NEXT: Multi-Grained Mixture of Experts via Text-Modulation for Multi-Modal Object Re-Identification

Shihao Li, Huaibo Huang, Junxian Duan, Aihua Zheng*, Jin Tang, and Jixin Ma

***Abstract*—Multi-modal object Re-IDentification (ReID) aims to obtain complete identity features across heterogeneous modalities. However, most existing methods rely on implicit feature fusion modules, making it difficult to model fine-grained recognition patterns under various challenges in real world. Benefiting from the powerful Multi-modal Large Language Models (MLLMs), the object appearances are effectively translated into descriptive captions. In this paper, we propose a reliable caption generation pipeline based on attribute confidence, which significantly reduces the unknown recognition rate of MLLMs and improves the quality of generated text. Additionally, to model diverse identity patterns, we propose a novel ReID framework, named NEXT, the Multi-grained Mixture of Experts via Text-Modulation for Multi-modal Object Re-Identification. Specifically, we decouple the recognition problem into semantic and structural branches to separately capture fine-grained appearance features and coarse-grained structure features. For semantic recognition, we first propose the Text-Modulated Semantic Experts (TMSE) module, which randomly samples high-quality captions to modulate experts capturing semantic features and mining inter-modality complementary cues. Second, to recognize structure features, we propose the Context-Shared Structure Experts (CSSE) module, which focuses on the holistic object structure and maintains identity structural consistency via a soft routing mechanism. Finally, we propose the Multi-Grained Features Aggregation (MGFA) module, which adopts a unified fusion strategy to effectively integrate multi-grained expert features into the final identity representations. Extensive experiments on two public person datasets and three vehicle datasets demonstrate the effectiveness of our method, showing that it significantly outperforms existing state-of-the-art methods.**

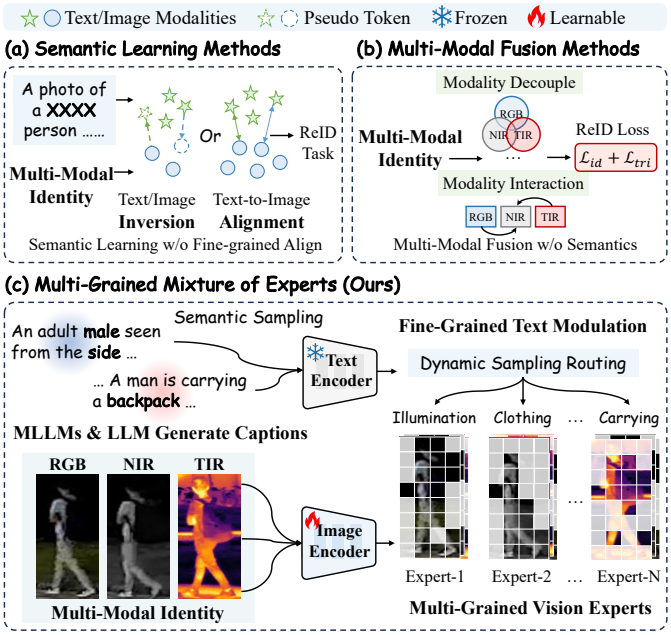

Fig. 1. Comparison with existing methods: (a) Previous semantic learning methods capture identity information via text/image prompt inversion or alignment strategies, which fail to achieve fine-grained alignment. (b) Existing multi-modal fusion methods fuse modality features through implicit fusion modules such as modal decoupling and interaction strategies, while the fusion process still lacks explicit semantic guidance. (c) Our proposed method modulates visual modalities via diverse text captions and decouples appearance and structure into multi-grained experts.

I. Introduction

OBJECT Re-IDentification (ReID) aims to construct discriminative identity representations by capturing object characteristics such as appearance, clothing, body shape, and posture [1]–[4]. However, in real-world scenarios, single-modal images are often vulnerable to perturbations caused by illumination changes, adverse weather, and occlusions [5]–[10], which hinders the performance of classical ReID methods. To address these challenges, multi-modal object ReID [11]–[14] has attracted increasing research interest in recent years. Benefiting from the complementary advantages of diverse spectral modalities, the identity features can be comprehensively captured. Despite this, intra-modal noise and inter-modal heterogeneity remain the key limitations for effective identity representation fusion. To overcome these challenges, existing methods focus on implicit fusion methods such as identity-level feature interaction [15]–[17], fusion feature decoupling [18], [19], data augmentation [20], [21], and frequency-domain interactions [22], [23]. However, explicit semantic-aware fusion [24], [25] of multi-modal identity features remains underexplored.

Recent studies in vision-language models [26]–[29] have advanced visual understanding tasks [30], [31]. Pioneering meth-

This research is supported in part by the National Natural Science Foundation of China under Grants 62372003, and the Natural Science Foundation of Anhui Province under Grants 2308085Y40. (*The corresponding author is Aihua Zheng.)

A. Zheng and S. Li are with the School of Artificial Intelligence, Anhui University, Hefei, 230601, China, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei, 230601, China, and the State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Hefei, 230000, China (e-mail: ahzheng214@foxmail.com; shli0603@foxmail.com).

H. Huang and J. Duan are with the State Key Laboratory of Multimodal Artificial Intelligence Systems and the New Laboratory of Pattern Recognition, CASIA, Beijing, 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China (e-mail: huaibo.huang@cripac.ia.ac.cn; junxian.duan@ia.ac.cn).

J. Tang is with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: tangjin@ahu.edu.cn).

J. Ma is with School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK (e-mail: j.ma@greenwich.ac.uk).

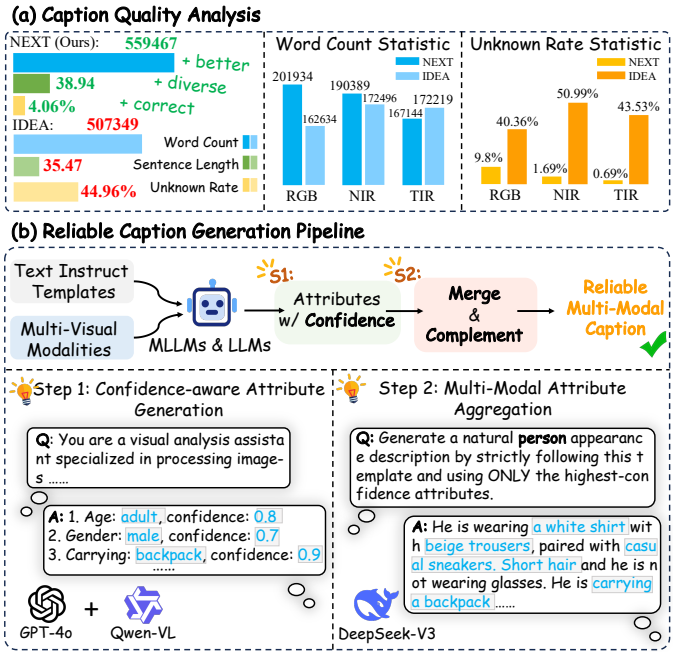

Fig. 2. Caption analysis and generation pipeline. We first report the word count and the rate of captions containing '*none*,' '*unknown*,' or '*unclear*' words to assess generation quality, and then show the reliable caption generation pipeline with MLLMs.

ods, such as IDEA [24], TVI-LFM [32], and MP-ReID [33], integrate MLLMs into ReID tasks by generating identity-relevant captions, thereby extending identity representations into the textual modality. However, low-quality spectral noise and style discrepancies introduce semantic noise and omissions into the text generation process of MLLMs. Information bias during long-context text generation serves as a major cause of hallucinations [34] in MLLMs, with longer texts increasing the likelihood of such errors. Additionally, guiding MLLMs to reason the fine-grained attributes across visible and infrared spectra is a highly labor-intensive process. Incorrect text caption labels cause ambiguity and semantic bias [35]–[37]. To tackle this, we propose a reliable caption generation framework, as shown in Fig. 2(b). Specifically, we decompose the caption generation process into two steps: confidence-aware attribute generation and multi-modal attribute aggregation. By enforcing the model to output a confidence score for each attribute, we encourage the MLLMs to perform fine-grained evaluation for each attribute. Based on the confidence scores, we quantify the attribute importance across different modalities, enabling us to complement missing semantic information through multi-modal attribute sets.

The high-quality captions effectively guide the model to focus on identity-relevant representations. Based on this, we propose a multi-grained mixture of experts framework named NEXT, as shown in Fig. 1(c), which models the identity recognition process through semantic-sampling and structure-aware experts. The framework consists of three main components: Text-Modulated Semantic Experts (TMSE), Context-Shared Structure Experts (CSSE), and Multi-Grained Features Aggregation (MGFA). Specifically, TMSE samples part-level fine-grained features within each modality. The text modulation mechanism encourages experts to focus on semantically relevant regions and capture intra-modality semantic identity features. Meanwhile, CSSE perceives inter-modality coarse-grained features of the object identity through a soft routing mechanism shared across modalities, which maintains the integrity of structural features. Finally, MGFA unifies the features of multi-grained experts into a comprehensive identity representation. Extensive experiments on five multi-modal object ReID benchmarks demonstrate the effectiveness of our proposed approach.

In summary, our contributions are as follows:

- To obtain accurate object appearance descriptions, we propose a reliable caption generation pipeline that leverages attribute confidence scores to effectively harness MLLMs in producing reliable high-quality captions.
- To model diverse identity patterns, we propose the Text-Modulated Semantic Experts (TMSE) and the Context-Shared Structure Experts (CSSE) to decouple the recognition problem into fine-grained semantic and coarse-grained structural branches. The Multi-Grained Features Aggregation (MGFA) is proposed to integrate multi-grained experts into a unified identity representation.
- Extensive experiments are conducted on five public benchmarks to validate the effectiveness of our method. The results demonstrate that the proposed method significantly outperforms the state-of-the-art methods.

II. Related Work

A. Multi-Modal Object Re-Identification

Multi-modal object ReID leverages the complementary imaging advantages of multi-spectra to enable robust identity recognition under adverse conditions. However, intra-modal noise and inter-modal spectral heterogeneity remain significant challenges. To fuse the spectral modalities, Wang *et al.* [15] propose the TOP-ReID which introduces a modality permutation and reconstruction mechanism to align multi-modal representations. Zhang *et al.* [16] propose a prompt-based token selection and fusion strategy to effectively integrate multi-modal features and suppress background interference. Wang *et al.* [17] propose the MambaPro, which presents a selective state-space fusion approach based on Mamba [38] to aggregate intra- and inter-modal features. Zheng *et al.* [14] propose the FACENet which utilizes illumination priors to enhance degraded modalities. Wang *et al.* [23] adopt frequency- and feature-based selection mechanisms to filter out background and low-quality noise. Yang *et al.* [22] propose the TIENet which introduces an amplitude-guided phase learning strategy for frequency-domain enhancement. Li *et al.* [25] propose the ICPL-ReID which applies an identity prototype-based conditional prompt learning method to guide semantic learning across modalities. Wang *et al.* [24] inverts textual features into the visual space to guide model learning, and captures discriminative features across multi-modal via deformable aggregation. Despite these advancements, existing methods still lack a deep exploration of identity semantics and intrinsic structures. To bridge this gap, we propose a multi-grained

mixture of experts via text-modulation that fuses features from different modalities through semantic and structural perspectives, aiming to enhance multi-modal identity representation based on explicit semantics.

B. Semantic Learning in Re-Identification

The large-scale Vision-Language foundation Models (VLM), such as CLIP [29], exhibit powerful visual perception and semantic understanding capabilities. Researchers employ the learnable prompt [39], [40] to effectively adapt these models to classification and retrieval recognition tasks. In ReID, methods like CLIP-ReID [30] and PromptSG [31] utilize learnable semantic prompts to transfer the CLIP [29] model to capture object identity semantic information [41]. With the rise of MLLMs [26]–[28], numerous studies demonstrate the effectiveness of MLLMs [42], [43] in generating high-quality text and robust generalization in vision-language generation. Current approaches, such as MP-ReID [33], TVI-LFM [32], and IDEA [24], exploit MLLMs to generate identity-level textual descriptions from existing ReID datasets. However, these methods still lack reliable text generation under multi-modal conditions. Unlike existing methods [24], [32], we further quantify each generated attribute and propose a confidence-aware attribute generation strategy. We incorporate a multi-modal merging and complementation mechanism to achieve high-quality and reliable caption generation. Additionally, we introduce the NEXT framework, which leverages semantic experts to extract fine-grained identity features, aiming to tackle diverse semantic challenges in multi-modal scenarios.

C. Mixture of Experts for Multi-Modal Learning

The Mixture-of-Experts (MoE) [44]–[46] paradigm decomposes a large and complex deep model into lightweight expert networks and a routing module. By intelligently routing inputs to different experts, the model enables flexible adaptation to diverse and complex tasks and environments [47], [48]. Recently, the MoE framework has been increasingly adopted in multi-modal learning [49]–[51]. Yun *et al.* [50] propose the Flex-MoE, which introduces a flexible framework that effectively incorporates arbitrary modality combinations and addresses the missing modality scenario across various domains. Fang *et al.* [52] recognize the modality importance varying across samples in fusion, and propose the EMoE which uses the predictive information of each modality to guide the fused features, preserving modality-specific characteristics in the multi-modal fusion. Peng *et al.* [53] introduce a novel top-down dynamic fusion mechanism that adaptively integrates multi-modal information, addressing the variation in data quality across different modalities and patients in various medical domains. Wang *et al.* [18] propose the DeMo, which first introduces the MoE framework to the multi-modal ReID task by combining fusion-specific experts to decouple shared and specific fusion features. Feng *et al.* [54] proposed MFRNet, which adopts an MoE-based modality generator for pixel-level modality feature generation and interaction, and MoE-based adapter to rebalance modality-shared and modality-specific representations. Despite this, these methods still lack multi-grained semantic understanding. To address this limitation, we propose text-modulated semantic experts and context-shared structure experts to jointly capture semantic and structural features, enabling more comprehensive identity representation of multi-modalities.

III. Methodology

In this section, we elaborate on the proposed framework, NEXT. As shown in Fig. 3, we propose the Multi-Grained Mixture-of-Experts framework that includes the Text-Modulated Semantic Experts (TMSE), the Context-Shared Structure Experts (CSSE), and the Multi-Grained Features Aggregation (MGFA) module.

A. Multi-Modal Features Extraction

Each multi-modal sample includes three modalities: Visible Light (RGB), Near Infrared (NIR), and Thermal Infrared (TIR), denoted as $\mathbf{X}_I = [X_{I,rgb}, X_{I,nir}, X_{I,tir}]$, along with detailed appearance captions $\mathbf{X}_T = [X_{T,rgb}, X_{T,nir}, X_{T,tir}]$ which are generated by our pipeline in Sec.IV. We feed the visual modality $\mathbf{X}_I$ into the CLIP visual encoder $\mathcal{V}(\cdot)$ to extract visual features $\mathbf{F}_I = [\mathbf{f}_I^{\text{cls}}; \mathbf{F}_I^{\text{tok}}] \in \mathbb{R}^{3\times(1+N)\times D}$, and input the appearance description $\mathbf{X}_T$ into the frozen CLIP text encoder $\mathcal{T}(\cdot)$ to extract textual features $\mathbf{F}_T = [\mathbf{f}_T^{\text{cls}}; \mathbf{F}_T^{\text{tok}}] \in \mathbb{R}^{3\times(1+L)\times D}$, where N denotes the number of visual patch tokens, L is the length of text tokens, D is the dimension length.

B. Text-Modulated Semantic Experts

In real world, challenges like illumination variations, background noise, and image degradation impact quality of local regions across modalities, making it essential to bridge modality gaps and extract fine-grained complementary features. To address this, we use text-modulated routing to sample semantically relevant local visual tokens within each modality. Therefore, TMSE focuses on identity-related local cues, such as clothing, accessories, body parts, vehicle components, and other appearance-level attributes. Specifically, we leverage the object-level textual semantics $\mathbf{X}_T$ extracted from MLLMs to guide a set of semantic sampling experts $\mathbf{E}_T = \{E_t^{(1)}, E_t^{(2)}, \ldots, E_t^{(N_T)}\}$ in dynamically select semantic-related local features across different modalities, to tackle various semantic challenges, where N_T denotes the number of semantic experts. The structure of each expert is defined as:

$$E_t^{(i)}(\mathbf{F}) = \text{Dropout}\left(\text{MLP}\left(\text{LN}(\mathbf{F})\right)\right) + \mathbf{F}, \tag{1}$$

where $\mathbf{F}$ denotes the input features, LN denotes a layer normalization, MLP refers to a multilayer perceptron, and Dropout with a rate of 0.1 is applied to prevent overfitting.

Dynamic Sampling Routing. To sample semantic features from each modality, we design a set of modality-specific sampling routes for each expert, denoted as $\mathbf{R} = [\mathbf{R}_{rgb}, \mathbf{R}_{nir}, \mathbf{R}_{tir}]$. The patch-level features $\mathbf{F}_{I,m}^{\text{tok}}$ are encoded

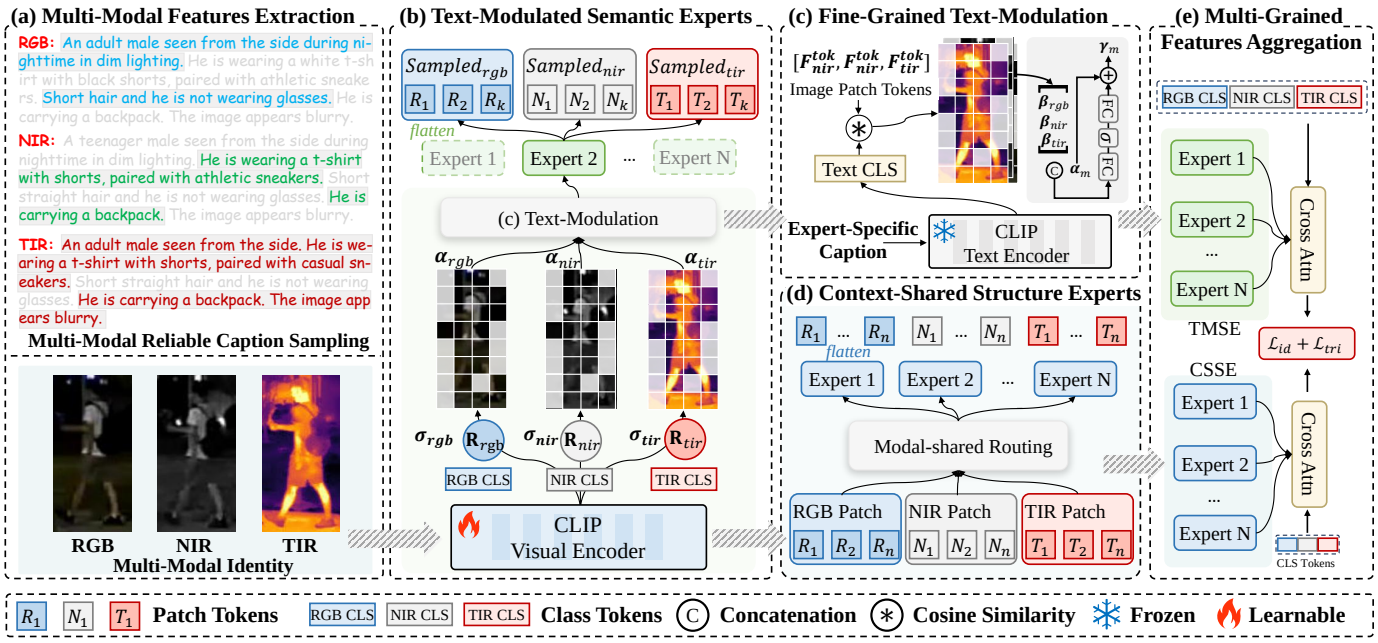

Fig. 3. The overview of our proposed framework. First, we divide multi-modal features into (a) reliable caption sampling and multi-modal identity representations. Then, we feed visual modalities into the CLIP visual encoder to obtain visual embeddings, which are passed into both (b) the Text-Modulated Semantic Experts (TMSE) and (d) Context-Shared Structure Experts (CSSE), decoupling identity recognition into fine-grained appearance recognition and coarse-grained structure recognition. For TMSE, we input randomly sampled captions into (c) the Fine-Grained Text-Modulation to guide the sampling process. Finally, (e) the Multi-Grained Features Aggregation (MGFA) efficiently integrates multi-grained expert features to form the final identity representation.

into the routing matrix $\boldsymbol{\alpha}_m$, and the class token $\mathbf{f}_{I,m}^{\text{cls}}$ are mapped into the dynamic threshold $\boldsymbol{\sigma}_m$ as follows:

$$\begin{aligned} \boldsymbol{\alpha}_m &= \text{FC}^2\left(\delta\left(\text{FC}^1(\mathbf{F}_{I,m}^{\text{tok}})\right)\right), \quad \boldsymbol{\alpha}_m \in \mathbb{R}^{H\times W} \\ \boldsymbol{\sigma}_m &= \text{FC}^2\left(\delta\left(\text{FC}^1(\mathbf{f}_{I,m}^{\text{cls}})\right)\right), \quad \boldsymbol{\sigma}_m \in \mathbb{R} \end{aligned} \tag{2}$$

where $\text{FC}^1 \in \mathbb{R}^{\text{D}\times\text{D}}$ and $\text{FC}^2 \in \mathbb{R}^{\text{D}\times 1}$ denotes fully connected layers, δ is GELU activation function [55], and $m \in \{rgb, nir, tir\}$. Then, we calculate the sampling matrix $\mathbf{M}_m$ to capture the key tokens for current expert,

$$\mathbf{M}_m(i,j) = \begin{cases} 1, & \text{if } \boldsymbol{\alpha}_m(i,j) > \boldsymbol{\sigma}_m \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

as formulated in Eq. (3), we assign 1 to entries in $\boldsymbol{\alpha}_m$ that exceed the threshold $\boldsymbol{\sigma}_m$ and 0 otherwise, to discard irrelevant tokens while keeping the operation differentiable.

Fine-Grained Text Modulation. To empower the semantic awareness of each expert, we employ random semantic text prompts during training to guide each expert to focus on distinct semantic challenges. Formally, the raw textual caption is decomposed into fine-grained sentences for each modality: $X_{T,m} = \{s_m^{(1)}, s_m^{(2)}, \ldots, s_m^{(n_m)}\}$, where n_m is the maximum number of sentences. For each expert $E_t^{(i)}$, we randomly select a subset $\hat{X}_{T,m}^{(i)}$ from $X_{T,m}$ as its modulation signal. The sampling process is defined as:

$$\begin{aligned} \hat{X}_{T,m}^{(i)} &\sim \textbf{Sampling}(X_{T,m}), \\ \hat{\mathbf{X}}_T^{(i)} &= [\hat{X}_{T,rgb}^{(i)}, \hat{X}_{T,nir}^{(i)}, \hat{X}_{T,tir}^{(i)}], \end{aligned} \tag{4}$$

where i denotes the i-th semantic expert, and **Sampling** is a random probability distribution which controls the diversity and consistency of modulated signals.

To modulate the sampling route, we first encode the randomly sampled high-quality semantic text into the visual-text latent space as text feature vector $\mathbf{f}_{\hat{T},m}^{\text{cls}}$. Second, we compute the cosine similarity between this text vector and the visual patch tokens to obtain the semantic matrix $\boldsymbol{\beta}_m$:

$$\boldsymbol{\beta}_m = \frac{\mathbf{f}_{\hat{T},m}^{\text{cls}} \cdot \mathbf{F}_{I,m}^{\text{tok}}}{\|\mathbf{f}_{\hat{T},m}^{\text{cls}}\|_2 \cdot \|\mathbf{F}_{I,m}^{\text{tok}}\|_2}, \quad \boldsymbol{\beta}_m \in \mathbb{R}^{H\times W} \tag{5}$$

where $\|\cdot\|_2$ means L_2 normalization. Then, the semantic guidance is integrated with the route matrix $\boldsymbol{\alpha}_m$ to produce the modulation matrix $\boldsymbol{\gamma}_m$:

$$\boldsymbol{\gamma}_m = \text{FC}^4\left(\delta\left(\text{FC}^3(\boldsymbol{\alpha}_m, \boldsymbol{\beta}_m)\right)\right) + \boldsymbol{\alpha}_m, \boldsymbol{\gamma}_m \in \mathbb{R}^{H\times W} \tag{6}$$

where $\text{FC}^3 \in \mathbb{R}^{2\times \text{D}/2}$ and $\text{FC}^4 \in \mathbb{R}^{\text{D}/2\times 1}$ denotes the fully connected layer within the modulation network whose structure is illustrated in Fig. 3(c).

Finally, we replace the route matrix $\boldsymbol{\alpha}_m$ in Eq. (3) with the modulated matrix $\boldsymbol{\gamma}_m$ to obtain the final modulation-based sampling matrix $\hat{\mathbf{M}}_m$.

$$\hat{\mathbf{F}}_{I,m}^{\text{tok}} = E_t(\mathbf{F}_{I,m}^{\text{tok}}), \hat{\mathbf{F}}_I = \text{Concat}\left(\hat{\mathbf{M}}_m \odot \hat{\mathbf{F}}_{I,m}^{\text{tok}}\right), \tag{7}$$

According to Eq. (7) above, $\hat{\mathbf{M}}_m$ dynamically samples the informative features for each semantic expert $\text{E}_t^{(i)}$ within a modality. These sampled features are concatenated to obtain the final multi-modal semantic feature $\hat{\mathbf{F}}_I$.

C. Context-Shared Structure Experts

While TMSE focuses on fine-grained part recognition under various semantic challenges, CSSE is designed for coarse-

grained structural recognition. Instead of using modality-specific semantic sampling, CSSE adopts a modality-shared routing mechanism over concatenated multi-modal patch tokens. Its goal is to preserve the holistic object structure and maintain structural consistency across modalities, such as viewpoint, body shape, vehicle contour, and global spatial configuration. To this end, we introduce the structure aware experts, denoted as $\mathbf{E}_C = \{E_c^{(1)}, E_c^{(2)}, \ldots, E_c^{(N_C)}\}$, to comprehensively perceive the holistic object structure across modalities, where N_C denotes the number of structure experts and $E_c^{(i)}$ shares the same architecture as $E_t^{(i)}$ in Eq. (1).

Specifically, we first concatenate multi-modal patch features as $\mathbf{F}_I^{\text{tok}} = [\mathbf{F}_{I,rgb}^{\text{tok}}, \mathbf{F}_{I,nir}^{\text{tok}}, \mathbf{F}_{I,tir}^{\text{tok}}]$, and then feed them into a modality-shared routing network $\mathbf{R}_s$ to compute the expert fusion weights as follows:

$$\boldsymbol{\omega} = \mathbf{Softmax}\left(\mathrm{FC}\left(\mathbf{F}_I^{\text{tok}}\right)\right), \quad \boldsymbol{\omega} \in \mathbb{R}^{N_C}, \tag{8}$$

where $\mathbf{Softmax}$ encodes the expert selection weights, resulting in the soft routing matrix $\boldsymbol{\omega}$.

Finally, we feed the concatenated multi-modal token features into the structure experts E_C and compute the final multi-modal structural representation by weighting each output of the expert using the soft routing matrix $\boldsymbol{\omega}$. The process is formulated as follows:

$$\begin{aligned} \tilde{\mathbf{F}}_I &= \sum_{i=1}^{N_C} \omega_i \cdot E_c^{(i)}(\mathbf{F}_I^{\text{tok}}), \\ \tilde{\mathbf{F}}_I^{\text{tok}} &= \left\{ E_c^{(i)}(\mathbf{F}_I^{\text{tok}}) \mid i = 1, 2, \ldots, N_C \right\}, \end{aligned} \tag{9}$$

where $\tilde{\mathbf{F}}_I$ represents the final multi-modal structure feature.

D. Multi-Grained Features Aggregation

Different types of experts mine identity features at varying levels of granularity. The semantic experts E_T extract fine-grained semantic features $\hat{\mathbf{F}}_I^{(i)}$, and the structural experts E_C capture coarse-grained structural consistency features $\tilde{\mathbf{F}}_I$. As illustrated in Fig. 3(e), we combine these features to form the complete object identity feature set $\mathbf{F}_I^{\text{exp}} = [\hat{\mathbf{F}}_I^{(1)}, \hat{\mathbf{F}}_I^{(2)}, \ldots, \hat{\mathbf{F}}_I^{(N_T)}; \tilde{\mathbf{F}}_I]$. The modality class tokens are concatenated to form the modality query features $\mathbf{f}_I^{\text{cls}} = [\mathbf{f}_{I,rgb}^{\text{cls}}, \mathbf{f}_{I,nir}^{\text{cls}}, \mathbf{f}_{I,tir}^{\text{cls}}]$. Following this, we treat $\mathbf{f}_I^{\text{cls}}$ as the query features Q and the expert features as the key features K and value features V. Through the Cross-Attention mechanism [56], we obtain the multi-modal representation for each expert and average them to form the final identity representation $\hat{\mathbf{f}}_I^{\text{cls}}$. The process is formulated as follows:

$$\hat{\mathbf{f}}_I^{\text{cls}} = \mathrm{Average}\left(\hat{\mathbf{f}}_I^{\text{cls}(1)}, \hat{\mathbf{f}}_I^{\text{cls}(2)}, \ldots, \hat{\mathbf{f}}_I^{\text{cls}(\mathrm{N_T})}; \tilde{\mathbf{f}}_I^{\text{cls}}\right), \tag{10}$$

$$\begin{aligned} \hat{\mathbf{f}}_I^{\text{cls(i)}} &= \mathbf{FFN}\Big(\mathrm{LN}\Big(\mathbf{CA}\big(\mathbf{f}_I^{\text{cls}}, \mathbf{F}_I^{\text{exp(i)}}\big)\Big)\Big), \\ &\text{for } i = 1, \ldots, N_T + 1 \end{aligned} \tag{11}$$

where $\mathbf{CA}$ is the Cross-Attention mechanism, and $\mathbf{FFN}$ is the Feed-Forward Network.

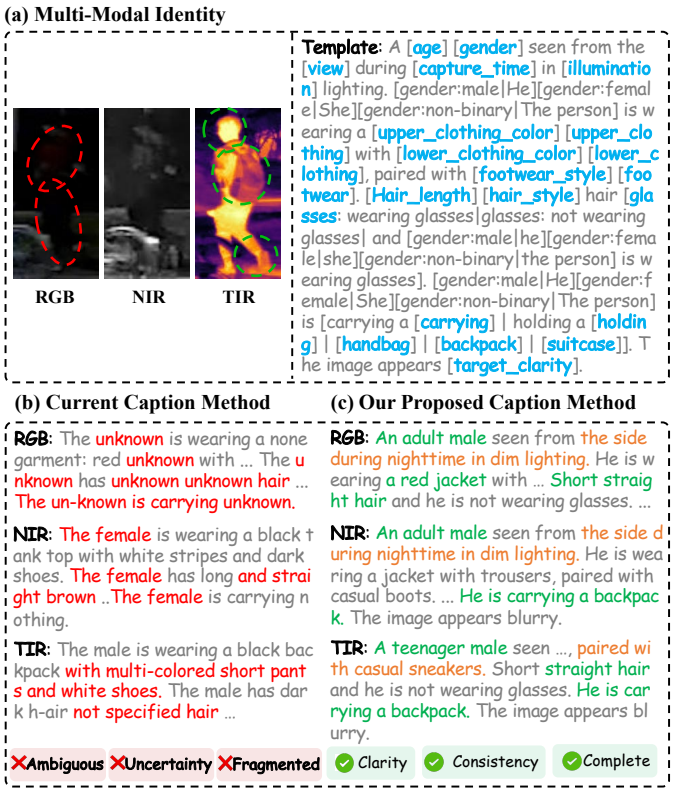

Fig. 4. Generated sample and quality comparisons on the person dataset.

E. Optimization and Inference

In line with prior works [3], [4], we use the object identity ID label as the ground-truth to supervise the classification loss $\mathcal{L}_{id}$, and adopt the triplet loss $\mathcal{L}_{tri}$ to enhance identity distribution compactness and separability. The final loss function is defined as:

$$\mathcal{L}_{final}(\hat{\mathbf{f}}_I^{\text{cls}}) = \mathcal{L}_{id}(\hat{\mathbf{f}}_I^{\text{cls}}) + \mathcal{L}_{tri}(\hat{\mathbf{f}}_I^{\text{cls}}). \tag{12}$$

During the inference, the fused feature serves as the final object representation for multi-modal retrieval.

IV. Reliable Caption Generation

Thanks to the powerful semantic understanding capability of Multi-modal Large Language Models (MLLMs), we leverage their out-of-the-box ability to generate comprehensive textual semantic descriptions of the object appearance. However, simple template-based text generation methods [24], [32], [33] face many practical challenges when dealing with multi-modal data, such as noise interference from low-quality images and significant style discrepancies in infrared modalities, making it hard to observe the object's fine-grained details. To overcome these challenges, we first employ multiple MLLMs to generate several attributes with confidence scores, and then leverage the confidence scores with a multi-modal attributes aggregation strategy to generate reliable and comprehensive identity captions.

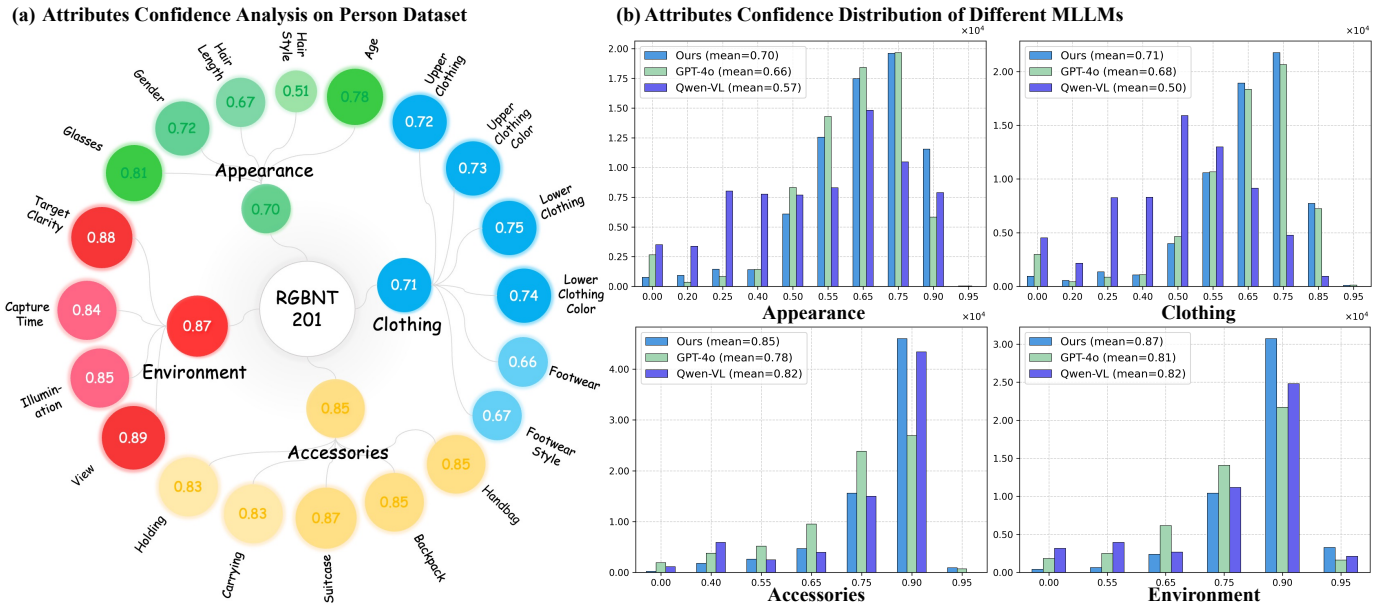

Fig. 5. In the person dataset RGBNT201, we divide attributes into four groups: Appearance, Clothing, Accessories, and Environment. (a) Most attributes are reliable and have a confidence level higher than 0.7. The size of the circle and the shade of the color represent the level of confidence. (b) The confidence distribution of our method is significantly higher than the GPT-4o and Qwen-VL native output on four different types of attributes.

A. Confidence-Aware Attribute Generation

To flexibly analyze the object appearance, we set the generation of identity attributes as the basic step in caption generation. We first define an instruction template to guide multiple MLLMs (*e.g.*, GPT-4o [26], Qwen-VL [27], etc.) in structuring the attribute set of the input image. However, simple structured attribute information cannot quantitatively measure the confidence level of each item, the model suffers from low-quality visual modality noise, which leads to incorrect attribute selection. To address this, we introduce the concept of *confidence* score in the instruction template. This allows us to observe the quality level of each MLLM assigns to its output for each attribute. Additionally, to avoid the model ignoring environment information, we not only structure the output of '*upper clothing*', '*lower clothing*', '*footwear*' and other appearance attributes, but also require the MLLMs to output '*view*', '*illumination*', '*capture time*', and '*target clarity*' attributes to capture environmental context.

B. Multi-Modal Attribute Aggregation

Based on the confidence-aware attribute set with confidence score, we are able to further generate reliable captions for the object. However, as shown in Fig. 2(a) and Fig. 4, the RGB modality is affected by lighting degradation, causing the MLLMs to fail to recognize the '*backpack*' attribute. In contrast, the thermal infrared modality for the same identity easily reveals additional '*backpack*' attribute. This leads to the model generating vague and incomplete textual captions such as '*none*,' '*unknown*,' or '*unclear*' based on a single modality. To address this problem, we employ multiple MLLMs to recognize attributes and output confidence scores. As shown in Fig. 5, GPT-4o produces higher confidence on fine-grained appearance and clothing attributes, while Qwen-VL focuses more on accessories and environment cues. Therefore, we propose a multi-modal attribute aggregation strategy based on confidence scores to preprocess complementary semantic information across different visual modalities.

Multi-Modal Merge and Complement. Specifically, we first use multiple MLLMs to identify object attributes, and for each identified attribute, we select the MLLMs with the highest confidence as their final value to improve the accuracy of each attribute within the modality. Then, taking the RGB modality as an example, its modality-specific low-frequency attributes, such as '*illumination*,' '*upper clothing color*,' and '*lower clothing color*,' are easy to recognize. In contrast, attributes requiring high-frequency information, such as '*carrying*,' '*holding*,' '*handbag*,' and '*backpack*,' are often incorrectly recognized as '*unknown*,' '*unclear*,' or '*not carrying*.' This significantly affects the clarity of the final generated captions. To address this, we rely on the confidence scores to select the highest-confidence attributes from the NIR and TIR modalities, which are unaffected by lighting conditions, for rechecking and complementing missing information. Similarly, we apply the same complementary strategy to each modality in order to obtain a complete and reliable set of attributes for each modality.

Caption Generation. Building on the above complete attribute set, we define the template-based instruction to guide the LLM [57] in combining the confidence-aware attribute set from each modality to generate the final text caption. Therefore, we have constructed a complete and reliable caption generation pipeline, and the same process is applied to vehicle datasets to obtain comprehensive semantic description of the vehicle appearance.

TABLE I
COMPARISON WITH EXISTING METHOD ON CAPTION GENERATION.

Aspect	IDEA [24]	NEXT (Ours)
Pipeline paradigm	Caption generation → attribute extraction	Attribute w/ confidence → aggregation → caption generation
Initial generation target	Free-form modality-specific description	Structured attribute set with confidence scores
MLLM usage	Reuse MLLMs for generation and extraction	Use multiple MLLMs for confidence-aware attributes
Attribute reliability modeling	No explicit attribute-level reliability estimation	Explicit confidence for each attribute
Handling hallucination noisy recognition	Prompt constraints and post-hoc extraction	Confidence-aware selection across MLLMs/modalities
Cross-modal semantic interaction	Independent modality-wise descriptions	RGB/NIR/TIR merge and complement
Main advantage	Structured and concise text annotation	Reliable, fine-grained and complete descriptions

C. Comparison of Caption Generation

As summarized in Table I, we compare our generation pipeline with existing generation method. IDEA [24] focuses on generating structured and concise captions through modality-specific description generation and post-hoc attribute extraction. In contrast, NEXT aims to produce reliable, fine-grained, and complete identity descriptions by directly generating confidence-aware attributes and aggregating complementary cues across RGB, NIR, and TIR modalities. This design enables explicit reliability estimation, cross-modal attribute selection, and semantic complementation for robust text-modulated multi-modal representation learning.

TABLE II
STATISTICAL ANALYSIS OF FIVE DATASETS.

Datasets	Type	# Samples	# IDs	# Cams	Cost (min)
RGBNT201 [13]	Person	4,787	201	4	14.4
Market-MM [58]	Person	32,668	1,501	6	51.7
MSVR310 [12]	Vehicle	2,087	310	8	14.1
RGBNT100 [11]	Vehicle	17,250	100	8	6.3
WMVEID863 [14]	Vehicle	4,709	863	8	98.0

V. EXPERIMENT

A. Datasets and Evaluation Protocols

1) Datasets: As shown in Table II, we evaluate our method on five multi-modal object ReID datasets: RGBNT201 [13], Market-MM [58], MSVR310 [12], RGBNT100 [11], and WMVEID863 [14]. To extend these datasets, we employ GPT-4o [26] and Qwen-VL [27] to automatically generate object attribute with confidence, and DeepSeek-V3 [57] to compose the final caption for each image modality. The caption generation costs for five datasets are 14.4, 51.7, 14.1, 6.3, and 98.0 minutes, respectively, with an average processing speed of 16.66 images/s. This process is conducted only once as offline preprocessing, and the generated captions can be cached and repeatedly reused for both training and testing without additional MLLM inference.

Person Datasets. RGBNT201 comprises 14,361 person images corresponding to 4,787 multi-modal samples of 201 identities, with 141 identities used for training, 30 for validation, and 30 for testing across four distinct viewpoints. Market-MM is a synthetic multi-modal person dataset derived from Market1501 [65], containing 32,668 multi-modal samples of 1,501 identities, of which 751 identities with 12,936 triples form the training set and the remaining 750 identities constitute the gallery and query partitions.

Vehicle Datasets. MSVR310 contains 6,261 images forming 2,087 multi-modal samples of 310 vehicle identities, with 155 identities and 1,032 samples used for training and the remaining 155 identities contributing 1,055 gallery samples and 591 query samples. RGBNT100 includes 100 vehicle identities and augments them with 17,250 thermal images, yielding 8,675 multi-modal triples for training and 8,575 triples for testing, from which 1,715 samples are selected as queries. WMVeID863 provides 4,709 image triplets of 863 identities captured from eight views, where 603 identities with 3,482 triplets are used for training and 260 identities with 1,226 triplets form the gallery and query sets, with 959 query samples selected.

2) Evaluation Protocols: Following the convention of community [2], [3], we use the Mean Average Precision (mAP) and Rank-K (K = 1, 5, 10) matching accuracy as evaluation metrics. As in previous works [11], [13], [14], we adopt the common evaluation protocol for RGBNT201, Market-MM, RGBNT100 and WMVEID863. For MSVR310, we filter out samples with the same identity and time span based on time labels to avoid easy matching [12].

B. Implementation Details

We resize each spectral image to 256×128 for person samples, and 128×256 for vehicle samples. Data augmentation includes random horizontal flipping, padding, cropping, and erasing [66]. All parameters in the CLIP text encoder [29] are frozen for text semantic projection, while the CLIP visual branch is fully trainable. Adam [67] optimizer is used with the learning rate of $3.5e$-4, weight decay of 0.0001, and momentum set to 0.9. All experiments are conducted on one NVIDIA RTX 4090 GPU using the PyTorch toolkits.

C. Comparison with State-of-the-Art Methods

Performance on Person Dataset. In Table III, we compare our NEXT with existing methods on the RGBNT201 and Market-MM datasets. Benefiting from the flexible fusion of semantic and structural experts, our method achieves a significant performance lead, reaching 82.4%/86.6% mAP and Rank-1 accuracy. Against IDEA [24], which utilizes semantic inversion and deformable offset sampling, our method leverages flexible text-modulation to guide expert sampling multi-modal semantics, improving mAP and Rank-1 by +2.2% and +4.5%, respectively. Compared with MFRNet [54], which employs the MoE structure to enhance cross-modal generation and representation learning, NEXT focuses more on multi-grained identity semantic and structural modeling, thereby achieving clear performance advantages. On the Market-MM dataset, NEXT also establishes clear advantages, reaching 85.8%/95.3% mAP and Rank-1 accuracy. The consistent improvements on the two datasets demonstrate the robustness of NEXT in learning discriminative identity features under diverse multi-modal conditions.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON PERSON DATASETS: RGBNT201, MARKET-MM ABOUT MAP(%) AND RANK-K(%). THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

	Methods	Venue	Backbone	RGBNT201				Market-MM			
				mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Single-Modal	HACNN [59]	CVPR'18	ResNet	21.3	19.0	34.1	42.8	42.9	69.1	86.6	92.2
	MLFN [60]	CVPR'18	ResNet	26.1	24.2	35.9	44.1	42.7	68.1	87.1	92.0
	OSNet [61]	ICCV'19	ResNet	25.4	22.3	35.1	44.7	39.7	69.3	86.7	91.3
	CAL [62]	ICCV'21	ResNet	27.6	24.3	36.5	45.7	-	-	-	-
	TransReID [3]	ICCV'21	ViT-B/16	63.8	65.8	78.5	83.9	73.0	88.9	95.8	97.6
Multi-Modal	HAMNet [11]	AAAI'20	ResNet	27.7	26.3	41.5	51.7	60.0	82.8	92.5	95.0
	PFNet [13]	AAAI'21	ResNet	38.5	38.9	52.0	58.4	60.9	83.6	92.8	95.5
	IEEE [58]	AAAI'22	ResNet	46.4	47.1	58.5	64.2	64.3	83.9	93.0	95.7
	HTT [63]	AAAI'24	ViT-B/16	71.1	73.4	83.1	87.3	67.2	81.5	95.8	97.8
	TOP-ReID [15]	AAAI'24	ViT-B/16	72.3	76.6	84.7	89.4	82.0	92.4	97.6	98.6
	EDITOR [23]	CVPR'24	ViT-B/16	66.5	68.3	81.1	88.2	77.4	90.8	96.8	98.3
	ICPL-ReID [25]	T-MM'25	CLIP-B/16	75.1	77.4	84.2	87.9	85.1	94.7	98.4	99.1
	MambaPro [17]	AAAI'25	CLIP-B/16	78.9	83.4	89.8	91.9	84.1	92.8	97.7	98.7
	PromptMA [16]	T-IP'25	CLIP-B/16	78.4	80.9	87.0	88.9	-	-	-	-
	DeMo [18]	AAAI'25	CLIP-B/16	79.0	82.3	88.8	92.0	83.6	93.1	97.5	98.7
	IDEA [24]	CVPR'25	CLIP-B/16	80.2	82.1	90.0	93.3	-	-	-	-
	MFRNet [54]	ICML'25	CLIP-B/16	80.7	83.6	91.9	94.1	83.8	93.4	97.7	98.7
	NEXT	**Ours**	**CLIP-B/16**	**82.4**	**86.6**	**92.0**	**94.7**	**85.8**	**95.3**	**98.7**	**99.4**
	Improvement	-	-	+↑1.7	+↑3.0	+↑0.1	+↑0.6	+↑0.7	+↑0.6	+↑0.3	+↑0.3

Performance on Vehicle Dataset. Table IV outlines the performance of our method on vehicle datasets. On the MSVR310 dataset, our approach significantly outperforms existing methods, achieving 60.8%/79.0% mAP and Rank-1 accuracy. MSVR310 contains high-quality vehicle images but suffers from severe cross-view variations, where textual semantics can guide NEXT to learn view-invariant identity cues and improve cross-modal feature fusion. For RGBNT100, the Rank-1 performance is nearly saturated due to many repetitive vehicle samples, while mAP better reflects the improved overall retrieval quality. For WMVeID863, strong illumination pollution and vehicle-part occlusion degrade both visual recognition and text generation. Nevertheless, NEXT still improves over FACENet [14] and MFRNet [54] by +1.1% mAP. These results confirm the strong generalization and robustness of NEXT in vehicle ReID under complex conditions.

D. Ablation Studies

On the RGBNT201 and MSVR310 datasets, we start from a baseline which is built on a three-branch CLIP visual encoder [29], and incrementally incorporate our components to evaluate their specific contributions to the model's overall performance. The three main components are: the Multi-Grained Features Aggregation (MGFA), the Text-Modulated Semantic Experts (TMSE), and the Context-Shared Structure Experts (CSSE).

Effectiveness of Key Modules. As shown in Table V, we conduct ablation studies on both the RGBNT201 and MSVR310 datasets to evaluate the effectiveness of each proposed component. In Table V(a), the baseline with a three-branch CLIP visual encoder achieves 71.0%/73.6% mAP and Rank-1 on RGBNT201 and 50.1%/68.9% on MSVR310. By introducing the MGFA in Table V(b), which fuses the modality-specific features as a unified expert, the performance improves to 74.2%/76.6% and 55.0%/73.3%, indicating that adaptive feature fusion benefits representation learning. In Table V(c), adding the TMSE further enhances performance to 78.9% mAP and 84.4% Rank-1 on RGBNT201, and brings an additional +4.7% mAP and +1.8% Rank-1 improvement on MSVR310, demonstrating the advantage of semantic modulation in sampling fine-grained features. Finally, Table V(d) combines the model with the CSSE and achieves the best results, 82.4%/86.6% mAP and Rank-1 on RGBNT201, and 60.8%/79.0% on MSVR310, validating that maintaining structural integrity and environment context awareness is beneficial for obtaining discriminative identity features. These results validate the complementary contributions and generalizability of the MGFA, TMSE, and CSSE modules, confirming their effectiveness in enhancing multi-modal fusion and discriminative representation learning across different datasets.

Semantic Modulation Diversity and Experts Token Distribution. As shown in Fig. 6, the sampling ratio in the modulation module plays a crucial role in balancing semantic diversity and consistency. In Sec.III.B, the **Sampling** is a random probability distribution which controls the diversity and consistency of modulated signals. Intuitively, the sampling probability determines whether a sentence is selected. A high probability causes the sentence to be sampled frequently, resulting in modulation signals that contain consistent content. Conversely, a lower sampling probability produces signals

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON VEHICLE DATASETS: MSVR310, RGBNT100, AND WMVEID863 ABOUT MAP(%) AND RANK-K(%). THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

	Methods	Venue	Backbone	MSVR310		RGBNT100		WMVEID863			
				mAP	R-1	mAP	R-1	mAP	R-1	R-5	R-10
Single-Modal	BoT [4]	CVPRW'19	ResNet	23.5	38.4	78.0	95.1	51.1	55.7	69.8	74.7
	OSNet [61]	ICCV'19	ResNet	28.7	44.8	75.0	95.6	42.9	46.8	61.9	69.4
	AGW [2]	T-PAMI'21	ResNet	8.9	46.9	73.1	92.7	30.3	35.3	43.3	46.5
	PFD [64]	AAAI'22	ResNet	23.0	39.9	67.5	92.6	50.2	55.3	69.8	75.3
	TransReID [3]	ICCV'21	ViT-B/16	33.4	48.9	75.6	92.9	67.0	74.7	79.5	82.4
Multi-Modal	HAMNet [11]	AAAI'20	ResNet	27.1	42.3	74.5	93.3	45.6	48.5	63.1	68.8
	PFNet [13]	AAAI'21	ResNet	23.5	37.4	68.1	94.1	50.1	55.9	68.7	75.1
	IEEE [58]	AAAI'22	ResNet	21.0	41.0	61.3	87.8	45.9	48.6	64.3	67.9
	CCNet [12]	INFFUS'23	ResNet	36.4	55.2	77.2	96.3	50.3	52.7	69.6	75.1
	TOP-ReID [15]	AAAI'24	ViT-B/16	35.9	44.6	81.2	96.4	67.7	75.3	80.8	83.5
	FACENet [14]	INFFUS'25	ViT-B/16	36.2	54.1	81.5	96.9	69.8	77.0	81.0	84.2
	EDITOR [23]	CVPR'24	ViT-B/16	39.0	49.3	82.1	96.4	65.6	73.8	80.0	82.3
	ICPL-ReID [25]	T-MM'25	CLIP-B/16	56.9	77.7	87.0	**98.6**	67.2	74.0	81.3	85.6
	MambaPro [17]	AAAI'25	CLIP-B/16	47.0	56.5	83.9	94.7	69.5	76.9	80.6	83.8
	PromptMA [16]	T-IP'25	CLIP-B/16	55.2	64.5	85.3	97.4	-	-	-	-
	DeMo [18]	AAAI'25	CLIP-B/16	49.2	59.8	86.2	97.6	68.8	77.2	81.5	83.8
	IDEA [24]	CVPR'25	CLIP-B/16	47.0	62.4	87.2	96.5	-	-	-	-
	MFRNet [54]	ICML'25	CLIP-B/16	50.6	64.8	**88.2**	97.4	69.8	**78.3**	83.1	**86.7**
	NEXT	**Ours**	**CLIP-B/16**	**60.8**	**79.0**	**88.2**	97.7	**70.9**	77.8	**84.3**	**86.7**
	Improvement	-	-	+↑3.9	+↑1.3	-	-	+↑1.1	-	+↑1.2	-

TABLE V
ABLATION STUDIES OF DIFFERENT MODULES ON PERSON DATASET RGBNT201 AND VEHICLE DATASET MSVR310.

Index	Modules			RGBNT201				MSVR310				Params	Flops
	MGFA	TMSE	CSSE	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	M	G
(a)	×	×	×	71.0	73.6	84.2	88.2	50.1	68.9	83.4	87.8	88.3	33.3
(b)	✓	×	×	74.2	76.6	87.8	91.3	55.0	73.3	86.8	91.7	88.5	33.5
(c)	✓	✓	×	78.9	84.4	91.1	93.2	59.7	75.1	88.2	92.0	92.5	43.7
(d)	✓	✓	✓	**82.4**	**86.6**	**92.0**	**94.7**	**60.8**	**79.0**	**89.2**	**92.2**	94.8	44.8

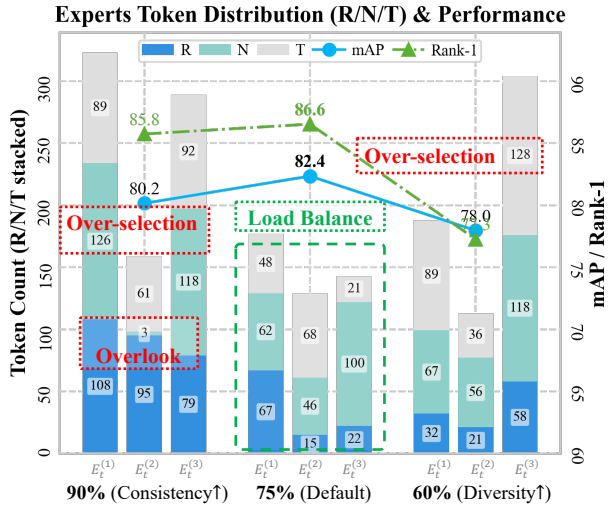

Fig. 6. Impact of semantic modulation diversity and consistency on expert load balancing and performance. R/N/T denote token counts from RGB, NIR, and TIR modalities for each semantic expert.

TABLE VI
EFFECTIVENESS OF CAPTION SOURCE AND CAPTION QUALITY.

Caption Setting	Quality	IDEA		NEXT (Ours)	
		mAP	R-1	mAP	R-1
Comparison of different caption sources					
IDEA-Text	100%	80.2	82.1	80.5	84.7
Qwen-VL	100%	78.0	81.1	80.7	86.4
GPT-4o	100%	79.3	80.9	81.8	86.7
Simple-Qwen-VL	100%	77.6	80.3	80.1	85.4
Simple-GPT-4o	100%	78.4	80.2	80.6	84.8
Ablation on caption quality using NEXT-Text					
NEXT-Text	35%	76.1	77.9	77.1	79.7
	70%	78.1	79.9	80.0	82.2
	100%	**80.2**	**84.0**	**82.4**	**86.6**

with greater diversity. A lower probability of 60% increases semantic diversity but leads to unstable expert routing and a notable decline in performance, reaching 78.0% mAP and

TABLE VII
EFFECTIVENESS OF SAMPLING STRATEGIES.

Index	Methods	Metrics		Params	Flops
		mAP	R-1	M	G
(a)	All-Token	79.0	83.0	94.8	43.9
(b)	Top-K	81.4	84.3	94.8	44.8
(c)	CNN	78.0	84.3	115.6	46.3
(d)	GCN	79.9	82.5	115.6	46.5
(e)	Attention	82.1	85.9	113.2	45.7
(f)	$\boldsymbol{\sigma}_m$ **(Ours)**	**82.4**	**86.6**	94.8	44.8

TABLE VIII
EFFECTIVENESS OF DIFFERENT ROUTING STRATEGIES. *Multi.* MEANS EACH MODALITY USES A DIFFERENT ROUTER. *Share.* DENOTES ALL MODALITIES USE A SHARED ROUTER.

Index	Router Type		Metrics	
	$\mathbf{E}_T$	$\mathbf{E}_C$	mAP	R-1
(a)	*Multi.*	*Multi.*	78.3	82.5
(b)	*Share.*	*Multi.*	78.7	83.4
(c)	*Share.*	*Share.*	80.3	84.3
(d)	***Multi.***	***Share.***	**82.4**	**86.6**

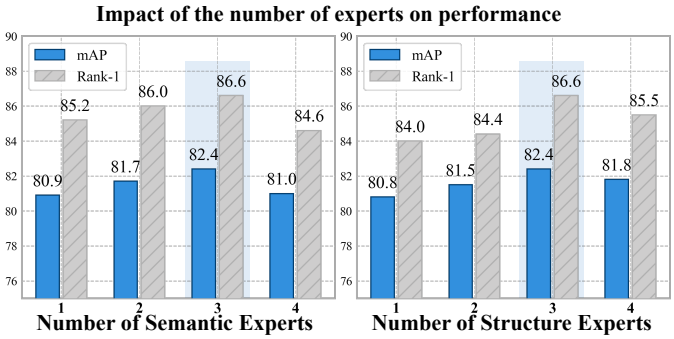

Fig. 7. Effectiveness of the number of semantic and structural experts.

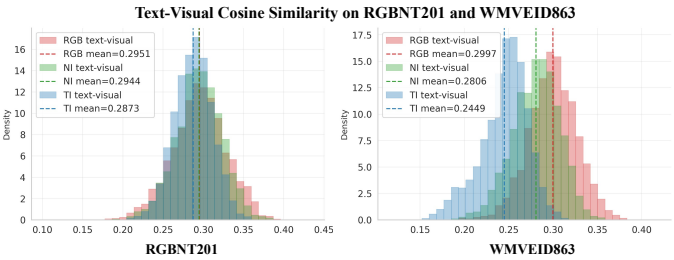

Fig. 8. Statistics of text-visual cosine similarity on RGBNT201 and WMVEID863 datasets.

77.3% Rank-1. In contrast, a higher probability of 90% improves feature consistency but results in slight performance degradation, with 80.2% mAP and 85.8% Rank-1, as token over-selection or overlook causes one modality to dominate or suppress in expert activation. The default probability of 75% achieves the most desirable trade-off, enabling the model to select semantically balanced tokens across modalities. Specifically, Experts $E_t^{(1)}$, $E_t^{(2)}$, and $E_t^{(3)}$ select 67, 15, and 22 tokens from RGB, 62, 68, and 100 from NIR, and 48, 68, and 21 from TIR, respectively. This balanced load preserves modal fairness while maintaining the diversity essential for stable and cooperative expert behavior.

Effectiveness of Caption Quality. As shown in Table VI, we evaluate the influence of caption source and quality on RGBNT201. When using 100% high-quality NEXT captions, our method achieves 82.4% mAP and 86.6% Rank-1, outperforming IDEA [24] by +2.2%/+2.6%. Replacing NEXT captions with IDEA-generated captions causes a 1.9%/1.9% drop in mAP and Rank-1, showing that richer and more fine-grained captions benefit text-visual alignment and modality fusion. We further compare captions generated by GPT-4o and Qwen-VL. GPT-4o performs slightly better than Qwen-VL due to its stronger fine-grained visual description ability, while NEXT with Qwen-VL generated captions still outperforms IDEA, suggesting that the proposed caption generation strategy improves caption reliability across different MLLMs.

We also evaluate a simpler confidence-aware strategy by directly filtering low-confidence attributes, denoted as Simple-Qwen-VL and Simple-GPT-4o. Both simplified variants perform worse than their full counterparts, suggesting that simple filtering may remove noisy attributes but also discards useful complementary cues. In contrast, our aggregation strategy preserves reliable attributes through complementation across MLLMs and modalities, leading to more complete captions.

When caption quality decreases to 70% and 35%, both methods suffer clear performance degradation, confirming that semantic-guided methods are sensitive to textual noise. Nevertheless, NEXT still outperforms IDEA under degraded captions, demonstrating better semantic robustness. We also note that high-confidence hallucination remains a general challenge in MLLM-based semantic generation. The proposed confidence-aware aggregation and sentence-level semantic sampling in TMSE are designed to mitigate the influence of unreliable textual cues, rather than fully eliminate them.

Effectiveness of Sampling Strategy. Table VII compares different semantic sampling strategies on RGBNT201. In Table VII(a), the All-Token strategy encodes all patch tokens $\mathbf{F}_{I,m}^{\text{tok}}$, but fails to identify key semantic regions, resulting in significant performance degradation. Table VII(b) adopts the Top-K strategy, replacing the dynamic routing threshold $\boldsymbol{\sigma}_m$ in $\mathbf{R}_m$ with the top 50% tokens, achieving a suboptimal result. In Table VII(c)-(e), we further compare more complex routing generators, including CNN [68], GCN [69], and Attention-based [70] routing. Although attention improves over CNN/GCN, these methods introduce more parameters and FLOPs, while still underperforming our dynamic threshold strategy. In contrast, our dynamic sampling strategy achieves optimal performance, confirming its effectiveness.

Effectiveness of Routing Strategy. As shown in Table VIII, different routing strategies are explored to investigate the impact of modality-specific or shared routing on performance. When both the semantic expert $\mathbf{E}_T$ and the structural expert $\mathbf{E}_C$ adopt modality-specific routing in Table VIII(a), the model achieves 78.3% mAP and 82.5% Rank-1. Sharing the routing for $\mathbf{E}_T$ slightly improves performance in Table VIII(b),

TABLE IX
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SIZE OF LEARNABLE PARAMETERS, FLOPS, AND THROUGHPUT ON RGBNT201. * DENOTES THE USE OF CLIP TEXT ENCODER IN THE INFERENCE STAGE.

Methods	Params	FLOPs	Throughput	Metrics	
	M	G	Images/s	mAP	R-1
EDITOR [23]	119.3	40.8	335.1	66.5	68.3
TOP-ReID [15]	324.5	35.5	398.9	72.3	76.6
ICPL-ReID [25]	45.4	39.8	358.4	75.1	77.4
PromptMA [16]	107.9	36.2	343.5	78.4	80.9
DeMo [18]	98.8	35.1	403.6	79.0	82.3
MambaPro [17]	74.8	52.4	243.2	78.9	83.4
NEXT (w/o Text)	94.8	36.0	354.0	79.2	85.5
IDEA* [24]	91.7	43.7	299.5	80.2	82.1
NEXT* (Ours)	94.8	44.8	275.1	**82.4**	**86.6**

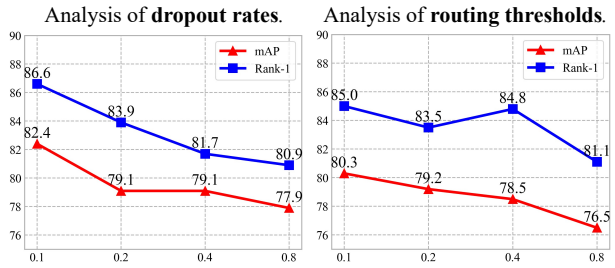

Fig. 9. Performance impact of the dropout rates and routing thresholds.

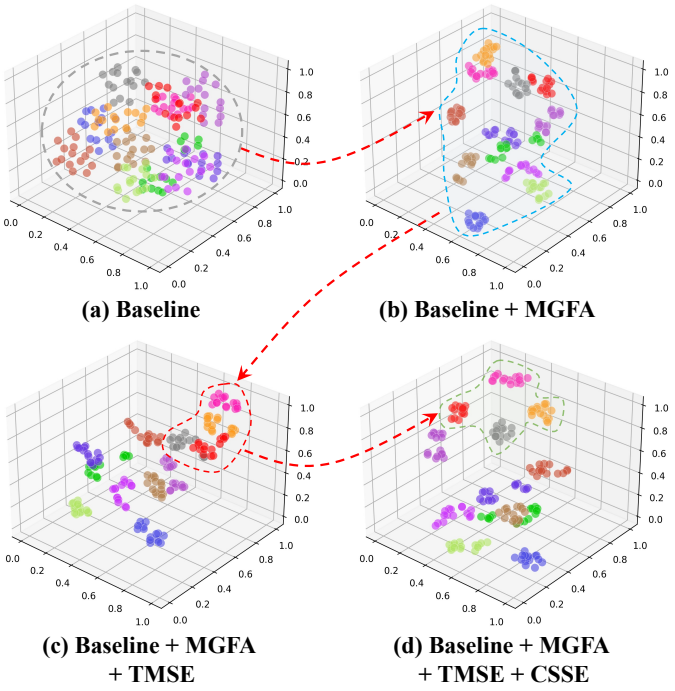

Fig. 10. T-SNE [71] visualization of the feature distribution on RGBNT201 (a) Baseline, (b) Baseline + MGFA, (c) Baseline + MGFA + TMSE, and (d) Baseline + MGFA + TMSE + CSSE (Ours).

while using shared routing for both experts in Table VIII(c) further boosts the results to 80.3% mAP and 84.3% Rank-1, suggesting that a certain level of cross-modal coupling is beneficial. The optimal configuration in Table VIII(d) is achieved when $\mathbf{E}_T$ uses modality-specific routers and $\mathbf{E}_C$ adopts a shared router, reaching 82.4% mAP and 86.6% Rank-1. Intuitively, semantic cues such as color, illumination, and temperature are modality-dependent and thus should be learned via modality-specific routing, while structural cues (*e.g.*, object shape, contour, or posture) are modality-invariant, making shared routing more suitable for consistent structural perception.

Effectiveness of Experts Number. As shown in Fig. 7, the impact of the number of experts on model performance is analyzed. Increasing the number of semantic experts from one to three progressively improves performance, with mAP rising from 80.9% to 82.4% and Rank-1 increasing from 85.2% to 86.6%. Introducing a fourth expert leads to a decline in both metrics, dropping to 81.0% in mAP and 84.6% in Rank-1. A similar trend is observed for structural experts. Expanding the expert count from one to three steadily boosts performance, with mAP growing from 80.8% to 82.4% and Rank-1 from 84.0% to 86.6%, while four experts introduce a clear performance drop. These results indicate that an excessive number of experts increases redundancy and weakens specialization, whereas three experts reach an effective balance between representational richness and expert diversity, yielding the most discriminative identity features.

Text-Visual Cosine Similarity. We compute the cosine similarity between generated captions and images in the CLIP embedding space. As shown in Fig. 8, RGBNT201 shows compact distributions across RGB, NIR, and TIR, with mean similarities of 0.2951, 0.2944, and 0.2873, indicating stable cross-modal text-visual alignment. In contrast, WMVEID863 exhibits a larger modality gap, with mean similarities of 0.2997, 0.2806, and 0.2449. The notably lower TIR similarity suggests that vehicle captions are more affected by modality degradation and limited fine-grained cues, supporting the necessity of confidence-aware multi-modal aggregation.

Effectiveness of Dropout Rates and Routing Thresholds. As shown in Fig. 9, we analyze expert dropout rates and fixed routing thresholds. The best performance is obtained with a dropout rate of 0.1, achieving 82.4% mAP and 86.6% Rank-1. Larger dropout rates gradually degrade performance, suggesting that excessive dropout weakens the representation capacity of semantic and structural experts. For routing thresholds, fixed thresholds lead to unstable results because they cannot adapt to modality degradation and sample-wise quality variations. In contrast, the learnable dynamic threshold σ adaptively adjusts the sampling criterion for each input, leading to better and more stable performance.

Training and Inference Efficiency Analysis. Table IX compares the efficiency of various methods on the RGBNT201 dataset in terms of learnable parameters, FLOPs, throughput, and retrieval metrics. Our method NEXT achieves the best performance 82.4%/86.6% mAP and Rank-1, while maintaining a moderate model size of 94.8M parameters. Compared to the MoE-based method DeMo [18] (98.8M), NEXT reduces the parameter count by 4M, demonstrating the higher efficiency. While adding the text encoder increases the FLOPs of NEXT by 8.8G relative to the version without text, it still maintains a clear computational advantage over MambaPro [17], which

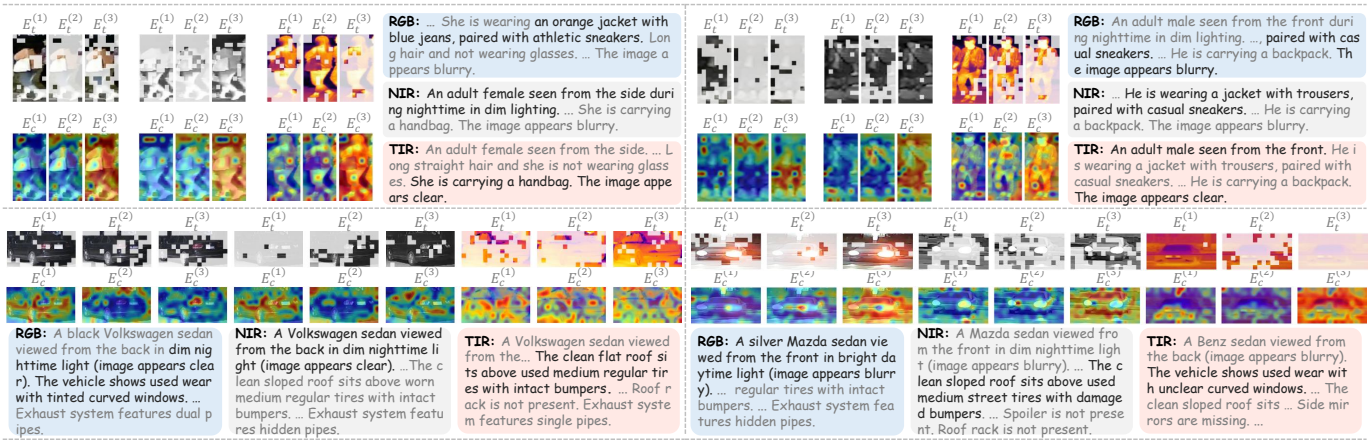

Fig. 11. Visualization of the sampled patch tokens and activated feature regions for experts on person and vehicle datasets.

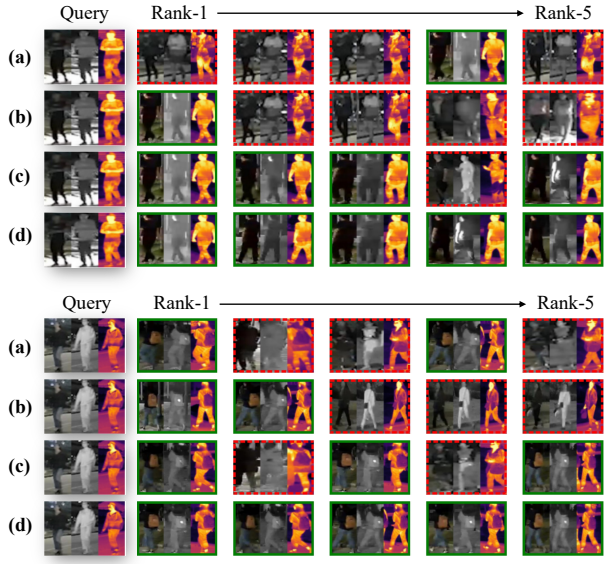

Fig. 12. Comparison of retrieval results for different modules on RGBNT201: (a) Baseline, (b) Baseline + MGFA, (c) Baseline + MGFA + TMSE, and (d) Baseline + MGFA + TMSE + CSSE (Ours).

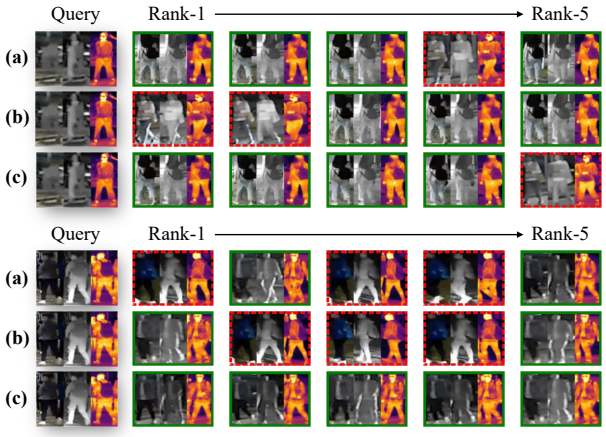

Fig. 13. Comparison of retrieval results for state-of-the-art methods on RGBNT201: (a) DeMo, (b) IDEA, and (c) NEXT (Ours).

consumes 52.4G. For inference throughput, NEXT achieves 275.1 Images/s, which is slightly lower than IDEA's 299.5 Images/s, but outperforms MambaPro's 243.2 Images/s. These results validate that NEXT obtains superior performance with a trade-off between model complexity and efficiency.

E. Visualization Analysis

Feature Distribution. As shown in Fig. 10, we utilize the T-SNE [71] to visualize the feature distributions on RGBNT201. Comparing Figs. 10(a) and (b), the introduction of MGFA effectively aggregates different instances of the same identity. In Fig. 10(c), incorporating the TMSE semantic experts improves the separability of samples across identities. In Fig. 10(d), the inclusion of CSSE structural experts leads to tighter clustering of hard samples, resulting in well-separated intra-class and inter-class distributions. These results validate the effectiveness of each proposed module.

Expert Feature Visualization. As shown in Fig. 11, we visualize the sampling masks of the semantic experts and the activation maps of the structural experts. We observe that different semantic experts $E_t^{(i)}$ exhibit distinct modality preferences: for instance in person dataset, $E_t^{(1)}$ tends to sample features from the RGB modality, $E_t^{(2)}$ from the TIR modality, and $E_t^{(3)}$ from the NIR modality. These sampling patterns are notably complementary across modalities, indicating that the text-modulated expert effectively learns complementary modality-specific features to enhance multimodal representations. On the other hand, structural experts maintain the integrity of coarse-grained identity structures and environmental context awareness. The parts ignored by expert $E_c^{(3)}$ are well captured by other structural experts. We further visualize several challenging samples, including cases with severe occlusion, strong illumination interference, and modality degradation. These examples show that different experts still attend to complementary semantic regions and structural cues under difficult conditions, while also revealing the remaining challenges when the visual quality is heavily degraded. These observations validate that NEXT successfully decouples identity recognition into semantic sampling and

structure modeling.

Retrieval Results. To verify the model's effectiveness in real world scenarios, we visualize the retrieval results of models with different module combinations on the RGBNT201 dataset and compare NEXT with state-of-the-art methods. As shown in Fig. 12, the introduction of different modules enables the model to recognize pedestrians with the same identity more effectively. Fig. 13 compares our method with existing state-of-the-art methods, demonstrating that our method achieves superior retrieval performance for the target person under various low-light conditions. These results validate the effectiveness of NEXT in leveraging multi-modal data and textual semantics for robust person ReID.

VI. Conclusion

In this paper, we focus on the explicit semantic learning of multi-modal object ReID. We first propose a reliable multi-modal caption generation pipeline based on MLLMs, which utilize template-based instructions to construct confidence-aware attribute sets, and merge multi-modal attributes to compose high-quality captions. Additionally, we propose the multi-modal ReID network, NEXT, a novel multi-grained MoE framework. Specifically, we propose the Text-Modulated Semantic Experts (TMSE) and the Context-Shared Structure Experts (CSSE) to decouple the recognition problem into fine-grained semantic sampling and coarse-grained structural perception. We incorporate a Multi-Grained Features Aggregation (MGFA) module to unify features from multi-grained experts and obtain complete identity representations. Extensive experiments on five popular datasets demonstrate the effectiveness of our method. In the future, we plan to further explore the reasoning capability of MLLMs to drive deeper semantic fusion, such as harder negative mining and chain-of-thought reasoning. We believe this direction will unlock new frontiers in interpretable and semantic-driven multi-modal ReID.

References

[1] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2020.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[3] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 993–15 002.

[4] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1487–1495.

[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[6] A. Lu, Z. Zhang, Y. Huang, Y. Zhang, C. Li, J. Tang, and L. Wang, "Illumination distillation framework for nighttime person re-identification and a new benchmark," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

[7] N. Dong, S. Yan, L. Zhang, and J. Tang, "Diverse semantics-guided feature alignment and decoupling for visible-infrared person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 12 245–12 259, 2025.

[8] R. Xi, Z. Fu, N. Huang, X. Zhao, Q. Zhang, and J. Han, "Csanet: Cross-modality self-paced association network for unsupervised visible-infrared person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 8672–8686, 2025.

[9] Y. Gong, L. Huang, and L. Chen, "Person re-identification method based on color attack and joint defence," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4313–4322.

[10] Y. Gong, Z. Zhong, Y. Qu, Z. Luo, R. Ji, and M. Jiang, "Cross-modality perturbation synergy attack for person re-identification," *Advances in Neural Information Processing Systems*, vol. 37, pp. 23 352–23 377, 2024.

[11] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 345–11 353.

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

[13] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 3529–3537.

[14] A. Zheng, Z. Ma, Y. Sun, Z. Wang, C. Li, and J. Tang, "Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification," *Information Fusion*, vol. 116, p. 102800, 2025.

[15] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "Top-reid: Multi-spectral object re-identification with token permutation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5758–5766.

[16] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, and Y. Zhang, "Prompt-based modality alignment for effective multi-modal object re-identification," *IEEE Transactions on Image Processing*, vol. 34, pp. 2450–2462, 2025.

[17] Y. Wang, X. Liu, T. Yan, Y. Liu, A. Zheng, P. Zhang, and H. Lu, "Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8150–8158.

[18] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Demo: Decoupled feature-based mixture of experts for multi-modal object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8141–8149.

[19] Z. Yu, Z. Huang, M. Hou, J. Pei, Y. Yan, Y. Liu, and D. Sun, "Representation selective coupling via token sparsification for multi-spectral object re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 4, pp. 3633–3648, 2025.

[20] Y. Gong, Y. Hou, C. Zhang, and M. Jiang, "Beyond augmentation: Empowering model robustness under extreme capture environments," in *International Joint Conference on Neural Networks*, 2024, pp. 1–8.

[21] Y. Gong, Y. Hou, J. Shi, K. L. Diep, and M. Jiang, "A theory-inspired framework for few-shot cross-modal sketch person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 6, 2026, pp. 4284–4292.

[22] X. Yang, W. Dong, D. Cheng, N. Wang, and X. Gao, "Tienet: A tri-interaction enhancement network for multimodal person reidentification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 9852–9863, 2025.

[23] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 117–17 126.

[24] Y. Wang, Y. Lv, P. Zhang, and H. Lu, "IDEA: inverted text with cooperative deformable aggregation for multi-modal object re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025, pp. 29 701–29 710.

[25] S. Li, A. Zheng, C. Li, J. Tang, and B. Luo, "Icpl-reid: Identity-conditional prompt learning for multi-spectral object re-identification," *IEEE Transactions on Multimedia*, 2025.

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

[27] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*.

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 34 892–34 916.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.

[30] S. Li, L. Sun, and Q. Li, "Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 1405–1413.

[31] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang, "A pedestrian is worth one prompt: Towards language guidance person re- identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 343–17 353.

[32] Z. Hu, B. Yang, and M. Ye, "Empowering visible-infrared person re-identification with large foundation models," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 117 363–117 387.

[33] Y. Zhai, Y. Zeng, Z. Huang, Z. Qin, X. Jin, and D. Cao, "Multi-prompts learning with cross-modal alignment for attribute-based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 6979–6987.

[34] Y. Fu, R. Xie, X. Sun, Z. Kang, and X. Li, "Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 16 563–16 577.

[35] C. Su, L. Peng, Y. Sun, D. Peng, X. Peng, and X. Wang, "Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations," *Advances in Neural Information Processing Systems*, vol. 38, pp. 108 581–108 603, 2026.

[36] C. Su, Y. Li, X. Wang, Y. Chen, H. Zheng, D. Peng, and Y. Sun, "Ambiguity-tolerant cross-modal hashing with partial labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 30, 2026, pp. 25 636–25 644.

[37] C. Su, H. Zheng, D. Peng, and X. Wang, "Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 20 610–20 618.

[38] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[39] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[40] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 816–16 825.

[41] S. He, W. Chen, K. Wang, H. Luo, F. Wang, W. Jiang, and H. Ding, "Region generation and assessment network for occluded person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 120–132, 2024.

[42] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 49 250–49 267.

[43] Y. Liu, P. Chen, J. Cai, X. Jiang, Y. Hu, J. Yao, Y. Wang, and W. Xie, "Lamra: Large multimodal model as your advanced retrieval assistant," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025, pp. 4015–4025.

[44] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181–214, 1994.

[45] K. Chen, L. Xu, and H. Chi, "Improved learning algorithms for mixture of experts in multiclass classification," *Neural Networks*, vol. 12, no. 9, pp. 1229–1252, 1999.

[46] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[47] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *Proceedings of the International Conference on Learning Representations*, 2024.

[48] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proceedings of the International Conference on Learning Representations*, 2017.

[49] C. Wu, Z. Shuai, Z. Tang, L. Wang, and L. Shen, "Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts," in *Proceedings of the International Conference on Learning Representations*, 2025.

[50] S. Yun, I. Choi, J. Peng, Y. Wu, J. Bao, Q. Zhang, J. Xin, Q. Long, and T. Chen, "Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 98 782–98 805.

[51] X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria, "Fusemoe: Mixture-of-experts transformers for fleximodal fusion," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 67 850–67 900.

[52] Y. Fang, W. Huang, G. Wan, K. Su, and M. Ye, "Emoe: Modality-specific enhanced dynamic emotion experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025, pp. 14 314–14 324.

[53] J. Peng, J. L. Ballard, M. Zhang, S. Yun, J. Xin, Q. Long, Y. Zhang, and T. Chen, "Modalities contribute unequally: Enhancing medical multi-modal learning through adaptive modality token re-balancing," in *Proceedings of the International Conference on Machine Learning*.

[54] Y. Feng, J. Li, C. Xie, L. Tan, and J. Ji, "Multi-modal object re-identification via sparse mixture-of-experts," in *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267, 2025, pp. 16 796–16 807.

[55] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[57] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[58] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2633–2641.

[59] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[60] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

[61] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3702–3712.

[62] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1005–1014.

[63] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5850–5858.

[64] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2540–2549.

[65] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015, pp. 1116–1124.

[66] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13 001–13 008.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[69] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[71] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.