STMI: Segmentation-Guided Token Modulation with Cross-Modal Hypergraph Interaction for Multi-Modal Object Re-Identification

Xingguo Xu[1*], Zhanyu Liu[1*], Weixiang Zhou[1*], Yuansheng Gao[2], Junjie Cao[1†], Yuhao Wang[3†], Jixiang Luo[4], Dell Zhang[4]

[1]School of Mathematical Sciences, Dalian University of Technology, China
[2]College of Computer Science and Technology, Zhejiang University, China
[3]School of Future Technology, Dalian University of Technology, China
[4]Institute of Artificial Intelligence (TeleAI), China Telecom, China
{xuxingguo, ramirez, s20201162006, 924973292}@mail.dlut.edu.cn, y.gao@zju.edu.cn, jjcao@dlut.edu.cn, luojx14@chinatelecom.cn, dell.z@ieee.org

Abstract

Multi-modal object Re-Identification (ReID) aims to exploit complementary information from different modalities to retrieve specific objects. However, existing methods often rely on hard token filtering or simple fusion strategies, which can lead to the loss of discriminative cues and increased background interference. To address these challenges, we propose **STMI**, a novel multi-modal learning framework consisting of three key components: (1) *Segmentation-Guided Feature Modulation* (SFM) module leverages SAM-generated masks to enhance foreground representations and suppress background noise through learnable attention modulation; (2) *Semantic Token Reallocation* (STR) module employs learnable query tokens and an adaptive reallocation mechanism to extract compact and informative representations without discarding any tokens; (3) *Cross-Modal Hypergraph Interaction* (CHI) module constructs a unified hypergraph across modalities to capture high-order semantic relationships. Extensive experiments on public benchmarks (i.e., RGBNT201, RGBNT100, and MSVR310) demonstrate the effectiveness and robustness of our proposed STMI framework in multi-modal ReID scenarios.

Introduction

In recent years, multi-modal object Re-Identification (ReID) has attracted increasing attention due to its wide range of applications in practical scenarios such as intelligent surveillance, cross-spectrum monitoring, and nighttime recognition. Unlike traditional RGB-based ReID (Liu et al. 2021; Zhang et al. 2021; Wang et al. 2021; Shi et al. 2024), multi-modal object ReID involves multiple visual modalities, including visible light (RGB), near-infrared (NIR), and thermal infrared (TIR), offering enhanced robustness under challenging conditions such as drastic illumination changes, low-light environments, or nighttime scenes (Zhao et al. 2023; He et al. 2023b; Zheng et al. 2025b; Tang et al. 2025). However, due to the significant distribution discrepancies across different modalities, achieving effective multi-modal representation learning remains a fundamental challenge.

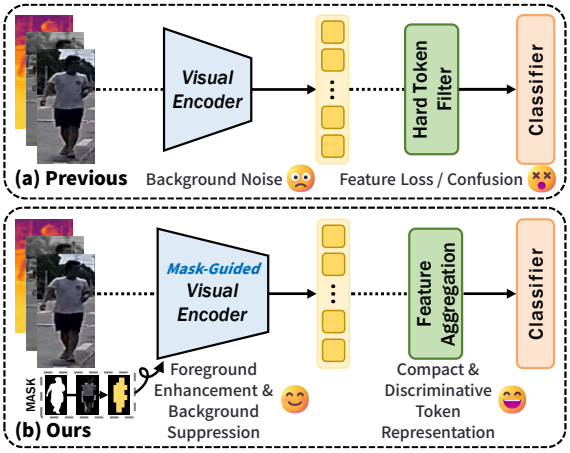

Figure 1: Motivation and intuitive comparison. (a) Existing methods suffer from background noise and information loss due to hard token filtering. (b) Our proposed STMI framework introduces segmentation-guided modulation module to enhance foreground and suppress background, enabling more discriminative feature learning across modalities.

Existing methods primarily focus on aligning and fusing visual features, typically leveraging strategies such as token selection, modality transformation, or attention mechanisms to process multi-modal images (Wan et al. 2025a; Li et al. 2025b; Wan et al. 2025b; Lin et al. 2025; Bian et al. 2025), as illustrated in Fig. 1. However, these approaches suffer from two major limitations. First, during token selection, "redundant" regions are often removed via hard cropping, which may inadvertently discard critical details and compromise discriminative performance. Second, in multi-modal feature fusion, the lack of effective modeling of high-order semantic relationships limits the ability to fully exploit complementary information across modalities, especially in complex scenes with background clutter or occlusions.

To address the aforementioned issues, we propose a fea-

*These authors contributed equally.
†Corresponding authors.

ture learning framework named **STMI**, which introduces **S**egmentation-guided **T**oken **M**odulation with cross-modal hypergraph **I**nteraction for multi-modal object ReID. First, we introduce the Segmentation-Guided Feature Modulation (SFM) module, which leverages foreground masks generated by the SAM segmentation model to guide attention learning. Specifically, we incorporate two learnable modulation parameters to adaptively reweight token features, emphasizing foreground regions while suppressing background noise. Second, we propose the Semantic Token Reallocation (STR) module to refine the token representation in a more structured manner. Rather than relying on hard filtering strategies, STR introduces multiple learnable query tokens that interact with patch tokens via a cross-attention mechanism. This enables the extraction of compact, informative semantic representations while preserving fine-grained visual details. Third, we design the Cross-Modal Hypergraph Interaction (CHI) module to capture high-order semantic relationships across different modalities. In this module, semantic tokens from RGB, NIR, and TIR images are treated as nodes within a unified hypergraph. Cross-modal hyperedges are constructed based on semantic similarity, allowing the model to learn structural correlations among local regions across modalities. By jointly leveraging segmentation priors, semantic token reconstruction, and high-order relational modeling, STMI effectively maintains token completeness and enhances feature discrimination, achieving superior performance in challenging multi-modal ReID scenarios. Our main contributions are summarized as follows:

- We propose STMI, a novel multi-modal ReID framework. To the best of our knowledge, it is the first work to incorporate segmentation masks for attention modulation in multi-modal object ReID.
- We introduce a Segmentation-Guided Feature Modulation (SFM) module that enhances foreground regions and suppresses background interference, preserving discriminative information without discarding any tokens.
- We design a Semantic Token Reallocation (STR) module based on cross-attention, which extracts structured and compact semantic tokens using learnable queries, avoiding information loss caused by hard token filtering.
- We present a Cross-Modal Hypergraph Interaction (CHI) module that models high-order semantic relationships across modalities by constructing a unified hypergraph, enabling rich inter-modal dependency modeling.
- Extensive experiments on three public multi-modal ReID datasets demonstrate that STMI achieves state-of-the-art performance, validating its effectiveness and robustness.

Related Work

Multi-Modal Object Re-Identification

Benefiting from the complementary information across modalities, multi-modal ReID has demonstrated superior stability and performance. Existing approaches primarily focus on modeling cross-modal interactions (Zhang et al. 2025; Yang et al. 2025a; Feng et al. 2025). For instance, TOP-ReID (Wang et al. 2024b) introduces a cyclic interaction mechanism via cross-attention to fuse tri-modal features. MambaPro (Wang et al. 2024a) adopts the Mamba architecture (Gu and Dao 2023) to capture both intra-modal and inter-modal dependencies. Furthermore, DeMo (Wang et al. 2024c) proposes an adaptive Mixture of Experts (MoE) framework to decouple modality-specific information and perform weighted cross-modal feature aggregation. However, most of these methods model all tokens across the entire image, making them vulnerable to background noise, which deteriorates feature quality and limits overall performance (Tian et al. 2018). To address this issue and better focus on informative regions, several works explore key feature extraction prior to cross-modal fusion. EDITOR (Zhang et al. 2024) leverages attention maps to select salient features, guiding the model to attend to important regions. IDEA (Wang et al. 2025) samples tokens from key spatial locations and adaptively learns positional shifts to capture fine-grained local details. NEXT (Li et al. 2025a) introduces textual cues to guide context-aware token sampling. While these sampling-based strategies help the model focus on salient regions, they also introduce new challenges. The adaptiveness of token selection does not always ensure the most informative features. This may result in **semantic loss** and **feature confusion** with hard token pruning, ultimately hindering performance. To overcome this limitation, we propose a token modulation strategy that **preserves critical information** more effectively, while also **mitigating noise and ambiguity** in the feature representation.

Semantic Segmentation for Feature Enhancement

Semantic segmentation has seen significant advancements in various visual tasks, particularly in handling complex image editing and instance segmentation, where it demonstrates powerful capabilities. In recent years, pre-trained models such as OpenPifPaf (Kreiss, Bertoni, and Alahi 2021), SAM (Kirillov et al. 2023), and SAM2 (Ravi et al. 2024) have demonstrated strong generalization capabilities across various vision scenarios. These models can generate masks containing rich semantic information, which can be effectively utilized in a wide range of visual tasks. For instance, VideoGrain (Yang et al. 2025b) leverages SAM for instance segmentation in multi-grained video editing tasks, generating precise masks to help control the editing of different parts of the video. Additionally, VoteSplat (Jiang et al. 2025) integrates SAM with a Hough voting mechanism to achieve accurate instance segmentation. SmartFreeEdit (Sun et al. 2025) uses SAM to generate reasoning segmentation masks, supporting image editing guided by natural language instructions. While in ReID tasks, for instance, Mask-Guided (Song et al. 2018) introduces binary masks to guide the generation of body and background attention maps for region-level learning. MaskReID (Qi et al. 2019) directly uses segmentation results along with RGB images for improved feature representations. Nonetheless, prior works incorporate segmentation masks merely as auxiliary inputs, lacking **fine-grained** and **token-level modulation**. In contrast, our proposed STMI framework embeds segmentation into the attention mechanism, enabling more precise and

Figure 2: Comparison with IDEA: (a) IDEA captions often include unknown or inconsistent attributes; (b) ours generates clearer and more accurate descriptions across modalities; (c) our method significantly reduces unknown attributes in both training and test sets.

consistent feature enhancement across modalities.

Proposed Method

As shown in Fig. 3, our proposed **STMI** consists of three components: Segmentation-Guided Feature Modulation (SFM) module, Semantic Token Reallocation (STR) module, and Cross-Modal Hypergraph Interaction (CHI) module. Below, we describe each component in detail.

Multi-Modal Caption Generation

In multi-modal object ReID tasks, introducing semantic descriptions as auxiliary guidance (Wang et al. 2025) has been shown to significantly improve model performance. However, existing text generation methods still suffer from several major limitations, as illustrated in Fig. 2: (1) **Modality inconsistency:** Most existing approaches generate textual descriptions based solely on a single modality (e.g., RGB), ignoring complementary semantic cues potentially present in other modalities such as NIR or TIR. This often leads to incomplete or biased descriptions; (2) **Semantic ambiguity:** Under challenging conditions such as occlusion, low light, or blur, multi-modal large language models (MLLMs) often fail to identify key attributes, resulting in vague responses like "unknown" or even refusal to answer. (3) **Lack of confidence estimation:** Most existing methods do not provide confidence scores for each generated attribute, making it difficult to assess the reliability of the semantic information.

To address the above issues, we propose two strategies to enhance the quality and reliability of multi-modal caption generation. First, we adopt an image concatenation-based input, where images of the same identity from three modalities are concatenated into a single composite image and fed into an MLLM. This design enables the model to perceive multi-modal information holistically and generate more complete and consistent natural language descriptions. Second, we introduce a structured attribute extraction and confidence-aware filling strategy, inspired by NEXT (Li et al. 2025a), which leverages attribute-level confidence to guide text generation. Specifically, we first use the MLLM to extract attribute–value–confidence triplets from each individual modality as well as the concatenated multi-modal image. These triplets are then fed into an LLM along with a predefined template, prompting it to select the most reliable attribute values based on confidence scores and generate the final description. **As shown in Fig. 2**, these two strategies significantly enhance the reliability and consistency of the generated textual descriptions, thereby providing high-quality semantic information for downstream multi-modal feature modeling.

Segmentation-Guided Feature Modulation

To enhance the model's ability to focus on foreground regions while suppressing background interference, we propose the *Segmentation-Guided Feature Modulation* (SFM) module. This module explicitly guides the attention maps in self-attention layers using semantic segmentation masks, enabling region-aware feature modeling. Specifically, given an input image $I \in \mathbb{R}^{3\times H\times W}$ and its corresponding binary segmentation mask $M \in \{0,1\}^{H\times W}$, we first divide the image into N patches through a vision encoder and extract $N+1$ token representations, including one class token as follows:

$$F = [f_{\text{cls}}, f_1, f_2, \ldots, f_N] \in \mathbb{R}^{(N+1)\times D}, \tag{1}$$

where f_{cls} denotes the class token, and the rest are patch tokens, with D being the feature dimension per token. Next, we construct a token-level binary mask based on the spatial overlap between each patch and the segmentation mask:

$$m = [1, m_1, m_2, \ldots, m_N] \in \{0,1\}^{N+1}, \tag{2}$$

where the first position is set to 1, treating the class token as part of the foreground. The remaining m_i indicate whether the i-th patch token lies within the foreground region.

At the l-th Transformer layer, the self-attention module (Vaswani et al. 2017) first computes the attention logits:

$$A^{(l)}_{\text{logit}} = Q^{(l)}(K^{(l)})^{\top}, \tag{3}$$

where $Q^{(l)}, K^{(l)} \in \mathbb{R}^{(N+1)\times D}$, and D is the attention dimension. We then construct positive and negative modulation matrices based on $A^{(l)}_{\text{logit}}$, used for enhancing foreground and suppressing background regions, respectively:

$$M^{\text{pos}} = \max(A^{(l)}_{\text{logit}}) - A^{(l)}_{\text{logit}}, \tag{4}$$

$$M^{\text{neg}} = A^{(l)}_{\text{logit}} - \min(A^{(l)}_{\text{logit}}), \tag{5}$$

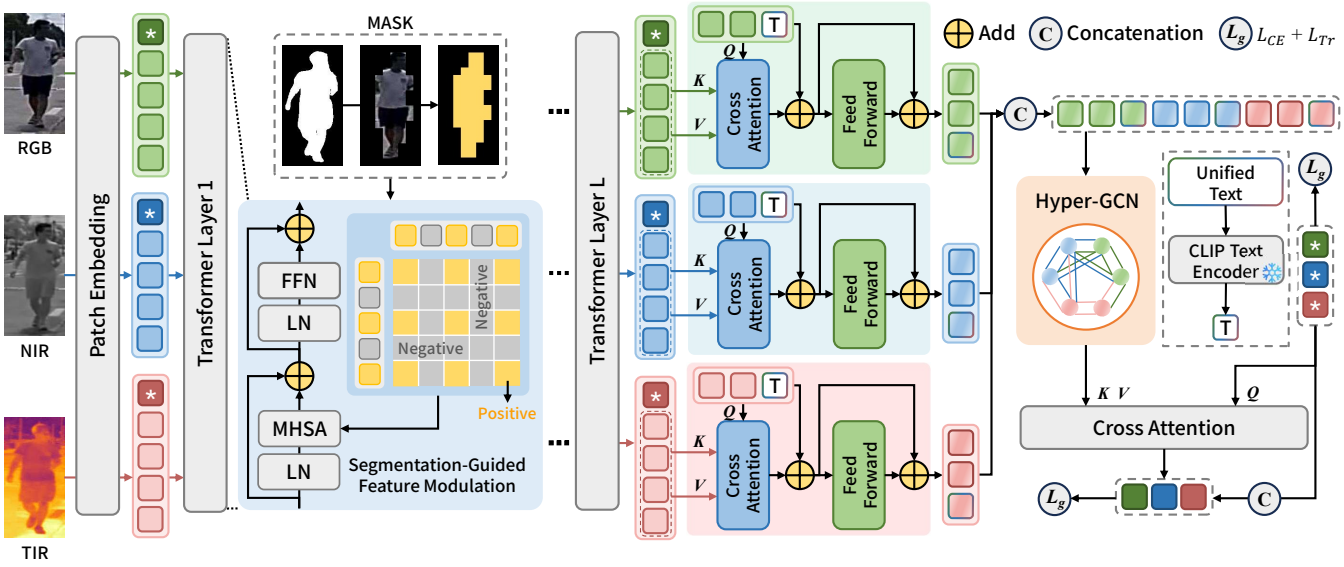

Figure 3: An overview of our proposed STMI framework, which consists of three main modules: (1) Segmentation-Guided Feature Modulation enhances foreground and suppresses background using SAM masks; (2) Semantic Token Reallocation extracts compact semantic tokens via cross-attention with learnable queries; (3) Cross-Modal Hypergraph Interaction builds a hypergraph across modalities for high-order semantic interaction.

where $\max(\cdot)$ and $\min(\cdot)$ are row-wise operations, broadcasting the maximum/minimum value of each row across the entire row. The resulting matrices $M^{\text{pos}}, M^{\text{neg}} \in \mathbb{R}^{(N+1)\times(N+1)}$ represent the relative deviation of each token pair. To guide the attention modulation, we construct a foreground interaction mask R based on m as follows:

$$R = m^{\top} \cdot m \in \{0,1\}^{(N+1)\times(N+1)}, \tag{6}$$

where $R[i,j] = 1$ if both tokens i and j are foreground. Otherwise, $R[i,j] = 0$, treating the pair as background.

Considering that the semantic segmentation mask may suffer from boundary ambiguity or incorrect segmentation, we introduce a mask perturbation mechanism during training to enhance model robustness. Specifically, only background tokens are perturbed with a probability p by flipping their label to foreground (i.e., $m_i := 1$ if $m_i = 0$), while foreground tokens remain unchanged. This strategy effectively mitigates the overfitting risk caused by mask guidance and improves the generalization ability of the model. Thus, the final modulation matrix S is defined as:

$$S = \alpha \cdot R \odot M^{\text{pos}} - \beta \cdot (1 - R) \odot M^{\text{neg}}, \tag{7}$$

where $\odot$ denotes element-wise multiplication, and α, β are learnable parameters that control the strength of foreground enhancement and background suppression. Then the modulated attention weights are computed as follows:

$$\hat{A}^{(l)} = \text{softmax}\left(\frac{A^{(l)}_{\text{logit}} + S}{\sqrt{D}}\right). \tag{8}$$

Through the above mechanism, SFM enhances the model's semantic understanding of foreground regions while effectively suppressing background noise during multi-modal feature modeling in the backbone.

Semantic Token Reallocation

To enhance semantic alignment across modalities, we propose the *Semantic Token Reallocation* (STR) module. It employs learnable, modality-specific query tokens, along with a shared global text feature, to guide the cross-attention process. This design enables a structured reconstruction of visual semantic tokens, leading to improved cross-modal consistency. Specifically, for each modality $m \in \{\text{RGB}, \text{NIR}, \text{TIR}\}$, the input is a sequence of patch tokens extracted from the backbone:

$$F^{(m)} = \left[f_1^{(m)}, f_2^{(m)}, \ldots, f_N^{(m)}\right] \in \mathbb{R}^{N\times D}. \tag{9}$$

For each modality, we introduce K independent learnable semantic query tokens, with the following equation:

$$Q^{(m)} = \left[q_1^{(m)}, \ldots, q_K^{(m)}\right] \in \mathbb{R}^{K\times D}. \tag{10}$$

To incorporate a cross-modal semantic prior, we further extract a shared global textual feature using the text encoder of CLIP (Radford et al. 2021), which captures the overall description of the image across all three modalities. This shared global feature is denoted as $T \in \mathbb{R}^D$. We concatenate it to the end of the semantic query token sequence to form an enhanced query sequence:

$$Q'^{(m)} = \left[Q^{(m)}; T\right] \in \mathbb{R}^{(K+1)\times D}. \tag{11}$$

Here, $Q'^{(m)}$ serves as the query, while $F^{(m)}$ serves as both key and value. A cross-attention operation followed by a Feed-Forward Network (FFN) (Vaswani et al. 2017) is applied to obtain the final semantic token representations:

$$Z^{(m)} = \text{CrossAttn}(Q'^{(m)}, F^{(m)}, F^{(m)}) + Q'^{(m)}, \tag{12}$$

$$\tilde{F}^{(m)} = \text{FFN}(Z^{(m)}) + Z^{(m)}. \tag{13}$$

The resulting $\tilde{F}^{(m)}$ is then used as input to the multi-modal semantic alignment and interaction modeling stage, serving as the basis for subsequent cross-modal feature fusion.

Cross-Modal Hypergraph Interaction

After obtaining the semantic token representations for each modality, we design the *Cross-Modal Hypergraph Interaction* (CHI) module to model high-order semantic relationships across modalities. Unlike traditional graph structures, hypergraphs can connect multiple nodes within a single hyperedge, making them naturally suitable for joint modeling of multi-modal information. Specifically, let the semantic tokens from the three modalities be denoted as $\tilde{F}^{(\text{RGB})}$, $\tilde{F}^{(\text{NIR})}$, and $\tilde{F}^{(\text{TIR})}$, each of shape $\mathbb{R}^{(K+1)\times D}$. We first concatenate them into a unified cross-modal semantic token set:

$$H = \left[\tilde{F}^{(\text{RGB})}; \tilde{F}^{(\text{NIR})}; \tilde{F}^{(\text{TIR})}\right] \in \mathbb{R}^{3(K+1)\times D}. \tag{14}$$

Based on this, we construct a cross-modal hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ corresponds to each token in H, i.e., $|\mathcal{V}| = 3(K+1)$. The hyperedge set $\mathcal{E}$ is dynamically generated based on intra- and inter-modal semantic similarities. Specifically, we compute a similarity matrix $S \in \mathbb{R}^{3(K+1)\times 3(K+1)}$ over all nodes. Two nodes i and j are connected within the same hyperedge if their similarity $s_{ij} \geq \tau$. This allows each hyperedge to capture high-order associations and enables effective cross-modal semantic propagation. To further model information flow within the hypergraph, we introduce a hypergraph convolution operation (Bai, Zhang, and Torr 2021). For the node features at the l-th layer $H^{(l)} \in \mathbb{R}^{3(K+1)\times D}$, their update process in the hypergraph is formulated as:

$$h_i^{(l+1)} = \sigma\left(\sum_{e\in\varepsilon(i)} w_e h_e^{(l)} + b_i\right). \tag{15}$$

Here, $h_e^{(l)}$ denotes the feature of hyperedge e at the l-th iteration. The set $\varepsilon(i)$ contains all hyperedges incident to node i. Each hyperedge e is associated with a learnable w_e. The term b_i is a node-specific bias. $\sigma(\cdot)$ is an activation function. Specifically, we adopt a node-to-hyperedge and hyperedge-to-node mechanism, where features are first aggregated from nodes to hyperedges, and then redistributed back to nodes. This operation allows each node to aggregate information from a group of connected nodes via hyperedges, thereby enhancing inter-modal interaction and fusion. Additionally, to preserve the independent semantic information of each original modality, we introduce a residual connection:

$$H^{(l+1)} = H^{(l+1)} + H^{(l)}. \tag{16}$$

Through the hypergraph convolution, the CHI module enables high-order semantic interactions across modalities, capturing complex and rich semantic dependencies between them, thus improving the final fused representation. The resulting multi-modal semantic tokens are denoted as $H' \in \mathbb{R}^{3(K+1)\times D}$.

Although the CHI module effectively models high-order semantic relationships across modalities, its output semantic tokens mainly focus on local regions and still lack explicit alignment with global semantic concepts. We further employ a cross-attention mechanism, using global image features as queries to selectively aggregate complementary information from the multi-modal semantic tokens. Specifically, we extract image-level global features from the three modalities, denoted as g^{RGB}, g^{NIR}, and g^{TIR}, respectively. Then, we concatenate these global features into a query, as follows:

$$G = \left[g^{\text{RGB}}; g^{\text{NIR}}; g^{\text{TIR}}\right] \in \mathbb{R}^{3\times D}. \tag{17}$$

Next, we use the fused semantic token representation H' from the CHI module as key and value pairs to construct a cross-modal cross-attention algorithm:

$$U = \text{CrossAttn}(G, H', H') + G, \tag{18}$$

where U represents the final fused global features. This process enables selective extraction of information relevant to global concepts from the multi-modal semantic tokens and feeds it back into the global representation.

Methods	RGBNT201			
	mAP	R-1	R-5	R-10
PFNet (Zheng et al. 2021)	38.5	38.9	52.0	58.4
IEEE (Wang et al. 2022)	47.5	44.4	57.1	63.6
DENet (Zheng et al. 2023a)	42.4	42.2	55.3	64.5
LRMM (Wu et al. 2025)	52.3	53.4	64.6	73.2
UniCat* (Crawford et al. 2023)	57.0	55.7	-	-
HTT* (Wang et al. 2024d)	71.1	73.4	83.1	87.3
TOP-ReID* (Wang et al. 2024b)	72.3	76.6	84.7	89.4
EDITOR* (Zhang et al. 2024)	66.5	68.3	81.1	88.2
RSCNet* (Yu et al. 2024)	68.2	72.5	-	-
WTSF-ReID* (Yu et al. 2025)	67.9	72.2	83.4	89.7
MambaPro† (Wang et al. 2024a)	78.9	<u>83.4</u>	89.8	91.9
DeMo† (Wang et al. 2024c)	79.0	82.3	88.8	<u>92.0</u>
IDEA† (Wang et al. 2025)	<u>80.2</u>	82.1	<u>90.0</u>	**93.3**
STMI†	**81.2**	**83.4**	**90.2**	91.6

Table 1: Performance comparison on RGBNT201. Best results are in bold, the second bests are underlined. † denotes CLIP-based methods, $*$ indicates ViT-based while others are CNN-based ones.

Objective Function

We apply supervision to key representations, including the concatenated global feature G, fused semantic feature U, and global text feature T. For each feature representation $\mathcal{F}$, we jointly employ a label-smoothed cross-entropy loss (Szegedy et al. 2016) and a triplet loss (Hermans, Beyer, and Leibe 2017), defined as follows:

$$\mathcal{L}_g(\mathcal{F}) = \mathcal{L}_{\text{CE}}(\mathcal{F}) + \mathcal{L}_{\text{Tri}}(\mathcal{F}), \tag{19}$$

where $\mathcal{L}_{\text{CE}}$ denotes the label-smoothed cross-entropy loss, and $\mathcal{L}_{\text{Tri}}$ represents the triplet loss. The final overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_g(G) + \mathcal{L}_g(U) + \mathcal{L}_g(T). \tag{20}$$

Methods	RGBNT100		MSVR310	
	mAP	R-1	mAP	R-1
PFNet (Zheng et al. 2021)	68.1	94.1	23.5	37.4
GAFNet (Guo et al. 2022)	74.4	93.4	-	-
GPFNet (He et al. 2023a)	75.0	94.5	-	-
CCNet (Zheng et al. 2023b)	77.2	96.3	36.4	55.2
LRMM (Wu et al. 2025)	78.6	96.7	36.7	49.7
GraFT* (Yin et al. 2023)	76.6	94.3	-	-
UniCat* (Crawford et al. 2023)	79.4	96.2	-	-
PHT* (Pan et al. 2023)	79.9	92.7	-	-
HTT* (Wang et al. 2024d)	75.7	92.6	-	-
TOP-ReID* (Wang et al. 2024b)	81.2	96.4	35.9	44.6
EDITOR* (Zhang et al. 2024)	82.1	96.4	39.0	49.3
FACENet* (Zheng et al. 2025a)	81.5	96.9	36.2	54.1
RSCNet* (Yu et al. 2024)	82.3	96.6	39.5	49.6
WTSF-ReID* (Yu et al. 2025)	82.2	96.5	39.2	49.1
MambaPro† (Wang et al. 2024a)	83.9	94.7	47.0	56.5
DeMo† (Wang et al. 2024c)	86.2	**97.6**	<u>49.2</u>	59.8
IDEA† (Wang et al. 2025)	<u>87.2</u>	96.5	47.0	<u>62.4</u>
STMI†	**89.1**	<u>97.1</u>	**64.8**	**76.1**

Table 2: Performance on RGBNT100 and MSVR310.

Index	Modules			Metrics	
	SFM	STR	CHI	mAP	Rank-1
A	×	×	×	70.3	72.1
B	✓	×	×	76.1	78.1
C	✓	✓	×	78.1	80.9
D	✓	✓	✓	**81.2**	**83.4**

Table 3: Ablation study of different modules in STMI.

Index	Model Variant	mAP	R-1	R-5	R-10
A	STMI (w/o CHI)	78.1	80.9	87.9	90.9
B	STMI (w/ MLP, w/o CHI)	78.0	82.4	87.3	90.0
C	STMI (w/ SA, w/o CHI)	78.4	81.5	86.2	89.2
D	**STMI (Full Model, w/ CHI)**	**81.2**	**83.4**	**90.2**	**91.6**

Table 4: Ablation of different fusion strategies in STMI.

Experiments

Datasets and Evaluation Protocols

Dataset Setup. To evaluate the effectiveness of the proposed method in complex multi-modal scenarios, we conduct experiments on three public multi-modal object ReID datasets. To improve annotation efficiency, we utilize the GPT-4o model (Hurst et al. 2024) provided by OpenAI to automatically generate one textual description for each image triplet. In addition, we employ the SAM2 (Ravi et al. 2024) to generate one high-quality segmentation mask per triplet. Specifically, *RGBNT201* (Zheng et al. 2021) contains 4,787 triplets and 4,787 textual descriptions, with an average length of 33.28 characters, covering 13 semantic attributes. The *MSVR310* (Zheng et al. 2023b) dataset consists of 2,087 triplets and corresponding descriptions, each with an average length of 31.51 characters, covering 6 attributes. *RGBNT100* (Li et al. 2020) is the largest dataset among them, containing 17,250 triplets and 17,250 textual annotations with an average description length of 31.90 characters, also covering 6 semantic attributes.

Evaluation Protocol. We adopt mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at ranks 1, 5, and 10 as evaluation metrics.

Implementation Details

The proposed model is implemented using the PyTorch framework and trained on an NVIDIA A800 GPU. For visual and textual encoding, we uniformly adopt the pre-trained CLIP model (Radford et al. 2021). The input images in the RGBNT201 dataset are resized to 256×128, while those in MSVR310 and RGBNT100 are resized to 128×256. Data augmentations include random horizontal flipping, random cropping, and random erasing (Zhong et al. 2020). The batch size is set to 72 for the RGBNT201 dataset and 64 for the MSVR310 dataset (Wang et al. 2025), with 8 images sampled per identity. For the RGBNT100 dataset, a larger batch size of 128 is used, with 16 images sampled per identity. We employ the Adam optimizer to fine-tune all learnable parameters in the model, with an initial learning rate of 3.5×10^{-6}, which is gradually decayed to 3.5×10^{-7}. Additional details on prompt template, hyperparameter settings, and training efficiency are provided in the **appendix**.

Comparison with State-of-the-Art Methods

Multi-Modal Person ReID. As shown in Tab. 1, STMI obtains **81.2%** mAP, achieving the best performance among all compared methods. Specifically, it surpasses the previous state-of-the-art IDEA by +1.0% in mAP. Compared with TOP-ReID (72.3%) and EDITOR (66.5%), STMI achieves substantial improvements of +8.9% and +14.7% in mAP, respectively. These results demonstrate the effectiveness of STMI in enhancing cross-modal semantic alignment and preserving token-level representation integrity.

Multi-Modal Vehicle ReID. As shown in Tab. 2, STMI achieves **89.1%** mAP on the RGBNT100 dataset, outperforming strong baselines such as IDEA (87.2%) and DeMo (86.2%). On the more challenging MSVR310 dataset, STMI achieves **64.8%** mAP, surpassing the previous best result by +17.8% over IDEA (47.0%). These results highlight the effectiveness and robustness of our method under complex conditions such as background clutter, occlusion, and modality inconsistency.

Ablation Studies

We conduct ablation studies on the RGBNT201 dataset to evaluate each component of our STMI framework. The baseline adopts a three-branch vision encoder, where retrieval is based on concatenated class tokens from three modalities.

Effects of Key Modules. Tab. 3 presents the performance of different combinations of the proposed modules. Model A serves as the baseline, achieving an mAP of 70.3% and Rank-1 accuracy of 72.1%. Model B incorporates the SFM module, which leverages SAM masks to enhance foreground regions and suppress background noise. This improves the mAP to 76.1%. Model C further introduces the STR module, improving the mAP to 78.1%. Finally, Model D integrates all three modules, including the CHI module for high-order semantic modeling via cross-modal hypergraphs, achieving the best performance with an 81.2% mAP.

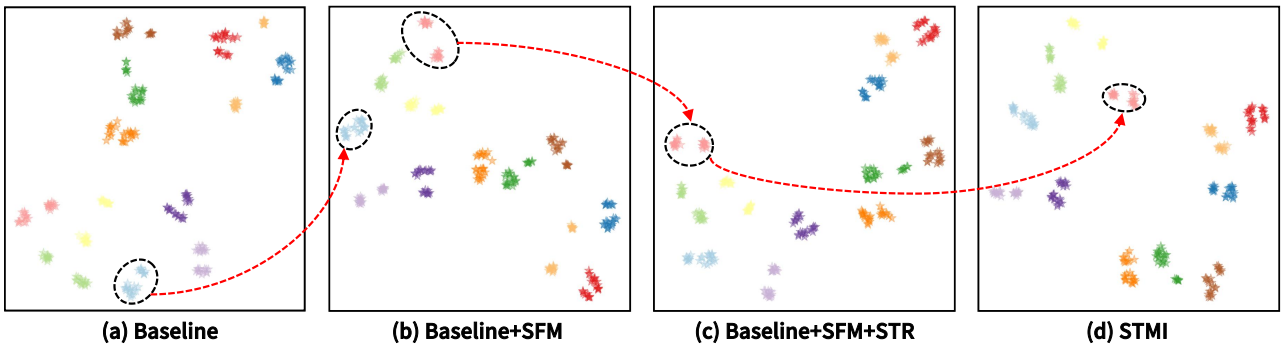

Figure 4: Visualization of the feature distributions with t-SNE (Van der Maaten and Hinton 2008). Different colors represent different identities.

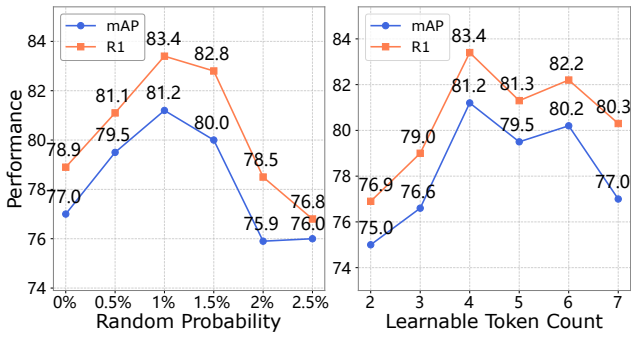

Figure 5: Comparison with different hyper-parameters.

Effects of CHI Configurations. To analyze the effectiveness of the CHI module, we conduct ablation experiments with different fusion strategies in the STMI framework. As shown in Tab. 4, Model A removes CHI entirely and directly concatenates modality-specific class tokens, achieving 78.1% mAP. Model B replaces CHI with a MLP to fuse features across modalities, but only obtains 78.0% mAP. Model C adopts a multi-head self-attention mechanism for cross-modal interaction, yielding 78.4% mAP. Finally, the full model (Model D) integrates the CHI module and achieves the best performance with 81.2% mAP. These results demonstrate that CHI effectively captures high-order semantic dependencies across modalities and provides more discriminative fusion representations compared to conventional fusion strategies.

Effects of SFM Configurations. We investigate parameter-sharing strategies in the SFM module. As shown in Tab. 5, sharing parameters across all layers (Model A) yields 77.2% mAP. Modulating only the early layers (Model B) or the late layers (Model C) results in suboptimal performance. Assigning head-wise parameters (Model D) gives 77.1% mAP. Layer-wise modulation with shared head parameters (Model E) achieves the best result (81.2% mAP), showing the benefits of hierarchical modeling and proper parameter sharing.

Effects of Randomness and Token Count. As shown in Fig. 5, introducing a small amount of randomness improves generalization, while higher levels degrade performance due to excessive noise. In terms of token count, using four learnable tokens achieves the best results. Adding more tokens leads to diminishing returns and potential overfitting.

Index	SFM Setting	mAP	R-1	R-5	R-10
A	Shared All Layers	77.2	80.9	86.6	90.3
B	Early Layers	73.5	74.3	83.9	87.9
C	Late Layers	76.7	78.9	87.1	91.3
D	Head-wise Parameters	77.1	79.9	86.6	91.0
E	**All Layers (Full Model)**	**81.2**	**83.4**	**90.2**	**91.6**

Table 5: Ablation study on different SFM configurations.

Visualization

Feature Distributions. We visualize the distribution of multi-modal features using t-SNE. As shown in Fig. 4, with the introduction of the SFM module, the features become more compact and identity clusters are better separated. The addition of STR further enhances intra-class compactness and inter-class separability. Finally, the full STMI model produces the clearest and most structured distribution, demonstrating the effectiveness of each module.

Conclusion

In this work, we propose STMI, a novel framework for multi-modal object ReID that addresses the limitations of token loss and weak semantic alignment in existing methods. Specifically, we introduce the Segmentation-Guided Feature Modulation (SFM) module to enhance foreground regions and suppress background noise based on SAM-generated masks. The Semantic Token Reallocation (STR) module extracts compact and informative semantic tokens via learnable queries and cross-attention, avoiding information loss from hard filtering. Furthermore, the Cross-Modal Hypergraph Interaction (CHI) module captures high-order semantic relationships across modalities through a unified hypergraph structure. Moreover, we construct a caption generation strategy that fuses multi-modal inputs to produce reliable textual descriptions. Extensive experiments on three public multi-modal ReID benchmarks demonstrate that STMI achieves state-of-the-art performance, validating its effectiveness and generalizability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62362051) and the Key Scientific and Technological R&D Program of Dalian (No. 2023YF11GX012).

References

Bai, S.; Zhang, F.; and Torr, P. H. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110: 107637.

Bian, Y.; Liu, M.; Yi, Y.; Wang, X.; Ma, Y.; and Wang, Y. 2025. Modality Unified Attack for Omni-Modality Person Re-Identification. *TIFS*.

Crawford, J.; Yin, H.; McDermott, L.; and Cummings, D. 2023. UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification. *arXiv preprint arXiv:2310.18812*.

Feng, Y.; Li, J.; Xie, C.; Tan, L.; and Ji, J. 2025. Multi-Modal Object Re-identification via Sparse Mixture-of-Experts. In *ICML*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Guo, J.; Zhang, X.; Liu, Z.; and Wang, Y. 2022. Generative and attentive fusion for multi-spectral vehicle re-identification. In *ICSP*, 1565–1572.

He, Q.; Lu, Z.; Wang, Z.; and Hu, H. 2023a. Graph-Based Progressive Fusion Network for Multi-Modality Vehicle Re-Identification. *TITS*, 1–17.

He, Z.; Shi, H.; Wu, Y.; and Tu, Z. 2023b. Low-rank fusion network for multi-modality person re-identification. In *ICSP*, 1578–1581. IEEE.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, M.; Jia, S.; Gu, J.; Lu, X.; Zhu, G.; Dong, A.; and Zhang, L. 2025. VoteSplat: Hough Voting Gaussian Splatting for 3D Scene Understanding. *arXiv preprint arXiv:2506.22799*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Kreiss, S.; Bertoni, L.; and Alahi, A. 2021. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *TITS*, 23(8).

Li, H.; Li, C.; Zhu, X.; Zheng, A.; and Luo, B. 2020. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, volume 34, 11345–11353.

Li, S.; Li, C.; Zheng, A.; Lu, A.; Tang, J.; and Ma, J. 2025a. NEXT: Multi-Grained Mixture of Experts via Text-Modulation for Multi-Modal Object Re-ID. *arXiv preprint arXiv:2505.20001*.

Li, S.; Li, C.; Zheng, A.; Tang, J.; and Luo, B. 2025b. ICPL-ReID: Identity-Conditional Prompt Learning for Multi-Spectral Object Re-Identification. *arXiv preprint arXiv:2505.17821*.

Lin, M.; Wang, S.; Wang, X.; Tang, J.; Fu, L.; Zuo, Z.; and Sang, N. 2025. DMPT: Decoupled Modality-aware Prompt Tuning for Multi-modal Object Re-identification. In *WACV*, 2103–2112. IEEE.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, 13334–13343.

Pan, W.; Huang, L.; Liang, J.; Hong, L.; and Zhu, J. 2023. Progressively Hybrid Transformer for Multi-Modal Vehicle Re-Identification. *Sensors*, 23(9): 4206.

Qi, L.; Huo, J.; Wang, L.; Shi, Y.; and Gao, Y. 2019. A mask based deep ranking neural network for person retrieval. In *ICME*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Shi, J.; Yin, X.; Zhang, Y.; Xie, Y.; Qu, Y.; et al. 2024. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *NeurIPS*, 37: 99715–99734.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2018. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 1179–1188.

Sun, Q.; Luo, J.; Zhang, D.; and Li, X. 2025. SmartFreeEdit: Mask-Free Spatial-Aware Image Editing with Complex Instruction Understanding. *arXiv preprint arXiv:2504.12704*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2025. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3): 1958–1974.

Tian, M.; Yi, S.; Li, H.; Li, S.; Zhang, X.; Shi, J.; Yan, J.; and Wang, X. 2018. Eliminating background-bias for robust person re-identification. In *CVPR*, 5794–5803.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wan, X.; Zheng, A.; Jiang, B.; Wang, B.; Li, C.; and Tang, J. 2025a. UGG-ReID: Uncertainty-Guided Graph Model for Multi-Modal Object Re-Identification. *arXiv preprint arXiv:2507.04638*.

Wan, X.; Zheng, A.; Wang, Z.; Jiang, B.; Tang, J.; and Ma, J. 2025b. Reliable Multi-Modal Object Re-Identification via Modality-Aware Graph Reasoning. *arXiv preprint arXiv:2504.14847*.

Wang, Y.; Liu, X.; Yan, T.; Liu, Y.; Zheng, A.; Zhang, P.; and Lu, H. 2024a. MambaPro: Multi-Modal Object Re-Identification with Mamba Aggregation and Synergistic Prompt. *arXiv preprint arXiv:2412.10707*.

Wang, Y.; Liu, X.; Zhang, P.; Lu, H.; Tu, Z.; and Lu, H. 2024b. Top-reid: Multi-spectral object re-identification with token permutation. In *AAAI*, volume 38, 5758–5766.

Wang, Y.; Liu, Y.; Zheng, A.; and Zhang, P. 2024c. Decoupled Feature-Based Mixture of Experts for Multi-Modal Object Re-Identification. *arXiv preprint arXiv:2412.10650*.

Wang, Y.; Lv, Y.; Zhang, P.; and Lu, H. 2025. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *CVPR*, 29701–29710.

Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; and Wang, D. 2021. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, 12026–12035.

Wang, Z.; Huang, H.; Zheng, A.; and He, R. 2024d. Heterogeneous Test-Time Training for Multi-Modal Person Re-identification. In *AAAI*, volume 38, 5850–5858.

Wang, Z.; Li, C.; Zheng, A.; He, R.; and Tang, J. 2022. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *AAAI*, volume 36, 2633–2641.

Wu, D.; Liu, Z.; Chen, Z.; Gan, S.; Tan, K.; Wan, Q.; and Wang, Y. 2025. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *ESWA*, 263: 125716.

Yang, X.; Dong, W.; Cheng, D.; Wang, N.; and Gao, X. 2025a. Tienet: A tri-interaction enhancement network for multimodal person reidentification. *TNNLS*.

Yang, X.; Zhu, L.; Fan, H.; and Yang, Y. 2025b. VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing. In *ICLR*.

Yin, H.; Li, J.; Schiller, E.; McDermott, L.; and Cummings, D. 2023. GraFT: Gradual Fusion Transformer for Multi-modal Re-Identification. *arXiv preprint arXiv:2310.16856*.

Yu, Z.; Huang, Z.; Hou, M.; Pei, J.; Yan, Y.; Liu, Y.; and Sun, D. 2024. Representation Selective Coupling via Token Sparsification for Multi-Spectral Object Re-Identification. *TCSVT*.

Yu, Z.; Huang, Z.; Hou, M.; Yan, Y.; and Liu, Y. 2025. WTSF-ReID: Depth-driven Window-oriented Token Selection and Fusion for multi-modality vehicle re-identification with knowledge consistency constraint. *ESWA*, 126921.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 516–525.

Zhang, P.; Wang, Y.; Liu, Y.; Tu, Z.; and Lu, H. 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, 17117–17126.

Zhang, S.; Luo, W.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; and Zhang, Y. 2025. Prompt-based modality alignment for effective multi-modal object re-identification. *TIP*.

Zhao, K.; Liu, X.; Sun, Z.; Wang, L.; Wang, X.; Cui, Q.; Li, X.; and Guo, Z. 2023. Multimodal consistency co-assisted training for person re-identification. In *MLCCIM*, 107–111. IEEE.

Zheng, A.; He, Z.; Wang, Z.; Li, C.; and Tang, J. 2023a. Dynamic Enhancement Network for Partial Multi-modality Person Re-identification. *arXiv preprint arXiv:2305.15762*.

Zheng, A.; Ma, Z.; Sun, Y.; Wang, Z.; Li, C.; and Tang, J. 2025a. Flare-aware cross-modal enhancement network for multi-spectral vehicle Re-identification. *Information Fusion*, 116: 102800.

Zheng, A.; Sun, Y.; Wang, Z.; Li, C.; and Tang, J. 2025b. Collaborative Enhancement Network for Low-quality Multi-spectral Vehicle Re-identification. *arXiv preprint arXiv:2504.14877*.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust multi-modality person re-identification. In *AAAI*, volume 35, 3529–3537.

Zheng, A.; Zhu, X.; Ma, Z.; Li, C.; Tang, J.; and Ma, J. 2023b. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100: 101901.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, volume 34, 13001–13008.