphasizes global constraints through iterative bipartite graph matching. It progressively reconstructs new graphs for the remaining unmatched clusters from the majority side, repeating this strict one-to-one matching across multiple rounds until all nodes find their correspondences. 2) Progressive Mutual-Nearest-Neighbor (PMNN): Focuses on strict noise isolation using a staged strategy. It sequentially extracts reliable pairs that are "mutually maximum" and "mutually Top-K with high relative reliability," and finally provides a fallback assignment for the remaining unmatched clusters from the majority side (handling the cluster number gap). 3) Argmax (Ours): A dense and unconditional hard assignment

that directly finds the minimum-cost target for each visual cluster. Fig.3 tracks the performance evolution trajectories of these strategies. The results reveal a counter-intuitive conclusion: PMNN and PGM both suffer from performance stagnation or degradation in the late training stage. This is primarily because PGM's strict one-to-one limitation and PMNN's multi-stage filtering mechanism essentially evade "Hard Positives," causing the network to suffer from cross-modal "information starvation." Conversely, the simplest Argmax strategy achieves the global optimum and maintains robust leadership throughout the entire cycle. This unconditional dense hard assignment provides drastic and sustained cross-domain pulling gradients for momentum updates. The resulting "Hard Regularization" effectively overcomes the model's optimization inertia toward hard samples, thereby thoroughly stimulating the network's potential to learn domain-invariant features.

Synergistic Effect of RCR-DAD. To verify the noise robustness of the RCR-DAD mechanism, we conduct decoupled analyses on two dimensions: confidence strategy evolution and dual-denoising synergy. (1) Confidence Strategy Evolution and Robustness: As shown in Fig.2, traditional absolute probability weighting suffers from a performance bottleneck due to excessive penalization of hard positives. Conversely, our Pure Relative Confidence Ratio (RCR) effectively balances noise suppression and valid gradient retention, achieving optimal performance at $\gamma = 2.0$ (mAP 52.04%, Rank-1 53.02%). Furthermore, the stable performance across a broad interval of $\gamma \in [1.0, 2.5]$ demonstrates strong hyperparameter robustness. (2) Synergistic Effect of Dual Denoising: As detailed in Tab.5, utilizing either Intra-Modal Denoising (IMD) or Cross-Modal Denoising (CMD) individually yields only marginal gains ($\sim$1%) over the baseline. However, their joint application produces a pronounced synergistic effect: IMD preserves the stability of the visual clustering manifold via temperature smoothing, while CMD prevents textual semantic pollution via dynamic soft-masking. This complementary mechanism successfully neutralizes the negative impacts of the modality gap, achieving the global best mAP of 52.04%.

5 Conclusion

In this paper, we present a novel and highly robust CLIP-based framework for Unsupervised Visible-Infrared Person Re-identification (USL-VI-ReID). By thoroughly analyzing the limitations of directly adapting large foundation models to highly noisy, cross-modal environments, we introduce an implicit adapter fine-tuning strategy and a decoupled multi-dimensional semantic prompting architecture. These designs effectively bridge the modality gap while rigorously preventing feature collapse and representation degradation. To conquer the inherent challenges of heterogeneous clustering, our proposed Cluster-Aware Cross-Modal Semantic Alignment (CCSA) leverages text priors as stable anchors to guide visual manifolds. More importantly, we challenge the conventional intuition of cross-modal matching by proposing Topology-Aware Optimal Transport (TOTO) and the Pure Relative Confidence Ratio (RCR). We explicitly demonstrate that replacing strict bipartite constraints (PGM) or confidence-based hi-

erarchical matching (PMNN) with unconditional dense Hard Regularization (via Argmax), coupled with relative confidence evaluation, are the keys to avoiding information starvation and rescuing hard positives from optimization inertia. Extensive evaluations on the SYSU-MM01 and RegDB benchmarks confirm that our framework sets a new state-of-the-art for unsupervised cross-modal person ReID. In the future, we plan to extend this robust semantic-topology alignment paradigm to other weakly-supervised multimodal matching tasks, such as video-based cross-modal ReID.

References

1. Chen, Z., et al.: Unveiling the power of clip in unsupervised visible-infrared person re-identification. In: ACM MM. pp. 3667–3675 (2023)
2. Cheng, D., He, L., Wang, N., Zhang, S., Wang, Z., Gao, X.: Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid. In: Proceedings of the 31st ACM international conference on multimedia. pp. 1325–1333 (2023)
3. Cheng, D., et al.: Hybrid dynamic contrast and probability distillation for unsupervised person re-id. IEEE TIP pp. 3334–3346 (2022)
4. Cheng, D., et al.: Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In: ACM MM. pp. 7085–7093 (2023)
5. Cho, Y., et al.: Part-based pseudo label refinement for unsupervised person re-identification. In: CVPR. pp. 7308–7318 (2022)
6. Dai, Z., et al.: Cluster contrast for unsupervised person re-identification. In: ACCV. pp. 1142–1160 (2022)
7. Fang, X., et al.: Visible-infrared person re-identification via semantic alignment and affinity inference. In: ICCV. pp. 11270–11279 (2023)
8. Kim, M., et al.: Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In: CVPR. pp. 18621–18632 (2023)
9. Li, S., et al.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: AAAI. pp. 1405–1413 (2023)
10. Li, Z., et al.: Inter-intra modality knowledge learning and clustering noise alleviation for unsupervised visible-infrared person re-identification. IEEE TKDE pp. 3934–3947 (2024)
11. Liang, W., et al.: Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. IEEE TIP pp. 6392–6407 (2021)
12. Lin, Y., et al.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI. pp. 8738–8745 (2019)
13. Liu, T., et al.: Unsupervised person re-identification with stochastic training strategy. IEEE TIP pp. 4240–4250 (2022)
14. Nguyen, X.B., et al.: Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition. In: CVPR. pp. 10847–10856 (2021)
15. Pang, Z., et al.: Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. IEEE TCSVT pp. 2706–2718 (2023)
16. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
17. Rao, Y., et al.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022)

18. Shi, J., et al.: Multi-memory matching for unsupervised visible-infrared person re-identification. In: ECCV. pp. 456–474 (2024)
19. Sun, R., et al.: Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network. IEEE TIFS pp. 2800–2813 (2024)
20. Teng, X., et al.: Relieving universal label noise for unsupervised visible-infrared person re-identification by inferring from neighbors. In: AAAI. pp. 7356–7364 (2025)
21. Wang, J., et al.: Optimal transport for label-efficient visible-infrared person re-identification. In: ECCV. pp. 93–109 (2022)
22. Wu, A., et al.: Rgb-infrared cross-modality person re-identification. In: ICCV. pp. 5380–5389 (2017)
23. Wu, L., et al.: Pseudo-pair based self-similarity learning for unsupervised person re-identification. IEEE TIP pp. 4803–4816 (2022)
24. Wu, Z., Ye, M.: Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In: CVPR. pp. 9548–9558 (2023)
25. Xi, R., et al.: Csanet: Cross-modality self-paced association network for unsupervised visible-infrared person re-identification. IEEE TIFS (2025)
26. Yang, B., et al.: Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In: ACM MM. pp. 2843–2851 (2022)
27. Yang, B., et al.: Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: ICCV. pp. 11069–11079 (2023)
28. Yang, M., et al.: Learning with twin noisy labels for visible-infrared person re-identification. In: CVPR. pp. 14308–14317 (2022)
29. Yang, Y., et al.: Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. IEEE TIFS pp. 1290–1304 (2025)
30. Yao, H., et al.: Unsupervised visible-infrared person re-identification under unpaired settings. In: ICCV. pp. 11916–11926 (2025)
31. Ye, M., et al.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: ECCV. pp. 229–247 (2020)
32. Ye, M., et al.: Channel augmented joint learning for visible-infrared recognition. In: ICCV. pp. 13567–13576 (2021)
33. Ye, M., et al.: Deep learning for person re-identification: A survey and outlook. IEEE TPAMI pp. 2872–2893 (2021)
34. Ye, M., et al.: Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. IEEE TPAMI (2025)
35. Zhang, G., et al.: Protohpe: Prototype-guided high-frequency patch enhancement for visible-infrared person re-identification. In: ACM MM. pp. 944–954 (2023)
36. Zhang, Y., et al.: Adaptive confidence-driven learning and cross-modal hard sample mining for unsupervised visible-infrared person re-identification. IPM p. 104346 (2026)
37. Zhang, Y., et al.: Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In: CVPR. pp. 2153–2162 (2023)
38. Zhou, K., et al.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
39. Zhou, K., et al.: Learning to prompt for vision-language models. IJCV pp. 2337–2348 (2022)