Chain-of-Thought Guided Multi-Modal Object Re-Identification

Supplementary Material

6. Introduction

In this supplementary material, we provide comprehensive experimental details, visual examples and extended analyses to support the findings of the main manuscript. To be specific, the supplementary material is organized as follows:

1. CoT-guided Multi-modal Caption Generation.
Detailed description of the CoT-guided reasoning-chains and descriptions generation pipeline.

- CoT-based logical reasoning process generation.
- CoT-guided object attribute feature generation.
- The examples of our captions with the objects.

2. Details of Modules and Experiments

- Discussions on CRE module.
- Discussions on CT-CMC module.

3. Module Validation and Analysis.

- The hyperparameter analysis plots of the number of negative samples n_0 and the coefficient α for RGBNT100, MSVR310, and RGBNT201 datasets, respectively.

4. Visualization Analysis of CoT-ReID.

- CAM maps of our method on person and vehicle datasets.
- Ranklist plots of our method on vehicle datasets.

These proposed analysis provide a deeper understanding of our caption generation pipeline, the proposed modules and experimental results, further elaborating on the experimental details and validating the effectiveness of our method.

7. CoT-guided Multi-modal Caption Generation

7.1. Details of the CoT-guided Caption Generation Pipeline.

To address the insufficient attention paid to semantic logical hierarchy in multi-modal object ReID, and the difficulty for models to capture and learn the logical hierarchy of visual features, we leverage the logical hierarchy of textual semantics to guide the entire visual training process. Specifically, the reasoning process of large language models (MLLMs) can generate chain-of-thought (CoT)-based reasoning procedures, which we adopt as the logical reasoning hierarchy of textual semantics. In other words, as shown in 8, we propose a novel CoT-based caption generation pipeline that utilizes the reasoning chain of MLLMs to generate two types of text: reasoning process texts with logical reasoning hierarchy, and object description caption texts based on logical reasoning. Concretely, our pipeline consists of two steps: **(1) CoT-based logical reasoning process generation, and (2) CoT-guided object attribute feature generation.** Without loss of generality, we take the WMVeID863 dataset as an example to illustrate the caption generation process. Because of each modality of multi-modal has its own unique advantages which are particularly helpful for multi-spectral joint decision-making. And when presented with images from multiple spectra, MLLMs usually focus on shared semantic information. Therefore, we generate MLLM-based captions for images of each spectrum separately.

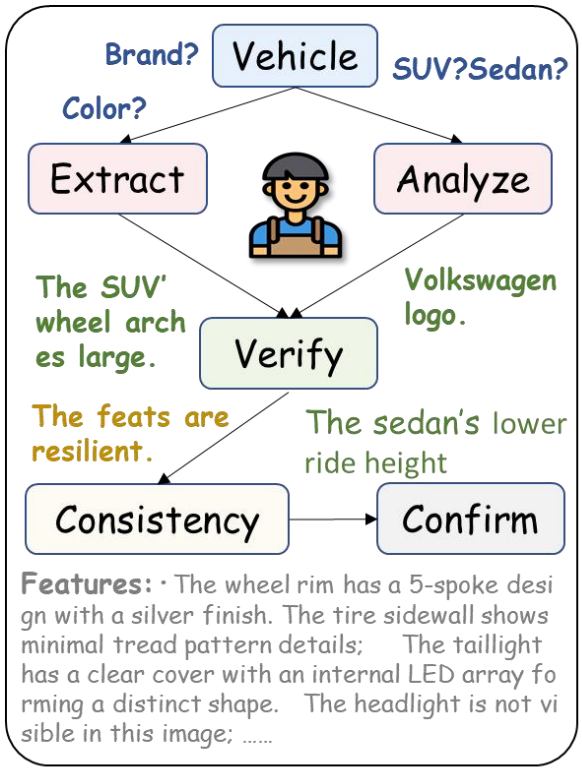
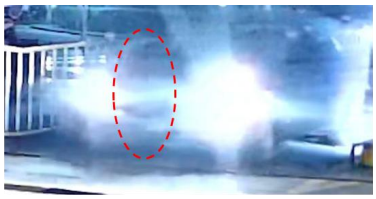


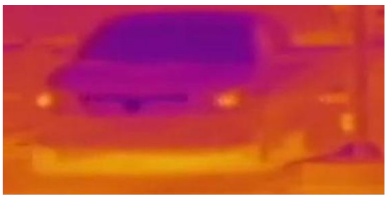
Figure 8. The CoT-guided multi-modal caption generation pipeline.

CoT-based logical reasoning process generation. Instruct MLLMs to focus on specific spectra (e.g., RGB, NIR, or TIR) and provide the reasoning process for feature description of objects in the corresponding spectrum.

1. Generic Annotation Template on WMVeID863: *"You are a vehicle Re-ID expert specializing in NI (Near-Infrared) images affected by flare. Analyze the NI image with these details: features:name: Flare Info, description: , reid−importance: High; name: View Angle (Flare−Adjusted), description:, reid−importance:*



1) Extract RGB-specific unique features (white painted body, oval headlights with vertical rounded-edge grille, tinted curved windows on used-wear body, clean sloped roof with no spoiler and no roof rack); 2) Verify consistency via feature relations: the contrast between white painted body and tinted curved windows enhances visibility under bright light despite blur, and the clean sloped roof aligns with the front-end styling indicating a consistent model trim; 3) Filter candidates mismatching critical unique features (e.g., non-white body, presence of spoiler or roof rack, non-oval headlights); 4) Leverage RGB's ability to retain key unique details (white body color, window tint level, grille shape) despite environmental conditions (bright light causing glare or blur); 5) Prioritize critical unique features (white painted body, oval headlights with vertical rounded-edge grille) for vehicle Re-ID accuracy."



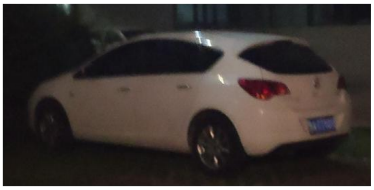
1) Extract TIR-specific unique features (cooled tinted curved windows with low thermal emission, horizontal grille with rounded edges exhibiting moderate thermal gradient, roof rack presence altering roof thermal uniformity, hidden exhaust system with suppressed rear thermal plume); 2) Verify consistency feature relations: the upper-body thermal disruption from roof rack aligns with window cooling patterns, and front grille gradient matches rear exhaust invisibility; 3) Filter candidates mismatching critical unique features (e.g., high-temperature engine zone suggesting active exhaust, sharp thermal edges on grille, or unbroken roof thermal field without rack); 4) Leverage TIR's ability to retain key unique details (window cooling signatures, roof rack anomalies, exhaust suppression) despite environmental conditions (e.g., darkness, smoke); 5) Prioritize critical unique features like cooled tinted curved windows for vehicle Re-ID accuracy."

1) Extract NIR-specific unique features: oval headlight contours with high NIR reflectance, horizontal grille with rounded edges and moderate NIR contrast, tinted curved windshield with low NIR transmittance, flat roof and intact bumper geometry with uniform NIR response; 2) Verify consistency via defined relations: headlight-grille symmetry reinforces frontal identity, windshield-roof/bumper contrast supports vertical structure stability; 3) Filter candidates mismatching critical unique features (e.g., non-oval headlights, high-transmittance windshield, deformed bumpers); 4) Leverage NIR's ability to retain key unique details such as material-based reflectance differences and geometric edge coherence despite environmental blur and nighttime lighting; 5) Prioritize critical unique features—specifically headlight contour reflectance and windshield transmittance—for vehicle Re-ID accuracy under challenging conditions."

1) Extract RGB-specific unique features (red body paint with used wear, tinted flat windows with reduced reflectivity, oval headlights and taillights with horizontal alignment, rounded-edge horizontal grille with clean sloped roof); 2) Verify consistency via feature relations: the low-reflectivity tinted windows complement the worn red paint under dim light, while the horizontal alignment of oval lights and grille reinforces frontal symmetry; 3) Filter candidates mismatching critical unique features (e.g., vehicles with chrome trims, curved windows, spoilers, or non-oval lighting); 4) Leverage RGB's ability to retain key unique details (faded red hue, window tint level, shape of lights and grille) despite environmental conditions (dim blur); 5) Prioritize critical unique features (red body paint with used wear, oval headlights and taillights) for vehicle Re-ID accuracy."

1) Extract TIR-specific unique features (Asymmetric off-round headlight thermal profile, Cooled tinted curved window thermal emission pattern, Horizontal straight-edged grille with sharp thermal gradient, Uniform roof-to-bumper thermal continuity with intact bumper integrity); 2) Verify consistency via feature relations: the grille's thermal symmetry must frame the asymmetric headlights, and the cooled window zone must seamlessly connect to the thermally uniform roof and bumpers; 3) Filter candidates mismatching critical unique features (e.g., symmetric headlight emissions, broken thermal continuity at bumper joints, diffuse grille edges); 4) Leverage TIR's ability to retain key unique details (asymmetric lighting profiles, cooled glass signatures, sharp thermal gradients on grilles) despite environmental conditions (e.g., darkness); 5) Prioritize critical unique features (Asymmetric off-round headlight thermal profile, Horizontal straight-edged grille with sharp thermal gradient) for vehicle Re-ID accuracy."

1) Extract NIR-specific unique features (Smooth, non-curved glass surfaces show uniform low transmittance and high reflectance, creating sharp boundary contrasts against the dim ambient light); 2) Verify consistency via feature relations: High-reflectance flat windows amplify ambient NIR interference, which is counterbalanced by the controlled emission pattern of off-oval headlights, improving feature isolation in low-light conditions ; 3) Filter candidates mismatching critical unique features (e.g., vehicles with low-reflectance headlights, broken bumper symmetry, or roof racks); 4) Leverage NIR's ability to retain key unique details (Grille structure exhibits strong edge-defined thermal contrast due to metal reflectance and rounded contour) despite environmental conditions (bright nighttime light causing blur); 5) Prioritize critical unique features like rectangular headlights with high NIR reflectance for vehicle Re-ID accuracy."



1) Extract RGB-specific unique features (white Mazda sedan body, tinted curved rear windows with angular taillights, rounded-edge horizontal grille with off oval headlights, single-pipe exhaust and no roof rack/spoiler); 2) Verify consistency feature relations (tinted windows enhance taillight contrast on white body, lack of spoiler aligns with soft grille design); 3) Filter candidates mismatching critical unique features (e.g., non-white body, presence of roof rack, dual exhaust, circular headlights); 4) Leverage RGB's ability to retain key unique details (angular taillights, window tint contrast, grille curvature) despite environmental conditions (dim blur); 5) Prioritize critical unique features (white Mazda sedan body, single-pipe exhaust and no roof rack/spoiler) for vehicle Re-ID accuracy."

1) Extract TIR-specific unique features (Asymmetric off-oval headlight thermal signature, Cooled tinted window thermal profile with low emissivity gradient, Horizontal grille with rounded-edge thermal contrast, Hidden exhaust thermal suppression at rear diffuser zone); 2) Verify consistency via feature relations: frontal asymmetry between headlights and cooled windows must coexist, and front grille cooling should align thermally with rear exhaust suppression; 3) Filter candidates mismatching critical unique features (e.g., symmetric headlight emissions, high-temperature rear bumper indicating exposed exhaust, sharp-edged grille with abrupt thermal drop); 4) Leverage TIR's ability to retain key unique details (asymmetric lighting signatures, window emissivity differences, exhaust heat absence) despite environmental conditions (e.g., darkness); 5) Prioritize critical unique features (Asymmetric off-oval headlight thermal signature, Hidden exhaust thermal suppression at rear diffuser zone) for vehicle Re-ID accuracy."

1) Extract NIR-specific unique features (off-oval headlight reflectance pattern, horizontal grille with rounded-edge NIR contrast, tinted curved windshield with low NIR transmittance, single-exhaust pipe thermal-NIR emission profile); 2) Verify consistency via feature relations: confirm that the off-oval headlight pattern aligns with grille edge contrast and that windshield reflectance enhances headlight definition; 3) Filter candidates mismatching critical unique features (e.g., circular headlights, non-tinted flat windshields, dual exhaust pipes); 4) Leverage NIR's ability to retain key unique details (off-oval headlight geometry, grille edge contrast, windshield reflectance behavior) despite environmental conditions (e.g., low light); 5) Prioritize critical unique features (off-oval headlight reflectance pattern, tinted curved windshield with low NIR transmittance) for vehicle Re-ID accuracy."

Figure 9. More example of the CoT-guided reasoning-chain text of WMVeID863.

Figure 10. More example of the CoT-guided reasoning-chain text of RGBNT201.

High; name: Flare-Resilient Features, description: , reid−importance: High; name: Modal Flare Mitigation, description: , reid−importance: Medium "reasoning−chain": Logical chain explaining NI modality advantages for ReID."

2. Anti-hallucination Instruction: Shehzaad et al. [3] combines CoT with a verification mechanism significantly enhances the factual accuracy of model outputs by generating intermediate verification steps. Building on this foundation, we further integrate anti-hallucination prompts into the instruction design. However the imaging of some spectra is blurred, MLLMs also struggle to capture detailed information. We use these prompt [30] to instruct MLLMs focus on the explicit relations:

"If certain attributes are invisible or the relationships among attributes are uncertain, such attributes and rela-

tionships should be ignored. Do not fabricate content that is not present in the images. Strictly follow the specified format, enhance the inference accuracy of attribute relationships, emphasize the impact of spectral advantages on attribute correctness and avoid making arbitrary assumptions."

CoT-guided object attribute feature generation. While enabling MLLMs to generate reasoning logic texts, we instruct them to generate reliable attribute descriptions of the target based on the CoT and the relationships between attributes. And, the prompt we provide to MLLMs is:

"Return ONLY valid JSON (no extra text, no comments, no code blocks). Strictly follow the structure below: "modal": "RGB", "features": "name": "Unique Feature Name 1", "description": "RGB-specific detail derived from the description (include unique elements)", "reid-importance": "[Critical/High/Medium], "name": "Unique Feature Name 2", "description": "RGB-specific detail derived from the description (include unique elements)", "reid-importance": "[Critical/High/Medium],
"name": "[Unique Feature Name 3]", "description": "RGB-specific detail derived from the description (include unique elements)", "reid-importance": "[Critical/High/Medium],
"name": "[Unique Feature Name 4]", "description": "RGB-specific detail derived from the description (include unique elements)", "reid-importance": "[Critical/High/Medium]",
"feature-relations": "relation": "[Feature A] vs.[Feature B]", "description": "Logical interaction between the two features, grounded in the description", "reid-value": "Practical Re-ID benefit",
"relation": "[Feature C] vs. [Feature D]", "description": "Logical interaction between the two features, grounded in the description", "reid-value": "Practical Re-ID benefit"

7.2. The examples of our captions with the images

Examples of multi-modal Vehicle ReID datasets. In the Fig. 9, we provide additional examples from the WMVeID863 dataset to demonstrate the reliability of our CoT-based logical reasoning subtitle generation pipeline. The generated text descriptions offer CoT-driven logical reasoning for each spectrum. As can be seen from the CoT-guided logical reasoning texts we presented, during the reasoning process, textual semantics first extract modal-specific information, second focus on key modal attribute information, then verify the consistency between attributes, and finally filter out reliable attribute descriptions to generate valid ones. For instance, when the vehicle color is visible, the textual semantics prompt attention to vehicles of the same color during logical reasoning. Such a textual logical reasoning process is progressive layer by layer, helping the model focus on semantic and logical information and improve recognition accuracy.

Examples of multi-modal person ReID datasets. In Fig. 13, we provide additional examples from the RGBNT201 dataset to demonstrate the reliability of our CoT-based logical reasoning subtitle generation pipeline. We present the CoT-based reasoning chain text generated for each spectrum, where we can observe the CoT-based reasoning process: it first extracts the spectrum-specific information of a person (e.g., focusing on clothing colors and other visual attributes), then identifies correlations between these attributes, verifies attribute consistency, and finally prompts the model to emphasize spectrum-specific information.

8. Details of Modules and Experiments

8.1. Discussion on the CRE Module.

Inspired by the register token concept from DINOv3 [23], we embed the reversed textual semantic logic token. This aims to guide visual features to focus on semantic logic in the early stage of visual feature extraction, thus enhancing the attention of visual modalities to reasoning logic.

The computational impact of embedded tokens on the backbone .

The core computational cost of the Transformer stems from the self-attention mechanism, with a time complexity of:

$$\text{Time} \propto B \cdot T^2 \cdot D + B \cdot T \cdot D^2, \tag{9}$$

where B = batch size, T = token numbers and D = dims.

Let the number of original tokens be T_0 (e.g.,$T_0 = 1(\text{CLS}) + 4(\text{Register}) + 196(\text{patch}) = 201$ in DINOv3-B). After adding one token, $T_1 = T_0 + 1 = 202$, and the increment in the self-attention computation amount is:$\frac{T_1^2 - T_0^2}{T_0^2} = \frac{(202^2 - 201^2)}{201^2} \approx 1.0\%$. Notably, the complexity grows quadratically with the number of tokens, while the increment introduced by a single token is almost negligible.

The different numbers of token reverse embeddings on the model performance. Through the experiments on the WMVeID863 dataset in the Table. 9, which compare different numbers of token reverse embeddings, we find that using one token for semantic inversion embedding is not only controllable in terms of parameter count but also achieves the optimal model performance.

The different embedding layer configurations on the model performance. Table 8 compares the impact of different embedding layer configurations on the model performance on the WMVeID863 dataset. We observe that embedding the cot token in every layer outperforms embedding only in the first 6 layers or only in the last 6 layers, achieving superior results in both mAP and Rank-1 metrics. In summary, these findings validate the effectiveness of our

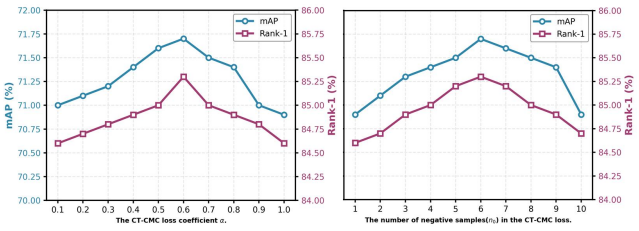


Figure 11. Analysis of hyperparameters on MSVR310 dataset of the method: coefficients α of the loss and the number of negative samples N_0.

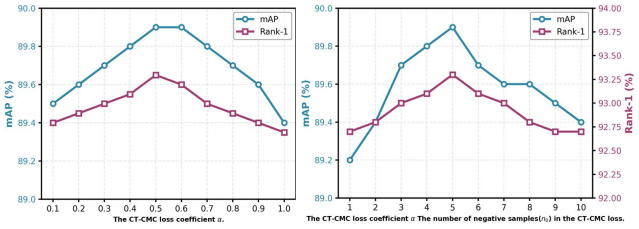


Figure 12. Analysis of hyperparameters on RGBNT100 dataset of the method: coefficients α of the loss and the number of negative samples N_0.

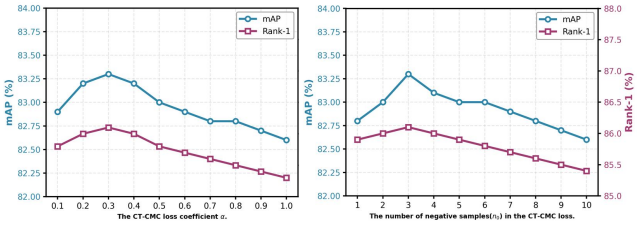


Figure 13. Analysis of hyperparameters on RGBNT201 dataset of the method: coefficients α of the loss and the number of negative samples N_0.

CRE module, leveraging CoT-based semantic reversion embedding and the reliability of our choice in the embedding configuration ablation study.

Token Num	Metric			
	mAP	R-1	R-5	R-10
1	**74.7**	**82.0**	**85.9**	**89.8**
2	73.0	80.2	83.5	87.6
3	71.9	79.9	83.0	86.9
4	71.0	79.3	82.8	86.7

Table 7. Ablation study on the number of cot tokens. "Token Num" denotes the number of additional cot tokens.

8.2. Effects of CoT-guided reasoning text and captions.

Table. 9 shows the comparison of replacing static text in IDEA with our CoT-based text on MSVR310 and RGBNT201. Together with Table. 3 in the main manuscript, the performance gain of our method from CoT-based text is demonstrated.

Layers	Metric			
	mAP	R-1	R-5	R-10
1∼6	73.5	79.5	83.2	87.2
7∼12	73.8	79.6	83.5	87.5
all	**74.7**	**82.0**	**85.9**	**89.8**

Table 8. Ablation study on the number of cot tokens. "Token Num" denotes the number of additional CoT tokens.

Methods	MSVR310		RGBNT201	
	mAP	R-1	mAP	R-1
IDEA†	47.0	62.4	80.2	82.1
IDEA†(CoT-Text)	47.8	63.8	80.9	82.8
IDEA°	67.0	82.4	81.3	83.2
IDEA°(CoT-Text)	68.4	84.1	81.5	84.5
CoT-ReID°	**71.7**	**85.3**	**83.3**	**86.1**

Table 9. Symbols: † (CLIP-based) and ° (DINOv3-based). Replacing the static text in the IDEA method with CoT-Text.

Methods	Metric			
	mAP	R-1	R-5	R-10
3M Loss [31]	69.7	82.7	92.4	94.3
CT-CMC	**71.7**	**85.3**	**94.3**	**96.5**

Table 10. Performance comparison of CT-CMC and $3M\,Loss$ [31] in our method on the WMVeID863 dataset.

8.3. Discussion of CT-CMC.

Table 10 validates the effectiveness of using the CoT-guided reasoning text for cross-modal consistency. To verify this, we conduct experiment on the MSVR310 dataset while maintaining the same framework but replacing our proposed CT-CMC with 3M Loss [31]. The 3M Loss aims to improve diversity of modal information by enlarging the distance of multi-modal feature centers. When compared to conventional cross-modal consistency constraints that do not incorporate semantic logical guidance, the results show a clear performance drop, with mAP/Rank-1 decreasing by **2.0%/2.6%** compared to our full model. This performance demonstrates the superiority of utilizing CoT-generated text as dual constraints to construct cross-modal consistency.

	RGBNT201	WMVeID863	MSVR310	RGBNT100
Train	171/3951	603/10446	155/1032	50/8675
Query	30/836	210/2904	52/591	50/1715
Gallery	30/836	272/3678	155/1055	50/8575
Challenges	Wide Views, Occlusions	Intense Flare	Longer Time Span, Complex Conditions	Different Views, Illumination Issue

Table 11. Details of the datasets partition settings and their corresponding challenges, */* represents ID/Sample.

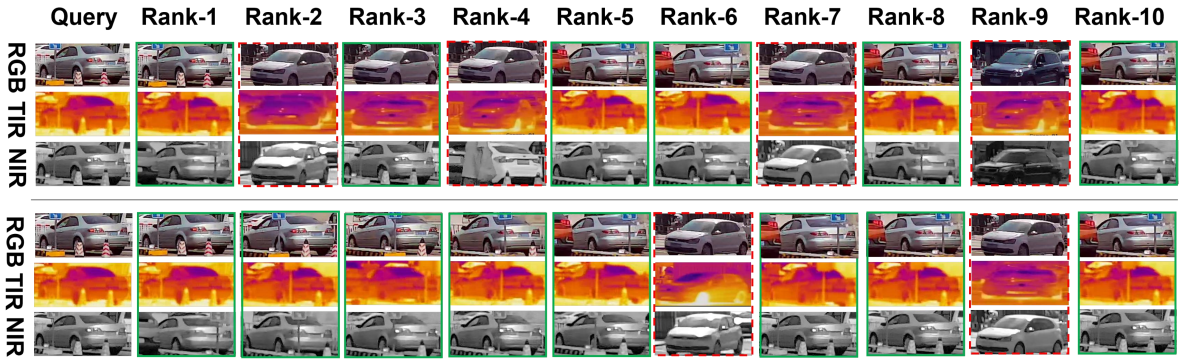


Figure 14. Visualization of the ranklist of the baseline method and our approach on the WMVeID863 dataset.

9. Implementation Details and Module Analysis

9.1. Datasets Information

To comprehensively evaluate the generalizability of the proposed CoT-ReID framework, we conduct experiments on four public object datasets. These include RGBNT201 [37], WMVeID863 [46], MSVR310 [45], and RGBNT100 [14]. These datasets collectively reflect a wide range of real-world scenarios and associated challenges. Tab. 11 summarizes the partition protocols and the specific challenges posed by each dataset. The challenges include flare pollution, large time, wide views, occlusions and viewpoint spans. Combined with the best performance of our method, it is demonstrated that CoT can handle most of the challenges in vehicle and person ReID.

9.2. Analysis of the experimental cost.

Table 12 compares our method with IDEA (which also uses a text branch for inference): with the same backbone, ours outperforms IDEA in both performance and GFLOPS.

9.3. Analysis of Model Parameters and Performance.

Fig. 11, 12 and 13 are hyperparameter analysis plots of the number of negative samples n_0 and the coefficient α for RGBNT100, MSVR310, and RGBNT201 datasets, respectively. As can be seen, the optimal choices of our proposed CT-CMC module are **5/0.5, 5/0.6, 3/0.3**.

Method	Paras(M)↓	FLOP(G)↓	mAP↑	R-1↑
TOP-ReID*	324.5	69.1	35.9	44.6
PromptMA†	107.4	67.4	55.2	64.5
PromptMA°	189.9	71.2	58.2	68.5
IDEA†	91.7	43.7	47.0	62.4
IDEA°	178.9	76.8	67.0	82.4
CoT-ReID°	**174.9**	**52.2**	**71.7**	**85.3**

Table 12. Symbols: * (ViT-based), † (CLIP-based), ° (DINOv3-based).

10. Visualization Analysis of CoT-ReID

10.1. Visualization of Channel Activation Maps on the Multi-modal Object ReID Dataset.

As shown in Fig. 15, we present the channel activation maps of different modalities in CoT-ReID on the person and vehicle ReID dataset. Context-aware textual guidance helps the model focus on more discriminative regions, enhancing feature robustness and improving interpretability. These activation maps effectively capture discriminative local information, highlighting the importance of leveraging interaction between global features and discriminative local information in multi-modal object ReID. The results validate the rationality of our method and discriminative region localization in complex ReID scenarios.

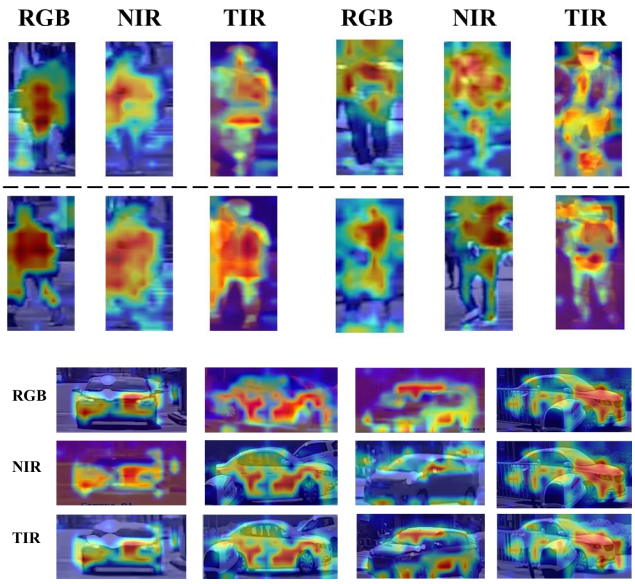


Figure 15. Visualization of channel activation maps of different modalities on the person ReID dataset and vehicle ReID dataset.

10.2. Visualization of Multi-modal Ranking List on the Multi-modal Vehicle ReID Dataset.

As shown in Fig. 14, we present a comparative analysis of multi-modal ranking lists across different module configurations that focuses on the diverse modalities of CoT-ReID on the WMVeID863 vehicle dataset. Our baseline performs effectively in many conventional scenarios but struggles significantly with hard cases involving visually indistinguishable vehicles. This limitation is particularly evident in its failure to consistently prioritize correct matches in the ranking list. In contrast, our CoT-ReID model leverages the logical reasoning chains to guide the visual features to learn, achieving a marked reduction in misrankings even for the most challenging cases. The comparative analysis validates the CoT's enhanced robustness in addressing practical challenges encountered in real-world vehicle ReID scenarios.