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

Ya Gao[1], Shihao Li[1], Zhaojun Liu[1], Aihua Zheng[1,2]*, Chenglong Li[1]*, Jin Tang[1]
[1]Anhui University, China
[2]Anhui Provincial Key Laboratory of Intelligent Detection and Diagnosis for Traffic Infrastructure
{gaoya615, shli0603, junmain.liu, ahzheng214, lcl1314}@foxmail.com, ahu_tj@163.com

Abstract

With the rise of visual-language models, multi-modal ReID retrieves specific targets by integrating different spectra and textual descriptions. Existing methods merely adopt descriptive representation learning for image-text, ignoring the relationships among the intrinsic logical hierarchies of semantic features. Since Chain-of-Thought (CoT) can provide textual logical context and enhance semantic perception in large-model reasoning, we propose ***CoT-ReID****, a CoT-guided framework that injects the Multi-modal Large Language Models (MLLMs) reasoning into multi-modal ReID. Specifically, we simulate human-like joint vision-text logical decision-making, leveraging CoT textual logical reasoning to guide visual feature learning at the early, late and decision-making levels: we first embed the semantic reversion of CoT hierarchical reasoning into visual features to calibrate bottom-level features and highlight visual hierarchical reasoning, then take CoT hierarchical reasoning text as an anchor condition to constrain the consistency of visual cross-modal semantics, and finally embed logically reasoned text attribute features into multi-modal decision-making via CoT's hierarchical reasoning to provide logical support for selecting discriminative identity features. By constructing CoT textual benchmarks and our proposed modules, our framework generates more robust multi-modal features in complex scenarios, and comprehensive experiments on four datasets (RGBNT100, MSVR310, WMVeID863, RGBNT201) demonstrate the superiority of our method over state-of-the-art approaches.*

1. Introduction

Object Re-Identification (ReID) aims to retrieve the same target from images captured by different camera views. During the past decade, RGB images have dominated the initial development stage [8, 18, 37, 48]. To address more real-world challenges such as darkness, occlusion, and strong light interference, Near-Infrared (NIR) and Thermal Infrared (TIR) spectra have been introduced [14, 27, 32, 45], expanding the coverage of object ReID scenarios, but introducing cross-modal discrepancies that complicate feature alignment and fusion. With the vigorous development of Multi-modal Large Language Models (MLLMs), existing methods [10, 30, 39] have introduced text annotations to supplement information such as attributes of objects and designed several text-image alignment methods to adaptively aggregate discriminative multi-modal information.

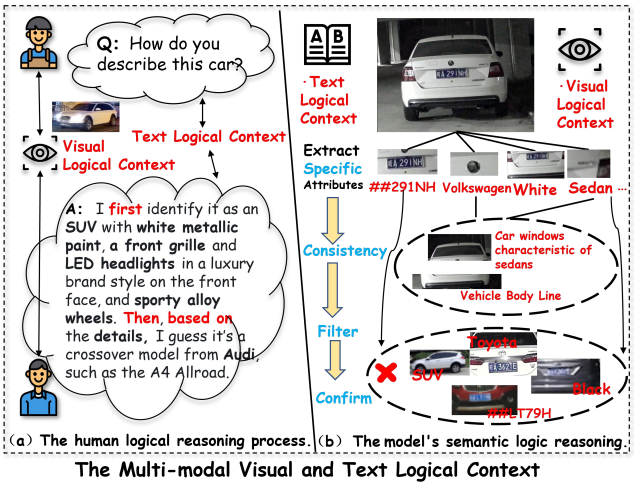

Figure 1. (a) The process of human reasoning about object information, including both visual and text semantic context. (b) During our model training, both visual and textual semantics contain the logical reasoning process.

While existing methods that incorporate semantic features offer descriptive representations of objects, they largely overlook the intrinsic hierarchical logic governing visual elements. Object visual features are usually composed of multiple interacting layers to form a precise context. As the example in Fig. 1 (b), a vehicle's brand and model imply its body line, just as a pedestrian's clothing suggests their gender. As shown in Fig. 1 (a), when humans [19] describe an object through conversational rea-

*corresponding author.

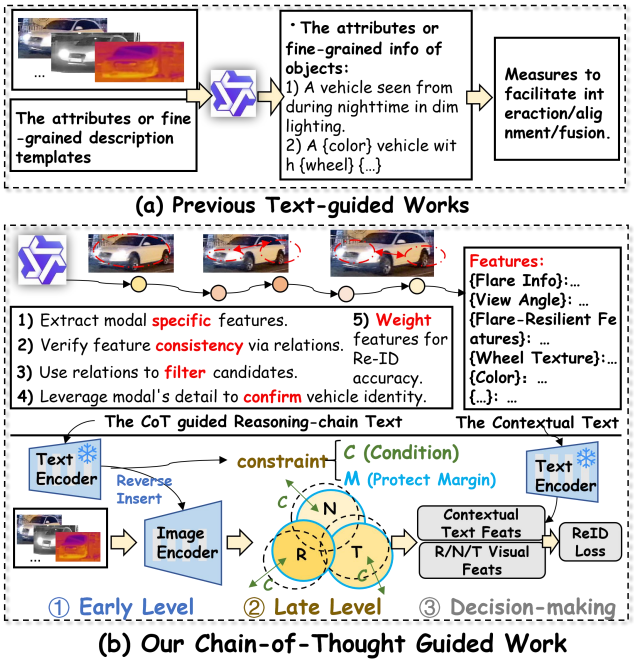

Figure 2. (a) The previous text-guided ReID methods based on text annotations. (b) The framework of our CoT-guided reasoning-chain text guide the model's full-process training.

soning, we first rely on visual observation and then verbal description, both of which contain visual contextual information and semantic contextual information. Since directly mining such hierarchical relationships purely within the visual domain remains challenging, whereas the interpretability and explicit logical reasoning of semantic contextual relationships make them more amenable to exploration. Notably, CoT [33], as a structured reasoning paradigm tailored for LLMs, guides LLMs to decompose complex visual attributes into sequential, interpretable steps.

Inspired by these observations and CoT reasoning mechanisms [26, 33, 36], we generate image attribute descriptions via MLLM reasoning to enhance visual feature understanding. To the best of our knowledge, no prior ReID work has used CoT as semantic context to boost the interpretability of visual context. As shown in Fig. 2 (b), our approach first identifies the main object, verifies attribute consistency, and weights reliable attributes instead of generating isolated, shallow annotations, which ensures CoT carries richer semantic context by encoding both attribute definitions and their logical relationships. Given inherent cross-modal disparities, previous methods have advanced in defining reasonable bounds for interactive alignment and disentanglement techniques [5, 6, 29, 31, 40], but still **rely on shallow semantic matching or low-constraint feature alignment**. To bridge the gap, we use the contextual semantics of CoT to simulate the hierarchical logic of human reasoning and we upgrade cross-modal consistency constraints from surface alignment based on static semantics to deep semantic binding and conditional marginal constraints based on dynamic reasoning.

Technically, we construct a **C**hain-**o**f-**T**hought guided multi-modal object ReID (CoT-ReID) framework. First, we develop a CoT-guided multi-modal object attribute generation pipeline, which leverages MLLM to generate both CoT text datasets that record the MLLMs' reasoning process and object attribute description datasets. Then, we leverage the CoT reasoning-chain text to guide the model's full-process training, to achieve logical reasoning and interpretable reasoning for vision. Specifically, as shown in the right part of Fig. 1, we first perform CoT semantic contextual reverse information embedding into the early visual level, aiming to guide the visual system to focus on contextual reasoning information from the early feature learning phase. Subsequently, we take CoT semantic text as a twofold constraint on cross-modal consistency: (1) it serves as a semantic anchor in the late visual level, serving as conditional prior information to constrain cross-modal information consistency; (2) it serves as marginal conditions for cross-modal interaction, enabling visual features to attend to contextual information while preventing over-alignment from undermining modality uniqueness, and ensuring effective information sharing.

In summary, the contributions of our work are as follows:

- We propose a novel CoT guided framework for multi-modal object ReID that simulates human-like joint visual-textual reasoning. By establishing full-process semantic guidance, our method steers visual feature learning toward logical semantics to achieve more interpretable and robust cross-modal representation.
- We propose three novel modules: CoT-guided Reverse Embedding (CRE) directs bottom-level visual features toward visual logical semantics; CoT-guided Cross-Modal Consistency (CT-CMC) achieves balanced fusion under logical semantic constraints and protection; and CoT-guided Text Features (CTF) injects reasoning into discriminative identity feature selection.
- We innovatively leverage MLLMs with CoT to generate benchmarks both object semantic reasoning processes and attribute annotations. Extensive experiments on four benchmarks demonstrate state-of-the-art performance, validating the effectiveness of our approach.

2. Related Work

2.1. Multi-modal Object Re-Identification

Multi-spectral object ReID leverages data from different spectra, including RGB, NIR, and TIR, to enhance ReID performance. The unique imaging characteristics of NIR and TIR bring new solutions to ReID tasks. Li et al. [14]

propose two multi-spectral vehicle datasets, RGBN300 and RGBNT100, and construct a baseline method HAMNet for fusing different modalities. To address inter-modal discrepancies inherent in multi-spectral data, Zheng et al. [45] propose CCNet, which improves fusion through a coherence constraint, and contribute the MSVR310 dataset. Zhang et al. [40] develop EDITOR to alleviate the impact of complex backgrounds on ReID performance.

The development of LLMs drives research on their applications in ReID. Existing methods integrate text semantics with multi-spectral data to supplement information and boost recognition accuracy. Building on CLIP [22]'s strong semantic-visual alignment, CLIP-ReID [16] incorporates text prompts and pioneers the application of text in ReID. To address limited text descriptions, some studies [10, 39] leverage MLLMs to generate fine-grained target captions. Li et al. [17] propose a prompt learning framework that harnesses the cross-modal alignment capabilities of vision-language pre-training models. Wang et al. [30] introduce a multi-expert strategy to couple shared information across different modalities. Despite these advances, current text-based approaches remain limited to static, fine-grained object descriptions that overlook inter-attribute relationships and lack semantic logical structure. To overcome these limitations, we propose CoT-ReID: a novel framework for multi-modal object ReID, accompanied by CoT reasoning-chain benchmarks and semantically enriched attribute descriptions.

2.2. Chain-of-Thought of Large Model

Inspired by human cognition, multi-modal rational construction enhances both accuracy and interpretability in multi-modal reasoning. Since the introduction of Chain-of-Thought (CoT) [33], various paradigms have emerged to improve multi-channel and multi-step reasoning, predominantly using text to encode multi-modal information for seamless integration with the reasoning mechanisms of LLMs or MLLMs. Several methods leverage CoT text to improve retrieval accuracy. For instance, vision-language-action models [42] generate intermediate reasoning steps to replace traditional direct action prediction, thereby enhancing model performance. X-CoT [20] employs an interpretable retrieval framework powered by LLM-based CoT reasoning, moving beyond embedding model-based similarity ranking. In image generation, PromptCoT [35] refines input prompts, PARM++ [7] optimizes reward mechanisms, and LayoutLLM-T2I [21] adopt text-based layout construction prior to synthesis, significantly improving output quality. In summary, CoT's simulation of human cognition has advanced multiple domains. However, previous multi-spectral ReID works have relied on static textual descriptions to supplement semantics, overlooking inherent semantic logical relationships. To address this, we propose

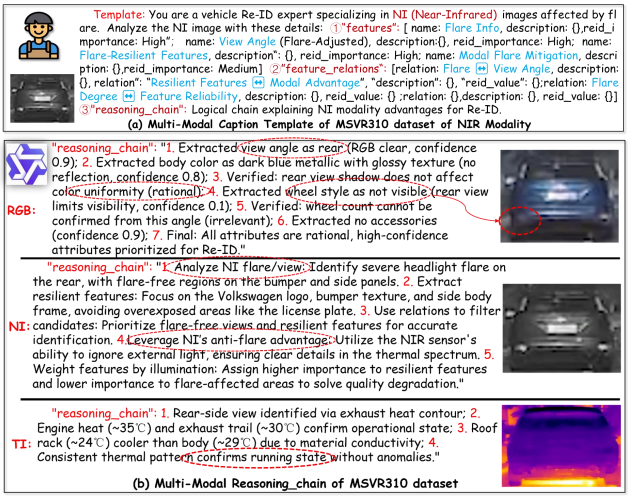

Figure 3. The template example of NIR modality and the multi-modal reasoning-chain of MSVR310 dataset.

a CoT-guided multi-modal object ReID framework that employs hierarchical reasoning through CoT semantic text to guide visual feature learning throughout the entire process.

3. Methodology

In this section, we present the process of CoT-guided reasoning-chains and captions generation, as well as our proposed modules. Fig. 3 illustrates the details of generation, including templates and examples. Fig. 4 presents the key modules of our proposed CoT: CoT-guided Reverse Embedding (CRE), CoT-guided Cross-Modal Consistency (CT-CMC) and CoT-guided Text Features (CTF). Details are described as follows.

3.1. Multi-modal Chain-of-Thought Guided Caption Generation

To obtain multi-modal semantic contextual information, we leverage the CoT of LLMs to generate two types of datasets: multi-modal objects logical reasoning-chains text dataset and objects attribute descriptions datasets. Specifically, as shown in Fig. 3, we enable the MLLM to generate unique semantic information for each modality based on the features of the respective modality, and require it to produce textual information detailing its logical reasoning process.

For multi-modal object ReID, each identity instance consists of three different modalities: visible light (RGB), near-infrared (NIR), and thermal infrared (TIR), denoted $X_i = [X_{rgb}, X_{nir}, X_{tir}]$. And we use $M(,)$ and T_{Pi} denote as MLLM and templates of {rgb, nir, tir}:

$$X_{T,i} = M(X_i, T_{Pi}),\ X_{T',i} = M(X_i, T_{Pi}), \qquad (1)$$

where $X_{T,i}$ and $X_{T',i}$, (i∈ (rgb, nir, tir)) denote as the CoT

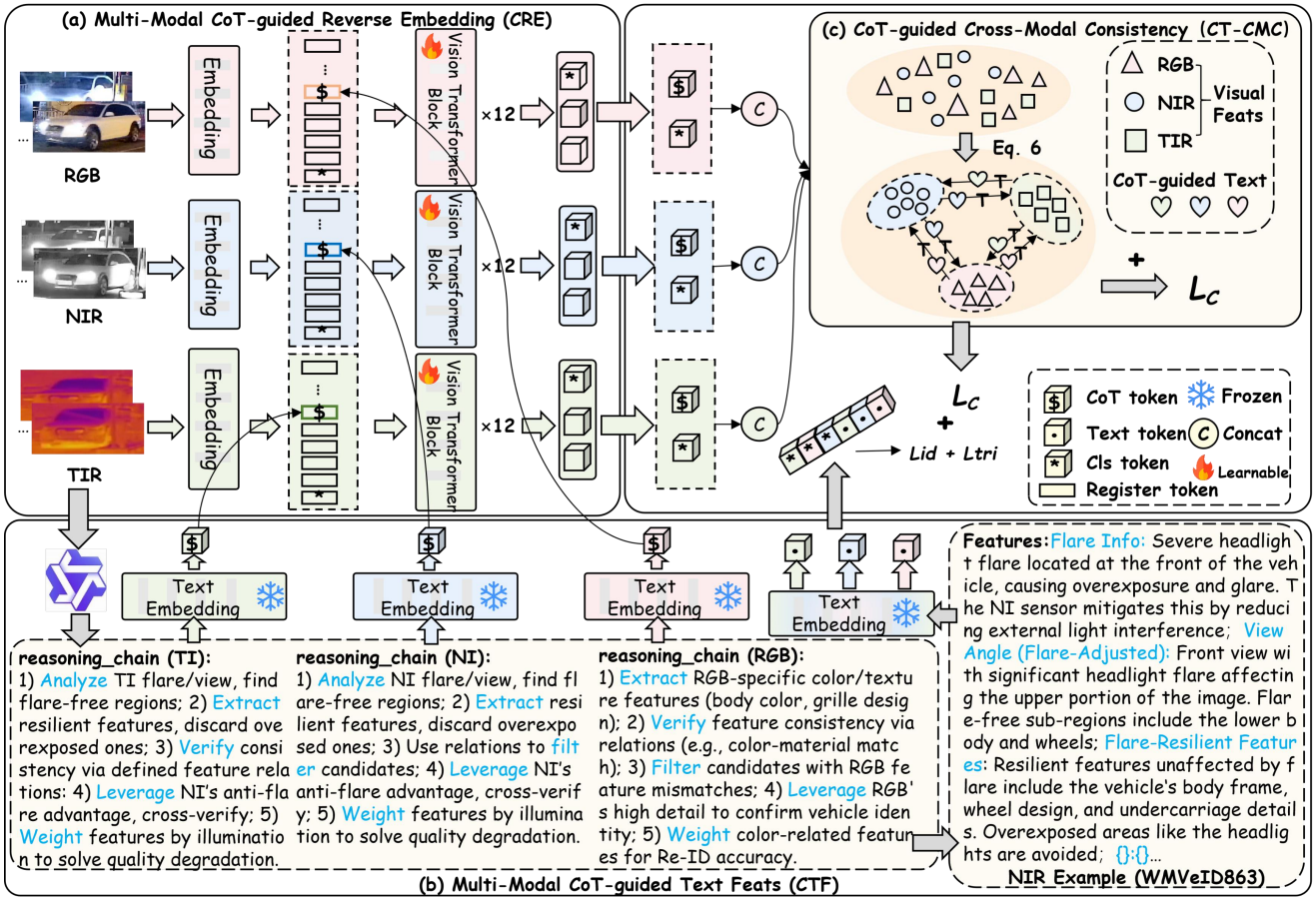

Figure 4. The framework of our CoT-ReID. We first employ MLLMs to generate both descriptive captions and logical reasoning texts based on the CoT process for target objects. The CoT reasoning text is then fed into the CoT-guided Reverse Embedding (CRE) module to guide bottom-level visual feature extraction. Next, the obtained tri-modal visual features are concatenated with their corresponding CoT texts and passed to the CoT-guided Cross-Modal Consistency (CT-CMC), where the text serves as a conditional anchor for bidirectional alignment. Finally, both visual and CoT-guided Features (CTF) serve as decision inputs to the ReID model and are jointly optimized under a unified ReID loss.

reasoning text and the CoT-guided text description of three modalities.

3.2. Multi-modal CoT-guided Reverse Embedding

Thanks to the powerful self-supervised pre-training capability of DINOv3 [23], we adopt it as our visual backbone, whose four register tokens serve as dynamic yet stable carriers for global semantic integration to overcome local patch limitations and enhance contrastive learning reliability. Inspired by this design, we embed the contextual semantic reversal of CoT reasoning into these register tokens, formally expressed as:

$$Q_T^T = \Phi(X_{Tt',i}),\ Q_T^V = Q_T^T \cdot W_{\text{proj}} + b_{\text{proj}}, \tag{2}$$

where Φ denotes the text embedding that is generated by a frozen CLIP [22] text encoder, Q_T^T denotes the CoT reasoning text of three modalities, and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{512\times C_v}$ is the projection weight matrix, $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{C_v}$ is the bias term, and $Q_T^V \in \mathbb{R}^{B\times C_v}$.

The register tokens are concatenated with the visual patch sequence to form the input to Transformer layers:

$$R_{\text{in}}^{(l)} = \left[Z_N^{(l)} \oplus r^{(l)} \oplus Q_T^{V(l)}\right] \in \mathbb{R}^{B\times(1+n+N)\times C_v}, \tag{3}$$

$$F_{\text{out}}^{(l)} = \alpha\left(R_{\text{in}}^{(l)}; \theta_{\text{blk}}^{(l)}\right) \in \mathbb{R}^{B\times(1+n+N)\times C_v}, \tag{4}$$

where $Z_N \in \mathbb{R}^{N\times C_v}$ is the visual patch sequence (N denotes the number of image patches), $\oplus$ represents sequence concatenation, n indicates the number of register tokens (set to 4), l ($1 \le l \le L$) indexes the Transformer layers, and $\alpha(;)$ denotes the l-th Transformer block in DINOv3 parameterized by $\theta_{\text{blk}}^{(l)}$.

3.3. CoT-guided Cross-Modal Consistency

Multi-modal learning needs to leverage the complementary information from different inputs. However, existing methods still face two significant challenges: **(1) ensuring effective cross-modal alignment without compromising modality-specific features, and (2) balancing the contributions of each modality to prevent one modality from dominating the fusion process.** When these representation imbalances are not properly addressed, the fusion process yields sub-optimal outcomes [4, 11, 15].

To achieve effective cross-modal alignment without over-fusion, we introduce semantic context as both a conditioning signal and a protective margin, inspired by InfoBridge [13]. This approach ensures a win-win effect of enhanced information sharing and preserved modality uniqueness.

To clarify the method: our positive pairs (F_i, F_j) refer to pairs of representations from different modalities that belong to the same ID, thus following the joint distribution $p^{ij}_{pos}(F_i, F_j|T)$. Negative pairs (F_i, F_j) are those where the representations come from different modalities and different IDs, which follows the product of marginals $p_{neg}(F_i|T) \cdot p_{neg}(F_j|T)$. In our contrastive learning setup, we denote the number of positive pairs as N_1 and the number of negative pairs as N_0.

Protective Margin Objective with Context. Li et al. [13] provide theoretical contribution, building upon the proven lower bound $I(F_i, F_j|T) \geq log(\frac{N_1}{N_0}) + L_{NCE}(h)$, which provides a principled way to control the degree of cross-modal fusion while maintaining modality-specific information. To prevent excessive fusion between modalities, we introduce a protective margin in our cross-modal consistency design: $\Delta = \log\left(\frac{N_1}{N_0}\right)$.

CoT-guided Semantic Anchor. Given the context condition $Q^{T'}_T$, we define the context-anchored cross-modal consistency loss $\mathcal{L}_C$:

$$\mathcal{L}_{i2j} = -log\left(E_{(F_i,F_j)\sim p^{ij}_{pos}} h(F_i, F_j, Q^{T'}_{T,i})\right) - \frac{N_1}{N_0} \cdot log\left(1 - E_{(F_i,F_j)\sim p^{ij}_{neg}} h(F_i, F_j, Q^{T'}_{T,i})\right), \tag{5}$$

where P^{ij}_{pos} represents positive pairs and P^{ij}_{neg} represents negative pairs, the critic function h implemented by MLPs, estimates the probability ($\in \{0, 1\}$) that a pair (F_i, F_j) is positive given reasoning-chain context $Q^{T'}_{T,i}$. The loss incorporates the protective margin $\frac{N_1}{N_0}$ as mentioned earlier.

To address over-alignment and modal disparity, we impose pairwise bidirectional constraints between the three modalities and the overall loss is:

$$\mathcal{L}_C(F_i, F_j, Q^{T'}_{T,i}) = \sum_{i,j\in(R,N,T)}^{n} \mathcal{L}_{i2j}. \tag{6}$$

In summary, CT-CMC not only promotes cross-modal consistency but also acts as a safeguard against over-fusion, thereby overcoming key limitations like modal dominance and insufficient consensus learning.

3.4. Objective Function

As shown in Fig. 4, we optimize CoT through losses applied to multiple features. To maintain consistency with prior works [29, 30], we separately extract image and text features, denoted by F^T_i and F^V_i, respectively. For each feature, we apply label smoothing cross-entropy loss [24] and triplet loss [9]:

$$\mathcal{L}_g(F) = \mathcal{L}_{CE}(F) + \mathcal{L}_{Tri}(F), \tag{7}$$

where F denotes the input features. Here, $\mathcal{L}_{CE}$ and $\mathcal{L}_{Tri}$ denote label smoothing cross-entropy loss and triplet loss. The overall objective function is then formulated as:

$$\mathcal{L} = \mathcal{L}_g(F^V_i) + \mathcal{L}_g(F^T_i) + \mathcal{L}_C(F^V_i, F^T_i, F^{T'}_i), \tag{8}$$

where F^V_i, F^T_i and $F^{T'}_i$ denote the multi-modal visual, text features and CoT-guided reasoning text, $\mathcal{L}_C$ denotes the loss from CT-CMC.

4. Experiment

4.1. Datasets and Evaluation Protocols

Datasets. We conduct experiments on three publicly available multi-spectral vehicle ReID datasets and a multi-spectral person dataset, including MSVR310 [45], RGBNT100 [14], WMVeID863 [46] and RGBNT201 [43]. To extend these datasets, we employ Qwen-VL [1] to automatically generate CoT-guided reasoning chain and object attribute for each image modality in the train and test sets. Our generated text benchmarks: We use the API-based Qwen-VL [1] to automatically generate textual descriptions. More details can be found in the supplementary material.

Evaluation Protocols. In line with the convention of the ReID community [8, 12], we use mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K (K = 1, 5, 10) to assess performance. The evaluation protocol is consistent with that used in the original dataset and baseline methods.

4.2. Implementation Details

Our model is implemented with the PyTorch toolbox. We conduct experiments on one NVIDIA RTX 3090 GPU. For data processing, we resize the images of each modality to 128x256 in vehicle datasets and 256x128 in person dataset to maintain the aspect ratio. We use random horizontal flipping, padding with 10 pixels and random cropping and erasing[47]. We use the pre-trained DINOv3-B [23] as the

	Methods	Venue	RGBNT100		WMVeID863				MSVR310			
			mAP	R-1	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
Single	BoT [18]	CVPRW'19	78.0	95.1	51.1	55.7	69.8	74.7	23.5	38.4	56.8	64.8
	OSNet [48]	ICCV'19	75.0	95.6	42.9	46.8	61.9	69.4	28.7	44.8	66.2	73.1
	AGW [37]	TPAMI'21	73.1	92.7	30.3	35.3	43.3	46.5	28.9	46.9	64.3	70.7
	TransReID* [8]	ICCV'21	75.6	92.9	67.0	74.7	79.5	82.4	26.9	43.5	62.4	70.7
Multi-modal	HAMNET [14]	CVPR'18	74.5	93.3	45.6	48.5	63.1	68.8	27.1	42.3	61.6	69.5
	PFNet [43]	AAAI'21	68.1	94.1	50.1	55.9	68.7	75.1	23.5	37.4	57.0	67.3
	IEEE [31]	AAAI'22	61.3	87.8	45.9	48.6	64.3	67.9	21.0	41.0	57.7	65.0
	CCNet [45]	INFFU'23	77.2	96.3	50.3	52.7	69.6	75.1	36.4	55.2	72.4	79.7
	TOP-ReID* [27]	AAAI'24	81.2	96.4	67.7	75.3	80.8	83.5	35.9	44.6	-	-
	EDITOR* [40]	CVPR'24	82.1	96.4	65.6	73.8	80.0	82.3	39.0	49.3	-	-
	HTT* [32]	AAAI'24	75.7	92.6	66.2	73.2	79.9	82.3	34.5	43.2	-	-
	FACENet* [46]	INFFU'25	81.5	96.9	69.8	77.0	81.0	84.2	36.2	54.1	-	-
	Mambapro† [28]	AAAI'25	83.9	94.7	69.5	76.9	80.6	83.8	47.0	64.0	-	-
	PromptMA† [41]	TIP'25	85.3	97.4	-	-	-	-	55.2	64.5	-	-
	DeMo† [29]	AAAI'25	86.2	97.6	68.8	77.2	81.5	83.8	49.2	59.8	-	-
	DeMo° [29]	AAAI'25	88.4	96.1	72.5	79.9	85.2	89.2	68.7	81.6	90.7	93.1
	ICPL† [17]	TMM'25	87.0	98.6	67.2	74.0	81.2	85.6	56.9	77.7	87.6	91.5
	IDEA† [30]	CVPR'25	87.2	96.5	-	-	-	-	47.0	62.4	-	-
	IDEA° [30]	CVPR'25	88.2	96.5	-	-	-	-	67.0	82.4	89.0	90.1
	DINOv3° [23]	-	87.0	98.5	70.1	80.8	84.3	87.7	68.2	83.3	93.5	94.4
	CoT°	**Ours**	**89.9**	**99.3**	**74.7**	**82.0**	**85.9**	**89.8**	**71.7**	**85.3**	**94.3**	**96.5**

Table 1. Performance comparison on multi-modal vehicle ReID datasets. The best results are in **bold** and the second in underlined. Symbols: † (CLIP-based), ∗ (ViT-based), ° (DINOv3-based), others (CNN-based). To facilitate a clear comparison, the "DINOv3°" method serves as our baseline.

visual encoder. We use Adam optimizer to optimize the network with the initial learning rate as 3.5×10^{-4} for WMVeID863, 1×10^{-5} for RGBNT100, 3.5×10^{-3} for MSVR310, and 3.5×10^{-3} for RGBNT201. And all of them with a momentum of 0.9 and a weight decays of 1×10^{-4} at total 120 epochs. Other details are provided in the supplementary material.

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

Performance on Multi-modal Vehicle Dataset. Table 1 compares our CoT° with existing multi-modal methods on three vehicle datasets. Our approach achieves significant improvements on WMVeID863 and MSVR310. For fair comparison, we reproduce baseline results using the DINOv3 backbone on DeMo† and IDEA†. Since IDEA† does not provide text for WMVeID863, we exclude its results on this dataset. In the same methodology, the DINOv3° backbone achieves higher performance compared to CLIP†. Specifically, on MSVR310, we improve mAP by **3.0%** and **4.7%** over DeMo° and IDEA°; on WMVeID863, we achieve gains of **2.2%**. These gains come from two factors: the stronger self-supervised initialization of DINOv3 and the additional CoT-based semantic guidance introduced in our framework.

Performance on Multi-modal Person Dataset. Table 2 outlines the performance of our method on RGBNT201 dataset. For a clear comparison, we reproduce the baseline of DINOV3 [23] and IDEA† [30] method on the DINOV3 backbone. On the one hand, it can be observed that the powerful self-supervised pre-training of DINOV3 provides a solid foundation for visual feature extraction. On the other hand, we conduct a fair comparison between our method and IDEA°, our approach achieves improvements of **2.0%/2.9%** in mAP/Rank-1 respectively. Through these experimental comparisons, we can conclude the advantages of our CoT-based reasoning semantic context text, as well as the superiority of our method in guiding visual feature learning throughout the entire ReID pipeline.

4.4. Ablation Studies

We evaluate the effectiveness of the proposed modules on the WMVeID863 dataset and discuss with the IDEA-text of our method on MSVR310 and RGBNT201 dataset due to the data limitation of the original paper.

Effects of Key Modules. Table 5 shows the performance of various combinations of our proposed modules, with Model A as the baseline, and Models B, C, D each add individually the module we designed CRE, CT-CMC, and CTF respectively. All of these achieve improved performance compared to Model A. Model E and F correspond to pairwise combinations of the modules, respectively. Finally, Model G integrates all three modules, enabling the perceptual guidance of CoT semantic text on visual features at the early, late, and decision-making level. The results fully validate

	Methods	RGBNT201			
		mAP	R-1	R-5	R-10
Multi-modal	HAMNet [14]	27.7	26.3	41.5	51.7
	PFNet [43]	38.5	38.9	52.0	58.4
	IEEE [31]	47.5	44.4	57.1	63.6
	DENet [44]	42.4	42.2	55.3	64.5
	LRMM [34]	52.3	53.4	64.6	73.2
	Unicat* [2]	57.0	55.7	-	-
	HTT* [32]	71.1	73.4	83.1	87.3
	TOP-ReID* [27]	72.3	76.6	84.7	89.4
	EDITOR* [40]	66.5	68.3	81.1	88.2
	RSCNet* [38]	68.2	72.5	-	-
	Mambapro† [28]	78.9	83.4	89.8	91.9
	ICPL† [17]	75.1	77.4	84.2	87.9
	DeMo† [29]	79.0	82.3	88.8	92.0
	IDEA† [30]	80.2	82.1	90.0	93.3
	IDEA° [30]	81.3	83.2	91.0	94.4
	DINOv3° [23]	77.5	78.9	85.8	88.9
	CoT°	**83.3**	**86.1**	**93.3**	**94.8**

Table 2. Performance comparison on RGBNT201 dataset. The best results are in **bold** and the second in underlined. Symbols: † (CLIP-based), ∗ (ViT-based), ° (DINOv3-based), others (CNN-based). To facilitate a clear comparison, the "DINOv3°" method serves as our baseline.

Text	Modules			Metric			
	CRE	CT-CMC	CTF	mAP	R-1	R-5	R-10
w/o CoT-guided	×	×	×	70.9	85.1	94.1	96.6
	✓	×	×	71.2	85.5	94.5	96.9
	✓	✓	×	71.8	85.8	94.8	97.0
	✓	✓	✓	**72.7**	**86.3**	**95.3**	**97.5**
w/o CoT-guided	×	×	×	80.1	84.4	92.8	93.7
	✓	×	×	80.9	84.9	92.9	93.9
	✓	✓	×	81.5	85.5	93.0	94.0
	✓	✓	✓	**83.3**	**86.1**	**93.3**	**94.8**

Table 3. Effectiveness of the CoT-guided text of MSVR310 dataset and RGBNT201 dataset: replace the text in our model with the text from the previous method IDEA [30] for comparison. "✓" denotes the text is from our CoT-guided, and "×" denotes not.

Methods	Metric			
	mAP	R-1	R-5	R-10
3M Loss [31]	72.2	80.1	83.2	87.6
CT-CMC	**74.7**	**82.0**	**85.9**	**89.8**

Table 4. Performance comparison of CT-CMC and $3MLoss$ [31] in our method on the WMVeID863 dataset.

the effectiveness and efficiency of the proposed modules.

Effects of Key Components in CT-CMC. Table 6 demonstrates the effectiveness of the components of our CT-CMC module. We conduct ablation experiments on variants with-

Index	Modules			WMVeID863			
	CRE	CT-CMC	CTF	mAP	R-1	R-5	R-10
A	×	×	×	70.1	80.8	84.3	87.7
B	×	×	✓	73.4	80.7	84.2	88.8
C	×	✓	×	72.8	80.0	84.3	88.6
D	✓	×	×	73.3	81.5	84.0	88.7
E	×	✓	✓	73.4	81.4	84.7	88.0
F	✓	✓	×	74.1	81.8	85.0	88.8
G	✓	✓	✓	**74.7**	**82.0**	**85.9**	**89.8**

Table 5. Performance comparison of different modules on WMVeID863 ReID dataset. CRE: CoT-guided Reverse Embedding, CT-CMC: CoT-guided Cross-Modal Consistency, CTF: CoT-guided Text Features.

Index	CT-CMC			Metric			
	CoT	Bi-Cross	Pro.Margin	mAP	R-1	R-5	R-10
A	✓	✓	×	73.4	81.0	85.8	89.3
B	✓	×	✓	74.5	80.9	85.4	88.9
C	×	✓	✓	73.2	80.9	85.4	89.0
D	✓	✓	✓	**74.7**	**82.0**	**85.9**	**89.8**

Table 6. Comparison of different components in CT-CMC of WMVeID863 dataset.

out text semantic protective margin (model A), without bidirectional guidance (model B), and without the constraints of CoT semantic text (model C), respectively. The experimental results show the effectiveness of our CoT semantic text in constraining cross-modal information consistency and the necessity of bidirectional semantic guidance and protection.

Discussion of CT-CMC. Table 10 validates the effectiveness of leveraging CoT-guided reasoning text for cross-modal consistency. To verify this, we conduct experiment on the WMVeID863 dataset while maintaining the same framework but replacing our proposed CT-CMC with 3M Loss [31]. 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 **1.8%/1.2%** compared to our full model. This result demonstrates the superiority of utilizing CoT-generated text as dual constraints to construct cross-modal consistency.

Effects of CoT-guided reasoning text and target captions. Table 3 verifies the superiority and necessity of our CoT-guided reasoning text. To fully demonstrate this, we replace our CoT-guided text with the text from IDEA [30] across the three CoT-guided training levels and conduct comparative experiments on the MSVR310 and RGBNT201 datasets, respectively. We observe that gradual text replacement experiments reveal the superiority of our MLLM-based CoT-guided full-process training method. In addition, as detailed in the supplementary material, we replace the static text of the IDEA [30] method with our CoT-

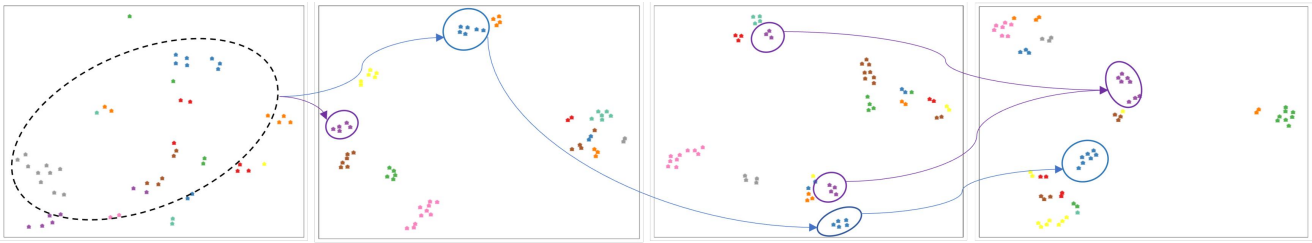

Figure 5. Visualization of the feature distributions with t-SNE [25] on WMVeID863 dataset. Different colors represent different identities.

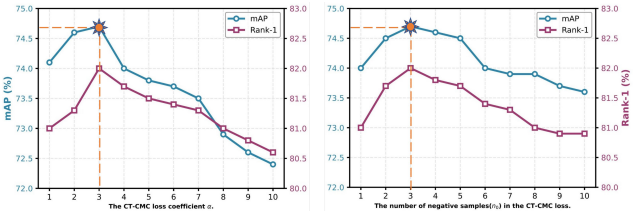

Figure 6. Analysis of hyperparameters of the method: coefficients α of the loss and the number of negative samples n_0.

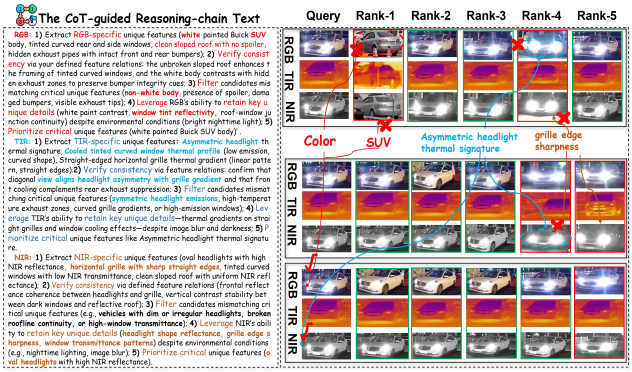

Figure 7. Visualization of the ranklist on WMVeID863 dataset. The left figure illustrates the CoT textual semantic logical reasoning process of the corresponding modality for our query, while the right figure, from top to bottom, shows the rank lists of our baseline, IDEA [30], and our method respectively.

guided text. We conduct comprehensive analysis across two experiments to validate the positive effect of our CoT-guided text.

Hyper-parameters Analysis. The line chart 6 illustrates the impact of the loss coefficient α and negative samples number n_0 of WMVeID863 dataset on the experimental results in CT-CMC. The optimal selection is α = 0.3 and n_0 =3. Hyperparameter analysis on other datasets can be found in the supplementary material.

4.5. Visualization Analysis

Multi-modal Feature Distributions. Fig. 5 visualizes the functional distribution of different modules. Compared with Fig. 5 (a) and (b), the use of CRE makes the feature distribution more discriminative. Finally, in Fig. 5 (c), CT-CMC further enhances feature discrimination. These visualizations demonstrate the effectiveness of our proposed modules.

Rank List Comparison. Fig. 7 presents the cross-camera rank lists of the WMVeID863 dataset. The left column displays the CoT reasoning semantics corresponding to the query, while the right columns show retrieval results from Baseline, IDEA [30], and our CoT-ReID (top to bottom). First, we analyze the generated CoT semantic text, which contains unique features of each specific modality. For instance, when RGB semantics indicate a "white vehicle", our method retrieves only white vehicles in the top-5, unlike Baseline which includes silver ones. Similarly, TIR semantics describing "Asymmetric headlights" lead to perfect top-5 alignment in our results, whereas IDEA selects vehicles with normal headlights. By comparison, we find that external logical semantics help filter mismatched candidates and improve recognition accuracy.

5. Conclusion

In this work, we propose CoT-ReID, a novel framework that leverages CoT reasoning semantics from MLLMs to guide the learning of visual logical semantics, thereby achieving more robust and resilient representations. At the early level, we introduce the CRE module, which embeds semantically reversed CoT logical text into visual features to calibrate low-level representations. Next, the CT-CMC module utilizes CoT logical text as a semantic anchor to constrain cross-modal consistency while preventing over-alignment. Finally, semantically reasoned features are integrated into the decision process, serving as joint conditions alongside visual features. Extensive experiments on four datasets validate the effectiveness of our method. In future work, we will further study which reasoning signals contribute most and how they interact with different modalities.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China(No. 62372003) and the Natural Science Foundation of Anhui Province(No. 2308085Y40).

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 5

[2] Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. Unicat: Crafting a stronger fusion baseline for multimodal re-identification. *arXiv preprint arXiv:2310.18812*, 2023. 7

[3] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578, 2024. 13

[4] Marco Federici, Anjan Dutta, Patrick Forr'e, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. 5

[5] Yingying Feng, Jie Li, Chi Xie, Lei Tan, and Jiayi Ji. Multi-modal object re-identification via sparse mixture-of-experts. In *Forty-second International Conference on Machine Learning*, 2025. 2

[6] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in neural information processing systems*, 33:11309–11321, 2020. 2

[7] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 3

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 1, 5, 6

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[10] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. *Advances in Neural Information Processing Systems*, 37: 117363–117387, 2024. 1, 3

[11] Changhee Lee and Mihaela Van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521. PMLR, 2021. 5

[12] Qiaomei Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2020. 5

[13] Chenxin Li, Yifan Liu, Panwang Pan, Hengyu Liu, Xinyu Liu, Wuyang Li, Cheng Wang, Weihao Yu, Yiyang Lin, and Yixuan Yuan. Infobridge: Balanced multimodal integration through conditional dependency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 393–404, 2025. 5

[14] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11345–11353, 2020. 1, 2, 5, 6, 7, 16

[15] Kunchi Li, Hongyang Chen, Jun Wan, and Shan Yu. Ckdf-v2: effectively alleviating representation shift for continual learning with small memory. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 5

[16] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1405–1413, 2023. 3

[17] Shihao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Icpl-reid: Identity-conditional prompt learning for multi-spectral object re-identification. *arXiv preprint arXiv:2505.17821*, 2025. 3, 6, 7

[18] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 6

[19] Wenjie Luo, Ruocheng Li, Shanshan Zhu, and Julian Perry. Coherent multimodal reasoning with iterative self-evaluation for vision-language models. *arXiv preprint arXiv:2508.02886*, 2025. 1

[20] Prasanna Reddy Pulakurthi, Jiamian Wang, Majid Rabbani, Sohail Dianat, Raghuveer Rao, and Zhiqiang Tao. X-cot: Explainable text-to-video retrieval via llm-based chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31172–31183, 2025. 3

[21] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 4

[23] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4, 5, 6, 7, 14

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 2818–2826, 2016. 5

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[26] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409, 2024. 2

[27] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5758–5766, 2024. 1, 6, 7

[28] Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8150–8158, 2025. 6, 7

[29] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8141–8149, 2025. 2, 5, 6, 7

[30] Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. Idea: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29701–29710, 2025. 1, 3, 5, 6, 7, 8, 13

[31] Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2633–2641, 2022. 2, 6, 7, 15

[32] Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5850–5858, 2024. 1, 6, 7

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 3

[34] Di Wu, Zhihui Liu, Zihan Chen, Shenglong Gan, Kaiwen Tan, Qin Wan, and Yaonan Wang. LRMM: low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *Expert Syst. Appl.*, 263:125716, 2025. 7

[35] Junyi Yao, Yijiang Liu, Zhen Dong, Mingfei Guo, Helan Hu, Kurt Keutzer, Li Du, Daquan Zhou, and Shanghang Zhang. Promptcot: Align prompt distribution via adapted chain-of-thought. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7027–7037, 2024. 3

[36] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023. 2

[37] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1, 6, 16

[38] Zhi Yu, Zhiyong Huang, Mingyang Hou, Jiaming Pei, Yan Yan, Yushi Liu, and Daming Sun. Representation selective coupling via token sparsification for multi-spectral object re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3633–3648, 2024. 7

[39] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6979–6987, 2024. 1, 3

[40] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17117–17126, 2024. 2, 3, 6, 7

[41] Shizhou Zhang, Wenlong Luo, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Prompt-based modality alignment for effective multi-modal object re-identification. *IEEE Transactions on Image Processing*, 2025. 6

[42] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 3

[43] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3529–3537, 2021. 5, 6, 7

[44] Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification. *arXiv preprint arXiv:2305.15762*, 2023. 7

[45] Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100:101901, 2023. 1, 3, 5, 6, 16

[46] Aihua Zheng, Zhiqi Ma, Yongqi Sun, Zi Wang, Chenglong Li, and Jin Tang. Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification. *Information Fusion*, 116:102800, 2025. 5, 6, 16

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 5

[48] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 1, 6