# Incentive Noise and Structural Prior Infusion for Multi-Modal Object Re-Identification

Weixiang Zhou[1*], Yuhao Wang[1*], Xingguo Xu[1], Weizhen Zhou[2], Zhixun Su[1†], Jinshan Pan[3], and Cong Wang[4]

[1] Dalian University of Technology, China
[2] New York University, USA
[3] Nanjing University of Science and Technology, China
[4] University of California, San Francisco, USA
{s20201162006, 924973292, xuxingguo}@mail.dlut.edu.cn
zxsu@dlut.edu.cn, {sdluran, supercong94}@gmail.com

**Abstract.** Multi-modal object Re-Identification (ReID) benefits from complementary information across heterogeneous imaging modalities. To further enrich semantic representation, text descriptions have recently been incorporated as an additional modality. However, recent vision-language approaches often treat text descriptions as clean, deterministic signals and overlook their inherent noise, including modality-mismatched phrases and semantically ambiguous expressions. Moreover, prevailing methods lack explicit mechanisms to reconcile fine-grained structural discrepancies between modalities, even after high-level semantic alignment. To address these challenges, we propose a novel framework centered on Positive-Incentive Noise ($\pi$-noise) and structured prompt modulation. First, the Semantic Cross-Modal Modulator harnesses task-aware $\pi$-noise, sampled from a distribution conditioned on both visual and text inputs, to perturb global tokens and enable semantics-guided cross-modal compensation. Second, the Structure-Aware Prompt Adapter injects learnable geometric priors via prompts to enhance spatial consistency. Third, the Context-Aware Sparse Fusion module distills structural context to guide adaptive fusion while shielding identity features from noisy local details. Experiments on three multi-modal ReID benchmarks demonstrate the effectiveness and robustness of our approach. The code is available at `https://github.com/zw-absin/INSPI`.

**Keywords:** Multi-modal object Re-Identification · Incentive noise · Structural prompt · Cross-modal modulation

## 1 Introduction

Object Re-Identification (ReID) aims to retrieve the same instance across non-overlapping camera views [9, 29, 32, 57, 58]. Most existing single-modality ap-

[*] Equal contribution.
[†] Corresponding author.

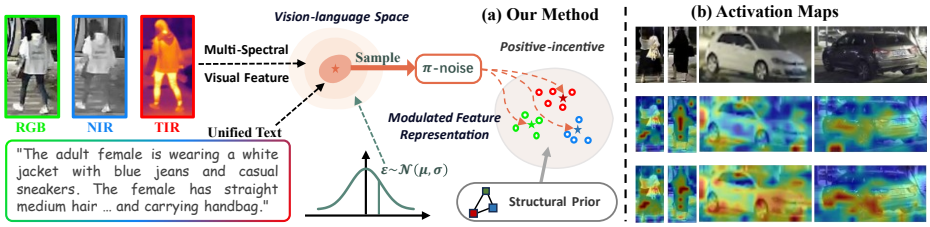



**Fig. 1:** Existing methods rely on uniform local modeling, such as key token selection, but ignore semantic guidance and structural information. In contrast, existing text-based methods directly fuse visual and textual features, assuming clean language input, yet propagate noise from ambiguous descriptions. (a) To address these, our method generates task-aware perturbations conditioned on multi-modal inputs, injects learnable structural priors via a shared adapter, and performs context-aware sparse fusion to enhance identity discrimination. (b) Feature activation before and after modulation: original images (top), raw activations (middle), enhanced heatmaps (bottom).

proaches rely solely on RGB imagery. These methods often perform poorly under adverse imaging conditions, such as occlusion, low illumination, or motion blur, where discriminative visual cues are significantly degraded or lost. To mitigate this limitation, multi-modal object ReID [5, 17–19, 42, 43, 52] has emerged as a promising direction by leveraging complementary information from multiple imaging modalities, such as infrared, depth, or event-based sensors, to learn more robust and discriminative feature representations in complex scenarios.

Recent advances in Contrastive Language-Image Pre-training (CLIP) [28] have further propelled multi-modal ReID by incorporating textual semantics [16, 21]. Moreover, the rise of foundation models and large language models has inspired new paradigms that incorporate auxiliary semantic signals, such as object masks or text descriptions, into ReID pipelines [6, 15, 40, 44] to enrich identity representation and facilitate semantic-aware cross-modal alignment. These methods expand the modality space and enhance robustness under degraded visual conditions. For instance, Wang et al. [40] integrate inverted text sequences into visual features via deformable aggregation, while Xu et al. [44] jointly leverage mask and text guidance for cross-modal modeling. Despite their effectiveness, these approaches treat text inputs as clean and deterministic anchors and ignore the inherent noise in natural language descriptions. Such noise includes modality-mismatched phrases, irrelevant padding tokens, and semantically ambiguous expressions, all of which can introduce harmful biases into representation learning and undermine alignment reliability.

Separately, multi-modal ReID faces another fundamental challenge arising from the heterogeneity of imaging sensors. The diversity across spectral modalities enables complementary perception but also introduces substantial background clutter from modality-specific artifacts and pixel-level inter-modal misalignment due to heterogeneous sensor characteristics [22]. Existing methods address these issues primarily through two strategies. The first refines feature representations via selective regional sampling, emphasizing discriminative regions

and suppressing unreliable ones [35, 50]. The second models modality-specific properties to preserve intrinsic discriminability while encouraging cross-modal complementarity [38,39]. For example, Zhang et al. [50] suppress background interference through token-level attention masking, while Wang et al. [38] decouple modality-specific features to retain individual characteristics. Nevertheless, these approaches typically assume uniform alignment across modalities and rely heavily on global constraints or model adaptivity, thereby overlooking fine-grained structural discrepancies that persist even after high-level semantic alignment.

To address these limitations, we propose a novel framework for multi-modal object ReID, centered on Positive-Incentive Noise ($\pi$-noise) [20, 49] and structured prompt modulation, as illustrated in Figure 1. Our core insight is that noisy textual semantics can serve as a source of beneficial uncertainty when properly regularized. In contrast to prior work [5] that suppresses unreliable visual regions by modeling visual uncertainty, our approach harnesses textual uncertainty to enhance feature learning. Specifically, we generate a task-aware noise distribution conditioned on both visual content and text captions. A sampled perturbation from this distribution is injected into the global token, yielding a semantically reinforced representation. This $\pi$-noise mechanism not only mitigates the adverse effects of textual noise but also actively promotes identity-discriminative feature refinement. Building on this principle, our framework integrates three synergistic components. First, the Semantic Cross-Modal Modulator (SCM) leverages $\pi$-noise-enhanced tokens as dynamic controllers to enable semantics-guided alignment and cross-modal complementarity. Second, the Structure-Aware Prompt Adapter (SPA) initializes modality-specific structural prompts to preserve intrinsic geometric cues and aligns them through a shared adapter to encourage cross-modal structural consistency. Third, the Context-Aware Sparse Fusion (CASF) distills self-supervised structural knowledge into a clean contextual signal, which governs sparse expert routing. This design achieves adaptive fusion while shielding identity features from direct exposure to noisy local details. Extensive experiments on three multi-modal ReID benchmarks demonstrate the effectiveness and generalization capability of our approach. Our main contributions are summarized as follows:

- We introduce Positive-Incentive Noise ($\pi$-noise), a novel representation learning paradigm that transforms noisy textual semantics into a beneficial perturbation for semantic reinforcement in multi-modal object ReID.
- We propose a unified multi-modal ReID framework integrating three components: the Semantic Cross-Modal Modulator (SCM) leveraging $\pi$-noise for semantic modulation; the Structure-Aware Prompt Adapter (SPA) injecting geometric priors via learnable prompts; and the Context-Aware Sparse Fusion (CASF) performing noise-resilient fusion via expert routing.
- Extensive experiments on three public multi-modal object ReID datasets, including RGBNT201, RGBNT100, and MSVR310, demonstrate the superiority of our method over state-of-the-art approaches.

## 2 Related Work

### 2.1 Multi-Modal Object Re-Identification

Multi-modal object ReID aims to match objects across camera views by fusing complementary cues from heterogeneous modalities. Unlike single-modality approaches, recent methods emphasize cross-modal interaction mechanisms that preserve modality-specific characteristics while enforcing semantic consistency [4, 37,39,45,51]. For example, Wang et al. [37] use Mamba to model intra- and inter-modal dependencies. Feng et al. [5] embed a Mixture-of-Experts (MoE) module in Vision Transformers to capture modality-specific features. Wang et al. [39] further decouple features under varying visual quality. Concurrently, alternative interaction strategies such as graph-based reasoning [36], parameter-efficient tuning [14], and hierarchical token alignment [23] have also been explored. Beyond fusing visual modalities, recent work has explored leveraging textual semantics to further bridge the semantic gap and provide high-level identity priors. The emergence of vision-language models like CLIP [28] has motivated the integration of text descriptions into multi-modal object ReID to enrich semantic context and align cross-modal representations [15,21,40,44]. These approaches typically treat text as a clean semantic anchor. However, they often ignore the inherent noise in text annotations, such as ambiguity, incompleteness, or misalignment, which can degrade representation learning when fused directly. To address this issue, our method embraces textual noise as a source of beneficial uncertainty through Positive-Incentive Noise, which actively promotes semantic reinforcement.

### 2.2 The Constructive Role of Noise

Contrary to the conventional assumption that noise is always harmful, recent studies suggest that certain types of noise can be beneficial. The concept of Positive-Incentive Noise was introduced to characterize perturbations that reduce task entropy, thereby simplifying the learning problem [20, 49]. Under this framework, even random noise may act as $\pi$-noise if it lowers the effective complexity of a task. This perspective has inspired methods that leverage noise as a regularizer or structural guide. For instance, Huang et al. [11] use structured noise to improve vision-language alignment. Jiang et al. [12] propose a mixture-of-noise mechanism for class incremental learning, where task-specific noise is injected into intermediate features to suppress irrelevant patterns and mitigate parameter drift in pre-trained backbones. Together, these works demonstrate that properly designed noise can serve as a constructive inductive bias. Nevertheless, the potential of $\pi$-noise in multi-modal object ReID, especially for harnessing noisy text inputs, has not been explored. In this work, we bridge this gap by generating $\pi$-noise through cross-modal fusion of visual and textual semantics and leveraging it to enhance feature representations.

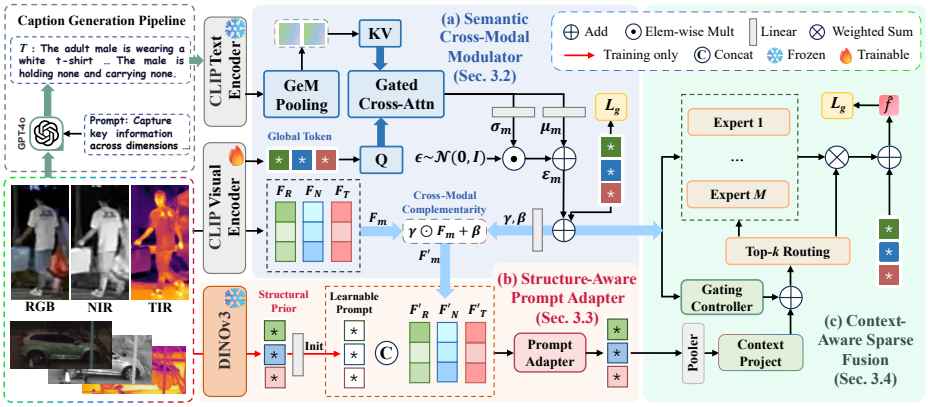



**Fig. 2:** Overview of the proposed framework. (a) The Semantic Cross-Modal Modulator (SCM) samples beneficial perturbations from a distribution conditioned on both visual and text inputs to enable semantics-guided cross-modal compensation. (b) The Structure-Aware Prompt Adapter (SPA) injects learnable structural priors to promote cross-modal structural alignment. (c) The Context-Aware Sparse Fusion (CASF) module performs multi-modal feature fusion via sparse routing guided by structural context. With these components, our method achieves robust identity matching under noisy text descriptions and heterogeneous imaging conditions.

## 3 Methodology

As illustrated in Figure 2, we propose a novel framework for multi-modal object ReID comprising three core components. First, the Semantic Cross-Modal Modulator (SCM) introduces the concept of Positive-Incentive Noise ($\pi$-noise) into ReID by sampling task-aware perturbations from a distribution conditioned on both visual and text inputs, thereby enhancing global representations and enabling semantics-guided cross-modal compensation. Second, the Structure-Aware Prompt Adapter (SPA) injects modality-specific, learnable structural priors to enrich geometric awareness while promoting cross-modal structural alignment through a shared adapter. Third, the Context-Aware Sparse Fusion (CASF) module leverages structural context to modulate sparse routing during multi-modal feature fusion. We detail each component of our framework below.

### 3.1 Feature Initialization

Following recent vision-language ReID approaches [40, 44], we employ the pre-trained CLIP encoder [28] to extract features from multi-spectral images and modality-agnostic text captions. This yields visual features $F_v = \{f_m, F_m\}$ and text features $F_t$, where $m \in \{R, N, T\}$ denotes the RGB, Near-Infrared (NIR), and Thermal Infrared (TIR) modalities, respectively. For each modality $m$, $f_m \in \mathbb{R}^D$ represents the visual global token, and $F_m \in \mathbb{R}^{N_p \times D}$ denotes the sequence of $N_p$ visual patch tokens. The text input is encoded as $F_t \in \mathbb{R}^{N_t \times D}$, where

$N_t$ is the number of text tokens. Then, these features are fed into our proposed modules for further refinement.

### 3.2 Semantic Cross-Modal Modulator

To mitigate the adverse effects of noisy text cues while enabling semantic-aware cross-modal alignment and complementarity, we propose the Semantic Cross-Modal Modulator (SCM). Inspired by the concept of Positive-Incentive Noise ($\pi$-noise) [11,20], we reinterpret noise in representation learning. Specifically, by leveraging both textual and visual features and applying proper regularization techniques, we can generate a structured perturbation. This perturbation is designed to reinforce identity-discriminative semantics, turning potential noise into a positive incentive for enhancing cross-modal alignment and complementarity.

Specifically, as illustrated in Figure 2(a), the input text sequence is first condensed into a global text feature $f_t \in \mathbb{R}^{N_t' \times D}$ via GeM pooling [27] after removing padding tokens. For each available visual modality $m \in \{R, N, T\}$ with global token $f_m \in \mathbb{R}^D$, we employ a Gated Attention-based Cross-Attention (GCA) module [26] to fuse textual and visual semantics with the following equation:

$$f_m' = \mathrm{GCA}(f_m, f_t) \in \mathbb{R}^D. \tag{1}$$

This fused representation is then projected to obtain a Gaussian distribution:

$$\mu_m = L_\mu(f_m'), \quad \sigma_m = L_\sigma(f_m') \in \mathbb{R}^D, \tag{2}$$

where $\mu_m$ represents the mean vector, and $\sigma_m$ denotes the standard deviation vector. Then, we sample structured noise via the reparameterization trick:

$$\varepsilon_m = \mu_m + \sigma_m \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where $\epsilon$ is a standard normal random vector. Crucially, this perturbation is not random but semantically conditioned. It is shaped by both visual content and textual semantics, actively reshaping features toward identity-discriminative structures. As a result, it establishes an information-gain mechanism rather than introducing task-irrelevant variations.

The resulting semantically enhanced global token is constructed as:

$$\tilde{f}_m = f_m + \varepsilon_m \in \mathbb{R}^D. \tag{4}$$

These enhanced tokens $\tilde{f}_m$ serve as semantic controllers for cross-modal modulation. Given a source modality's local features $F_i \in \mathbb{R}^{N_p \times D}$ and a target modality's enhanced token $\tilde{f}_j$, SCM applies an affine transformation to $F_i$:

$$\gamma_j = \mathrm{MLP}_\gamma(\tilde{f}_j), \quad \beta_j = \mathrm{MLP}_\beta(\tilde{f}_j) \in \mathbb{R}^D, \tag{5}$$

$$F_i' = F_i \odot \gamma_j + \beta_j \in \mathbb{R}^{N_p \times D}, \quad i, j \in \{R, N, T\}. \tag{6}$$

This operation reinterprets the content of modality $i$ through the semantic lens of modality $j$, enabling semantics-guided feature compensation. Notably, since modulation relies only on the enhanced tokens of available modalities, our design naturally supports robust inference under arbitrary modality-missing patterns.

### 3.3 Structure-Aware Prompt Adapter

In multi-modal ReID, multi-spectral images exhibit substantial appearance discrepancies yet share consistent high-level structural semantics. However, prevailing vision encoders such as CLIP, while offering strong discriminative power, inadequately model cross-modal structural consistency, particularly in non-visible modalities where geometric priors are prone to degradation. This limitation arises from the training objective of large-scale contrastive models, which emphasizes global semantic alignment but often sacrifices fine-grained structural fidelity in dense representations [31].

To address this issue, we propose the Structure-Aware Prompt Adapter (SPA), a training-inference decoupled module that injects structure-aware global priors into learnable prompts to guide cross-modal feature alignment and interaction within a unified semantic space. Specifically, during training, we initialize modality-specific learnable prompts using the global tokens extracted from DINOv3 [31], which are known to encode rich structural information. Critically, DINOv3 is used only once at initialization and is entirely detached during both training and inference. This design leverages the frozen model's superior structural awareness without incurring additional computational overhead. As optimization proceeds, these prompts internalize structural priors and serve dual roles: acting as modality identifiers and structural anchors. These prompts are then linearly projected into the CLIP feature space and prepended to the visual token sequence $F'_m$ of each modality, forming an augmented sequence:

$$\overline{F}_m = [\tilde{f}_m, F'_m] \in \mathbb{R}^{(N_p+1)\times D}, \tag{7}$$

where $[\cdot]$ denotes concatenation. The three modality-specific sequences are then stacked into a unified input and processed by a shared adapter block $\mathcal{A}$ as follows:

$$\overline{F} = [\overline{F}_R, \overline{F}_N, \overline{F}_T], \quad \hat{F} = \mathcal{A}(\overline{F}) \in \mathbb{R}^{3(N_p+1)\times D}, \tag{8}$$

where $\mathcal{A}$ consists of Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN) [3]. This shared adapter jointly performs cross-modal fusion and dynamic prompt adaptation, producing context-rich representations that preserve structural consistency while enabling semantic interaction across modalities.

### 3.4 Context-Aware Sparse Fusion

Omitting structural modeling leads to the loss of geometric and contour cues, while directly fusing raw local features risks contamination from low-quality images and compromises robustness. To reconcile this dilemma, we propose Context-Aware Sparse Fusion (CASF), which distills self-supervised structural priors into a compact context representation that is both structure-aware and semantically coherent. This context does not contribute directly to the final identity embedding. Instead, it serves as a routing signal to dynamically modulate the fusion of multi-modal global tokens.

Specifically, we first extract the global components $\hat{F}_g$ from $\hat{F}$ and apply average pooling $\mathcal{P}$ to obtain a coarse structural summary. The context signal $\hat{f}_c \in \mathbb{R}^D$ is then computed as:

$$\hat{f}_c = L_{\text{CP}}(\mathcal{P}(\hat{F}_g)) + L_{\text{GC}}(\tilde{f}_g), \quad \tilde{f}_g = [\tilde{f}_R, \tilde{f}_N, \tilde{f}_T], \tag{9}$$

where $L_{\text{CP}}$ (Context Projection) and $L_{\text{GC}}$ (Gating Controller) are linear projections. This context modulates the gating logits of a sparsely-gated Mixture-of-Experts (MoE) layer [30]. The input to the MoE is the global tokens $\tilde{f}_g \in \mathbb{R}^{3D}$. Only the top-$k$ experts are activated per sample, and their outputs are combined via learned weights to produce a fused representation $f_{\text{fuse}} \in \mathbb{R}^D$. Because expert selection is guided by structural context rather than raw features, the fusion process becomes inherently resilient to noise and modality degradation. Moreover, since the identity representation is constructed without direct exposure to local details, CASF achieves a principled balance between discriminability and robustness. The final discriminative feature for each modality $m \in \{R, N, T\}$ is obtained by residual addition:

$$\hat{f}_m = f_m + f_{\text{fuse}}, \quad \hat{f} = [\hat{f}_R, \hat{f}_N, \hat{f}_T]. \tag{10}$$

### 3.5 Objective Functions

As shown in Figure 2, the network is trained with the following loss function:

$$\mathcal{L} = \mathcal{L}_g([f_R, f_N, f_T]) + \mathcal{L}_g(\hat{f}) + 0.01 \cdot \mathcal{L}_{\text{aux}}, \tag{11}$$

where $\mathcal{L}_g(\cdot)$ denotes the sum of label smoothing cross-entropy loss [34] and triplet loss [10]. This formulation applies $\mathcal{L}_g$ separately to the modality-specific features $[f_R, f_N, f_T]$ and to the CASF-fused representation $\hat{f}$, providing dual supervision that promotes discriminative representations at both stages. $\mathcal{L}_{\text{aux}}$ is an auxiliary load-balancing loss from the MoE layer, which encourages balanced expert utilization. The above losses ensure effective joint optimization of feature learning.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

We evaluate our method on 3 public multi-modal ReID benchmarks. RGBNT201 [54] is a large-scale person ReID dataset containing 201 identities captured across four non-overlapping cameras with synchronized RGB, Near-Infrared (NIR), and Thermal Infrared (TIR) images for each identity. It includes 4787 aligned images per modality and exhibits significant challenges such as viewpoint changes, occlusion, cluttered backgrounds, and imaging degradations due to adverse lighting or weather conditions. For vehicle ReID, we use RGBNT100 [13], which comprises 17250 aligned RGB–NIR–TIR triplets across 100 vehicles and features viewpoint variation, low illumination, and partial occlusion. Additionally, we

**Table 1:** Performance comparison on RGBNT201. The best and second-best results are shown in **bold** and underlined, respectively. Methods with † are CLIP-based, those with ∗ are ViT-based, and the rest are CNN-based.

| | Methods | RGBNT201 | | | |
|---|---|---|---|---|---|
| | | mAP | R-1 | R-5 | R-10 |
| Single | $OSNet_{2019}$ [57] | 25.4 | 22.3 | 35.1 | 44.7 |
| | $CAL_{2021}$ [29] | 27.6 | 24.3 | 36.5 | 45.7 |
| | $PCB_{2018}$ [33] | 32.8 | 28.1 | 37.4 | 46.9 |
| Multi-Modal | $HAMNet_{2020}$ [13] | 27.7 | 26.3 | 41.5 | 51.7 |
| | $PFNet_{2021}$ [54] | 38.5 | 38.9 | 52.0 | 58.4 |
| | $IEEE_{2022}$ [42] | 47.5 | 44.4 | 57.1 | 63.6 |
| | $DENet_{2023}$ [52] | 42.4 | 42.2 | 55.3 | 64.5 |
| | $LRMM_{2025}$ [43] | 52.3 | 53.4 | 64.6 | 73.2 |
| | $UniCat^{*}_{2023}$ [1] | 57.0 | 55.7 | - | - |
| | $HTT^{*}_{2024}$ [41] | 71.1 | 73.4 | 83.1 | 87.3 |
| | $TOP\text{-}ReID^{*}_{2024}$ [38] | 72.3 | 76.6 | 84.7 | 89.4 |
| | $EDITOR^{*}_{2024}$ [50] | 66.5 | 68.3 | 81.1 | 88.2 |
| | $RSCNet^{*}_{2024}$ [47] | 68.2 | 72.5 | - | - |
| | $WTSF\text{-}ReID^{*}_{2025}$ [48] | 67.9 | 72.2 | 83.4 | 89.7 |
| | $DESANet^{*}_{2025}$ [2] | 74.6 | 77.6 | 87.1 | 91.3 |
| | $PromptMA^{\dagger}_{2025}$ [51] | 78.4 | 80.9 | 87.0 | 88.9 |
| | $MambaPro^{\dagger}_{2025}$ [37] | 78.9 | 83.4 | 89.8 | 91.9 |
| | $DeMo^{\dagger}_{2025}$ [39] | 79.0 | 82.3 | 88.8 | 92.0 |
| | $IDEA^{\dagger}_{2025}$ [40] | 80.2 | 82.1 | 90.0 | 93.3 |
| | **Ours**† | **80.6** | **83.9** | **91.6** | **93.4** |

**Table 2:** Performance comparison on RGBNT100 and MSVR310.

| | Methods | RGBNT100 | | MSVR310 | |
|---|---|---|---|---|---|
| | | mAP | R-1 | mAP | R-1 |
| Single | $PCB_{2018}$ [33] | 57.2 | 83.5 | 23.2 | 42.9 |
| | $OSNet_{2019}$ [57] | 75.0 | 95.6 | 28.7 | 44.8 |
| | $TransReID^{*}_{2021}$ [9] | 75.6 | 92.9 | 18.4 | 29.6 |
| Multi-Modal | $HAMNet_{2020}$ [13] | 74.5 | 93.3 | 27.1 | 42.3 |
| | $PFNet_{2021}$ [54] | 68.1 | 94.1 | 23.5 | 37.4 |
| | $GAFNet_{2022}$ [7] | 74.4 | 93.4 | - | - |
| | $GPFNet_{2023}$ [8] | 75.0 | 94.5 | - | - |
| | $CCNet_{2023}$ [55] | 77.2 | 96.3 | 36.4 | 55.2 |
| | $LRMM_{2025}$ [43] | 78.6 | 96.7 | 36.7 | 49.7 |
| | $GraFT^{*}_{2023}$ [46] | 76.6 | 94.3 | - | - |
| | $UniCat^{*}_{2023}$ [1] | 79.4 | 96.2 | - | - |
| | $PHT^{*}_{2023}$ [25] | 79.9 | 92.7 | - | - |
| | $HTT^{*}_{2024}$ [41] | 75.7 | 92.6 | - | - |
| | $TOP\text{-}ReID^{*}_{2024}$ [38] | 81.2 | 96.4 | 35.9 | 44.6 |
| | $EDITOR^{*}_{2024}$ [50] | 82.1 | 96.4 | 39.0 | 49.3 |
| | $RSCNet^{*}_{2024}$ [47] | 82.3 | 96.6 | 39.5 | 49.6 |
| | $FACENet^{*}_{2025}$ [53] | 81.5 | 96.9 | 36.2 | 54.1 |
| | $WTSF\text{-}ReID^{*}_{2025}$ [48] | 82.2 | 96.5 | 39.2 | 49.1 |
| | $DESANet^{*}_{2025}$ [2] | 82.1 | 97.4 | 39.2 | 47.8 |
| | $PromptMA^{\dagger}_{2025}$ [51] | 85.3 | 97.4 | 55.2 | 64.5 |
| | $MambaPro^{\dagger}_{2025}$ [37] | 83.9 | 94.7 | 47.0 | 56.5 |
| | $DeMo^{\dagger}_{2025}$ [39] | 86.2 | 97.6 | 49.2 | 59.8 |
| | $IDEA^{\dagger}_{2025}$ [40] | 87.2 | 96.5 | 47.0 | 62.4 |
| | **Ours**† | **89.9** | **98.2** | **65.0** | **77.2** |

include MSVR310 [55], a compact but highly challenging vehicle dataset with 2087 triplets collected under extreme conditions, making it a stringent test of robustness. Following standard practice in cross-modal person ReID, we report mean average precision (mAP) and Rank-$K$ accuracy for $K \in \{1, 5, 10\}$.

### 4.2 Implementation Details

Our framework is implemented in PyTorch and trained on an NVIDIA A6000 GPU, using CLIP [28] as both the visual and text backbone. Input images are resized to 256×128 for RGBNT201 and to 128×256 for RGBNT100 and MSVR310. Data augmentation includes random horizontal flipping, random cropping, and random erasing [56] to enhance model robustness. The mini-batch size is set to 64 for RGBNT201 and MSVR310, and 128 for RGBNT100, with 8 images per identity for RGBNT201 and MSVR310 and 16 images per identity for RGBNT100. All models are trained for 50 epochs. We use the Adam optimizer to fine-tune all learnable parameters, starting with an initial learning rate of $3.5 \times 10^{-6}$, which is linearly decayed to $3.5 \times 10^{-7}$ over the course of training. Text prompts are generated using the caption generation pipeline proposed by Xu et al. [44].

**Table 3:** Incremental ablation study of the main modules.

| Index | Modules | | | RGBNT201 | | MSVR310 | | FLOPs | Params |
|---|---|---|---|---|---|---|---|---|---|
| | SCM | SPA | CASF | mAP | R-1 | mAP | R-1 | (G) | (M) |
| A | ✗ | ✗ | ✗ | 70.3 | 71.9 | 40.4 | 56.0 | 34.27 | 86.41 |
| B | ✗ | ✗ | ✔ | 76.5 | 80.3 | 55.8 | 66.3 | 34.29 | 92.73 |
| C | ✗ | ✔ | ✔ | 78.6 | 82.2 | 61.1 | 72.9 | 35.66 | 95.75 |
| D | ✔ | ✔ | ✔ | **80.6** | **83.9** | **65.0** | **77.2** | 38.68 | 96.23 |

### 4.3 Comparison with State-of-the-Art Methods

**Multi-Modal Person ReID.** Table 1 compares our method (Ours$^{\dagger}$) with existing multi-modal approaches on RGBNT201. The results confirm that leveraging complementary modalities consistently improves upon both single-modality baselines and CNN-based methods. Our method achieves 80.6% mAP and 83.9% Rank-1 accuracy, outperforming TOP-ReID$^{*}$ by 8.3% in mAP and 7.3% in Rank-1. Notably, compared to recent CLIP-based methods such as MambaPro$^{\dagger}$ and DeMo$^{\dagger}$, our approach demonstrates consistent gains across all metrics. This advantage stems from two key aspects. First, the $\pi$-noise mechanism regularizes unreliable textual semantics into beneficial perturbations rather than treating them as deterministic anchors. Second, the structure-aware prompt modulation enforces geometric consistency across heterogeneous sensors, which is critical in scenes where thermal signatures lack fine-grained textures but retain shape cues.
**Multi-Modal Vehicle ReID.** We further evaluate our method on vehicle ReID benchmarks RGBNT100 and MSVR310, as shown in Table 2. The strong performance on the large-scale RGBNT100 dataset validates the generalization capability of our framework. On RGBNT100, Ours$^{\dagger}$ attains 89.9% mAP, surpassing TOP-ReID$^{*}$ by 8.7% and EDITOR$^{*}$ by 7.8%. On the more challenging MSVR310 dataset, our method achieves 65.0% mAP and 77.2% Rank-1 accuracy, representing substantial improvements of 18.0% and 14.8% over IDEA$^{\dagger}$. These gains highlight the robustness of our framework in handling real-world complexities. Specifically, the CASF component plays a crucial role in MSVR310 by adaptively routing features through experts conditioned on structural context, thereby suppressing background interference from dense urban scenes. Meanwhile, the SCM ensures that even when textual prompts contain irrelevant attributes, the injected $\pi$-noise prevents feature corruption and maintains identity fidelity. These mechanisms allow our model to generalize well across datasets exhibiting diverse degrees of modality heterogeneity and environmental noise.
**Effect of Key Modules.** As shown in Table 3, we conduct ablation studies on RGBNT201 and MSVR310 to evaluate the contribution of our three core components proposed in Sec. 3, all of which play a critical role in boosting overall performance. Taking RGBNT201 as an example, Exp. B validates the role of CASF in mitigating local noise. By routing features through sparse experts conditioned on structural context rather than fusing raw patch tokens directly, CASF avoids propagating unreliable details from cluttered backgrounds. This

**Table 4:** Comparison of computational efficiency on RGBNT201.

| Model | FLOPs (G) | mAP |
|---|---|---|
| TOP-ReID* [38] | 35.51 | 72.3 |
| IDEA† [40] | 43.73 | 80.2 |
| **Ours†** | **38.68** | **80.6** |

**Table 5:** Comparison of interaction mechanisms in SCM on RGBNT201.

| Methods | mAP | R-1 |
|---|---|---|
| w/o $\pi$-noise modulation | 78.6 | 82.1 |
| w/ shared noise | 79.2 | 80.6 |
| w/ CA | 79.4 | 80.3 |
| **w/ gated CA (Ours)** | **80.6** | **83.9** |

**Table 6:** Ablation study on stochastic perturbation mechanisms.

| Dataset | w/o $\pi$ | | w/ Gauss | | w/ uncond. | | w/ dropout | | Ours | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| RGBNT201 | 78.6 | 82.1 | 76.1 | 79.6 | 74.8 | 77.5 | 75.7 | 79.2 | **80.6** | **83.9** |
| MSVR310 | 61.4 | 73.4 | 60.1 | 73.1 | 59.3 | 70.1 | 60.9 | 73.3 | **65.0** | **77.2** |

design yields a substantial improvement of +6.2% mAP and +8.4% Rank-1, confirming that adaptive fusion guided by clean structural signals is critical for robust identity representation. Building on this, Exp. C demonstrates that the SPA further refines cross-modal alignment. SPA injects DINOv3-derived geometric priors via learnable prompts and aligns them across modalities through a shared adapter. The resulting gains (78.6% mAP, 82.2% Rank-1) highlight that explicit structural guidance compensates for texture degradation, enabling more consistent feature matching under modality heterogeneity. Finally, Exp. D shows that the SCM unlocks the full potential of textual semantics. Instead of treating text as a fixed anchor, SCM leverages Positive-Incentive Noise ($\pi$-noise) to regularize textual uncertainty into a beneficial perturbation. This leads to the best overall performance (80.6% mAP, 83.9% Rank-1), proving that semantic signals, when properly modulated, can actively enhance discriminability rather than introduce bias. Similar consistent improvements are also observed on MSVR310.

**Computational Efficiency and FLOPs Analysis.** We analyze the computational efficiency of our method on RGBNT201 in Table 3 and Table 4. Our model achieves the best performance (80.6% mAP) with 38.68 GFLOPs. Specifically, our text branch configuration and FLOPs calculation follow the protocol established in prior work [40, 44]. The reduced FLOPs compared to IDEA arise from our shared text encoder across all three modalities, whereas IDEA employs a separate text encoder for each spectral modality. This efficiency stems from our lightweight cross-modal interaction design: the majority of computation is dominated by the shared CLIP backbone, and we deliberately avoid introducing additional heavy fusion modules or modality-specific transformers.

### 4.4 Further Analysis

**Semantic Cross-Modal Modulator.** We present a detailed ablation study on the internal design choices of SCM in Table 5. The results show that removing $\pi$-noise modulation (w/o $\pi$-noise modulation) leads to 78.6% mAP and 82.1%

**Table 7:** Comparison of prompt configurations in SPA on RGBNT201.

| Methods | mAP | R-1 |
|---|---|---|
| w/ linear | 78.0 | 80.5 |
| w/o adapter | 79.1 | 81.3 |
| w/o prompt | 77.0 | 81.2 |
| prompt w/o DINO init | 78.0 | 81.3 |
| **prompt w/ DINO init (Ours)** | **80.6** | **83.9** |

**Table 8:** Comparison of routing strategies in CASF on RGBNT201.

| Methods | mAP | R-1 |
|---|---|---|
| w/ CA | 78.1 | 79.2 |
| w/o MoE | 78.5 | 80.6 |
| input w/ $[f_R, f_N, f_T]$ | 78.6 | 81.8 |
| **input w/ $\tilde{f}_g$ (Ours)** | **80.6** | **83.9** |

**Table 9:** Arbitrary modality-missing patterns on RGBNT201 and MSVR310. R, N, and T respectively denote RGB, NIR, and TIR. M(X) means missing modality X.

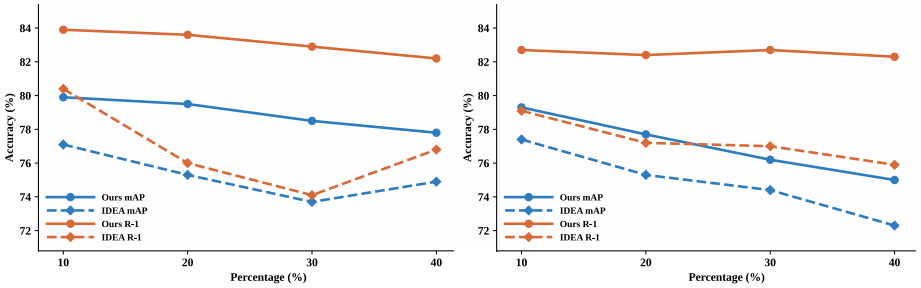
| Dataset | Method | M(R) | | M(N) | | M(T) | | M(RN) | | M(RT) | | M(NT) | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 | mAP | R-1 |
| RGBNT201 | DeMo | 63.3 | 65.3 | **72.6** | **75.7** | 56.2 | 54.1 | **45.6** | **46.5** | 26.3 | 24.9 | **40.3** | 38.5 | 50.7 | 50.8 |
| | Ours | **64.5** | **65.9** | 71.5 | 74.4 | **62.9** | **62.7** | 43.2 | 46.2 | **37.8** | **37.7** | 39.5 | **41.2** | **53.2** | **54.7** |
| MSVR310 | DeMo | 35.9 | 50.3 | **42.0** | **58.4** | 43.1 | **59.7** | **15.3** | **28.8** | 27.2 | 43.5 | 35.3 | 49.6 | 33.1 | 48.4 |
| | Ours | **43.8** | **57.2** | 40.1 | 54.7 | **45.0** | 58.7 | 11.3 | 21.3 | **32.2** | **49.1** | **41.6** | **56.9** | **35.7** | **49.6** |

Rank-1, which are 2.0% and 1.8% lower than our full model. We also test a variant with a shared $\pi$-noise across modalities (w/ shared noise). While it benefits from noise-induced perturbation, it cannot adapt modulation to each modality's characteristics. Moreover, using standard cross-attention (w/ CA) instead of the gated mechanism yields inferior performance, indicating that gating conditioned on visual global features enables more effective text-image feature fusion.

**Stochasticity Analysis.** Table 6 verifies that the gain of $\pi$-noise stems from task-aware structure rather than mere stochasticity. Compared to the deterministic baseline (w/o $\pi$), naive Gaussian noise (w/ Gauss), unconditional sampling (w/ uncond.), and dropout (w/ dropout) all degrade performance, confirming that unstructured randomness harms the aligned CLIP feature space. In contrast, our semantically-conditioned $\pi$-noise achieves the best results.

**Structure-Aware Prompt Adapter.** To examine the effectiveness of structural prior injection in SPA, we compare different designs in Table 7. Replacing the adapter with a simple linear layer (w/ linear) leads to worse performance. This indicates that a naive linear transformation fails to properly modulate features with the structural prior and may even introduce harmful interference. Bypassing the adapter and directly using the prior for routing (w/o adapter) reduces mAP and Rank-1 by 1.5% and 2.6%, respectively, highlighting the adapter's role in stable feature transformation. Using only CLIP patch tokens without structural prompts (w/o prompt) results in 77.0% mAP, while initializing prompts randomly instead of with DINOv3 (prompt w/o DINO init) achieves 78.0% mAP, both below the full model (Ours). This confirms that lightweight DINOv3 initialization effectively injects structural information.

**Context-Aware Sparse Fusion.** As shown in Table 8, we evaluate the routing-based fusion mechanism in CASF by comparing it against two variants: (1)




**(a) Text Mismatch** **(b) Text Corruption**

**Fig. 3:** Comparison under text corruption and mismatch on RGBNT201.

**Table 10:** Number of GeM pooling branches on RGBNT201.

| $\lvert N_t' \rvert$ | mAP | R-1 | R-5 | R-10 |
|---|---|---|---|---|
| 1 | 79.7 | 83.9 | 92.0 | 93.8 |
| 2 | 80.6 | 83.9 | 91.6 | 93.4 |
| 3 | 80.4 | 83.3 | 91.0 | 93.4 |
| 4 | 80.3 | 83.9 | 91.3 | 94.6 |
| 5 | 79.1 | 81.3 | 89.2 | 92.0 |

**Table 11:** Comparison of the number of trainable parameters among different methods.

| Methods | Params (M) | RGBNT201 | | MSVR310 | |
|---|---|---|---|---|---|
| | | mAP | R-1 | mAP | R-1 |
| CLIP Baseline† | 86.41 | 70.3 | 71.9 | 40.4 | 56.0 |
| TOP-ReID* [38] | 324.53 | 72.3 | 76.6 | 35.9 | 44.6 |
| EDITOR* [50] | 118.55 | 66.5 | 68.3 | 39.0 | 49.3 |
| WTSF-ReID* [48] | 143.60 | 67.9 | 72.2 | 39.2 | 49.1 |
| DeMo† [39] | 98.79 | 79.0 | 82.3 | 49.2 | 59.8 |
| **Ours†** | **96.23** | **80.6** | **83.9** | **65.0** | **77.2** |

replacing the MoE router with a linear fusion layer (w/o MoE), and (2) using cross-attention to directly fuse modality-specific features (w/ CA). Both variants underperform the full model, indicating that direct feature blending without selective routing introduces cross-modal interference and degrades discriminative cues. Additionally, restricting the MoE inputs to modality-specific global tokens (w/ $[f_R, f_N, f_T]$) while excluding incentive noise from routing yields lower mAP and Rank-1. This validates our routing design and the efficacy of incorporating noise sampled from semantic-aware distributions.

**Robustness to Missing Modalities.** Table 9 evaluates arbitrary modality-missing patterns on RGBNT201 and MSVR310. Our method achieves a higher average performance than DeMo. Notably, under the challenging M(RT) setting, our method achieves 37.8% mAP while DeMo achieves 26.3% on RGBNT201. This demonstrates stronger compensation when critical modalities are absent.

**Robustness to Textual Degradation.** We further evaluate robustness under two textual degradation protocols: mismatch (random description swaps) and corruption (partial attribute replacement with Unknown). Figure 3 shows that our method consistently outperforms IDEA across all ratios. Under 20% corruption, the baseline drops 18.6% mAP while ours drops only 3.6%, confirming that $\pi$-noise disproportionately benefits high-noise scenarios.

**Number of GeM Pooling Branches.** Table 10 reports the effect of varying the number of pooled tokens $|N_t'|$ in the text feature $f_t$. The model achieves the best mAP and Rank-1 at $|N_t'| = 2$, with marginal drops for other values. This

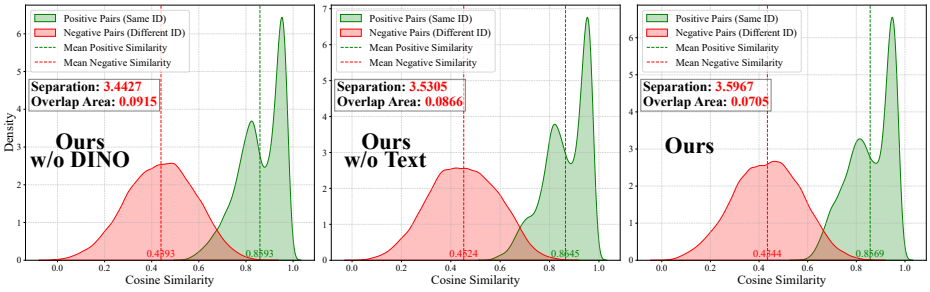



**Fig. 4:** Cosine similarity distributions of positive and negative feature pairs on RGBNT201 for our full model and ablated variants (w/o Text and w/o DINO).

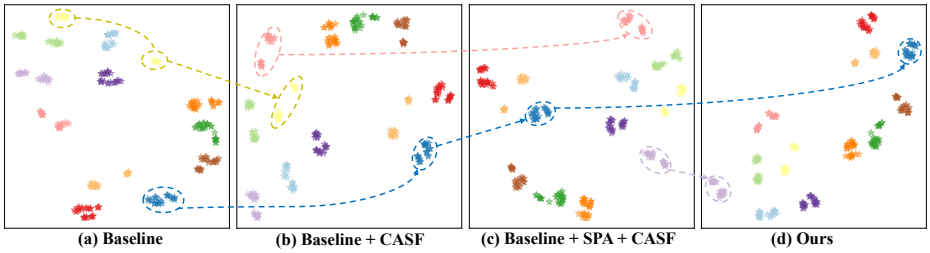



**Fig. 5:** Visualization of multi-modal ReID features on RGBNT201 using t-SNE [24]. Ours means the full model (baseline+CASF+SPA+SCM).

indicates that two tokens strike a good balance between semantic preservation and compactness, and the model is robust to the pooling granularity.

**Trainable Parameter Count.** Table 3 shows the parameter increase from each module, and Table 11 lists the total. Compared to the CLIP baseline†, our model adds only a few parameters while gaining +10.3% mAP on RGBNT201 and +24.6% mAP on MSVR310. Our approach also surpasses established methods such as EDITOR* [50] and DeMo† [39] while using fewer trainable parameters.

### 4.5 Visualization Analysis

**Cosine Similarity Distributions.** To examine the impact of text-guided noise and structural priors on feature similarity distributions, we present the cosine similarity histograms in Figure 4. We adopt two lightweight metrics to quantify feature discriminability: the separation metric (higher is better) and the overlap area between positive and negative similarity distributions (lower is better). A higher separation metric (**Separation** ↑) and a lower overlap area (**Overlap Area** ↓) jointly indicate stronger intra-class compactness and inter-class separability. Ablations without learnable structural priors (w/o DINO) or without textual guidance (w/o Text) both lead to reduced separation and increased overlap, confirming the effectiveness of our design. Specifically, the results show

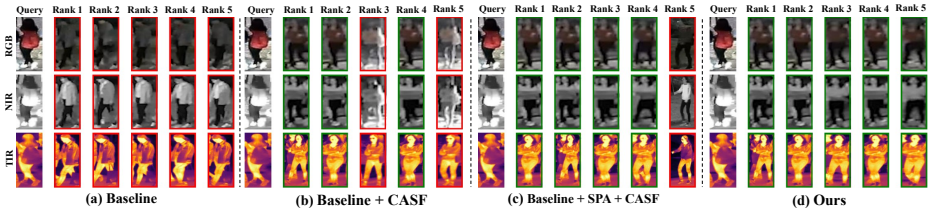



**Fig. 6:** Rank list comparison between the baseline and our approach.

that incorporating DINOv3-initialized structural priors and text-guided beneficial noise improves feature separation, yielding a separation metric of 3.5967 and an overlap area of 0.0705. When combined, these components produce the clearest inter-class separation and intra-class compactness among all variants, confirming their complementary roles in shaping a discriminative feature space.

**Multi-Modal Feature Distributions.** Figure 5 displays t-SNE [24] embeddings of features from several model configurations. Comparing Figure 5(a) and (b), the addition of CASF reduces ambiguity in regions such as the yellow circle, where samples from the same identity exhibit less overlap with neighboring classes. Further comparisons between (b) and (c), and between (c) and (d), reveal that clusters in the pink and purple circles become increasingly distinct after introducing SPA and SCM, respectively. Additionally, the cluster highlighted in blue becomes progressively more compact as each component is added. These observations reflect a consistent refinement in feature geometry across stages.

**Rank List Comparison.** Figure 6 compares cross-camera retrieval results from the baseline and our method. Our approach returns more accurate rank lists, with relevant matches consistently ranked higher and fewer irrelevant items in the top positions. In contrast, the baseline shows greater variability and includes more false matches, especially under challenging conditions such as pose changes or low-quality queries. This qualitative comparison supports the conclusion that our model learns more discriminative and robust cross-modal representations.

## 5 Conclusion

In this paper, we present a novel multi-modal object ReID framework that rethinks textual noise and structural context as constructive signals for cross-modal learning. The Semantic Cross-Modal Modulator (SCM) leverages $\pi$-noise for semantics-driven interaction between vision and language. The Structure-Aware Prompt Adapter (SPA) injects learnable geometric priors via prompts to align cross-modal structures. The Context-Aware Sparse Fusion (CASF) leverages structural context to guide sparse fusion while preserving identity discrimination from noisy local features. Experiments on standard benchmarks validate the effectiveness and generalization capability of our approach.

**Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China under Grant 62476041.

## References


1. Crawford, J., Yin, H., McDermott, L., Cummings, D.: Unicat: Crafting a stronger fusion baseline for multimodal re-identification. arXiv:2310.18812 (2023)
2. Dong, W., Yang, X., Cheng, D., Wang, N., Gao, X.: Escaping modal interactions: An efficient desanet for multi-modal object re-identification. TIP (2025)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
4. Feng, Y., Li, J., Hu, J., Zhang, Y., Tan, L., Ji, J.: Mdreid: Modality-decoupled learning for any-to-any multi-modal object re-identification. In: NeurIPS (2025)
5. Feng, Y., Li, J., Xie, C., Tan, L., Ji, J.: Multi-modal object re-identification via sparse mixture-of-experts. In: ICML (2025)
6. Gao, Y., Li, S., Liu, Z., Zheng, A., Li, C., Tang, J.: Chain-of-thought guided multi-modal object re-identification. In: CVPR (2026)
7. Guo, J., Zhang, X., Liu, Z., Wang, Y.: Generative and attentive fusion for multi-spectral vehicle re-identification. In: ICSP (2022)
8. He, Q., Lu, Z., Wang, Z., Hu, H.: Graph-based progressive fusion network for multi-modality vehicle re-identification. T-ITS (2023)
9. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: ICCV (2021)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
11. Huang, S., Zhang, H., Li, X.: Enhance vision-language alignment with noise. In: AAAI (2025)
12. Jiang, K., Shi, Z., Zhang, D., Zhang, H., Li, X.: Mixture of noise for pre-trained model-based class-incremental learning. arXiv preprint arXiv:2509.16738 (2025)
13. Li, H., Li, C., Zhu, X., Zheng, A., Luo, B.: Multi-spectral vehicle re-identification: A challenge. In: AAAI (2020)
14. Li, H., Liu, G., Wang, X., Liang, B., Luo, Y.: PEFT-BoA: Parameter-efficient fine-tuning with bag-of-adapters for multi-modal object re-identification. In: AAAI (2026)
15. Li, S., Li, C., Zheng, A., Lu, A., Tang, J., Ma, J.: Next: Multi-grained mixture of experts via text-modulation for multi-modal object re-id. arXiv:2505.20001 (2025)
16. Li, S., Li, C., Zheng, A., Tang, J., Luo, B.: Icpl-reid: Identity-conditional prompt learning for multi-spectral object re-identification. arXiv:2505.17821 (2025)
17. Li, S., Leng, J., Gan, J., Mo, M., Gao, X.: Shape-centered representation learning for visible-infrared person re-identification. Pattern Recognition (2025)
18. Li, S., Leng, J., Kuang, C., Tan, M., Gao, X.: Video-level language-driven video-based visible-infrared person re-identification. TIFS (2025)
19. Li, S., Leng, J., Kuang, C., Tan, M., Yuan, Y., Gao, X.: Causal bootstrapped alignment for unsupervised video-based visible-infrared person re-identification. arXiv:2604.15631 (2026)
20. Li, X.: Positive-incentive noise. TNNLS (2022)
21. Lin, M., Wang, S., Wang, X., Tang, J., Fu, L., Zuo, Z., Sang, N.: Dmpt: Decoupled modality-aware prompt tuning for multi-modal object re-identification. In: WACV (2025)
22. Liu, Y., Wang, Y., Zhang, P.: Signal: Selective interaction and global-local alignment for multi-modal object re-identification. arXiv:2511.17965 (2025)

23. Liu, Y., Wang, Y., Zhang, P.: Signal: Selective interaction and global-local alignment for multi-modal object re-identification. In: AAAI (2026)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
25. Pan, W., Huang, L., Liang, J., Hong, L., Zhu, J.: Progressively hybrid transformer for multi-modal vehicle re-identification. Sensors (2023)
26. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv:2505.06708 (2025)
27. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. TPAMI (2018)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: ICCV (2021)
30. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv:1701.06538 (2017)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv:2508.10104 (2025)
32. Somers, V., Alahi, A., Vleeschouwer, C.D.: Keypoint promptable re-identification. In: ECCV (2024)
33. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018)
34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
35. Wan, X., Zheng, A., Wang, Z., Jiang, B., Tang, J., Ma, J.: Reliable multi-modal object re-identification via modality-aware graph reasoning. arXiv:2504.14847 (2025)
36. Wan, X., Zheng, A., Wang, Z., Jiang, B., Tang, J., Ma, J.: Reliable multi-modal object re-identification via modality-aware graph reasoning. TIFS (2026)
37. Wang, Y., Liu, X., Yan, T., Liu, Y., Zheng, A., Zhang, P., Lu, H.: Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In: AAAI (2025)
38. Wang, Y., Liu, X., Zhang, P., Lu, H., Tu, Z., Lu, H.: Top-reid: Multi-spectral object re-identification with token permutation. In: AAAI (2024)
39. Wang, Y., Liu, Y., Zheng, A., Zhang, P.: Decoupled feature-based mixture of experts for multi-modal object re-identification. In: AAAI (2025)
40. Wang, Y., Lv, Y., Zhang, P., Lu, H.: Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In: CVPR (2025)
41. Wang, Z., Huang, H., Zheng, A., He, R.: Heterogeneous test-time training for multi-modal person re-identification. In: AAAI (2024)
42. Wang, Z., Li, C., Zheng, A., He, R., Tang, J.: Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In: AAAI (2022)
43. Wu, D., Liu, Z., Chen, Z., Gan, S., Tan, K., Wan, Q., Wang, Y.: Lrmm: Low rank multi-scale multi-modal fusion for person re-identification based on rgb-ni-ti. ESWA (2025)

44. Xu, X., Liu, Z., Zhou, W., Gao, Y., Cao, J., Wang, Y., Luo, J., Zhang, D.: Stmi: Segmentation-guided token modulation with cross-modal hypergraph interaction for multi-modal object re-identification. In: AAAI (2026)
45. Yang, Y., Li, S., Ye, J., Dong, N., Li, F., Li, H.: Dinov2 driven gait representation learning for video-based visible-infrared person re-identification. In: ACM MM (2025)
46. Yin, H., Li, J., Schiller, E., McDermott, L., Cummings, D.: Graft: Gradual fusion transformer for multimodal re-identification. arXiv:2310.16856 (2023)
47. Yu, Z., Huang, Z., Hou, M., Pei, J., Yan, Y., Liu, Y., Sun, D.: Representation selective coupling via token sparsification for multi-spectral object re-identification. TCSVT (2024)
48. Yu, Z., Huang, Z., Hou, M., Yan, Y., Liu, Y.: Wtsf-reid: Depth-driven window-oriented token selection and fusion for multi-modality vehicle re-identification with knowledge consistency constraint. ESWA (2025)
49. Zhang, H., Huang, S., Guo, Y., Li, X.: Variational positive-incentive noise: How noise benefits models. TPAMI (2025)
50. Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H.: Magic tokens: Select diverse tokens for multi-modal object re-identification. In: CVPR (2024)
51. Zhang, S., Luo, W., Cheng, D., Xing, Y., Liang, G., Wang, P., Zhang, Y.: Prompt-based modality alignment for effective multi-modal object re-identification. TIP (2025)
52. Zheng, A., He, Z., Wang, Z., Li, C., Tang, J.: Dynamic enhancement network for partial multi-modality person re-identification. arXiv:2305.15762 (2023)
53. Zheng, A., Ma, Z., Sun, Y., Wang, Z., Li, C., Tang, J.: Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification. Information Fusion (2025)
54. Zheng, A., Wang, Z., Chen, Z., Li, C., Tang, J.: Robust multi-modality person re-identification. In: AAAI (2021)
55. Zheng, A., Zhu, X., Ma, Z., Li, C., Tang, J., Ma, J.: Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. Information Fusion (2023)
56. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
57. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)
58. Zhu, K., Guo, H., Liu, Z., Tang, M., Wang, J.: Identity-guided human semantic parsing for person re-identification. In: ECCV (2020)