ALIGNING WITHOUT HOMOGENIZING: JOINTLY EXPLAINABLE COMMON-SPECIFIC DECOMPOSITION FOR MULTI-MODAL OBJECT RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal object Re-Identification aims to reduce cross-modal discrepancies while preserving identity-relevant modality-specific cues. Previous methods often align representations from heterogeneous modalities uniformly; however, treating common and modality-specific information alike may over-smooth useful modality-specific cues and cause information loss. To address this issue, we propose JESD-ReID, which first distinguishes these two types of evidence using a joint-explainability criterion and then coordinates each according to its role. Jointly Explainable Semantic Decomposition (JESD) uses shared semantic queries to form compact observations with shared query-slot indexing and reconstructs them jointly from a single latent code through constrained modality-conditioned dictionaries. The reconstructed evidence forms the common components, while the remaining token-level evidence is retained as modality-specific residuals. Low-volume regularization organizes the common span, whereas relational optimal transport coordinates local relational structure in the residuals without forcing feature values to coincide. Six lightweight modality–role experts use sample-dependent gates to control how strongly each branch contributes, while a jointly predicted shared grid controls where complementary evidence is sampled for aggregation. JESD-ReID achieves 82.4%, 88.0%, and 66.6% mAP on RGBNT201, RGBNT100, and MSVR310, respectively.

1 INTRODUCTION

Multi-modal object Re-Identification (ReID) retrieves the same object across non-overlapping cameras using heterogeneous observations from visible (RGB), near-infrared (NIR), and thermal infrared (TIR) sensors. These modalities provide complementary color, reflectance, and thermal cues, but also introduce substantial representation discrepancies. Existing studies address this challenge through feature fusion (Li et al., 2020; Zheng et al., 2021; 2023; Crawford et al., 2023), token interaction and selection (Wang et al., 2024; Zhang et al., 2024; Yu et al., 2025), prompt learning (Lin et al., 2025; Wang et al., 2025a), correlation-based and global-local alignment (Andrew et al., 2013; Liu et al., 2026), and adaptive aggregation (Ye et al., 2020; Wang et al., 2025b;c; Xu et al., 2026; Feng et al., 2025b). Despite considerable progress, many approaches align or aggregate heterogeneous representations without explicitly distinguishing the roles of shared and modality-dependent evidence.

Reducing modality discrepancy and preserving complementary information are compatible goals; the difficulty is deciding which differences should be removed and which should be retained. Cross-modal variation mixes sensing-induced nuisance with useful modality-dependent evidence. Applying one alignment rule to both can improve consistency while suppressing discriminative color, reflectance, or thermal cues, as illustrated in Fig. 1. For example, Signal (Liu et al., 2026) applies global volume contraction and local geometric alignment without explicitly distinguishing these representation roles. These operations can improve cross-modal coherence, but they do not first decide whether a discrepancy is nuisance or useful sensor-dependent evidence. Consequently, the same smoothing and alignment pressure can act on both representation roles. Shared-private representation learning (Bousmalis et al., 2016; Hazarika et al., 2020) and common-specific ReID methods (Choi et al., 2020; Zhang et al., 2022; Yu et al., 2023; Shi et al., 2024; Wang et al., 2025b;

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.

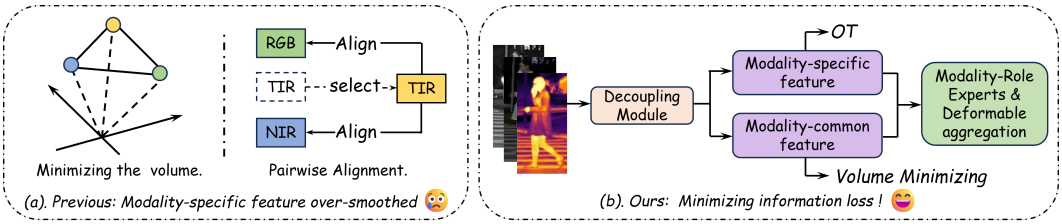

Figure 1: **Motivation of our method.** (a) Conventional global contraction and pairwise alignment operate without distinguishing representation roles and may over-smooth discriminative modality-specific cues. (b) We explicitly decouple modality-shared and modality-specific representations and adaptively aggregate their complementary information, reducing cross-modal discrepancies while retaining complementary modality-specific evidence.

Feng et al., 2025a; Qian et al., 2026) have explored the separation of shared and modality-dependent information, often through branch architectures and auxiliary losses. We instead define the shared component through joint reconstruction and use this criterion to guide the subsequent role-specific constraints. We therefore propose JESD-ReID under one principle: *decompose by joint explainability and coordinate by representation role*. Its core, Jointly Explainable Semantic Decomposition (JESD), starts from shared semantic queries that convert heterogeneous visual tokens into compact observations with shared query-slot indexing. One latent code, jointly inferred from these observations, reconstructs the jointly explained common component through restricted modality-conditioned dictionaries; the remaining token-level evidence becomes the modality-specific residual. This gives commonness a testable criterion rather than assigning it through a branch label.

The two components require different constraints. Low-volume regularization organizes the common span without enforcing pairwise equality, whereas relational residual transport (Cuturi, 2013; Peyré & Cuturi, 2019) coordinates local relational structure while allowing feature-level differences. Finally, role-conditioned aggregation uses six lightweight experts to determine how strongly common and specific evidence contributes, while a shared deformation grid determines where cross-modal evidence is sampled.

JESD-ReID achieves 82.4%, 88.0%, and 66.6% mAP on RGBNT201, RGBNT100, and MSVR310, respectively. The ablations examine the contribution of the decomposition, role-specific coordination, and adaptive aggregation.

Our contributions are summarized as follows:

- **Jointly Explainable Semantic Decomposition (JESD).** We address the alignment–information-loss tension by defining common information as evidence reconstructed from one latent code across shared semantic observations. Unreconstructed token-level evidence is retained as the modality-specific residual rather than being smoothed away.
- **Role-specific coordination.** We organize the jointly explained component with low-volume regularization and coordinate the residual with relational transport, so the two roles do not receive the same alignment objective.
- **A unified retrieval framework.** JESD-ReID integrates semantic decomposition, role-specific coordination, and adaptive aggregation to preserve and use complementary identity evidence. Modality–role experts with shared deformable aggregation connect the decomposed representations to the final retrieval descriptor. Experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate the effectiveness of the complete framework.

2 Related Work

2.1 Multi-modal object Re-Identification

Multi-modal object ReID exploits complementary RGB, NIR, and TIR observations under heterogeneous sensing conditions. Prior work reduces modality gaps through feature fusion (Li et al.,

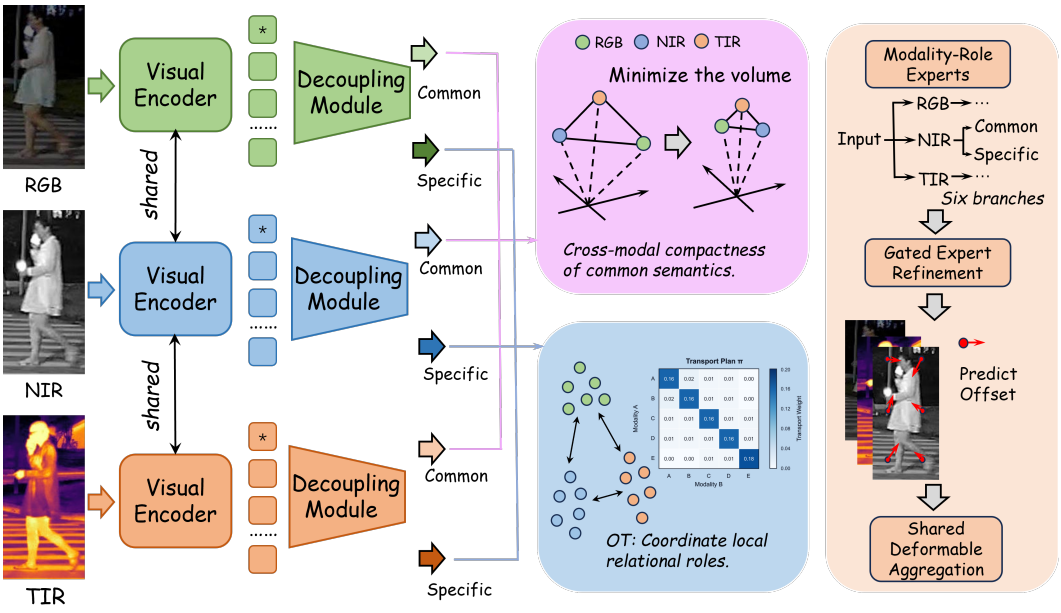

Figure 2: Overview of JESD-ReID. RGB, NIR, and TIR observations are decomposed into jointly explained components and modality-specific token residuals. Volume minimization encourages a low-volume joint organization of the common representations, while relational optimal transport coordinates local relational roles in the residuals. Gated modality–role experts control how strongly each branch contributes, and a jointly predicted shared grid controls where evidence is sampled for cross-modal aggregation.

2020; Zheng et al., 2021; 2023; Crawford et al., 2023), token interaction and selection (Wang et al., 2024; Zhang et al., 2024; Yu et al., 2025), prompt learning, correlation-based and global-local alignment (Andrew et al., 2013; Liu et al., 2026), and adaptive aggregation (Ye et al., 2020; Lin et al., 2025; Wang et al., 2025a;c;b; Xu et al., 2026; Feng et al., 2025b). Signal minimizes a global geometric volume and performs shift-aware local feature alignment without explicitly distinguishing representation roles. Its objectives therefore smooth and align cross-modal features without separate common-specific constraints; the same pressure can affect both jointly supported evidence and useful modality-dependent cues. JESD addresses this decision first. Its same-code criterion assigns evidence to a jointly explained component or a modality-specific residual; only the former receives low-volume organization, while the latter is coordinated through local relational roles rather than feature-wise matching. The contribution is consequently not volume minimization alone, but role-specific coordination enabled by an explicit decomposition criterion.

2.2 Common-Specific Representation Learning

Common-specific learning separates shared and domain-dependent factors, as in Domain Separation Networks (Bousmalis et al., 2016) and MISA (Hazarika et al., 2020). In cross-modal ReID, this idea appears in modality compensation, unified-specific learning, common-divergent learning, and modality decoupling (Choi et al., 2020; Zhang et al., 2022; Yu et al., 2023; Shi et al., 2024; Feng et al., 2025a; Qian et al., 2026). Most methods separate branches architecturally or through auxiliary losses. JESD instead adopts a same-code explanatory criterion: common evidence is the component reconstructed when one latent code, jointly inferred from all modalities, explains the three modality-conditioned semantic observations under a restricted shared-plus-low-rank dictionary family. The remaining token-level evidence is the modality-specific residual. This constrains the decomposition criterion itself, rather than defining sharedness primarily through branch architecture or auxiliary objectives.

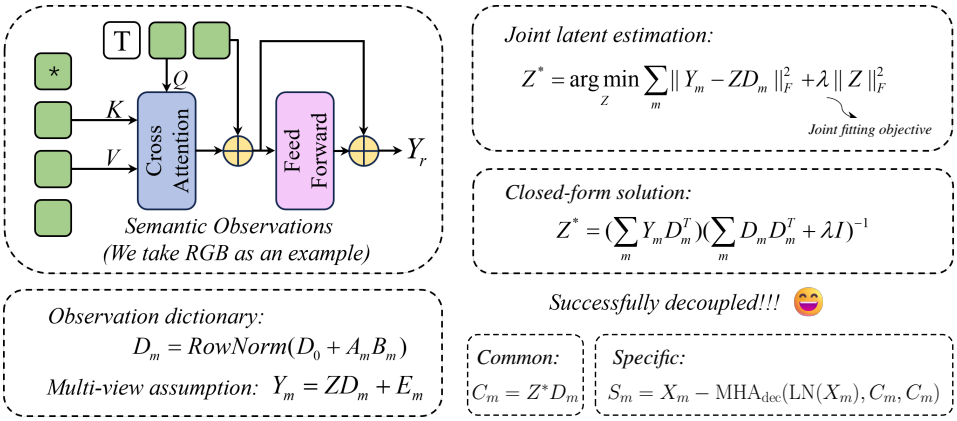

Figure 3: Illustration of Jointly Explainable Semantic Decomposition (JESD). Shared semantic queries transform heterogeneous visual tokens into compact semantic observations. Row-normalized modality-conditioned dictionaries support the estimation of one latent code from all modalities. The code generates jointly explained semantic components, which are decoded into token space to obtain modality-specific residuals.

3 METHOD

3.1 FRAMEWORK OVERVIEW

As illustrated in Fig. 2, JESD-ReID has three steps. Shared semantic queries first map RGB, NIR, and TIR tokens to observations with shared query-slot indexing. JESD then separates what one latent code can explain jointly from the remaining modality-specific token evidence. We organize the two outputs by role: low-volume regularization acts on the jointly explained component, while relational optimal transport coordinates local residual relations. Six gated modality–role experts control branch contributions; the resulting modality streams jointly predict a shared sampling grid before cross-modal interaction. The role-specific regularizers are used only during training.

3.2 JOINTLY EXPLAINABLE SEMANTIC DECOMPOSITION

JESD supplies an operational criterion for common-specific representation learning. It does not designate a branch as "common" by architecture alone: common evidence is the part of the semantic observations jointly reconstructed from one shared code under constrained modality dictionaries. Evidence outside this joint reconstruction is retained as a token-level residual. The distinction is therefore made before the two outputs receive different role-specific constraints.

Semantic observation model. Given relevance-masked visual tokens $X_m \in \mathbb{R}^{N\times d}$ that retain the spatial grid of modality $m \in \{R, N, T\}$, shared queries $Q \in \mathbb{R}^{K\times d}$ construct compact observations:

$$Y_m = \mathrm{Attn}(Q + e_m + \phi(t_m), X_m, X_m), \qquad Y_m \in \mathbb{R}^{K\times d}, \tag{1}$$

where $t_m \in \mathbb{R}^d$ is the modality's global text feature and ϕ is a learned projection. Here, Attn denotes a cross-attention block with residual and feed-forward updates. The shared queries provide shared slot indexing across modalities. The learnable $e_m \in \mathbb{R}^{1\times d}$ is a modality-specific offset broadcast to every query slot, allowing sensor-aware adaptation without changing this shared indexing.

One shared code can produce different modality features through different dictionaries. We therefore use a shared basis with a constrained modality correction:

$$D_m = \mathrm{RowNorm}(D_0 + A_m B_m)\,, \tag{2}$$

where $D_0 \in \mathbb{R}^{p\times d}$ is a globally shared basis and $A_m B_m$ has rank at most r. D_0 anchors a common latent coordinate system; the low-rank term models systematic differences in how sensors express the same latent factors. It is therefore neither an independent per-modality dictionary nor an unrestricted capacity increase. The rank constraint applies before row normalization, which removes

the trivial row-wise scale ambiguity between the code and dictionary rows. This family preserves shared structure across observation maps without requiring them to be identical. In particular, an unconstrained D_m would allow each modality to learn an unrelated observation map, whereas setting $A_m B_m = 0$ would impose the same map on all modalities. Appendix A details this constrained parameterization.

Joint latent estimation. For each multi-modal sample group, a shared latent code $Z \in \mathbb{R}^{K\times p}$ is estimated by

$$Z^* = \arg\min_{Z} \sum_{m\in\{R,N,T\}} \|Y_m - ZD_m\|_F^2 + \lambda \|Z\|_F^2 , \tag{3}$$

which admits the closed-form solution

$$Z^* = \left(\sum_m Y_m D_m^\top\right)\left(\sum_m D_m D_m^\top + \lambda I\right)^{-1}. \tag{4}$$

Here, Z^* is jointly estimated from the multi-modal observations, while D_m maps shared latent coordinates to modality m. Requiring the same code couples the fit across the three observation maps. The dictionaries are learned across the training set, whereas Z^* is inferred for each multi-modal group. With $\lambda > 0$, the ridge estimator is well posed. This is an operational definition of jointly explained evidence; it does not assert identifiability of physical factors.

Common-residual decomposition. The jointly explained semantic component is first obtained in the compact observation space:

$$C_m = Z^* D_m, \qquad C_m \in \mathbb{R}^{K\times d}. \tag{5}$$

We then decode it into the original token space and define the modality-specific residual as

$$\widehat{X}_m^c = \mathrm{MHA}_{\mathrm{dec}}(\mathrm{LN}(X_m), C_m, C_m), \qquad S_m = X_m - \widehat{X}_m^c. \tag{6}$$

The decoder uses multi-head attention without residual or feed-forward updates. S_m retains token-level evidence not reconstructed from the jointly estimated code. Thus, "specific" is defined relative to this explanatory model, not as a claim that every residual dimension is identity-discriminative. Residuals may contain useful modality-dependent evidence and nuisance variation; retrieval supervision, role-specific coordination, and gated experts determine their contribution. The differentiable ridge solve infers the code in each forward pass; no additional reconstruction loss is added to the training objective.

3.3 Role-Specific Representation Coordination

The two JESD outputs need different constraints. The common component should have a compact joint span, whereas forcing residual values to match could erase useful sensing cues. We therefore regularize local residual relations to discourage unconstrained drift while retaining modality-dependent feature differences.

Low-volume common organization. We pool the common components into normalized descriptors $c_m = \mathrm{Norm}(\mathrm{Pool}(C_m))$, with layer and ℓ_2 normalization, and stack them as $A = [c_R^\top; c_N^\top; c_T^\top]$. Their regularized joint-span volume is

$$\mathcal{L}_{\mathrm{com}} = \sqrt{\det\left(AA^\top + \epsilon I\right)}. \tag{7}$$

This expression measures the regularized joint-span volume of the three normalized descriptors. Minimizing it encourages a low-volume joint configuration without directly minimizing pairwise distances or forcing the descriptors to coincide. Combined with the retrieval objectives, this regularizer promotes cross-modal coordination without requiring identical modality descriptors.

Relational residual transport. Directly matching residual feature values may suppress useful sensor-dependent variation. We instead characterize each residual token i by its sorted K_g nearest-neighbor distances:

$$g_{m,i} = \frac{d_{m,i}}{\max(\frac{1}{K_g}\mathbf{1}^\top d_{m,i}, \epsilon)}, \qquad d_{m,i} = \mathrm{Sort}_{K_g}\left\{\|\bar{s}_{m,i} - \bar{s}_{m,j}\|_2\right\}_{j\neq i}, \tag{8}$$

where $\bar{s}_{m,i} = S_{m,i}/\max(\|S_{m,i}\|_2, \epsilon)$ denotes a normalized residual token. The descriptor $g_{m,i}$ characterizes its local relational role without relying on spatial or index-wise correspondence. Dividing by the local mean reduces sensitivity to uniform scaling of within-modality distances.

For a modality pair (m, n), we define the relational cost as

$$C_{ij}^{m,n} = \|g_{m,i} - g_{n,j}\|_2^2. \tag{9}$$

This cost induces the entropy-regularized transport plan

$$T_{m,n}^* = \arg\min_{T\in\Pi(a,b)} \langle T, C^{m,n}\rangle + \tau \sum_{i,j} T_{ij}\left(\log T_{ij} - 1\right), \tag{10}$$

where $\Pi(a, b)$ denotes the set of transport plans with marginals a and b; we use uniform token marginals, so neither modality is treated as a reference. With a Sinkhorn estimate $\widehat{T}$ and stop-gradient operator sg, the implemented residual loss is

$$\mathcal{L}_{\mathrm{spec}} = \frac{1}{|\mathcal{P}|}\sum_{(m,n)\in\mathcal{P}} \left\langle \mathrm{sg}(\widehat{T}_{m,n}), C^{m,n}\right\rangle, \qquad \mathcal{P} = \{(R,N),(R,T),(N,T)\}. \tag{11}$$

This objective constrains local relational profiles, discouraging unconstrained drift while allowing modality-dependent feature values. It requires neither one-to-one token correspondence nor exact spatial alignment, and does not guarantee preservation of the complete residual geometry.

3.4 Modality–Role Experts and Deformable Aggregation

Six modality–role experts take decoded common maps $U_{m,c} = \widehat{X}_m^c$ and residuals $U_{m,s} = S_m$ from the three modalities as inputs:

$$\mathcal{U} = \{\widehat{X}_R^c, \widehat{X}_N^c, \widehat{X}_T^c, S_R, S_N, S_T\}. \tag{12}$$

Each expert has a separate token-wise bottleneck adapter and a sample-wise gate:

$$\begin{aligned} h_{m,p}(U) &= \mathrm{Up}_{m,p}(\mathrm{GELU}(\mathrm{Down}_{m,p}(\mathrm{LN}(U)))), \\ \delta_{m,p} = g_{m,p}\, h_{m,p}(U_{m,p}), &\qquad g_{m,p} = \sigma(b_{m,p} + r_{m,p}(\mathrm{Pool}(\mathrm{LN}(U_{m,p})))). \end{aligned} \tag{13}$$

Here $p \in \{c, s\}$ indexes the role; each bottleneck adapter uses $d \to 64 \to d$ Down/Up maps. Mean pooling and a scalar predictor $r_{m,p}$ set the gate. The six corrections refine three modality streams with residual scale η:

$$F_m = (1-\zeta)\left[X_m + \eta(\delta_{m,c} + \delta_{m,s})\right] + \zeta\widehat{X}_m^c, \tag{14}$$

where ζ optionally moves the refined features toward the decoded common map. These streams jointly predict a shared sampling offset field:

$$\Delta P = \Gamma \odot \tanh(f_{\mathrm{off}}([F_R; F_N; F_T])), \qquad \widetilde{F}_m = \mathrm{GridSample}(F_m, P + \Delta P), \tag{15}$$

where P is the reference grid and Γ bounds displacement. The shared grid samples all streams into one key–value bank. Modality-specific query and output projections update each stream. Three global descriptors then query the refined tokens via cross-attention and a feed-forward block. Their concatenation receives gated global and common residuals (Appendix A). Gates control contribution strength; offsets control readout locations.

3.5 Optimization Objective

The core training objective is

$$\mathcal{L} = \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{tri}}\mathcal{L}_{\mathrm{tri}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{id}}^{\mathrm{com}} + \alpha\mathcal{L}_{\mathrm{com}} + \beta\mathcal{L}_{\mathrm{spec}}, \tag{16}$$

Here $\mathcal{L}_{\mathrm{id}}$ and $\mathcal{L}_{\mathrm{id}}^{\mathrm{com}}$ are label-smoothed identity losses for the fused and common descriptors; $\mathcal{L}_{\mathrm{tri}}$ is the triplet loss (Hermans et al., 2017). The common head pools C_m and predicts training identities (Appendix A.3). We use $\lambda_{\mathrm{id}} = 0.25$ and $\lambda_{\mathrm{tri}} = 1.0$; auxiliary weights are in Appendix A. $\mathcal{L}_{\mathrm{com}}$ regularizes the common span and $\mathcal{L}_{\mathrm{spec}}$ local residual relations; both regularizers are used only during training.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION PROTOCOLS

Datasets and metrics. We evaluate JESD-ReID on RGBNT201 (Zheng et al., 2021), RGBNT100 (Li et al., 2020), and MSVR310 (Zheng et al., 2023). RGBNT201 is a person ReID dataset containing 4,787 RGB–NIR–TIR triplets, while RGBNT100 and MSVR310 are vehicle ReID datasets containing 17,250 and 2,087 triplets, respectively. We follow the official splits and report mAP and CMC scores at Rank-1, Rank-5, and Rank-10 (Zheng et al., 2015). Baseline results are taken from the cited papers; our evaluations omit reranking.

4.2 IMPLEMENTATION DETAILS

We implement JESD-ReID in PyTorch (Paszke et al., 2019) with a CLIP-pretrained ViT-B/16 backbone (Radford et al., 2021; Dosovitskiy et al., 2021). Training and inference use per-image captions and camera metadata. Training uses standard flipping, cropping, and random erasing (Zhong et al., 2020). We optimize with Adam (Kingma & Ba, 2015), label-smoothed identity loss (Szegedy et al., 2016), and triplet loss (Hermans et al., 2017). Evaluation uses features extracted directly from the model, without reranking. All ablation variants share the same backbone, CLIP initialization, data augmentation, optimizer, and non-reranked inference protocol; they differ only in the component under study. Dataset-specific input sizes, learning rates, schedules, batch settings, JESD/OT hyperparameters, test views, and the complexity analysis are given in Appendix A and Appendix B.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

Person ReID. As shown in Table 1, JESD-ReID achieves 82.4% mAP and 87.2% Rank-1 on RGBNT201 under the direct-distance, no-reranking protocol. The gains also persist at Rank-5 and Rank-10, indicating improved retrieval quality beyond the top-ranked match.

Table 1: Comparison on RGBNT201. Bold: best; underlined: second best.

Method	mAP	R-1	R-5	R-10
TOP-ReID (Wang et al., 2024)	72.3	76.6	84.7	89.4
EDITOR (Zhang et al., 2024)	66.5	68.3	81.1	88.2
MambaPro (Wang et al., 2025a)	78.9	83.4	89.8	91.9
DeMo (Wang et al., 2025b)	79.0	82.3	88.8	92.0
IDEA (Wang et al., 2025c)	80.2	82.1	90.0	93.3
STMI (Xu et al., 2026)	81.2	83.4	90.2	91.6
DMPT (Lin et al., 2025)	78.5	81.3	<u>90.4</u>	<u>93.5</u>
MDReID (Feng et al., 2025a)	<u>82.1</u>	<u>85.2</u>	90.3	92.6
JESD-ReID (Ours)	**82.4**	**87.2**	**91.1**	**93.9**

Vehicle ReID. Table 2 shows 88.0% mAP and 97.7% Rank-1 on RGBNT100, the highest Rank-1 among the listed methods. On MSVR310, the reported 66.6% mAP exceeds the strongest listed baseline by 1.8 points.

Table 2: Comparison on RGBNT100 and MSVR310. Best and second-best results are bold and underlined, respectively.

Method	RGBNT100		MSVR310	
	mAP	R-1	mAP	R-1
TOP-ReID (Wang et al., 2024)	81.2	96.4	35.9	44.6
EDITOR (Zhang et al., 2024)	82.1	96.4	39.0	49.3
MambaPro (Wang et al., 2025a)	83.9	94.7	47.0	56.5
DeMo (Wang et al., 2025b)	86.2	<u>97.6</u>	49.2	59.8
IDEA (Wang et al., 2025c)	87.2	96.5	47.0	62.4
STMI (Xu et al., 2026)	**89.1**	97.1	<u>64.8</u>	<u>76.1</u>
DMPT (Lin et al., 2025)	81.7	94.1	36.6	52.1
MDReID (Feng et al., 2025a)	85.3	95.6	51.0	68.9
JESD-ReID (Ours)	<u>88.0</u>	**97.7**	**66.6**	**76.5**

4.4 Ablation Studies

Unless otherwise specified, all ablations are conducted on RGBNT201.

Main components. We progressively add JESD, the two role-specific regularizers, and gated expert aggregation to a common baseline. All rows use the same backbone and training protocol.

Table 3: Main-component ablation on RGBNT201. Vol.: low-volume common regularization; Rel.: relational residual transport; Experts: modality–role experts with deformable cross-modal aggregation.

Index	JESD	Vol.	Rel.	Experts	mAP	R-1
A	×	×	×	×	77.4	81.2
B	✓	×	×	×	79.0	83.0
C	✓	✓	×	×	80.4	84.1
D	✓	✓	✓	×	81.3	86.6
E	✓	✓	✓	✓	**82.4**	**87.2**

JESD defines common and residual branches through joint explanation; the two regularizers and expert aggregation then yield monotonic gains. The successive mAP gains are 1.6, 1.4, 0.9, and 1.1 points, and Rank-1 also increases at every step, supporting the complementary roles of decomposition, coordination, and expert aggregation.

Property assignment. We vary the regularizers assigned to the common components and modality-specific residuals. Volume regularizes the joint span, whereas relational transport coordinates local distance-profile distributions.

Table 4: Property-assignment ablation on RGBNT201 with expert aggregation fixed. Vol. and Rel. denote low-volume joint-span regularization and relational transport, respectively.

Index	Common	Residual	mAP	R-1
A	Vol.	Vol.	80.8	84.4
B	Rel.	Rel.	80.1	83.7
C	Rel.	Vol.	80.4	84.9
D	Vol.	Rel.	**82.4**	**87.2**

Changing either regularizer lowers performance, supporting the role-specific assignment rather than a generic combination of losses. Uniform volume and relational assignments reduce mAP by 1.6 and 2.3 points, respectively, while swapping their roles costs 2.0 points. These results support matching each constraint to its intended representation.

Cross-modal correspondence. This diagnostic tests sensitivity to input identity correspondence with a fixed model. We replace NIR and/or TIR observations with other identities before estimating Z^*. Table 5 compares correct and disrupted input groups using pre-aggregation common slots; the correct-triplet row is the reference for this intervention.

Table 5: Cross-modal correspondence diagnostic for JESD on RGBNT201. Retrieval uses pre-aggregation common semantic slots; "Change" is $1 - \cos$ between descriptors from correct and disrupted groups.

Grouping	mAP	R-1	Recon. error	Change
Correct triplet	**58.80**	**65.43**	0.12365	–
Shuffled NIR	50.10	58.01	0.12483	0.211
Shuffled TIR	34.36	46.53	0.12458	0.315
Shuffled NIR+TIR	22.30	29.90	0.12531	0.564

Shuffling both modalities reduces common-only mAP from 58.80% to 22.30%, while reconstruction error increases only from 0.12365 to 0.12531. This supports the importance of identity-consistent cross-modal observations for JESD's common representation: low reconstruction error alone does not ensure that the jointly reconstructed evidence is identity-discriminative.

4.5 Visualization and Analysis

Figure 4 shows that identity clusters become progressively more separated after decomposition, role-specific coordination, and expert aggregation. Figure 5 complements this embedding view with retrieval examples, where JESD-ReID returns more cross-modal positives among the top ranks. Additional similarity diagnostics and Top-10 cases are provided in Appendix B.

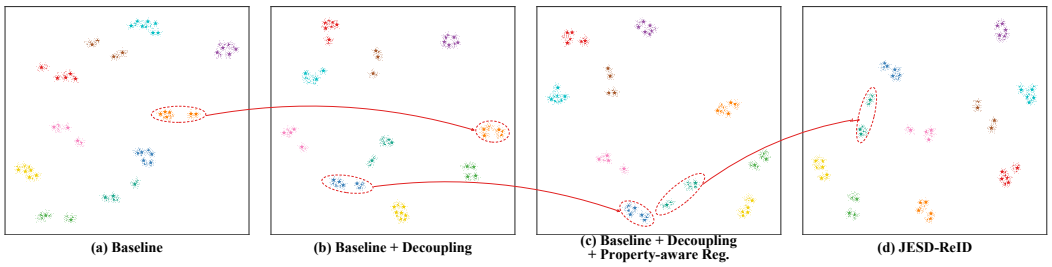

Figure 4: t-SNE visualization on RGBNT201. Colors denote identities; circles and arrows identify selected clusters in adjacent panels.

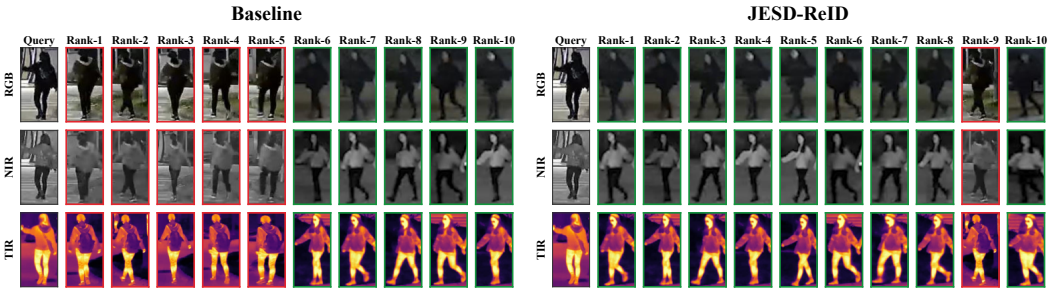

Figure 5: Retrieval comparison on RGBNT201. RGB, NIR, and TIR results are arranged in three rows; green and red borders denote correct and incorrect matches.

5 Conclusion

We presented JESD-ReID for Multi-modal object Re-Identification, motivated by the observation that uniformly aligning heterogeneous features without distinguishing their roles can over-smooth useful modality-specific evidence. JESD defines jointly explained components through a shared latent code and constrained modality-conditioned dictionaries, while preserving unreconstructed token-level evidence as modality-specific residuals. The common components and residuals are then coordinated according to their distinct properties: low-volume regularization organizes the common semantic span, and relational transport coordinates local relations in the residuals without enforcing feature-value equality. Finally, role-conditioned experts control how strongly each branch contributes, while a shared deformation grid controls where complementary evidence is sampled. Experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate the effectiveness of the complete decomposition, coordination, and aggregation framework.

AI USE STATEMENT

Generative AI tools assisted with language editing and translation, and with literature search and related-work discovery. The authors are responsible for verifying the final text and references.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 1247–1255, 2013.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10257–10266, 2020.

Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. UniCat: Crafting a stronger fusion baseline for multimodal re-identification. *arXiv preprint arXiv:2310.18812*, 2023.

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

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

Yingying Feng, Jie Li, Jie Hu, Yukang Zhang, Lei Tan, and Jiayi Ji. MDReID: Modality-decoupled learning for any-to-any multi-modal object re-identification. In *Advances in Neural Information Processing Systems*, 2025a.

Yingying Feng, Jie Li, Chi Xie, Lei Tan, and Jiayi Ji. Multi-modal object re-identification via sparse mixture-of-experts. In *Proceedings of the International Conference on Machine Learning*, pp. 16796–16807, 2025b.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

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

Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11345–11353, 2020.

Minghui Lin, Shu Wang, Xiang Wang, Jianhua Tang, Longbin Fu, Zhengrong Zuo, and Nong Sang. DMPT: Decoupled modality-aware prompt tuning for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2103–2112, 2025.

Yangyang Liu, Yuhao Wang, and Pingping Zhang. Signal: Selective interaction and global-local alignment for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7359–7367, 2026. doi: 10.1609/aaai.v40i9.37674.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026.

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

Jiangming Shi, Xiangbo Yin, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 99715–99734, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. TOP-ReID: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5758–5766, 2024.

Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. MambaPro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8150–8158, 2025a.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. DeMo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8141–8149, 2025b.

Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. IDEA: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29701–29710, 2025c.

Xingguo Xu, Zhanyu Liu, Weixiang Zhou, Yuansheng Gao, Junjie Cao, Yuhao Wang, Jixiang Luo, and Dell Zhang. STMI: Segmentation-guided token modulation with cross-modal hypergraph interaction for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11433–11441, 2026. doi: 10.1609/aaai.v40i14.38125.

Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pp. 229–247, 2020.

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*, pp. 11185–11195, 2023.

Zhi Yu, Zhiyong Huang, Mingyang Hou, Jiaming Pei, Yan Yan, Yushi Liu, and Daming Sun. Representation selective coupling via token sparsification for multi-spectral object re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3633–3648, 2025.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17117–17126, 2024.

Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. FMCNet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7349–7358, 2022.

Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3529–3537, 2021.

Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100:101901, 2023.

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

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020.

APPENDIX

This appendix clarifies the operational interpretation of JESD, describes its implementation and reproducibility details, and provides qualitative and efficiency evidence that complements the main experiments.

A ADDITIONAL METHOD DETAILS

A.1 JOINT ESTIMATION AS AN EXPLANATORY MODEL

For each modality, shared semantic queries map visual tokens to K compact observations $Y_m \in \mathbb{R}^{K\times d}$. Corresponding query slots give the observations shared indexing, while e_m and the projected global text feature $\phi(t_m)$ provide modality-aware conditioning. JESD then uses

$$Y_m = ZD_m + E_m, \qquad D_m = \mathrm{RowNorm}(D_0 + A_m B_m), \tag{17}$$

where Z is sample-specific, D_0 is a dataset-level shared basis, and E_m is unexplained evidence. Z fits the observations jointly; D_m describes their modality-dependent expression. The shared D_0 anchors latent coordinates, while $A_m B_m$ has rank at most r before row normalization. Row normalization removes a trivial scale trade-off between the code and dictionary rows.

For every aligned group, one code is inferred by ridge regression:

$$Z^* = \arg\min_{Z} \sum_{m\in\{R,N,T\}} \|Y_m - ZD_m\|_F^2 + \lambda\|Z\|_F^2 = \left(\sum_m Y_m D_m^\top\right)\left(\sum_m D_m D_m^\top + \lambda I\right)^{-1}. \tag{18}$$

The ridge term stabilizes correlated dictionaries, and the differentiable estimate permits retrieval supervision of trainable components. All modality-conditioned reconstructions therefore depend on a jointly estimated code, while their expressions can differ across sensors.

The explained slots $C_m = Z^* D_m$ are decoded into token space, giving $S_m = X_m - \widehat{X}_m^c$. This defines commonness under the learned model, without claiming unique physical factors. Residuals can contain both identity cues and nuisance variation; retrieval supervision and role-specific processing determine their contribution. This reconstruction residual is not an orthogonal projection, and common and specific representations may remain statistically dependent.

A.2 WHAT ROLE-SPECIFIC COORDINATION CONSTRAINS

For pooled and normalized common descriptors c_m, JESD minimizes

$$\mathcal{L}_{\mathrm{com}} = \sqrt{\det(AA^\top + \epsilon I)}, \qquad A = [c_R^\top; c_N^\top; c_T^\top]. \tag{19}$$

This term contracts the *joint span* without imposing pairwise equality or guaranteeing small pairwise distances. Normalization controls descriptor scale. Retrieval losses supply discrimination; the volume objective alone does not prevent collapsed common representations.

For the residual branch, each token is represented by sorted distances to its K_g nearest residual neighbors, normalized by their mean distance. Optimal transport compares these local relation profiles across modalities under uniform marginals, so no modality is chosen as a reference and no one-to-one token match is presumed. It therefore aligns distributions of local roles described by relative neighbor distances, rather than raw values, token indices, or complete global geometry. This softer constraint leaves room for modality-dependent appearance while discouraging incompatible local organization. In implementation, 20 log-domain Sinkhorn iterations approximate the plan; its gradient is stopped when computing the transport-cost loss. Gradients still propagate through the relational costs.

A.3 EXPERT IMPLEMENTATION AND AUXILIARY SUPERVISION

Token experts. In Eq. 13, each modality–role pair has its own bottleneck adapter $h_{m,p}$ and gate predictor. The Down/Up maps in Eq. 13 are affine layers with bottleneck width 64 and output width

$d = 512$; each adapter shares its weights across spatial positions. It learns a correction in the original feature space from its assigned common or residual input; its output has the same $N \times d$ shape as X_m. Pool averages tokens, $r_{m,p}$ is an affine scalar predictor, $b_{m,p}$ is a learned scalar bias, and σ is sigmoid. Each sample thus has six separately parameterized scalar gates, broadcast over tokens and channels; the gates need not sum to one. The adapter's final linear layer is initialized to zero, so the expert correction initially vanishes.

Purpose of the residual update. Let $\bar{X}_m = X_m + \eta(\delta_{m,c} + \delta_{m,s})$, with η the configured residual scale. Equation 14 is equivalently $F_m = \bar{X}_m + \zeta(\hat{X}^c_m - \bar{X}_m)$. The two role-specific experts refine a direct path from the original modality features, while the optional $\zeta \in [0, 1]$ mixture moves the result toward its decoded common map before shared sampling. Setting $\zeta = 0$ leaves the expert refinement intact. The refinement affects both the predicted offsets and the features read at those locations; it does not impose equality between modalities.

Common identity supervision. The auxiliary classifier uses the explained slots $C_m = Z^* D_m$, before decoding and expert aggregation. For one aligned group, its descriptor is

$$a_m = \mathrm{LN}_o\left(K^{-1}\sum_{k=1}^{K} C_{m,k}\right), \qquad u_c = \mathrm{LN}_o((a_R + a_N + a_T)/3)\,. \tag{20}$$

Both occurrences of LN_o use the same output LayerNorm. Unlike the volume descriptors c_m, a_m is not ℓ_2-normalized before modality averaging. A common-branch BatchNorm and a bias-free linear classifier produce probabilities over the M training identities:

$$p_i = \mathrm{softmax}(W^{\mathrm{com}}_{\mathrm{id}}\,\mathrm{BN}_c(u_{c,i}))\,, \qquad q_{ij} = 0.9\,\mathbf{1}[j = y_i] + 0.1/M. \tag{21}$$

Using the same identity label y_i as the main retrieval task, the loss is

$$\mathcal{L}^{\mathrm{com}}_{\mathrm{id}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} q_{ij}\log p_{ij}, \tag{22}$$

where n is the batch size. This label-smoothed cross-entropy directly encourages the jointly explained representation to retain identity information; it uses no additional annotations. Equation 16 weights it once by $\lambda_{\mathrm{aux}} = 0.25/0.10/0.25$ on RGBNT201/RGBNT100/MSVR310. The classifier and loss are used only in training; the common descriptors also feed the separate retrieval residual paths described in Appendix A.

Distinction from the correspondence diagnostic. Table 5 uses $f_{\mathrm{diag}} = \mathrm{Norm}_2([a_R; a_N; a_T])$: the three pooled common descriptors are concatenated and ℓ_2-normalized for distance-based retrieval. This diagnostic uses neither modality averaging, BatchNorm, nor a classifier, and precedes expert aggregation. At a fixed checkpoint, disrupted groups replace selected modalities and their associated text with observations from other identities before re-estimating JESD common slots. RGB observations and retrieval labels remain fixed.

A.4 Implementation and Inference

We use a CLIP-pretrained ViT-B/16 backbone, Adam, label-smoothed identity classification, and a margin-free triplet loss. The identity and triplet terms have weights 0.25 and 1.0, respectively, and Adam uses weight decay 10^{-4}. Image normalization uses $(0.5, 0.5, 0.5)$ for both the mean and standard deviation. Training augmentation consists of horizontal flipping with probability 0.5, padding by 10 pixels followed by random cropping, and random erasing with probability 0.5. We use identity-balanced sampling. Image sizes below are reported as height $\times$ width. Training and inference both use per-image modality-specific captions and camera metadata, following the text-augmented setting of IDEA (Wang et al., 2025c).

At test time, descriptors are ℓ_2-normalized and query–gallery rankings are computed by Euclidean distance; test-view combinations are stated with the relevant experiments. We do not use reranking, query expansion, gallery-neighbor refinement, pseudo-labeling, or test-time parameter updates. The two role-specific regularizers and auxiliary identity classifiers contribute only to training. The common-descriptor identity coefficient λ_{aux} is 0.25 on RGBNT201 and MSVR310 and 0.10 on RGBNT100.

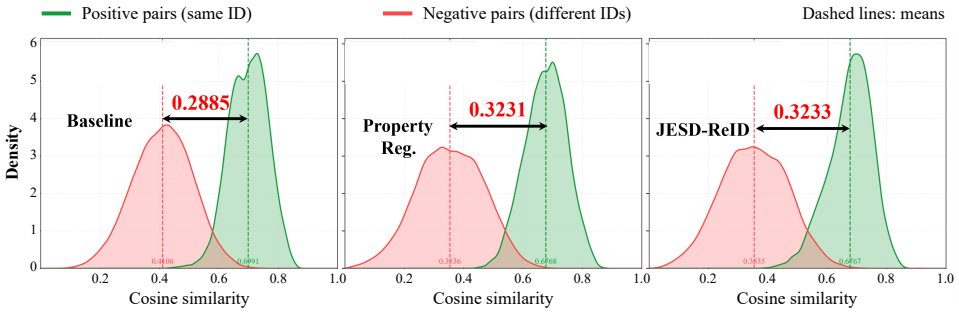

Figure 6: Positive and negative cosine-similarity distributions on RGBNT201. Their mean separation increases from 0.2885 to 0.3233.

RGBNT201. Inputs are resized to 256×128; each batch contains 72 aligned groups with eight instances per identity. The released five-epoch adaptation updates only JESD at a nominal learning rate of 10^{-5}, with the remaining parameters and BatchNorm statistics frozen. We use $K = 4$ shared slots, adapter rank $r = 2$, latent dimension 64, and ridge coefficient $\lambda = 10^{-2}$. Both the common volume term and relational OT have weight 5×10^{-3}. OT uses entropy coefficient 5×10^{-2}, 32-neighbor relation profiles, at most 80 residual tokens, and 20 Sinkhorn iterations. The OT cap uniformly subsamples residual positions. Separately, the spatial selector unions intra- and cross-modal relevance masks with an intra-modal budget of 80, preserving the full grid rather than shortening it to 80 tokens.

RGBNT100. Inputs are resized to 128×256; each batch contains 128 aligned groups with 16 instances per identity. Global-feature training is followed by interaction-enabled adaptation, with auxiliary global ID/triplet supervision weighted by 0.25. A final two-epoch compact-fusion adaptation at 10^{-5} updates only the interaction block and visual gate, with other parameters and BatchNorm statistics frozen. Its one-epoch warmup uses initial and minimum learning-rate factors of 0.5 and 0.2. Each modality supplies four common slots and eight refined spatial tokens, selected at positions with the largest absolute residual–global cosine similarities. The reported result averages normalized pre-BN descriptors over scales 0.95, 1.0, and 1.05, each with its horizontal flip, followed by ℓ_2 normalization.

MSVR310 training configuration. For the documented run, we initialize from CLIP and train for 50 epochs with seed 1111, frozen text encoder and updated BN statistics. Inputs are 128×256; batches contain 64 groups, eight per ID. Base/common rates are $3.5 \times 10^{-4}/3 \times 10^{-4}$, with 10-epoch warmup and cosine decay. JESD uses four slots, latent 64, rank 32, ridge 0.01 and volume/OT weights 0.005. We use token-wise softmax and average original/flip pre-BN features without reranking. The supporting package supplies the full configuration and evaluation routine.

Descriptor paths. RGBNT201/RGBNT100 gate the interaction output against a global-query path using zero visual tokens; MSVR310 uses the interaction output directly. Two gated common-descriptor projections, before and after BatchNorm, provide retrieval residuals. Post-BN test gates are $0.10/0.25/0.10$ for RGBNT201/RGBNT100/MSVR310. At test time, the pre-sampling gate ζ is 0.01 for RGBNT201/RGBNT100 and disabled for MSVR310.

B Additional Qualitative and Efficiency Results

B.1 Representation Visualization

Figure 6 shows cosine similarities for negative (red) and positive (green) pairs; dashed lines mark their means. Their increasing separation from the baseline to the complete model is a descriptive diagnostic of identity discrimination, rather than a training objective.

B.2 Qualitative Retrieval Cases

The following figures present four independent queries on RGBNT201 and three on MSVR310 using JESD-ReID. Figure 7 places difficult and failure cases above easy and correct cases. Each

block contains a query and its Top-10 gallery ranking across RGB, NIR, and TIR; green and red borders denote correct and incorrect gallery identities.

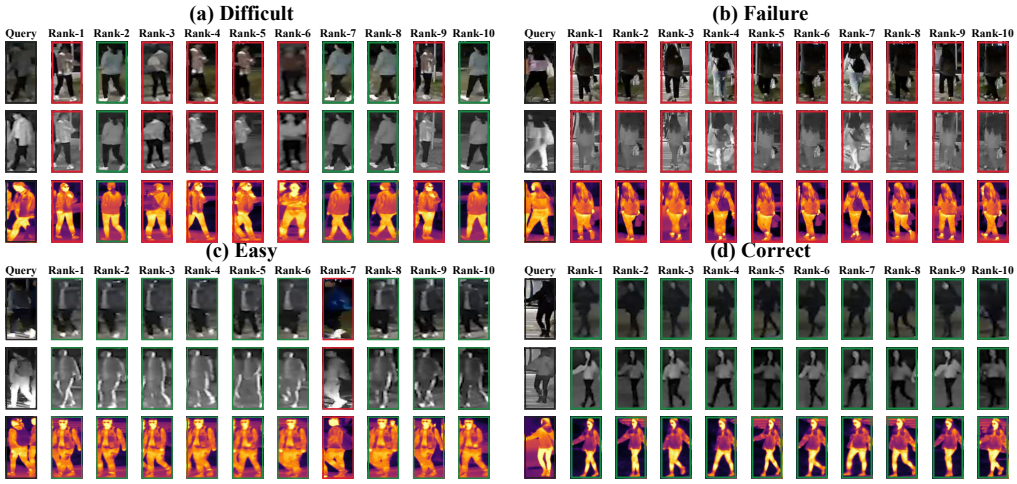

Figure 7: RGBNT201 retrieval examples from JESD-ReID. Top: difficult and failure cases; bottom: easy (9/10 correct) and correct (10/10 correct) cases. Each block shows RGB, NIR, and TIR from top to bottom; green and red borders mark correct and incorrect gallery identities.

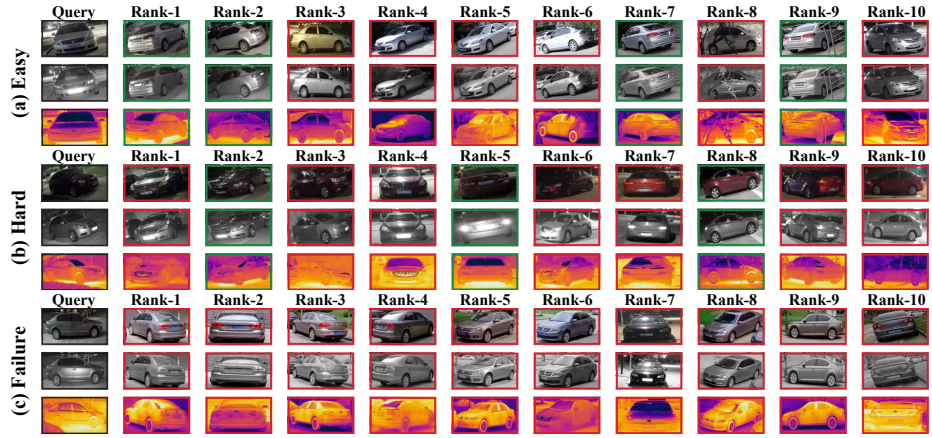

Figure 8: Additional MSVR310 retrieval results from the final JESD-ReID model, covering easy, hard, and failure cases. Green and red borders denote correct and incorrect vehicle identities.

B.3 Component Contribution

Under the main result's flip-augmentation and direct-distance protocol, we evaluate descriptors from the final $r = 2$ checkpoint on matched input groups. Common slots and specific residuals are separately pooled and concatenated across RGB, NIR, and TIR; the last row uses the fused descriptor. This compares descriptor choices, rather than disrupting input correspondence. Both branches retain identity information, and fusion outperforms either branch alone.

Table 6: Contribution of the JESD common and modality-specific components on RGBNT201 without reranking.

Descriptor	mAP	Rank-1
Common semantic slots only	66.16	71.89
Modality-specific residual only	75.63	79.07
Full retrieval descriptor	**82.45**	**87.20**

B.4 Image-Stream Removal

To assess sensitivity to visual input availability, we mask selected RGBNT201 image streams without retraining or reranking, while retaining the original per-image text descriptions. Table 7 reports performance for each combination of unmasked image streams.

Table 7: RGBNT201 image-stream removal with original text retained, without reranking.

Available image streams	mAP	Rank-1	Rank-5	Rank-10
RGB+NIR+TIR	82.45	87.20	91.15	93.90
NIR+TIR (missing RGB)	62.88	63.04	73.09	77.75
RGB+TIR (missing NIR)	76.06	77.39	88.28	91.63
RGB+NIR (missing TIR)	57.52	56.94	70.45	77.03
TIR only	42.10	42.94	60.17	68.06
NIR only	22.39	17.94	31.22	43.78
RGB only	37.47	35.89	48.09	56.94

B.5 Hyperparameter Sensitivity

We vary the number of shared slots $K \in \{2, 4, 8, 16, 32\}$ and adapter rank $r \in \{1, 2, 4, 8, 16\}$ using the same training and evaluation protocol, direct distances, and flip augmentation, without reranking. The final setting is $K = 4, r = 2$; performance remains within 81.94–82.42 mAP across the K sweep and 82.03–82.45 mAP across the r sweep.

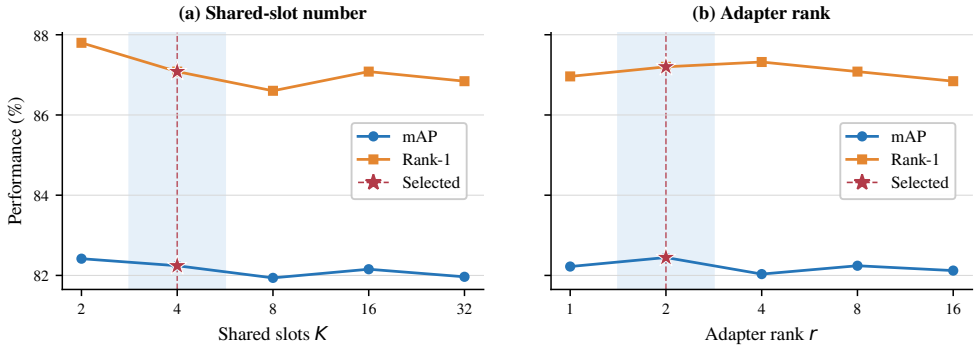

Figure 9: Sensitivity to the number of shared slots and adapter rank on RGBNT201. Stars and dashed lines mark the selected coordinate in each sweep; the shaded bands are visual guides, not confidence intervals.

B.6 Efficiency

Table 8 compares trainable parameter counts on RGBNT201 under each method's training configuration.

Table 8: Accuracy and trainable parameters on RGBNT201. Baseline counts follow the comparison reported by IDEA (Wang et al., 2025c); JESD-ReID is counted before its final JESD-only adaptation.

Method	Trainable params (M)	mAP
UniCat (Crawford et al., 2023)	259.02	57.0
TOP-ReID (Wang et al., 2024)	324.53	72.3
EDITOR (Zhang et al., 2024)	118.55	66.5
RSCNet (Yu et al., 2025)	124.10	68.2
IDEA (Wang et al., 2025c)	91.67	80.2
DeMo (Wang et al., 2025b)	98.79	79.0
MambaPro (Wang et al., 2025a)	74.20	78.9
JESD-ReID	100.42	**82.4**

The baseline counts follow IDEA's published trainable-parameter comparison. For JESD-ReID, 100,421,393 parameters are trainable before the final five-epoch adaptation restricts updates to JESD; that final stage trains 3,458,432 parameters. The reported mAP includes this adaptation. Freezing policies differ between methods, so these counts describe their training configurations rather than total model size or training cost.

Decoupling-module accounting. The counts in Table 8 cover each method's trainable components and thus do not isolate representation decomposition itself. At the module level, DeMo's Hierarchical Decoupling Module (HDM) has 7.358M parameters, whereas the JESD decomposition module has 3.458M, a reduction of 3.900M (53.0%). These counts include the respective decomposition blocks and exclude CLIP encoders, downstream descriptor projections, classifiers, and expert fusion. The difference follows from DeMo instantiating seven separate attention branches, whereas JESD shares semantic observation formation across modalities and parameterizes modality-conditioned dictionaries with a shared 64×512 base and rank-2 corrections. Relational OT introduces no learnable parameters.

Limitations and scope. JESD defines commonness under a learned explanatory model; residuals may retain nuisance variation. It assumes synchronized RGB–NIR–TIR groups, leaving severe desynchronization and unseen sensors beyond the scope of this work.