MoMI-RC: Identity-Aware Reliability Calibration and Multi-Order Modality Interaction for Multi-Modal Object Re-Identification

Anonymous authors
Paper under double-blind review

Abstract

Multi-modal object Re-Identification (ReID) exploits complementary RGB, near-infrared, and thermal-infrared information to retrieve the same person or vehicle under challenging imaging conditions. However, fusion weights predicted solely from individual inputs may lack a stable identity-discriminative reference, while merging interactions from different modality combinations can obscure their distinct sources. To address these issues, we propose MoMI-RC, a multi-order modality interaction network with identity-aware reliability calibration. The reliability calibration module learns global modality anchors and bounded sample-wise corrections under supervision based on intra-class compactness and inter-class separation. The anchors provide a stable reference, while the corrections adapt modality contributions to individual inputs. The multi-order modality interaction module constructs compact pairwise and three-way interactions in a shared low-rank space, preserving each interaction in a separate coordinate block alongside calibrated unimodal features. Modality-subset simulation and consistency regularization further stabilize the descriptor during training. Extensive experiments demonstrate the effectiveness of MoMI-RC, which achieves 84.2%, 89.6%, and 70.9% mAP on RGBNT201, RGBNT100, and MSVR310, respectively, without task-specific text or segmentation priors.

1 Introduction

Multi-modal object Re-Identification (ReID) aims to retrieve the same person or vehicle across cameras by exploiting complementary observations from multiple modalities. RGB, near-infrared (NIR), and thermal-infrared (TIR) images provide complementary cues under illumination changes, occlusion, and imaging degradation. Recent methods strengthen multi-modal representations through text-guided deformable aggregation (Wang et al., 2025c) and segmentation-guided hypergraph modeling (Xu et al., 2026). Our focus is on two related aspects of fusion: determining the contribution of each modality and retaining evidence from different modality combinations.

A modality's usefulness for retrieval depends both on its overall discriminative ability across identities and on the quality of individual observations. Adaptive fusion methods such as DeMo (Wang et al., 2025b) address input-dependent variation by dynamically weighting features. This input-adaptive formulation, however, does not explicitly separate a global modality-level reference from sample-specific adjustments. We argue that this reference should be tied to the retrieval task: features from a reliable modality should exhibit small intra-identity distances and clear inter-identity separation. This motivates supervising a global reference with identity relationships across training samples, while allowing bounded corrections for individual inputs.

Reliability weighting alone does not explicitly encode multiplicative relationships among modality features. Tensor Fusion Network (TFN) models unimodal, bimodal, and trimodal terms through outer products (Zadeh et al., 2017), while Low-rank Multimodal Fusion (LMF) reduces fusion costs through modality-specific factors (Liu et al., 2018). For retrieval, the question is how to retain interaction evidence in a compact descriptor. Merging products from different modality combinations removes their explicit separation, whereas full outer-product expansion incurs substantial dimensional growth.

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.

Our goal is to preserve evidence from individual modalities and their combinations while controlling the cost of higher-order modeling.

To address these issues, we propose MoMI-RC, a multi-order modality interaction network with identity-aware reliability calibration. Reliability Calibration (RC) learns global modality anchors under identity-discriminative supervision and combines them with bounded sample-wise corrections to calibrate unimodal features. Multi-order Modality Interaction (MoMI) constructs pairwise and three-way element-wise products in a shared low-rank space. Each interaction occupies a separate coordinate block and is concatenated with the calibrated unimodal features at their original dimensionality. This descriptor brings together reliability-adjusted first-order evidence and higher-order evidence from distinct modality combinations. During training, modality-subset simulation and consistency regularization encourage consistent identity representations across complete and subset inputs.

Experiments on RGBNT201, RGBNT100, and MSVR310 demonstrate the effectiveness of MoMI-RC, with mAP scores of 84.2%, 89.6%, and 70.9%, respectively, without task-specific text or segmentation priors. Controlled ablations further support the benefits of explicit reliability supervision and separately retained interaction blocks. Our main contributions are summarized as follows:

- We introduce identity-aware reliability calibration that uses inter-sample identity relationships to supervise global modality anchors and combines them with bounded input-dependent corrections.
- We develop a compact retrieval descriptor that preserves pairwise and three-way products in separate blocks after a shared low-rank projection, alongside calibrated unimodal features at their original dimensionality.
- We integrate modality-subset simulation and consistency regularization into training, and validate the framework through extensive experiments on person and vehicle ReID.

2 Related Work

2.1 Multi-Modal Object Re-Identification

Multi-modal object ReID combines RGB, NIR, and TIR information to compensate for single-sensor limitations. UniCat (Crawford et al., 2023) identifies *modality laziness* in joint training and establishes independent learning followed by concatenation as a strong baseline, emphasizing the need to preserve unimodal discrimination.

Vision Transformers (Dosovitskiy et al., 2021; He et al., 2021) enable fine-grained cross-modal token interaction. TOP-ReID (Wang et al., 2024) uses cyclic token permutation and complementary reconstruction, while EDITOR (Zhang et al., 2024) combines spatial-frequency token selection with hierarchical aggregation. MambaPro (Wang et al., 2025a) adapts CLIP (Radford et al., 2021) with Mamba-based aggregation. IDEA (Wang et al., 2025c) uses generated text and deformable aggregation to enhance local visual features.

Complementary approaches preserve modality-specific information through interaction and margin learning (Wang et al., 2022), progressive fusion (Yin et al., 2023), and selective token coupling (Yu et al., 2025). DENet (Zheng et al., 2023a) and All-in-One (Li et al., 2024) address incomplete modality inputs. DeMo (Wang et al., 2025b) dynamically weights decoupled features, Signal (Liu et al., 2026) aligns global and local representations, and STMI (Xu et al., 2026) combines segmentation priors with hypergraph modeling. Our method instead calibrates reliability from training-identity metric structure and separately retains multi-order interaction evidence.

2.2 Multi-Modal Fusion and Interaction Learning

Early fusion mixes modalities during encoding, whereas late fusion independently encodes and then concatenates or weights them. Interactive methods use attention, token exchange, or graph propagation to model cross-modal dependencies; graph-based progressive fusion (He et al., 2023) and generative attentive fusion (Guo et al., 2022) are examples in vehicle ReID. GMU (Arevalo et al., 2017) learns multiplicative gates for modality contributions, while DeMo (Wang et al., 2025b)

dynamically weights decoupled ReID features. RC instead explicitly supervises a global reference using intra-/inter-identity structure and combines it with a bounded sample-wise correction.

For interaction modeling, bilinear pooling (Lin et al., 2015) and low-rank bilinear interactions (Kim et al., 2017) model feature relationships through multiplicative interactions. Tensor Fusion Network (TFN) (Zadeh et al., 2017) uses tensor outer products to jointly encode unimodal, bimodal, and trimodal information, while Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) reduces fusion costs through modality-specific low-rank factors. Another line of work models cross-modal relations using graphs or hypergraphs; for example, STMI (Xu et al., 2026) captures higher-order semantic relations through hypergraph propagation. For multi-modal identity retrieval, MoMI separately constructs pairwise and three-way interactions in a shared low-dimensional space and retains them in corresponding coordinate blocks alongside unimodal features at their original dimensionality. This design requires no iterative message passing and incorporates identity-supervised reliability calibration to learn the contributions of individual modalities.

3 METHOD

As illustrated in Figure 1, MoMI-RC comprises two core modules: Identity-Aware Reliability Calibration (RC) and Multi-Order Modality Interaction (MoMI). RC calibrates the contributions of individual modalities, while MoMI captures complementary information through compact pairwise and three-way interactions. We further introduce modality-subset simulation and consistency regularization during training to improve descriptor stability. The following subsections describe the overall framework, the two modules, and the training objectives.

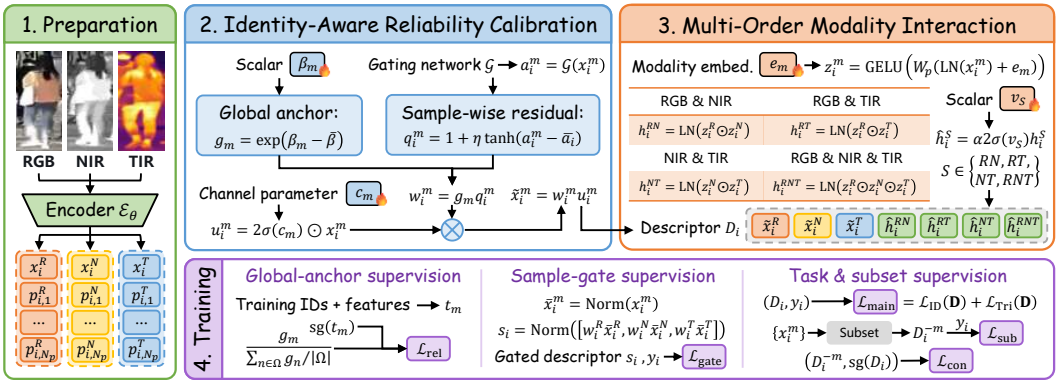

Figure 1: Overview of MoMI-RC. A shared DINOv3-B encoder extracts global features from RGB, NIR, and TIR images. RC calibrates modality contributions, while MoMI captures pairwise and three-way interactions in a shared low-rank space. The calibrated first-order features and higher-order interaction blocks are concatenated into the retrieval descriptor. The bottom panel summarizes the training objectives. LN, Norm, and sg denote LayerNorm, L2 normalization, and stop-gradient, respectively.

3.1 OVERALL FRAMEWORK

Given synchronized RGB (R), NIR (N), and TIR (T) images of the same object, $\mathcal{I}_i = \{I_i^R, I_i^N, I_i^T\}$, a parameter-shared DINOv3-B (Siméoni et al., 2025) encoder $\mathcal{E}_\theta$ produces a class (CLS) token and a sequence of N_p patch tokens p_i^m for each modality. We select the CLS token as the global feature for subsequent calibration and interaction:

$$x_i^m = \mathrm{CLS}(\mathcal{E}_\theta(I_i^m)), \qquad m \in \Omega = \{R, N, T\}. \tag{1}$$

Here, $\mathrm{CLS}(\cdot)$ selects the class token, i indexes samples, y_i is the identity label, θ denotes the shared encoder parameters, and $x_i^m \in \mathbb{R}^C$, with $C = 768$. RC calibrates first-order modality features; MoMI constructs low-rank higher-order blocks. Their concatenation is optimized with identity classification and metric learning.

3.2 IDENTITY-AWARE RELIABILITY CALIBRATION

RC does not prescribe a relative ordering among the sensors; instead, it learns reliability from training identities. We decompose reliability into a global anchor g_m and a sample-wise residual q_i^m. The former describes relatively stable modality discriminability across samples in the training distribution, whereas the latter responds to the quality of the current input.

3.2.1 GLOBAL RELIABILITY ANCHORS

For each modality, we define a learnable scalar β_m and remove an unidentifiable common scale by centering across modalities:

$$g_m = \exp\left(\beta_m - \frac{1}{|\Omega|}\sum_{n\in\Omega}\beta_n\right). \tag{2}$$

All β_m are initialized to zero and jointly optimized with the remaining modules under identity-discriminative supervision.

To make g_m reflect identity-retrieval capability, we construct a stop-gradient reliability target from each training batch. Let $\bar{x}_i^m = x_i^m/\|x_i^m\|_2$. The distance between samples i and j in modality m is

$$d_{ij}^m = \|\bar{x}_i^m - \bar{x}_j^m\|_2. \tag{3}$$

For every valid anchor i, we select the farthest sample with the same identity and the nearest sample with a different identity:

$$d_{i,+}^m = \max_{\substack{j:y_j=y_i\\ j\neq i}} d_{ij}^m, \qquad d_{i,-}^m = \min_{j:y_j\neq y_i} d_{ij}^m. \tag{4}$$

The modality ranking difficulty is defined as

$$\ell_m^{\mathrm{mod}} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathrm{softplus}(d_{i,+}^m - d_{i,-}^m), \tag{5}$$

where $\mathcal{V}$ is the set of anchors for which both positive and negative samples exist. A smaller ℓ_m^{mod} indicates a more compact same-identity distribution and clearer inter-identity margins. We convert the difficulty into a soft target with mean one:

$$t_m = |\Omega|\,\frac{\exp(-\ell_m^{\mathrm{mod}}/T_r)}{\sum_{n\in\Omega}\exp(-\ell_n^{\mathrm{mod}}/T_r)}, \tag{6}$$

where $T_r > 0$ is the reliability temperature controlling the sharpness of the modality-level target distribution. We then supervise the global anchors using

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{|\Omega|}\sum_{m\in\Omega}\left(\frac{g_m}{\frac{1}{|\Omega|}\sum_{n\in\Omega}g_n} - \mathrm{sg}(t_m)\right)^2. \tag{7}$$

Here, $\mathrm{sg}(\cdot)$ denotes stop-gradient. The target t_m depends only on the current training batch and its identity labels; it does not access a validation set, queries, or gallery neighborhoods. As training proceeds, g_m aggregates identity structures across batches into a stable reliability anchor.

3.2.2 SAMPLE-WISE BOUNDED RESIDUAL GATING

Global anchors cannot describe occlusion and imaging degradation in individual inputs. We therefore use a lightweight gating network $\mathcal{G}$, shared across the three modalities, to predict $a_i^m = \mathcal{G}(x_i^m)$. The network comprises LayerNorm, two linear layers, and GELU. Its outputs are centered across modalities and converted into bounded residual coefficients:

$$q_i^m = 1 + \eta\tanh\left(a_i^m - \frac{1}{|\Omega|}\sum_{n\in\Omega}a_i^n\right), \tag{8}$$

where η is a magnitude bound shared by all datasets and modalities, and the final layer of $\mathcal{G}$ is zero-initialized. Each modality also has a channel parameter $c_m \in \mathbb{R}^C$, initialized to zero, that selects

discriminative first-order dimensions without mixing modality sources. The final first-order reliability coefficient and calibrated feature are

$$w_i^m = g_m q_i^m, \qquad u_i^m = 2\sigma(c_m) \odot x_i^m, \qquad \widetilde{x}_i^m = w_i^m u_i^m. \tag{9}$$

Here, $\sigma(\cdot)$ and $\odot$ denote the sigmoid function and element-wise multiplication, respectively. To directly optimize the sample gate for retrieval, we construct a normalized first-order gated descriptor

$$s_i = \mathrm{Norm}\big([w_i^R \bar{x}_i^R, w_i^N \bar{x}_i^N, w_i^T \bar{x}_i^T]\big), \tag{10}$$

where $\mathrm{Norm}(\cdot)$ denotes L2 normalization. We compute on s_i the same batch-hard ranking objective as in Equations 4–5, denoted by $\mathcal{L}_{\mathrm{gate}}$. Thus, $\mathcal{L}_{\mathrm{rel}}$ supervises stable cross-sample anchors, whereas $\mathcal{L}_{\mathrm{gate}}$ optimizes input-specific residual adjustments; both are driven by identity discrimination.

3.3 Multi-Order Modality Interaction

Reliability calibration controls the contribution of first-order evidence but does not explicitly represent combinatorial relationships among modalities. MoMI uses a shared low-rank projection and element-wise multiplication to construct all non-empty interaction orders of three variables at modest cost. We first map each modality into an r-dimensional latent space:

$$z_i^m = \mathrm{GELU}(W_p(\mathrm{LN}(x_i^m) + e_m)), \qquad z_i^m \in \mathbb{R}^r, \tag{11}$$

where $\mathrm{LN}(\cdot)$ denotes LayerNorm, W_p is shared across modalities, and e_m is a learnable modality embedding. The three second-order modality pairs are

$$h_i^{mn} = \mathrm{LN}(z_i^m \odot z_i^n), \qquad (m,n) \in \{(R,N),(R,T),(N,T)\}, \tag{12}$$

and the third-order joint interaction is

$$h_i^{RNT} = \mathrm{LN}(z_i^R \odot z_i^N \odot z_i^T). \tag{13}$$

Each higher-order block is modulated by an independent learnable gate:

$$\widehat{h}_i^S = \alpha\, 2\sigma(v_S) h_i^S, \qquad S \in \{RN, RT, NT, RNT\}, \tag{14}$$

where α is shared by all four blocks and v_S is a learnable scalar initialized to zero. The complete descriptor is

$$D_i = \left[\widetilde{x}_i^R, \widetilde{x}_i^N, \widetilde{x}_i^T, \widehat{h}_i^{RN}, \widehat{h}_i^{RT}, \widehat{h}_i^{NT}, \widehat{h}_i^{RNT}\right]. \tag{15}$$

Reliability coefficients calibrate only the three first-order blocks and are not repeatedly propagated to the higher-order blocks, which learn their utility through separate gates. This decoupling prevents reliability from being amplified twice in first- and higher-order evidence. Rather than forming full bilinear interactions in the original C-dimensional space, MoMI performs element-wise products, normalization, and four scalar gates after a shared low-rank projection. RC adds only a lightweight shared gating network, while the training-only subset branch contains the shared reconstruction layer in Equation 16.

3.4 Modality-Subset Simulation and Consistency Regularization

Higher-order multiplicative interactions can be sensitive to single-modality noise. Whereas ModDrop (Neverova et al., 2016) randomly drops modalities during training, we enumerate three leave-one-modality-out branches and impose complete–subset consistency. Suppose modality m is masked and the remaining modalities are p and q. We reconstruct the first-order feature of the missing modality from the mean of the other calibrated features and their low-rank interaction:

$$\widehat{x}_i^m = \frac{1}{2}(\widetilde{x}_i^p + \widetilde{x}_i^q) + b_m + \boldsymbol{\gamma} \odot W_r \left[\frac{1}{2}(z_i^p + z_i^q), z_i^p \odot z_i^q\right], \tag{16}$$

where W_r is a shared reconstruction layer, b_m is a zero-initialized modality bias, and $\boldsymbol{\gamma} \in \mathbb{R}^C$ is a channel-wise residual scale shared by all three reconstruction branches. Every second- or third-order block involving the masked modality is zeroed, while interactions not involving that modality are preserved, yielding a subset descriptor D_i^{-m}. We apply the same identity classification and triplet supervision to all three subset descriptors and align their directions with the complete descriptor. Let

$\mathcal{B}$ denote the current mini-batch, and let $\mathbf{D} = \{D_i\}_{i\in\mathcal{B}}$ and $\mathbf{D}^{-m} = \{D_i^{-m}\}_{i\in\mathcal{B}}$ collect its complete and subset descriptors. Following the implementation, the auxiliary branches are averaged:

$$\mathcal{L}_{\text{sub}} = \frac{1}{|\Omega|}\sum_{m\in\Omega}\left(\mathcal{L}_{\text{ID}}(\mathbf{D}^{-m}) + \mathcal{L}_{\text{Tri}}(\mathbf{D}^{-m})\right), \tag{17}$$

$$\mathcal{L}_{\text{con}} = \frac{1}{|\Omega|}\sum_{m\in\Omega}\left[1 - \cos(\mathbf{D}^{-m}, \text{sg}(\mathbf{D}))\right], \tag{18}$$

where $\cos(\mathbf{D}^{-m}, \mathbf{D})$ denotes the mean cosine similarity between corresponding descriptors in the mini-batch. Under the standard complete-input protocol, the subset branches are disabled. Their learned reconstruction is reused only for the missing-modality analysis in Section A.2.

3.5 Optimization Objective and Training Procedure

The complete descriptor batch $\mathbf{D}$ is passed through BNNeck (Luo et al., 2019) and a linear classifier. The main-branch loss combines label-smoothed cross-entropy (Szegedy et al., 2016) and soft-margin triplet loss (Hermans et al., 2017):

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{ID}}(\mathbf{D}) + \mathcal{L}_{\text{Tri}}(\mathbf{D}). \tag{19}$$

For concise notation, we first combine the retrieval losses of the complete and modality-subset branches:

$$\mathcal{L}_{\text{reid}} = (1 - \lambda_{\text{sub}})\mathcal{L}_{\text{main}} + \lambda_{\text{sub}}\mathcal{L}_{\text{sub}}.$$

During the second-stage joint optimization, the complete objective is

$$\mathcal{L} = \lambda_{\text{reid}}\mathcal{L}_{\text{reid}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}. \tag{20}$$

Here, λ_{sub} controls the contribution of modality-subset supervision, while λ_{reid}, λ_{con}, λ_{gate}, and λ_{rel} weight the retrieval, consistency, sample-gating, and global-reliability objectives, respectively.

Training follows a warm-up-then-joint-optimization strategy. We initialize from a trained DINOv3-B ReID baseline and first warm up the channel gates, MoMI, and modality-subset branches with reliability calibration disabled, preventing newly introduced modules from perturbing the learned representation. We then freeze the first-order channel gates learned in the warm-up stage, reset the global anchors and the output layer of the sample gate to the equal-reliability state, enable global and sample-wise reliability calibration, and jointly optimize the complete shared backbone and the remaining MoMI-RC parameters using Equation 20.

Under the standard complete-input protocol, the full tri-modal descriptor is L2-normalized for Euclidean-distance retrieval.

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. We evaluate MoMI-RC on three public multi-modal object ReID benchmarks. RGBNT201 (Zheng et al., 2021) is a multi-spectral person ReID dataset containing 4,787 aligned RGB-NIR-TIR image groups from 201 identities. RGBNT100 (Li et al., 2020) is a large-scale multi-spectral vehicle ReID dataset with 17,250 tri-modal groups from 100 vehicle identities under diverse imaging conditions. MSVR310 (Zheng et al., 2023b) contains 2,087 tri-modal groups from 310 vehicle identities and presents challenges such as cross-environment observations and long temporal spans.

Evaluation protocols. Following the standard complete tri-modal retrieval setting, we evaluate mean Average Precision (mAP) and the Cumulative Matching Characteristic (CMC) at Rank-1, Rank-5, and Rank-10. We report single-model results without test-time adaptation or re-ranking.

Table 1: Comparison on RGBNT201, RGBNT100, and MSVR310 without task-specific external priors. Bold and underlined values denote the best and second-best results. Backbone symbols are defined in Section 4.3.

Method	RGBNT201				RGBNT100		MSVR310	
	mAP	R-1	R-5	R-10	mAP	R-1	mAP	R-1
HAMNet (Li et al., 2020)	27.7	26.3	41.5	51.7	74.5	93.3	27.1	42.3
PFNet (Zheng et al., 2021)	38.5	38.9	52.0	58.4	68.1	94.1	23.5	37.4
CCNet (Zheng et al., 2023b)	–	–	–	–	77.2	96.3	36.4	55.2
LRMM (Wu et al., 2025)	52.3	53.4	64.6	73.2	78.6	96.7	36.7	49.7
UniCat (Crawford et al., 2023)*	57.0	55.7	–	–	79.4	96.2	–	–
TOP-ReID (Wang et al., 2024)*	72.3	76.6	84.7	89.4	81.2	96.4	35.9	44.6
EDITOR (Zhang et al., 2024)*	66.5	68.3	81.1	88.2	82.1	96.4	39.0	49.3
MambaPro (Wang et al., 2025a)†	78.9	83.4	89.8	91.9	83.9	94.7	47.0	56.5
DeMo (Wang et al., 2025b)†	79.0	82.3	88.8	92.0	86.2	97.6	49.2	59.8
Signal (Liu et al., 2026)†	80.3	85.2	91.4	93.7	86.3	97.6	53.6	71.9
FUSE (Qi et al., 2026)†	81.4	**86.1**	91.5	93.8	<u>88.5</u>	96.9	50.1	65.7
MDReID (Feng et al., 2025)†	82.1	85.2	90.3	92.6	85.3	95.6	51.0	68.9
ProxyTTT (Zheng et al., 2026)†	82.3	84.7	90.6	92.7	88.4	97.9	62.1	71.7
ProxyTTT (Zheng et al., 2026)‡	76.7	78.9	87.4	90.6	87.0	97.6	<u>63.6</u>	79.2
PEFT-BoA (Li et al., 2026)†	<u>82.7</u>	**86.1**	**92.3**	**94.7**	85.1	97.8	57.4	74.5
PEFT-BoA (Li et al., 2026)‡	79.8	82.4	88.9	91.0	86.1	<u>98.3</u>	63.0	<u>82.7</u>
MoMI-RC (ours)‡	**84.2**	<u>85.5</u>	<u>92.0</u>	<u>94.6</u>	**89.6**	**98.4**	**70.9**	**83.9**

4.2 Implementation Details

We implement MoMI-RC in PyTorch (Paszke et al., 2019) and train it on NVIDIA RTX 4090D GPUs. The visual backbone is DINOv3 ViT-B/16 (Siméoni et al., 2025), pre-trained on LVD-1689M, with all backbone parameters shared across modalities. The latent dimension of MoMI is 128, the shared interaction scale α is set to 0.125, the sample residual bound η is set to 0.5, and the hidden dimension of the sample gating network is 64. Training images are augmented by random horizontal flipping, random cropping, and random erasing (Zhong et al., 2020), and batches are constructed with an identity-balanced sampler.

All datasets share the same model architecture and three-stage optimization schedule. We train the DINOv3-B ReID baseline for 30 epochs, followed by 2 epochs of interaction warm-up and 20 epochs of full-backbone joint optimization. All stages use AdamW (Loshchilov & Hutter, 2019) with cosine learning-rate scheduling (Loshchilov & Hutter, 2017); their base learning rates are 3.5×10^{-4}, 2.5×10^{-4}, and 1.25×10^{-4}, respectively. We use automatic mixed precision. Stage A enables only the main/subset retrieval terms and consistency regularization. Stage B additionally enables complete reliability calibration. The reliability temperature T_r is set to 0.10 for all datasets, while λ_{reid}, λ_{sub}, λ_{con}, λ_{gate}, and λ_{rel} are set to 0.10, 0.05, 0.05, 0.25, and 1.00, respectively. Complete configurations and hyperparameters are provided with the code.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares methods without task-specific external priors. Symbols *, †, and ‡denote other ViT, CLIP, and DINOv3 backbones, respectively; unmarked methods use CNNs. For ProxyTTT and PEFT-BoA, we additionally replace the visual encoder with DINOv3-B in the official implementations while retaining their core designs. Both ProxyTTT variants use the setting without test-time training (w/o TTT).

Main results. Table 1 reports person and vehicle ReID results on all three benchmarks. MoMI-RC achieves 84.2%, 89.6%, and 70.9% mAP on RGBNT201, RGBNT100, and MSVR310, improving the respective second-best results by 1.5, 1.1, and 7.3 points. Rank-1 reaches 85.5%, 98.4%, and

Table 2: Ablation of the overall components on RGBNT201.

MoMI	RC	Subset	Cons.	mAP	R-1	R-5	R-10
–	–	–	–	73.2	73.7	84.0	90.0
✓	–	–	–	77.9	78.5	86.4	90.8
–	✓	–	–	80.5	82.3	90.0	93.3
✓	✓	–	–	80.8	82.2	89.8	94.0
✓	✓	✓	–	82.4	84.1	91.1	93.4
✓	✓	✓	✓	**84.2**	**85.5**	**92.0**	**94.6**

83.9%. On RGBNT201, Rank-5 and Rank-10 are only 0.3 and 0.1 points below the best results, respectively. Compared with methods using task-specific external priors, MoMI-RC achieves the highest RGBNT201 mAP in Table 6; the full comparison is given in Appendix A.1.

4.4 Ablation Studies

We ablate the principal designs of MoMI-RC on RGBNT201. Unless otherwise stated, all variants use the same backbone, data split, and optimization setting as the complete model.

Overall components. Table 2 evaluates MoMI, RC, modality-subset training, and consistency regularization. Adding either MoMI or RC to the baseline yields a clear improvement, demonstrating their independent value. Once they are combined, modality-subset training and consistency regularization further increase mAP from 80.8% to 84.2%, validating the complementarity of the four components. Additional analyses of interaction orders and key hyperparameters are provided in Appendix A.

Interaction organization. Table 3 compares three fusion designs with the same training protocol, RC module, low-rank dimension, and $4r$-dimensional interaction output. MLP fusion projects concatenated low-rank modality features. Merged interactions compute the three pairwise products and the three-way product, then average them before projection. MoMI instead retains the four products in separate coordinate blocks. Explicit products followed by merging improve mAP from 82.5% to 83.1%. Retaining the products separately further increases mAP to 84.2%, a gain of 1.1 points over merged interactions and 1.7 points over MLP fusion. These results support preserving interaction sources beyond merely introducing nonlinear fusion at the same output dimension.

Table 3: Interaction organization on RGBNT201 with the same $4r$-dimensional interaction output.

Interaction organization	mAP	R-1	R-5	R-10
MLP fusion	82.5	84.2	**92.0**	94.5
Merged interactions	83.1	85.0	91.5	94.3
Separate interaction coordinates (MoMI)	**84.2**	**85.5**	**92.0**	**94.6**

4.5 Reliability Calibration

We fix MoMI, modality-subset training, and consistency regularization and vary the global anchor, sample gate, and explicit calibration supervision in Table 4. Either the global anchor or sample gate outperforms fixed equal weighting, while their combination achieves the best result in the complete RC. Removing explicit calibration supervision decreases mAP to 76.8%, showing that the retrieval objective alone is insufficient to stably learn two-level reliability and that calibration from training-identity structure is essential.

Table 4: Ablation of reliability calibration on RGBNT201.

Global anchor	Sample gate	Calib. sup.	mAP	R-1	R-5	R-10
–	–	–	78.2	79.9	87.3	91.4
✓	–	✓	81.2	82.9	90.8	93.4
–	✓	✓	82.0	85.3	90.2	93.5
✓	✓	–	76.8	78.5	86.1	90.4
✓	✓	✓	**84.2**	**85.5**	**92.0**	**94.6**

4.6 Model Complexity

Table 5: Model complexity and RGBNT201 retrieval accuracy.

Model	Complexity		RGBNT201			
	Params (M)	FLOPs (G)	mAP	R-1	R-5	R-10
IDEA (CLIP-B/16) (Wang et al., 2025c)	91.67	43.70	80.2	82.1	90.0	93.3
DINOv3-B ReID baseline	86.07	35.13	73.2	73.7	84.0	90.0
PEFT-BoA (DINOv3-B) (Li et al., 2026)	92.82	36.26	79.8	82.4	88.9	91.0
MoMI-RC	86.91	35.13	**84.2**	**85.5**	**92.0**	**94.6**

Table 5 compares model complexity and retrieval accuracy. Compared with IDEA, MoMI-RC uses 4.76M fewer parameters and 8.57G fewer inference FLOPs while improving mAP by 4.0 points. With the same DINOv3-B backbone, MoMI-RC adds only 0.84M parameters and less than 0.01G FLOPs over the ReID baseline while improving mAP by 11.0 points. Compared with PEFT-BoA, it reduces parameters and FLOPs by 5.91M and 1.13G, respectively, while improving mAP by 4.4 points.

4.7 Feature Visualization

Figure 2 uses t-SNE (van der Maaten & Hinton, 2008) to compare four model variants on RGBNT201, showing 25 images per identity for ten identities. For these examples, MoMI yields tighter clusters than the baseline, RC further improves compactness, and the full model produces the most compact clusters.

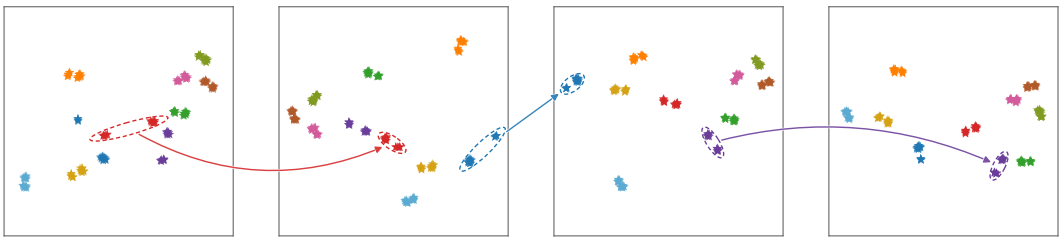

Figure 2: Visualization of the feature distributions with t-SNE on RGBNT201. Different colors represent different identities.

5 Conclusion

We presented MoMI-RC for multi-modal object ReID. Identity-supervised global anchors and bounded sample gates calibrate modality reliability, while low-rank interactions retain pairwise and three-way evidence in separate coordinates. Modality-subset simulation and consistency regularization stabilize the descriptor. Experiments on three benchmarks demonstrate the method's effectiveness, supported by component ablations and robustness and efficiency analyses.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and literature discovery. For writing assistance, they helped refine grammar, word choice, and sentence structure to improve clarity and readability. For literature discovery, they helped locate related studies on multi-modal object Re-Identification and identify relevant publications for comparison and citation. The authors take full responsibility for the final manuscript, including its content, claims, and references.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16×16 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*, 2025.

Ya Gao, Shihao Li, Zhaojun Liu, Aihua Zheng, Chenglong Li, and Jin Tang. Chain-of-thought guided multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37705–37714, 2026.

Jinbo Guo, Xiaojing Zhang, Zhengyi Liu, and Yuan Wang. Generative and attentive fusion for multi-spectral vehicle re-identification. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1565–1572, 2022.

Qiaolin He, Zefeng Lu, Zihan Wang, and Haifeng Hu. Graph-based progressive fusion network for multi-modality vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):12431–12447, 2023.

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

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

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations*, 2017.

He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17459–17469, 2024.

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*, pp. 11345–11353, 2020.

Hongchao Li, Guangxing Liu, Xixi Wang, Baihe Liang, and Yonglong Luo. PEFT-BoA: Parameter-efficient fine-tuning with bag-of-adapters for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6127–6135, 2026.

Shihao Li, Huaibo Huang, Junxian Duan, Aihua Zheng, Jin Tang, and Jixin Ma. NEXT: Multi-grained mixture of experts via text-modulation for multi-modal object re-identification. *arXiv preprint arXiv:2505.20001*, 2025.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1457, 2015.

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*, pp. 7359–7367, 2026.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Xuanhao Qi, Tom H. Luan, Yukang Zhang, Jinkai Zheng, Zhou Su, Shuwei Li, and Lei Tan. FUSE: Frequency-domain unification and spectral energy alignment for multi-modal object re-identification. In *Proceedings of the International Conference on Machine Learning*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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.

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

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*, 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*, 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*, 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.

Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and EnlargE: Boosting modality-specific representations for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2633–2641, 2022.

Di Wu, Zhihui Liu, Zihan Chen, Shenglong Gan, Kaiwen Tan, Qin Wan, and Yaonan Wang. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *Expert Systems with Applications*, 263:125716, 2025.

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*, pp. 11433–11441, 2026.

Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. GraFT: Gradual fusion transformer for multimodal re-identification. *arXiv preprint arXiv:2310.16856*, 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.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

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.

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*, pp. 3529–3537, 2021.

Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification. *arXiv preprint arXiv:2305.15762*, 2023a.

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, 2023b.

Aihua Zheng, Zhaojun Liu, Xixi Wan, Chenglong Li, Jin Tang, and Yan Yan. ProxyTTT: Proxy-driven test-time training for multi-modal re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13342–13350, 2026.

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

SUPPLEMENTARY MATERIALS

A ADDITIONAL EXPERIMENTS

A.1 COMPARISON WITH TASK-SPECIFIC EXTERNAL PRIORS

We further compare on RGBNT201 with methods that depend on task-specific external priors, defined as textual descriptions, segmentation masks, or structured reasoning information beyond visual-backbone pre-training. Without task-specific text, segmentation masks, or chain-of-thought semantics, MoMI-RC achieves the highest mAP of 84.2%. Under the same DINOv3-B backbone, it improves over IDEA by 2.9 points in mAP and 2.3 points in Rank-1. Compared with CoT-ReID, MoMI-RC improves mAP by 0.9 points while trailing Rank-1 by only 0.6 points. These results show that identity-aware reliability calibration and multi-order interaction can learn competitive identity representations directly from tri-modal visual data.

Table 6: Comparison with methods using task-specific external priors on RGBNT201.

Method	Visual backbone	Extra prior	mAP	R-1
IDEA (Wang et al., 2025c)	CLIP-B/16	Text	80.2	82.1
IDEA (DINOv3-B) (Gao et al., 2026)	DINOv3-B	Text	81.3	83.2
NEXT (Li et al., 2025)	CLIP-B/16	Text	82.4	**86.6**
STMI (Xu et al., 2026)	CLIP-B/16	Text + Mask	81.2	83.4
CoT-ReID (Gao et al., 2026)	DINOv3-B	CoT Text	<u>83.3</u>	<u>86.1</u>
MoMI-RC (ours)	DINOv3-B	None	**84.2**	85.5

A.2 ROBUSTNESS TO MISSING MODALITIES

For each dataset, we evaluate all seven non-empty combinations of RGB (R), NIR (N), and TIR (T) using its complete-input checkpoint, without retraining or fine-tuning. Higher-order blocks involving an unavailable modality are suppressed while the descriptor dimension remains fixed. When one modality is missing, we reuse the corresponding singleton reconstruction learned by the modality-subset branches; when only one modality is available, all unavailable singleton and interaction blocks are set to zero. Table 7 reports the full results. Among the incomplete inputs, R+T consistently yields the highest mAP, retaining 79.4%, 89.4%, and 68.5% on RGBNT201, RGBNT100, and MSVR310, respectively. This indicates that visible appearance and thermal structure provide the strongest complementary pair. Performance decreases as only one sensor remains, as expected, but the same model still produces meaningful retrieval results without adaptation. On MSVR310, R+T slightly exceeds the complete input in Rank-1 while remaining lower in mAP, suggesting that NIR improves the overall ranking even when it does not improve the first match. Overall, MoMI-RC degrades gracefully with two available modalities and remains functional in the more extreme single-sensor setting.

Table 7: Missing-modality robustness using each dataset's complete-input checkpoint. No retraining or fine-tuning is used.

Available input	RGBNT201		RGBNT100		MSVR310	
	mAP	R-1	mAP	R-1	mAP	R-1
R+N+T	**84.2**	**85.5**	**89.6**	**98.4**	**70.9**	83.9
R+N	60.0	58.1	79.1	95.3	66.2	81.2
R+T	79.4	79.9	89.4	97.7	68.5	**84.3**
N+T	67.6	69.5	84.9	95.6	63.7	79.2
R only	44.5	42.1	72.5	91.7	59.9	75.8
N only	28.2	24.3	65.9	86.0	55.8	74.3
T only	56.1	56.0	61.8	86.2	36.2	57.9

A.3 Interaction Orders

Table 8 compares the interaction orders retained as separate coordinates in the final descriptor. The first-order variant achieves 82.5% mAP. The first-plus-second-order and first-plus-third-order variants obtain 81.1% and 83.5% mAP, respectively. Among the reported runs, retaining both higher-order interactions in separate coordinates yields the highest mAP of 84.2%. The complete model also achieves the highest Rank-5 and Rank-10, while the first-plus-third-order variant achieves the highest Rank-1.

Table 8: Ablation of interaction orders on RGBNT201.

Interaction orders	mAP	R-1	R-5	R-10
First-order	82.5	84.0	91.1	93.7
First + second	81.1	82.9	90.4	93.4
First + third	83.5	**86.6**	91.5	94.0
First + second + third	**84.2**	85.5	**92.0**	**94.6**

A.4 Parameter Analysis

Figure 3 analyzes the low-rank dimension r and sample residual magnitude η on RGBNT201. We evaluate $r \in \{32, 64, 128, 256, 512\}$ with $\eta = 0.5$, and $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $r = 128$; all remaining settings match the complete model. The model obtains the highest mAP at $r = 128$. Although $r = 64$ yields slightly higher Rank-1, its mAP is 1.0 point lower. For the sample gate, $\eta = 0.5$ achieves the highest mAP and Rank-1. We therefore adopt $r = 128$ and $\eta = 0.5$.

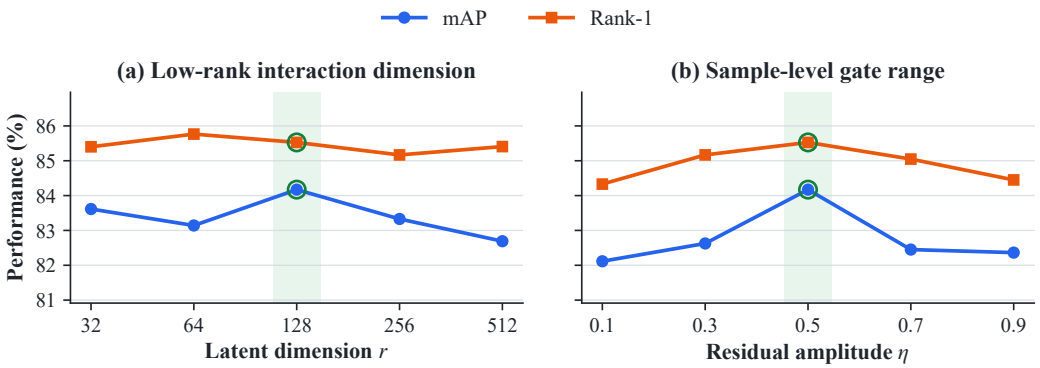

Figure 3: Analysis of the low-rank dimension and sample residual magnitude on RGBNT201. Green markers indicate the adopted configuration.

A.5 Training Cost

We measure training cost on RGBNT201 with batch size 64 using one NVIDIA RTX 4090D GPU. Table 9 reports the median batch time, end-to-end wall time, and peak GPU memory for each training phase. Wall time includes initialization, evaluation, and checkpoint I/O, with inter-stage idle time excluded. Stage A and Stage B increase the median batch time by at most 4.4% and do not raise the measured peak memory beyond the baseline's 13.0 GiB.

A.6 Training-Budget Control

To assess whether the improvement comes merely from additional training, we compare the baseline and MoMI-RC under the same $30 + 2 + 20$-epoch schedule in Table 10. All proposed components are disabled in the budget-matched baseline. MoMI-RC achieves 84.2% mAP and 85.5% Rank-1, outperforming this baseline by 8.7 and 8.9 percentage points, respectively. These results indicate that the gains cannot be explained by the longer training schedule alone.

Table 9: Training cost on RGBNT201. The complete pipeline includes the 30-epoch baseline initialization.

Training phase	Epochs	Training cost		
		Time/batch (s)	Wall time (min)	Peak memory (GiB)
DINOv3-B ReID baseline	30	0.307	11.7	13.0
MoMI-RC Stage A	2	0.317	0.9	13.0
MoMI-RC Stage B	20	0.320	8.1	12.9
Full MoMI-RC pipeline	$30 + 2 + 20$	–	20.6	13.0

Table 10: Training-budget control on RGBNT201. Both models follow the same $30 + 2 + 20$-epoch schedule; all proposed components are disabled in the baseline. MoMI-RC retains an 8.7-point mAP advantage.

Method	Epochs	mAP	R-1	R-5	R-10
DINOv3-B ReID baseline (budget-matched)	$30 + 2 + 20$	75.5	76.6	85.0	90.2
MoMI-RC	$30 + 2 + 20$	**84.2**	**85.5**	**92.0**	**94.6**

A.7 Rank-list Visualization

Person rank-list. Figure 4 compares the Top-10 retrieval results of the DINOv3-B ReID baseline and MoMI-RC on RGBNT201. Following the visualization protocol of IDEA, every query and retrieved gallery sample is shown with its synchronized RGB, NIR, and TIR observations. For the two examples, MoMI-RC moves the first correct match from ranks 7 and 8 to rank 1, increasing per-query Average Precision (AP) from 13.5% and 23.1% to 100.0% and 98.9%, respectively. All displayed Top-10 results are correct after applying MoMI-RC, and the consistent corrections across the three gallery modalities show that the improvement is not tied to a particular sensor view.

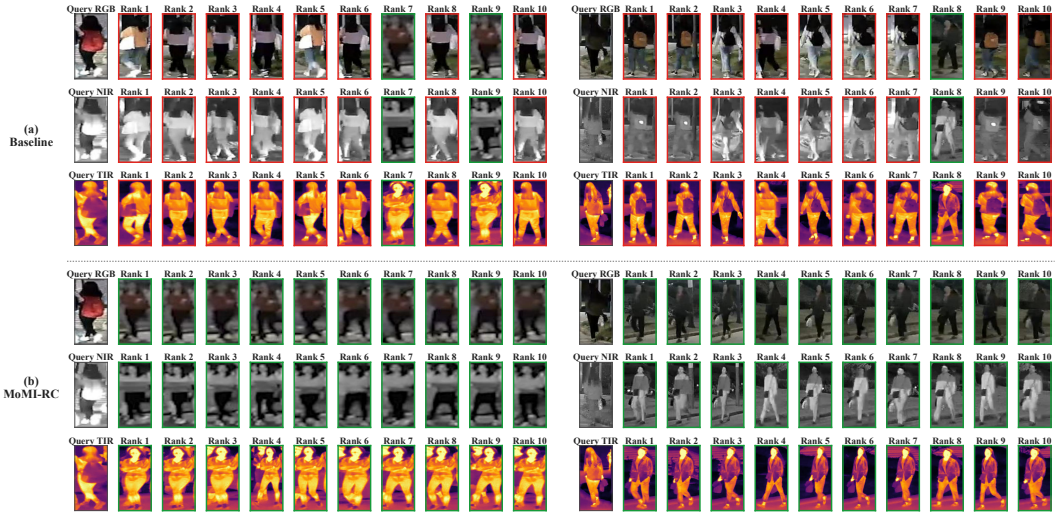

Figure 4: Multi-modal rank-list comparison on RGBNT201. The two query examples are arranged side by side: (a) DINOv3-B ReID baseline and (b) MoMI-RC. Green and red borders denote correct and incorrect identities, respectively.

Vehicle rank-list. Figure 5 further compares Top-5 retrieval results on MSVR310 under its official scene-filtered protocol. As in the person visualization, each vehicle is represented by synchronized RGB, NIR, and TIR observations. For the two vehicle queries, the baseline retrieves only three and two correct identities within the Top-5, whereas all five results are correct with MoMI-RC. Query AP increases from 66.9% and 49.1% to 93.7% and 74.0%, respectively. The corrections cover different viewpoints and thermal appearances, complementing the person examples and showing that the learned interactions generalize across object categories.

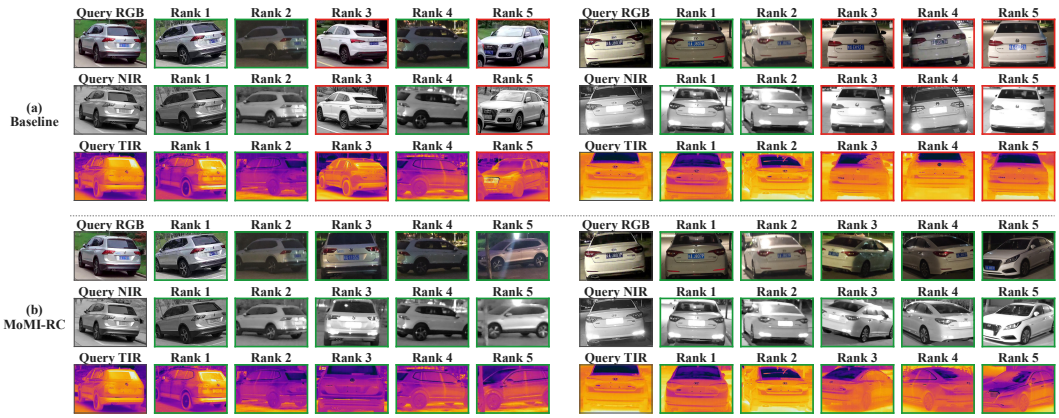

Figure 5: Multi-modal vehicle rank-list comparison on MSVR310. The two query examples are arranged side by side: (a) DINOv3-B ReID baseline and (b) MoMI-RC. Green and red borders denote correct and incorrect identities, respectively.

A.8 Feature Distribution Visualization

Figure 6 compares cosine-similarity distributions over the same balanced set of positive and negative query–gallery pairs. The positive–negative mean gap $\Delta_{\cos}$ increases from 0.250 for the baseline to 0.332 after adding MoMI and RC, and reaches 0.403 for the complete model, indicating greater separation of positive and negative pairs in the original descriptor space for the evaluated pairs.

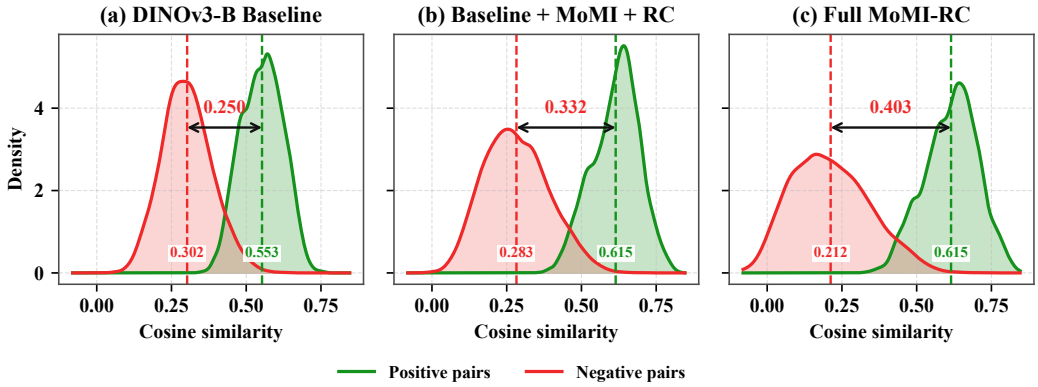

Figure 6: Cosine-similarity distributions of positive and negative query–gallery pairs on RGBNT201. Dashed lines indicate distribution means, and $\Delta_{\cos}$ denotes their difference. The same pair indices are used for all variants.

A.9 RC Weight Visualization

We visualize both RC levels on RGBNT201. The statistics are computed over all query samples using the final checkpoint, without retraining or test-time adaptation. Figure 7 shows RGB/NIR/TIR global anchors of $[1.026, 0.827, 1.178]$ on RGBNT201, with TIR receiving the largest global weight and NIR the smallest. The 5th–95th percentile intervals of the sample residual corrections are $[-12.0\%, 27.8\%]$, $[-14.4\%, 15.7\%]$, and $[-19.8\%, 8.6\%]$ for RGB, NIR, and TIR, respectively. Each interval contains both positive and negative corrections, showing that sample-wise adjustments can increase or decrease the contribution of a modality relative to its global anchor on RGBNT201.

Figure 8 compares two person examples from RGBNT201 before and after synthetic occlusion. A centered gray rectangle covers 75% of the RGB image for ID 275 or the TIR image for ID 280, while the other two modalities remain unchanged. For ID 275, the RGB residual correction changes from $+38.6\%$ to -8.3%, and its final weight decreases from 1.42 to 0.94. For ID 280, the TIR residual correction changes from $+13.9\%$ to -8.9%, and its final weight decreases from 1.34 to 1.07. In both examples, the weights of the unoccluded modalities increase. These paired comparisons illustrate that RC can downweight a synthetically occluded modality in response to changes in the input.

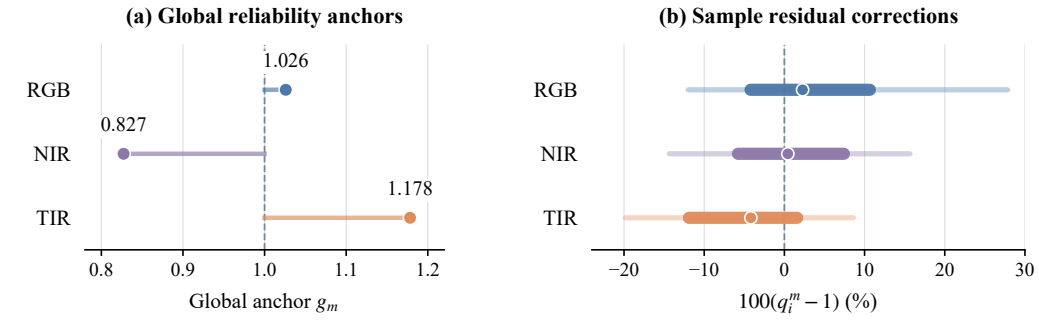

Figure 7: Two-level RC statistics on RGBNT201. (a) Global anchors learned from equal initialization; each segment starts at one. (b) Sample residual corrections over all query samples, where thin and thick segments denote the 5–95 and 25–75 percentile ranges, respectively, and circles denote medians.

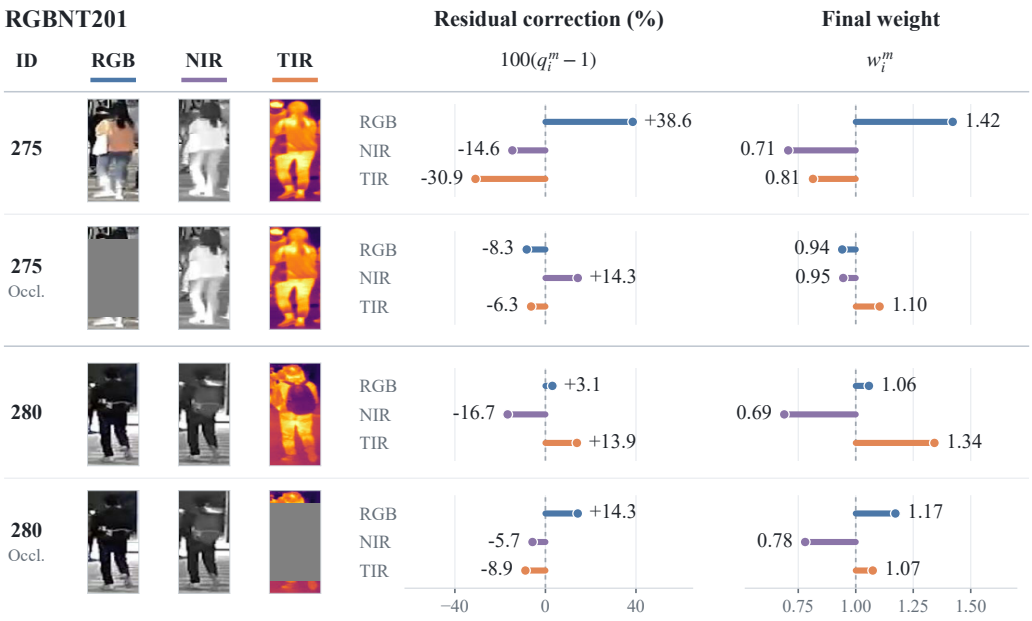

Figure 8: RC weight responses to synthetic occlusion on RGBNT201. Each original sample is followed by its occluded counterpart (Occl.): a central 75% gray mask on RGB for ID 275 or TIR for ID 280, with the other inputs unchanged. Each row shows the RGB/NIR/TIR inputs, sample residual corrections $100(q_i^m - 1)$, and final weights w_i^m. Dashed lines indicate zero correction and unit weight, respectively.