EPIC: Semantic Inverse Prompting and Evidential Corroboration for Multi-modal Object Re-Identification

Xingan Ma[1], Yuhao Wang[2,†], Jinhui Yi[1], Juergen Gall[1,†]
[1]University of Bonn [2]Dalian University of Technology
maxingan1993@gmail.com, 924973292@mail.dlut.edu.cn, {yij, gall}@iai.uni-bonn.de

***Abstract*—Multi-modal object Re-Identification (ReID) aims to match object identities by leveraging complementary information from multiple modalities. However, existing methods primarily focus on fusing heterogeneous features, while overlooking modality-specific noise that can severely degrade feature quality. Moreover, most fusion strategies treat modalities uniformly, failing to adaptively weight them under varying imaging conditions. To address these issues, we propose EPIC, a novel framework that tackles modality-specific noise through semantic governance and performs reliability-aware fusion via evidential corroboration. Specifically, we propose Semantic Inverse Prompting and Consensus (SIPC), which converts textual semantics into effective visual guidance via three cascaded sub-modules. Minimum Consensus Gating (MCG) filters unreliable semantics through cross-modal consensus, Semantically-Anchored Inverse Prompting (SAIP) projects the gated semantics into visual prompt tokens, and Multi-grained GeM Aggregation (MGA) aggregates multi-grained representations. We further introduce Evidential Corroborated Reliability (ECR), which estimates modality reliability from evidential theory and strengthens representations with adaptive residual rectification. Together, SIPC and ECR form a coherent pipeline from semantic denoising to reliability-aware fusion, yielding robust multi-modal representations in complex scenarios. Extensive experiments on three multi-modal object ReID benchmarks demonstrate the effectiveness of EPIC.**

***Index Terms*—Multi-modal Object ReID, Semantic Governance, Evidential Corroboration, Inverse Prompting**

I. Introduction

Object Re-Identification (ReID) aims to match the same object across different camera views. While single-modal methods based on RGB images have achieved significant progress [1]–[4], they are susceptible to adverse conditions such as darkness and glare. Multi-modal imaging, which integrates complementary information from RGB, Near Infrared (NIR) and Thermal Infrared (TIR) modalities, has emerged as a promising solution to enhance feature robustness [5]–[7]. However, as shown in Fig. 1 (a), existing multi-modal methods often encode modality-specific noise together with discriminative cues and then apply fixed fusion, which compromises feature quality. Meanwhile, modality reliability varies substantially across conditions, yet most fusion strategies treat modalities uniformly and fail to assign reliability-aware weights.

Recently, textual semantics generated by Multi-modal Large Language Models (MLLMs) have been introduced to enhance multi-modal ReID [8]. Text descriptions offer modality-agnostic semantic anchors, making it possible to suppress modality-specific noise by enforcing cross-modal consensus. Motivated by this observation, we advocate semantic governance. We use trusted semantics to actively regulate feature extraction. This contrasts with prior practice that aligns representations after encoding. Meanwhile, evidential theory [9] provides a way to quantify uncertainty, which we exploit to estimate modality reliability and enable adaptive fusion.

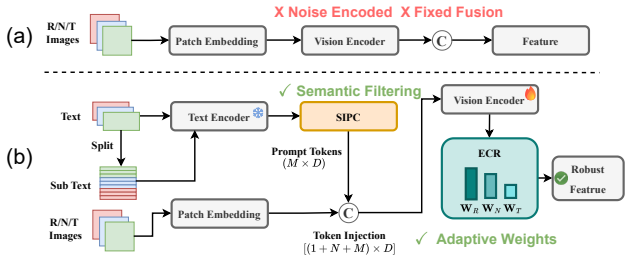

Fig. 1. (a) Existing methods directly encode noisy features and adopt fixed fusion. (b) Our proposed EPIC employs semantic governance to guide visual encoding and uses evidential theory for reliability-aware adaptive fusion.

Based on the above analysis, we propose EPIC, a novel framework that consists of two tightly coupled modules. The Semantic Inverse Prompting and Consensus (SIPC) module transforms textual semantics into visual guidance via three cascaded sub-modules: Minimum Consensus Gating (MCG) filters unreliable semantics through cross-modal consensus, Semantically-Anchored Inverse Prompting (SAIP) projects the gated semantics into visual prompt tokens, and Multi-grained GeM Aggregation (MGA) aggregates multi-grained representations. The Evidential Corroborated Reliability (ECR) module performs reliability-aware fusion, where Evidential Reliability Weighting (ERW) computes modality weights based on evidential theory and cross-modal corroboration, and Adaptive Residual Rectification (ARR) enhances features via residual learning. Extensive experiments validate the effectiveness of our method. Overall, our main contributions are as follows:

- We propose EPIC, a novel framework that addresses modality-specific noise via semantic governance and performs reliability-aware fusion via evidential corroboration for robust multi-modal object ReID.
- We introduce the Semantic Inverse Prompting and Consensus (SIPC), which filters unreliable semantics by min-

†Corresponding authors.

imum cross-modal consensus and injects semantically-anchored visual prompt tokens to govern visual encoding.
- We propose the Evidential Corroborated Reliability (ECR), which estimates modality reliability from evidential theory and refines features with corroboration-aware weighting and adaptive residual rectification.
- Extensive experiments on three multi-modal ReID benchmarks demonstrate the effectiveness of our method.

II. Related Work

A. Multi-Modal Object Re-Identification

Multi-modal object ReID leverages complementary spectra to maintain robust identity recognition under adverse conditions. However, intra-modal noise and inter-modal spectral heterogeneity still hinder feature learning and fusion. To mitigate these issues, existing studies design diverse fusion and alignment strategies. Li [6] introduce a coherence loss to guide multi-modal feature fusion, while Zheng [5] propose progressive fusion to learn discriminative representations. Wang [10] design an interact-embed-enlarge framework to facilitate cross-modal knowledge expansion, and Zheng [11] address the modal-missing setting via pixel-level reconstruction. With Vision Transformers [12], Wang [7] introduce token permutation with complementary reconstruction to reduce modality discrepancy. Zhang [13] adopt frequency-driven token selection to suppress background noise and enhance object-centric cues. Wang [14] leverage test-time training to mine inter-modal relations on unseen data. Despite these advances, most methods model uncertainty only implicitly and lack a principled mechanism to estimate modality reliability for adaptive fusion. This gap motivates our ECR module, which computes reliability weights under evidential theory to support reliability-aware fusion for robust multi-modal representations.

B. Semantic Guidance in Re-Identification

Large-scale vision-language foundation models provide strong visual perception and semantic understanding for ReID. Radford [15] introduce contrastive vision–language pretraining on web-scale image–text pairs, enabling transferable semantic priors. Building on this foundation, prompt learning offers a parameter-efficient way to adapt vision-language models to ReID. Li [3] introduce learnable prompts to transfer CLIP to ReID without full fine-tuning, and Zhang [16] extend prompt learning to multi-modal scenarios via modality alignment. In addition, MLLMs generate descriptive texts that serve as semantic cues. Wang [8] inject MLLM-generated semantics into the visual encoder and leverage deformable aggregation to highlight localized discriminative regions. However, most existing methods use text mainly for alignment or auxiliary supervision, and rarely exploit the cross-modal consensus in semantics to actively filter modality-specific noise. Moreover, inverse prompting remains underexplored. Unlike conventional prompt learning that adapts text to vision, inverse prompting projects textual semantics into the visual space and reverses the modulation direction. These limitations motivate our SIPC module, which purifies text features via cross-modal consensus and then transforms them into visual prompt tokens to enable semantic governance during visual feature extraction.

III. Proposed Method

As illustrated in Fig. 2, EPIC comprises two components: Semantic Inverse Prompting and Consensus (SIPC) for cross-modal feature enhancement and Evidential Corroborated Reliability (ECR) for adaptive multi-modal fusion. Given tri-modal input images $\mathbf{I}_m$ where $m \in \{R, N, T\}$ denotes RGB, NIR, and TIR, together with their text descriptions $\mathbf{T}_m$, SIPC injects textual semantics into the visual encoding process and outputs aggregated features $\mathbf{h}_m \in \mathbb{R}^D$. ECR then estimates modality reliability and fuses $\{\mathbf{h}_m\}$ into the final representation $\mathbf{f} \in \mathbb{R}^{3D}$. Details are presented as follows.

A. Semantic Inverse Prompting and Consensus

We extract text features at two granularities from $\mathbf{T}_m$ to provide both global semantics and fine-grained cues. Global features $\mathbf{L}_m \in \mathbb{R}^D$ are obtained by encoding the complete description $\mathbf{T}_m$. We split $\mathbf{T}_m$ at periods into J sentences $\{T_{m,1}, \ldots, T_{m,J}\}$, randomly sample K of them to form $\mathbf{T}_m^s = \{T_{m,1}^s, \ldots, T_{m,K}^s\}$, and encode each independently to obtain local sentence features $\mathbf{L}_m^s \in \mathbb{R}^{K\times D}$. SIPC then converts these text features into aggregated visual features $\mathbf{h}_m$ through three cascaded sub-modules: MCG enhances $\mathbf{L}_m$ to obtain $\tilde{\mathbf{L}}_m$, SAIP projects the gated semantics into visual prompt tokens $\mathbf{P}_m \in \mathbb{R}^{K\times D_v}$ and injects them into the vision encoder, and MGA aggregates the encoder outputs.

Minimum Consensus Gating. MCG enhances $\mathbf{L}_m$ by incorporating cross-modal consensus information. We first compute the consensus score a_m as the minimum pairwise similarity with other modalities, then enhance the feature via soft gating:

$$\tilde{\mathbf{L}}_m = \mathbf{L}_m + \sigma\left(\frac{a_m - \theta}{\tau}\right)\cdot\alpha\cdot\mathbf{v}, \quad a_m = \min_{m'\neq m}\cos(\mathbf{L}_m, \mathbf{L}_{m'}) \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, θ and τ are learnable threshold and temperature, α is a scaling coefficient, and $\mathbf{v} = \mathrm{MLP}([\mathbf{L}_R; \mathbf{L}_N; \mathbf{L}_T])$ is the shared consensus vector. The minimum criterion ensures enhancement only when all modalities agree, with the gating score controlling the enhancement magnitude, thus providing reliable semantic anchors for SAIP.

Semantically-Anchored Inverse Prompting. SAIP projects textual semantics into the visual space as prompt tokens. For each sentence feature $\mathbf{L}_{m,k}^s$ from $\mathbf{L}_m^s$, we compute:

$$\mathbf{P}_{m,k} = \mathrm{MLP}([\tilde{\mathbf{L}}_m; \mathbf{L}_{m,k}^s; \mathbf{L}_m]) \in \mathbb{R}^{D_v} \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation and the MLP projects the $3D$-dimensional input into the D_v-dimensional visual space. Stacking yields $\mathbf{P}_m = [\mathbf{P}_{m,1}; \ldots; \mathbf{P}_{m,K}]$. The dual anchor design with both $\tilde{\mathbf{L}}_m$ and $\mathbf{L}_m$ prevents semantic drift. We then inject $\mathbf{P}_m$ into the visual encoder. After patch embedding on $\mathbf{I}_m$, we obtain $\mathbf{F}^{v,in} \in \mathbb{R}^{N\times D_v}$, where N is the number of patches. After concatenating the CLS token and prompt tokens, the vision encoder input sequence becomes $[\mathbf{F}^{cls}; \mathbf{F}^{v,in}; \mathbf{P}_m] \in \mathbb{R}^{(1+N+K)\times D_v}$. This injection mechanism

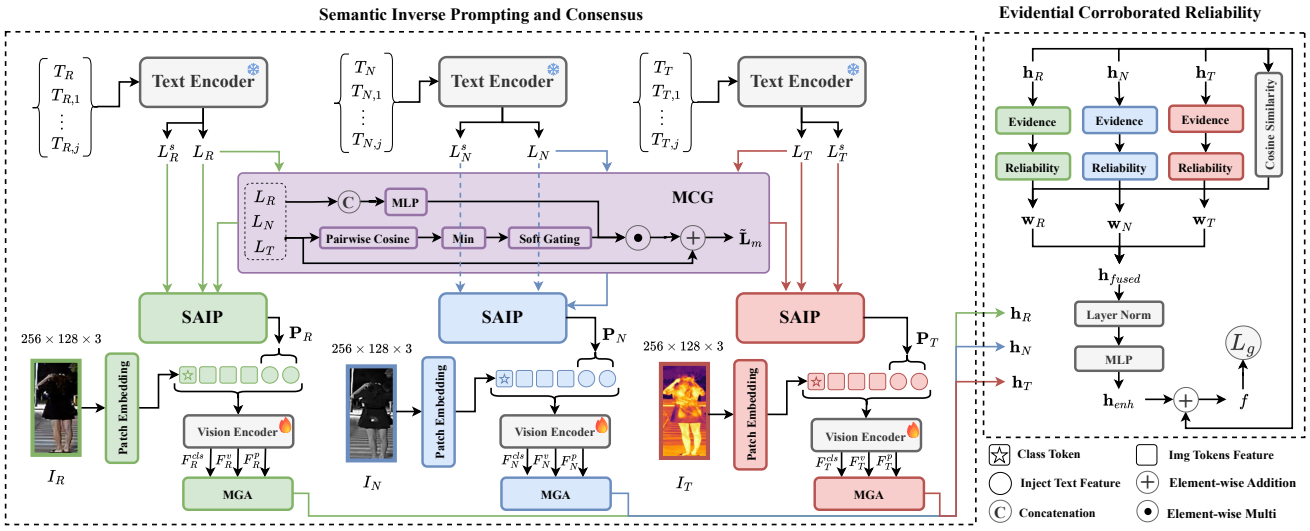

Fig. 2. Overall framework of EPIC. SIPC performs semantic-governed feature enhancement by (i) suppressing unreliable semantics via Minimum Consensus Gating (MCG), (ii) projecting gated semantics into the visual space as learnable prompt tokens via Semantically-Anchored Inverse Prompting (SAIP), and (iii) aggregating class, patch, and prompt cues with Multi-grained GeM Aggregation (MGA). ECR performs reliability-aware adaptive fusion by (i) estimating evidential reliability and cross-modal corroboration to derive modality weights (ERW) and (ii) refining representations through adaptive residual rectification (ARR). With the proposed modules, EPIC unifies semantic governance and evidential corroboration for robust multi-modal object ReID.

enables textual semantics to actively govern visual feature extraction through self-attention for enhanced representations.

Multi-grained GeM Aggregation. After vision encoder encoding and projection, we obtain the class token $\mathbf{F}_m^{cls} \in \mathbb{R}^D$, patch features $\mathbf{F}_m^v \in \mathbb{R}^{N\times D}$, and prompt tokens $\mathbf{F}_m^p \in \mathbb{R}^{K\times D}$. MGA aggregates these three granularities using Generalized Mean Pooling with the following equation:

$$\mathbf{h}_m = \mathrm{LN}\left(\mathbf{W}[\mathbf{F}_m^{cls}; \mathrm{GeM}(\mathbf{F}_m^v); \mathrm{GeM}(\mathbf{F}_m^p)]\right) \tag{3}$$

where $\mathrm{GeM}(\mathbf{F}) = (\frac{1}{L}\sum_{i=1}^{L}\mathbf{F}_i^\rho)^{1/\rho}$ denotes Generalized Mean Pooling with $\mathbf{F} \in \mathbb{R}^{L\times D}$, L being the sequence length (N for patches, K for prompt tokens), the power operation applied element-wise, and ρ being a learnable exponent. $\mathbf{W}$ is a projection matrix and $\mathrm{LN}(\cdot)$ denotes layer normalization. By fusing class, patch, and prompt cues, MGA yields compact $\mathbf{h}_m$ for reliability-aware fusion in ECR, completing SIPC's semantic-governed feature enhancement.

B. Evidential Corroborated Reliability

The ECR module fuses the tri-modal aggregated features $\{\mathbf{h}_m\}$ into the final representation $\mathbf{f}$ through two sub-modules. Evidential Reliability Weighting (ERW) estimates modality reliability and incorporates cross-modal corroboration to produce reliability-aware weights, and Adaptive Residual Rectification (ARR) refines the representations with an adaptive residual enhancement. Detailed formulations are as follows.

Evidential Reliability Weighting. ERW first computes the intrinsic reliability for each modality as follows:

$$\mathbf{e}_m = \mathrm{softplus}(\mathbf{W}_e\mathbf{h}_m), \quad r_m = \frac{\sum_d e_{m,d}}{\sum_d e_{m,d} + \kappa} \tag{4}$$

where $\mathbf{W}_e$ is a learnable matrix, $\kappa > 0$ is the prior strength that controls the conservativeness of reliability estimation (a larger κ yields lower r_m, making the model more cautious), and $r_m \in (0, 1)$. To leverage cross-modal corroboration, we compute the pairwise consistency and corroboration score:

$$c_{m,m'} = \frac{\cos(\mathbf{h}_m, \mathbf{h}_{m'}) + 1}{2}, \quad s_m = \sum_{m'\neq m} r_{m'} \cdot c_{m,m'} \tag{5}$$

where $c_{m,m'} \in [0, 1]$ is the normalized cosine similarity, and s_m aggregates reliability and consistency from other modalities. The fusion weight combines reliability and corroboration, yielding the fused feature with the following equation:

$$w_m = \mathrm{softmax}(r_m(1 + \beta s_m)), \quad \mathbf{h}_{fused} = \sum_m w_m\mathbf{h}_m \tag{6}$$

where $\beta > 0$ is the corroboration coefficient. This mechanism assigns higher weights to modalities that are both reliable and consistent with others, providing $\mathbf{h}_{fused}$ as the reliability-aware basis for the subsequent refinement in ARR.

Adaptive Residual Rectification. ARR generates an enhancement vector from weighted fusion and applies residual rectification to each modality with the following equation:

$$\mathbf{f} = [\mathbf{h}_R + \mathbf{h}_{enh};\ \mathbf{h}_N + \mathbf{h}_{enh};\ \mathbf{h}_T + \mathbf{h}_{enh}] \tag{7}$$

where $\mathbf{h}_{enh} = \mathrm{MLP}(\mathrm{LN}(\mathbf{h}_{fused}))$ is the enhancement vector. This residual rectification enables each modality to benefit from cross-modal complementary information while preserving its original characteristics. Through ERW and ARR, ECR achieves reliability-aware adaptive multi-modal fusion, producing the robust final representation $\mathbf{f}$ for robust retrieval.

C. Objective Functions

We train the EPIC framework with cross-entropy loss $\mathcal{L}_{\text{id}}$ for identity classification and triplet loss $\mathcal{L}_{\text{tri}}$ for metric learning on the final representation $\mathbf{f}$ with the following equation:

$$\mathcal{L} = \mathcal{L}_{\text{id}}(\mathbf{f}) + \lambda\mathcal{L}_{\text{tri}}(\mathbf{f}) \tag{8}$$

where λ represents the weighting coefficient of the triplet loss.

TABLE I
PERFORMANCE COMPARISON ON THREE BENCHMARKS. BEST RESULTS ARE IN BOLD, SECOND BESTS ARE UNDERLINED. † DENOTES CLIP-BASED METHODS, ∗ INDICATES VIT-BASED, OTHERS ARE CNN-BASED.

Methods	RGBNT201				RGBNT100		MSVR310	
	mAP	R-1	R-5	R-10	mAP	R-1	mAP	R-1
HAMNet [6]	27.7	26.3	41.5	51.7	74.5	93.3	27.1	42.3
PFNet [5]	38.5	38.9	52.0	58.4	68.1	94.1	23.5	37.4
IEEE [10]	47.5	44.4	57.1	63.6	-	-	-	-
DENet [11]	42.4	42.2	55.3	64.5	-	-	-	-
CCNet [17]	-	-	-	-	77.2	96.3	36.4	55.2
LRMM [18]	52.3	53.4	64.6	73.2	78.6	96.7	36.7	49.7
UniCat* [19]	57.0	55.7	-	-	-	-	-	-
PHT* [20]	-	-	-	-	79.9	92.7	-	-
HTT* [14]	71.1	73.4	83.1	87.3	75.7	92.6	-	-
TOP-ReID* [7]	72.3	76.6	84.7	89.4	81.2	96.4	35.9	44.6
EDITOR* [13]	66.5	68.3	81.1	88.2	82.1	96.4	39.0	49.3
RSCNet* [21]	68.2	72.5	-	-	82.3	96.6	39.5	49.6
DeMo† [22]	79.0	82.3	88.8	92.0	86.2	97.6	49.2	59.8
MambaPro† [23]	78.9	83.4	89.8	91.9	83.9	94.7	47.0	56.5
PromptMA† [16]	78.4	80.9	87.0	88.9	85.3	97.4	**55.2**	64.5
IDEA† [8]	80.2	82.1	90.0	93.3	87.2	96.5	47.0	62.4
EPIC†	**80.3**	**85.8**	**92.8**	**94.5**	**87.5**	**98.1**	53.2	**71.2**

TABLE II
COMPARISON OF PARAMETERS AND FLOPS.

Methods	Params (M)	FLOPs (G)	RGBNT201 mAP	RGBNT201 R-1	RGBNT100 mAP	RGBNT100 R-1	MSVR310 mAP	MSVR310 R-1
TOP-ReID [7]	324.5	35.5	72.3	76.6	81.2	96.4	35.9	44.6
EDITOR [13]	119.3	40.8	66.5	68.3	82.1	96.4	39.0	49.3
MambaPro [23]	74.8	52.4	78.9	83.4	83.9	94.7	47.0	56.5
DeMo [22]	98.8	35.1	79.0	82.3	86.2	97.6	49.2	59.8
IDEA [8]	91.7	43.7	80.2	82.1	87.2	96.5	47.0	62.4
PromptMA [16]	107.9	36.2	78.4	80.9	85.3	97.4	**55.2**	64.5
EPIC	94.2	44.5	**80.3**	**85.8**	**87.5**	**98.1**	53.2	**71.2**

IV. EXPERIMENTS

A. Datasets and Evaluation Protocols

We evaluate our method on three multi-modal object ReID benchmarks. RGBNT201 [5] is a multi-modal person ReID dataset, consisting of 4,787 aligned RGB, NIR, and TIR image triplets from 201 identities. RGBNT100 [6] is a large-scale multi-modal vehicle ReID dataset with 17,250 image triplets, covering a wide range of challenging visual conditions. MSVR310 [17] is a small-scale multi-modal vehicle ReID dataset with 2,087 image triplets, featuring diverse environments and time spans. We report mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K ($K = 1, 5, 10$) for comprehensive evaluation.

B. Implementation Details

We implement our model in PyTorch and adopt the pre-trained CLIP [15] as the visual encoder. We resize images in triplets to 256×128 for RGBNT201 and 128×256 for RGBNT100 and MSVR310. For data augmentation, we apply random horizontal flipping, cropping, and random erasing [24]. For RGBNT201 and MSVR310, we set the mini-batch size to 64 and sample 8 images per identity. For RGBNT100, we set the mini-batch size to 128 and sample 16 images per identity.

TABLE III
PERFORMANCE COMPARISON WITH DIFFERENT MODULES.

Index	Modules SIPC	Modules ECR	Params (M)	Metrics mAP	Metrics Rank-1
A	×	×	86.7	72.4	76.1
B	×	✓	87.5	78.3	83.1
C	✓	×	93.6	77.3	81.2
D	✓	✓	94.2	**80.3**	**85.8**

TABLE IV
ABLATION STUDY OF INTERNAL COMPONENTS.

Index	SIPC MCG	SIPC SAIP	SIPC MGA	ECR ERW	ECR ARR	Metrics mAP	Metrics Rank-1
A	✓	✓	×	✓	✓	71.7	73.4
B	×	✓	✓	✓	✓	75.0	76.7
C	✓	×	✓	✓	✓	75.5	78.9
D	✓	✓	✓	×	✓	76.4	81.3
E	✓	✓	✓	✓	×	76.9	80.1
Full	✓	✓	✓	✓	✓	**80.3**	**85.8**

We use the Adam optimizer to optimize the proposed modules with a learning rate of 3.5×10^{-4}, and set a lower learning rate of 5×10^{-6} for the visual encoder. We train the model for 50 epochs. In SAIP, we set the number of sampled sentences K to 2. We initialize the GeM exponent ρ to 3.0.

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

Multi-modal Person ReID. As shown in Tab. I, CLIP-based methods generally outperform ViT-based approaches, benefiting from semantic priors learned via vision–language pre-training. EPIC achieves the best performance across all metrics, reaching 80.3% mAP and 85.8% R-1, and outperforming MambaPro by 2.4% in R-1. It also attains 92.8% R-5 and 94.5% R-10, both the best results.

Multi-modal Vehicle ReID. On the large-scale RGBNT100 dataset, EPIC ranks first with 87.5% mAP and 98.1% R-1. On the more challenging MSVR310 dataset, EPIC achieves 71.2% R-1, surpassing the second-best PromptMA by 6.7%. Although PromptMA attains a slightly higher mAP (55.2% vs. 53.2%), EPIC leads substantially in R-1 (71.2% vs. 64.5%), indicating stronger first-match precision. These results confirm EPIC's generalizability across diverse scenarios.

Parameter and Computation Comparison. As shown in Tab. II, EPIC achieves the best performance on all three datasets with comparable parameters (94.2M vs. 91.7M) and FLOPs (44.5G vs. 43.7G) as IDEA. Compared to TOP-ReID with 324.5M parameters, EPIC achieves better performance using only 29% as many parameters.

D. Ablation Studies

We evaluate the effectiveness of our method on the RGBNT201 dataset. Our baseline uses a three-branch encoder and retrieves with concatenated class tokens across modalities.

Effects of Key Modules. Tab. III presents the performance of various module combinations. Model A serves as the baseline, achieving only 72.4% mAP. Model B introduces ECR alone and improves mAP by 5.9%, validating the effectiveness

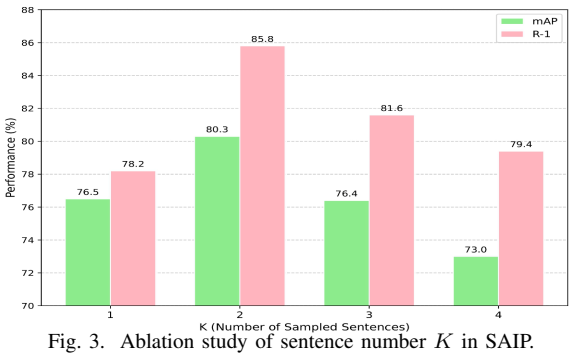

Fig. 3. Ablation study of sentence number K in SAIP.

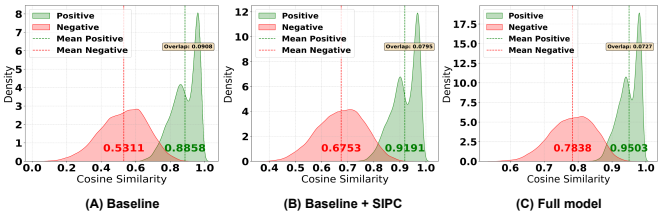

Fig. 4. Visualization of cosine similarity distributions.

of evidential corroboration for modality fusion. Model C introduces SIPC alone and improves mAP by 4.9%, confirming the value of semantic inverse prompting. Model D combines both modules and achieves 80.3% mAP, a 7.9% gain over the baseline, demonstrating strong complementarity.

Effects of Internal Components. Tab. IV reports ablations on internal components, where each setting removes one component and replaces it with its baseline counterpart. Model A replaces MGA with a single CLS token, yielding the largest drop (8.6% mAP) and highlighting the importance of multi-grained feature aggregation. Removing MCG (Model B) leads to a 5.3% mAP drop, while removing SAIP (Model C) causes a 4.8% mAP drop, confirming the benefit of semantic guidance. Model D removes ERW, where fusion weights degrade to a reliability-only softmax without corroboration, causing a 3.9% mAP drop. Model E removes ARR, which skips residual rectification and concatenates the modality features, leading to a 3.4% mAP drop. These results demonstrate the importance of evidential corroboration and adaptive rectification.

Effect of Sentence Number K. Fig. 3 shows the ablation results of the sampled sentence number K in SAIP. $K = 2$ achieves the best performance. When $K = 1$, the semantic information is insufficient, while $K \geq 3$ introduces noise that degrades performance.

E. Visualization Analysis

Cosine Similarity Distributions. Fig. 4 visualizes the cosine similarity distributions under three configurations. The baseline (A) exhibits an overlap ratio of 9.08%. After incorporating SIPC (B), the overlap decreases to 7.95%, showing that semantic consensus gating effectively filters modality-specific noise. With the full model (C), the overlap further reduces to 7.27%, a 19.9% relative reduction from baseline. The progressively sharper peaks indicate that same-identity features become increasingly compact, supporting complementarity.

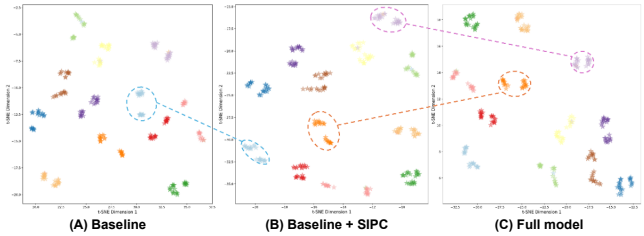

Fig. 5. t-SNE visualization of features. Each color represents an identity.

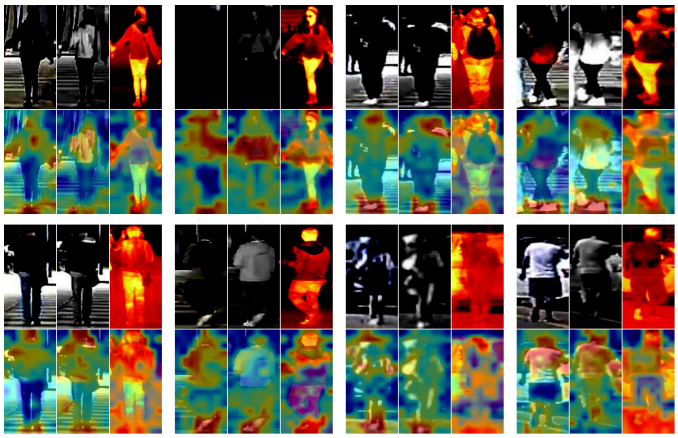

Fig. 6. Channel activation maps on person ReID.

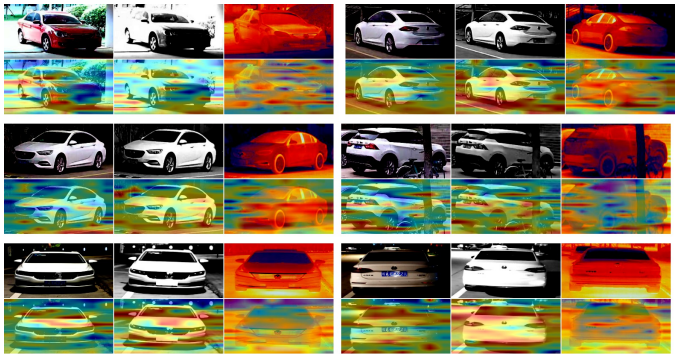

Fig. 7. Channel activation maps on vehicle ReID.

t-SNE Feature Visualization. Fig. 5 presents the t-SNE visualization of learned features, where each color denotes a distinct identity. In the baseline (A), samples from the same identity are scattered across the feature space, as highlighted by the dashed circle. After adding SIPC (B), intra-identity clusters become more compact, and the inter-identity separation improves. With the full model (C), same-identity samples form tight clusters while maintaining clear boundaries between different identities. This progression demonstrates that our method learns increasingly discriminative representations.

Channel Activation Maps. Fig. 6 and Fig. 7 illustrate the channel activation maps across different modalities. Each modality exhibits distinct activation patterns and highlights semantically meaningful regions, indicating that our modules effectively exploit complementary multi-modal cues.

Fig. 8. Ranking list on person ReID. (A) Baseline. (B) Baseline + SIPC. (C) Full model. Green/red boxes indicate correct/incorrect matches.

Fig. 9. Ranking list on vehicle ReID. (A) Baseline. (B) Baseline + SIPC. (C) Full model. Green/red boxes indicate correct/incorrect matches.

Ranking List Visualization. Fig. 8 and Fig. 9 present qualitative ranking list comparisons on person and vehicle datasets, respectively. From (A) to (C), the proportion of correct matches progressively increases, validating the complementary effects of SIPC and ECR for robust retrieval.

V. Conclusion

This paper presents EPIC, a novel framework for multi-modal object ReID that mitigates modality-specific noise through semantic governance. The SIPC module transforms textual semantics into visual guidance through three cascaded sub-modules: MCG filters noise via cross-modal consensus, SAIP projects text features into visual prompt tokens, and MGA aggregates multi-grained representations. The ECR module performs reliability-aware fusion through ERW, which computes modality weights based on evidential corroboration, and ARR, which applies adaptive residual rectification. Together, SIPC and ECR establish a complete pipeline from semantic gating to reliability-aware fusion. Extensive experiments on three multi-modal ReID benchmarks validate the effectiveness and superiority of our proposed method for feature learning.

References

[1] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *ICCV*, 2021.

[2] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *CVPR*, 2021.

[3] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *AAAI*, 2023.

[4] K. Zhu, H. Guo, S. Zhang, Y. Wang, J. Liu, J. Wang, and M. Tang, "Aaformer: Auto-aligned transformer for person re-identification," *TNNLS*, 2023.

[5] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *AAAI*, 2021.

[6] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *AAAI*, 2020.

[7] Y. Wang, X. Liu, P. Zhang, H. Lu, Z. Tu, and H. Lu, "Top-reid: Multi-spectral object re-identification with token permutation," in *AAAI*, 2024.

[8] Y. Wang, Y. Lv, P. Zhang, and H. Lu, "Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification," in *CVPR*, 2025.

[9] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[10] 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 *AAAI*, 2022.

[11] A. Zheng, Z. He, Z. Wang, C. Li, and J. Tang, "Dynamic enhancement network for partial multi-modality person re-identification," *arXiv preprint arXiv:2305.15762*, 2023.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *CVPR*, 2024.

[14] Z. Wang, H. Huang, A. Zheng, and R. He, "Heterogeneous test-time training for multi-modal person re-identification," in *AAAI*, 2024.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[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 TIP*, 2025.

[17] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark," *arXiv preprint arXiv:2208.00632*, 2022.

[18] D. Wu, Z. Liu, Z. Chen, S. Gan, K. Tan, Q. Wan, and Y. Wang, "Lrmm: Low rank multi-scale multi-modal fusion for person re-identification based on rgb-ni-ti," *ESWA*, 2025.

[19] J. Crawford, H. Yin, L. McDermott, and D. Cummings, "Unicat: Crafting a stronger fusion baseline for multimodal re-identification," *arXiv preprint arXiv:2310.18812*, 2023.

[20] W. Pan, L. Huang, J. Liang, L. Hong, and J. Zhu, "Progressively hybrid transformer for multi-modal vehicle re-identification," *Sensors*, 2023.

[21] 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," *TCSVT*, 2024.

[22] Y. Wang, Y. Liu, A. Zheng, and P. Zhang, "Decoupled feature-based mixture of experts for multi-modal object re-identification," in *AAAI*, 2025.

[23] 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 *AAAI*, 2025.

[24] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.