CTRL: CONSISTENCY-AWARE TEXT-DRIVEN REPRESENTATION LEARNING FOR MULTIMODAL VEHICLE RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal vehicle Re-Identification (ReID) aims to retrieve the same vehicle by utilizing complementary information across heterogeneous modalities. Recent studies further generate corresponding vehicle descriptions for images in each modality and leverage these textual semantics to guide visual feature learning and supply identity information. However, existing generation pipelines often emphasize descriptive completeness while overlooking the distinct requirements of text as visual guidance and as identity supplementation. Moreover, prior methods provide limited control over how textual semantics guide visual feature formation during encoding, and lack selective supplementation for each modality based on complementary identity cues retained by other modalities. To address these issues, we propose CTRL, a Consistency-Aware Text-Driven Representation Learning framework. Its text generation pipeline unifies attribute values and models modality-specific visibility to generate descriptions supported by the corresponding images and consistent in attribute values across modalities. Intra-modal Text-guided Asymmetric Propagation (ITAP) uses text to regulate information propagation during visual encoding, limits the information that identity-discriminative features aggregate from regions with weak identity information, and updates these regions in a controlled manner. Cross-modal Text-guided Semantic Completion (CTSC) supplements each modality with cues from other modality descriptions that are selected as receiving sufficient source-visual support while being insufficiently represented in the target text and visual features. Experiments on three multimodal vehicle ReID datasets show that CTRL significantly outperforms existing methods.

1 INTRODUCTION

Vehicle re-identification (Liu et al., 2016; Lou et al., 2019) aims to retrieve the same vehicle images from a gallery captured across different camera views. Traditional methods rely primarily on RGB images and lack robustness to variations in illumination, weather, and occlusion. Multimodal vehicle re-identification (Li et al., 2020; Zheng et al., 2023; 2025) combines RGB, near-infrared (NIR), and thermal infrared (TIR) images to exploit complementary information from different imaging mechanisms and learn more complete, robust vehicle identity representations in complex environments. Existing studies (Wang et al., 2025b; Li et al., 2026b; Su et al., 2026) show that combining multimodal vehicle ReID with vision-language learning can substantially improve model performance. Given the lack of textual annotations for images and the high cost of manual annotation, MLLM-based text generation pipelines (Hu et al., 2024; Han et al., 2024) have received increasing attention. Text generally serves two non-exclusive roles in existing methods, namely guiding visual feature learning as a semantic reference and providing identity information for visual modalities. The former requires intra-modal image-text consistency, with each description supported by its corresponding image; otherwise, it may guide the model toward visual evidence absent from that image. The latter requires cross-modal attribute-value consistency. Modality descriptions can be complementary, but must agree on shared identity attributes to avoid conflicting information.

However, as shown in Figure 1, prior work falls short in both text generation and text-driven representation learning. Existing generation pipelines (Hu et al., 2024; Li et al., 2026b) emphasize descriptive completeness without explicitly enforcing both consistency requirements outlined above,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

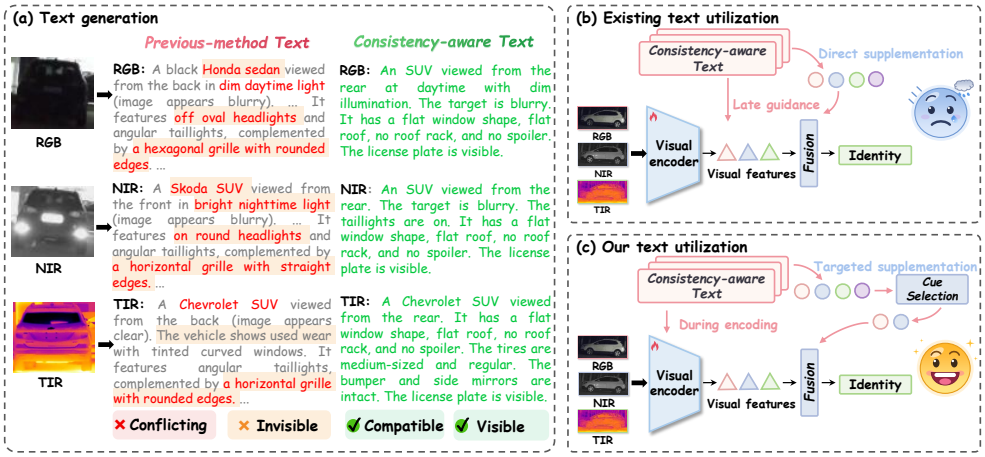


Figure 1. Motivation of CTRL. (a) Red text indicates conflicting descriptions of the same vehicle across modalities, orange shading marks visually unsupported content, and green text denotes consistent, visually grounded descriptions. (b) Limitations of existing text utilization. (c) Our solution.

and representation-learning approaches remain limited in using text in each of these two roles. When serving as a semantic reference, existing methods lack effective mechanisms to use text to regulate information propagation during visual encoding, restricting its influence on visual feature formation. When serving as an identity source, modality descriptions reveal complementary cues across imaging mechanisms, but directly incorporating them into a target representation may introduce redundant semantics without adequately strengthening identity cues underrepresented in that modality.

To address these issues, we propose CTRL, a Consistency-Aware Text-Driven Representation Learning framework for multimodal vehicle ReID. It comprises a consistency-aware text-generation pipeline and two collaborative modules, namely, Intra-modal Text-guided Asymmetric Propagation (ITAP) and Cross-modal Text-guided Semantic Completion (CTSC). The pipeline processes images of the same vehicle from all three modalities with MLLMs to generate candidate attribute values and modality-specific visibility judgments with ordinal confidence levels. It then aggregates these outputs to unify attribute values across modalities and constructs each modality description according to attribute visibility. ITAP uses the corresponding modality description to regulate information propagation during visual encoding and evaluates each patch's identity score to distinguish identity-discriminative patches from weak-identity patches. Through asymmetric propagation, the CLS token and identity-discriminative patches aggregate information solely among themselves, while weak-identity patches absorb score-modulated global context to refine their representations. CTSC aggregates identity semantics from the three modality descriptions and selectively supplements each target visual representation with cues that are estimated to have source visual support but are insufficiently represented in the target text and visual features. Our main contributions are as follows.

- To generate descriptions satisfying intra-modal image-text and cross-modal attribute-value consistency, we propose a text generation pipeline that leverages ordinal confidence levels and modality-specific visibility to harness MLLMs in producing reliable textual descriptions.
- We propose CTRL, a consistency-aware text-driven multimodal representation learning framework with ITAP and CTSC. ITAP suppresses interference from regions with weak identity information during visual encoding, while CTSC selectively uses other modalities' complementary cues to strengthen each modality's underrepresented identity semantics.
- Extensive experiments on three benchmarks show our method's superior effectiveness.

2 Related Work

2.1 Vision-Language Learning in Re-Identification

Large-scale vision-language foundation models such as CLIP (Radford et al., 2021) possess rich visual-semantic knowledge. Building on this, CLIP-ReID (Li et al., 2023) pioneered the adaptation

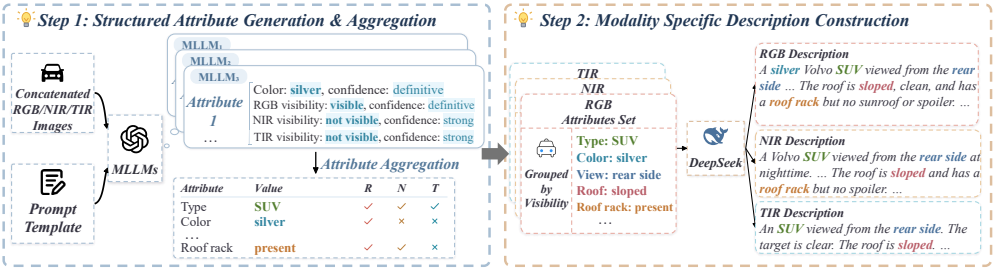


Figure 2. Our consistency-aware text generation pipeline. Multiple MLLMs jointly process the RGB, NIR, and TIR images to generate confidence-aware attribute candidates, aggregate unified attribute values and modality visibility, and construct consistency-aware, modality-specific descriptions using visibility-based selection.

of CLIP to ReID and used identity-specific learnable text prompts to constrain visual representations. PromptSG (Yang et al., 2024) further maps images into pseudo-tokens and inserts them into text templates; neither method produces readable instance-level descriptions. Subsequently, CLIP-SCGI (Han et al., 2024) and TVI-LFM (Hu et al., 2024) use MLLMs to generate image descriptions, though free-form outputs can be redundant and structurally inconsistent. IDEA (Wang et al., 2025b) generates structured attribute descriptions using spectrum-specific prompts and key attribute extraction. NEXT (Li et al., 2026b) further improves description completeness through attribute confidence and cross-modal completion. However, existing text generation methods do not explicitly enforce both intra-modal image-text consistency and cross-modal attribute-value consistency. Consequently, even more complete descriptions may remain unsuitable for downstream ReID use.

2.2 Multimodal Vehicle Re-Identification

Multimodal vehicle ReID exploits complementary multispectral information to improve retrieval reliability in complex environments. HAMNet (Li et al., 2020) pioneered multispectral vehicle ReID through adaptive fusion. GPFNet (He et al., 2023) uses graph structures to model modality relationships and progressively fuse cross-modal information. IDEA (Wang et al., 2025b) maps modality-specific textual semantics into pseudo-image tokens and inserts them into corresponding visual branches, thereby incorporating textual identity information into visual representations. NEXT (Li et al., 2026b) uses text to guide semantic feature sampling. CCL (Su et al., 2026) adaptively refines visual representations with text semantics according to estimated image-text correspondence difficulty. Despite these advances, existing methods provide limited control over how textual semantics regulate layer-wise information propagation during visual encoding. When used as a general identity supplement, text may introduce redundant semantics without adequately strengthening identity cues underrepresented in the target modality.

3 Method

3.1 Consistency-Aware Text Generation Pipeline

Most existing MLLM-based text generation pipelines (Wang et al., 2025b; Hu et al., 2024) for multimodal ReID take a single-modality image as input and generate a description for each modality separately. To improve descriptive completeness, some pipelines (Li et al., 2026b) further use high-confidence attributes from other modalities to fill in missing information. Although this strategy enriches the descriptions, determining attribute values independently for each modality can cause cross-modal conflicts, and direct supplementation may introduce attributes that are not visible in the target image. To address these issues, as shown in Figure 2, we jointly input RGB, NIR, and TIR images of the same vehicle into multiple MLLMs. For each predefined attribute, each MLLM outputs a candidate value and its confidence level, together with visibility judgments for that value in the RGB, NIR, and TIR images and the confidence level of each judgment.

Based on these outputs, we aggregate candidate values by their confidence levels to determine a unified value for each attribute across the three modalities. We then retain only the outputs supporting the selected value and aggregate their modality-specific visibility judgments. Using the resulting visibility estimates, we select attributes judged to be verifiable from each modality image from the

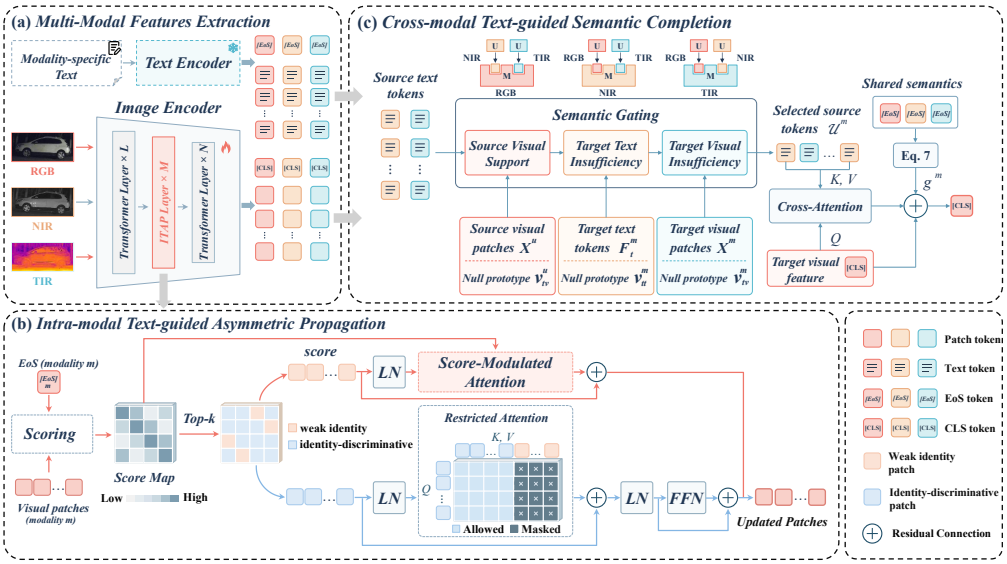


Figure 3. Overview of CTRL. (A) Multimodal feature extraction with a frozen text encoder and an image encoder equipped with ITAP layers. (B) Intra-modal Text-guided Asymmetric Propagation. (C) Cross-modal Text-guided Semantic Completion.

unified attribute set, forming three modality-specific subsets that are subsequently converted into textual descriptions. Since all subsets inherit their attribute values from the same unified set, shared attributes remain consistent across modality descriptions. Meanwhile, visibility-based selection restricts each description to information judged to be verifiable from its corresponding image.

3.2 Intra-modal Text-guided Asymmetric Propagation

Existing methods (Wang et al., 2025b; Li et al., 2026b; Su et al., 2026) do not explicitly use text to regulate layer-wise information propagation between visual regions. During layer-wise visual encoding, image patches and global class token exchange information through self-attention. Interference from regions with weak identity information, such as the background and occlusions, also propagates, entangling discriminative identity information with noise. To this end, as shown in Figure 3, we introduce textual semantics into the visual encoding process to assess patch identity relevance and constrain information propagation, restricting the CLS token and identity-discriminative patches to interact solely among themselves while blocking interference from weak-identity patches.

Text-aware Identity Routing. Let $m \in \{R, N, T\}$ index the RGB, NIR, and TIR modalities. For modality m, let $x_i^{m,l}$ denote the feature of the i-th image patch received by the l-th ITAP layer, where $i = 1, \ldots, n$, and let $c^{m,l}$ denote the CLS token feature. The complete visual sequence is $X^{m,l} = [c^{m,l}; x_1^{m,l}, \ldots, x_n^{m,l}]$. The text encoder's EoS feature is mapped into the visual space and denoted by $\bar{t}^m$. The visual evidence and text relevance of the i-th image patch are given by

$$p_i^{m,l} = \mathrm{MLP}\left(x_i^{m,l}\right), \qquad q_i^{m,l} = \cos\left(x_i^{m,l}, \bar{t}^m\right). \tag{1}$$

Here, $p_i^{m,l}$ represents the visual evidence of the image patch itself, while $q_i^{m,l}$ measures the relevance of the image patch to the identity text. The notation $[\cdot\,;\cdot]$ denotes feature concatenation, and Norm_i denotes within-sample standardization along the image-patch dimension. The final identity routing score is computed as

$$r_i^{m,l} = \mathrm{Sigmoid}\Big(\mathrm{Norm}_i\Big[p_i^{m,l} + \mathrm{MLP}\Big([p_i^{m,l}; q_i^{m,l}]\Big)\Big]\Big). \tag{2}$$

Given a retention ratio κ, we select the $\lfloor \kappa n + 1/2 \rfloor$ image patches with the highest identity scores $r_i^{m,l}$, and denote their index set by $\Omega^{m,l}$. The patches in $\Omega^{m,l}$ are treated as identity-discriminative patches, while the remaining patches are treated as weak-identity patches.

Asymmetric Information Propagation. Based on the routing results, let $\mathcal{A}^{m,l}$ denote the set consisting of the CLS token and the identity-discriminative patches, and let $\mathcal{B}^{m,l}$ denote the set of

weak-identity patches. The two sets are defined as

$$\mathcal{A}^{m,l} = \{c^{m,l}\} \cup \{x_i^{m,l} \mid i \in \Omega^{m,l}\}, \qquad \mathcal{B}^{m,l} = \{x_i^{m,l} \mid i \notin \Omega^{m,l}\}. \tag{3}$$

Weak-identity patches contain both partial identity cues and substantial noise. To improve their identity representations, they are allowed to aggregate global contextual information from the complete visual sequence, with the contributions modulated by the identity routing scores. Let $\mathrm{MHSA}(x_i^{m,l}, X^{m,l}; s^{m,l})$ denote multi-head self-attention that takes patch i as the query and the complete visual sequence as the keys and values. The controlled increment is given by

$$\Delta x_i^{m,l} = r_i^{m,l}\,\mathrm{MHSA}\Big(x_i^{m,l}, X^{m,l}; s^{m,l}\Big), \qquad x_i^{m,l} \in \mathcal{B}^{m,l}. \tag{4}$$

Here, $s^{m,l} = [1, r_1^{m,l}, \ldots, r_n^{m,l}]$ denotes the attention modulation coefficients for the visual tokens. After the standard attention weights are computed, $r_j^{m,l}$ adjusts the contribution of the j-th visual token to the aggregated output, while $r_i^{m,l}$ scales the update magnitude of the current patch. The routing score of the CLS token is fixed to 1, and the modulated attention weights are not renormalized.

For the CLS token and identity-discriminative patches, we restrict their attention scope to the identity-discriminative set $\mathcal{A}^{m,l}$. Let Block_l denote the visual Transformer encoding block used at layer l, and let Q, K, and V denote the query, key, and value, respectively. The CLS token and identity-discriminative patches aggregate information only from $\mathcal{A}^{m,l}$, producing

$$\left\{\widetilde{c}^{m,l+1}, \widetilde{x}_i^{m,l+1}\right\}_{x_i^{m,l} \in \mathcal{A}^{m,l}} = \mathrm{Block}_l\left(Q \leftarrow \mathcal{A}^{m,l},\ K, V \leftarrow \mathcal{A}^{m,l}\right). \tag{5}$$

Since both the keys and values are drawn from $\mathcal{A}^{m,l}$, weak-identity patches cannot propagate information to the CLS token or identity-discriminative patches. Combining this restricted branch with the controlled increments of weak-identity patches, the output of the l-th ITAP layer is

$$c^{m,l+1} = \widetilde{c}^{m,l+1}, \qquad x_i^{m,l+1} = \begin{cases} \widetilde{x}_i^{m,l+1}, & x_i^{m,l} \in \mathcal{A}^{m,l}, \\ x_i^{m,l} + \Delta x_i^{m,l}, & x_i^{m,l} \in \mathcal{B}^{m,l}. \end{cases} \tag{6}$$

Thus, the CLS token and identity-discriminative patches aggregate information only from the identity-discriminative set, while weak-identity patches update their representations using routing-modulated global context, forming asymmetric information propagation. The updated visual sequence is used as the input to the next layer. Subsequent ITAP layers recompute the routing scores, whereas the remaining layers perform standard visual Transformer encoding. After visual encoding, modality m produces the global visual feature c^m and the image-patch feature sequence $X^m = [x_1^m, \ldots, x_n^m]$.

3.3 Cross-modal Text-guided Semantic Completion

Although ITAP suppresses intra-modal interference, it operates within each modality's visual encoding and cannot incorporate identity cues more effectively captured by other modalities. However, directly incorporating other modalities' descriptions may introduce redundant semantics without adequately strengthening identity cues underrepresented in that modality. We therefore formulate cross-modal identity completion as selecting identity semantics from source-modality text that have source visual support but are insufficiently represented in both the target text and target visual features.

Cross-modal Identity Semantic Aggregation. The three modality descriptions retain identity information visible in their corresponding images, including many shared cues. For modality m, let t^m denote the EoS feature produced by its text encoder and $F_t^m = [f_{t,1}^m, \ldots, f_{t,L_m}^m]$ the content token sequence, where L_m is the number of content tokens. Let ρ^m denote the nonnegative cosine similarity between the global visual feature c^m and the corresponding EoS text feature, and let λ denote the strength of shared identity semantic supplementation. We compute

$$\rho^m = \max(0, \cos(c^m, t^m)), \qquad g^m = \lambda \rho^m\,\mathrm{Align}_{c^m}\left(\mathrm{Linear}\left([t^R; t^N; t^T]\right)\right). \tag{7}$$

Here, $\mathrm{Linear}(\cdot)$ denotes a linear mapping, and $\mathrm{Align}_{c^m}(\cdot)$ preserves the direction of its input feature while rescaling its ℓ_2 norm to match that of the global visual feature c^m. This process aggregates overall identity semantics from the three modality descriptions but cannot select modality-specific identity semantics retained exclusively by particular imaging mechanisms.

Source Support and Target Representation Sufficiency Assessment. For target modality m, the other two modalities $u \in \{R, N, T\} \setminus \{m\}$ serve as source modalities. The text sequence F_t^u of source modality u provides candidate semantics for completing the identity representation of target modality m. We assess whether these semantics receive sufficient visual support from the source modality while being insufficiently represented in both the text and visual features of the target modality. To this end, we match each content token in F_t^u against the source image-patch sequence X^u, the target text sequence F_t^m, and the target image-patch sequence X^m, retaining only tokens that satisfy all three conditions.

To unify these matching operations, let z denote the content token feature currently being assessed from the source text sequence F_t^u, and Z the reference sequence, which is X^u, F_t^m, or X^m. Let $h \in \{tt, tv\}$ denote text-text and text-visual matching, respectively, and let $\mathrm{Linear}_{h,Q}$ and $\mathrm{Linear}_{h,K}$ be linear mappings that project the candidate and reference features into the same matching space. Let $\tau_h > 0$ denote the aggregation temperature, y a feature in the reference sequence, and $|Z|$ the sequence length. The matching score is

$$a_h(z, Z) = \tau_h \log \left[\frac{1}{|Z|} \sum_{y \in Z} \exp \left(\frac{\cos\left(\mathrm{Linear}_{h,Q}(z), \mathrm{Linear}_{h,K}(y)\right)}{\tau_h} \right) \right]. \tag{8}$$

To determine whether the reference features sufficiently support the candidate semantics, we introduce a learnable null prototype ν_h^v for each matching type h and reference modality v, serving as a matching reference for the insufficient-support state. Here, v denotes the modality of the reference sequence and takes the value u for source visual matching or m for target text or target visual matching. The cosine similarity between the candidate feature and the null prototype gives the null-match score

$$b_h(z, v) = \cos\left(\mathrm{Linear}_{h,Q}(z), \nu_h^v\right). \tag{9}$$

Let $\gamma_h > 0$ denote the selection temperature and $d_h(z, Z, v) \in \{0, 1\}$ the matching indicator. Hard Gumbel-Softmax (Jang et al., 2017) selects between the match and null-match states

$$[d_h(z, Z, v); 1 - d_h(z, Z, v)] = \mathrm{GumbelSoftmax}_{\mathrm{hard}} \left(\frac{[a_h(z, Z); b_h(z, v)]}{\gamma_h} \right). \tag{10}$$

Here, $d_h = 1$ and $d_h = 0$ indicate the match and null-match states, respectively. For the k-th content token of source modality u ($k = 1, \ldots, L_u$), the indicator $\mu_{m \leftarrow u,k}$ specifies whether the token is selected to supplement the identity representation of target modality m and is given by

$$\mu_{m \leftarrow u,k} = d_{tv}\left(f_{t,k}^u, X^u, u\right) \left[1 - d_{tt}\left(f_{t,k}^u, F_t^m, m\right)\right] \left[1 - d_{tv}\left(f_{t,k}^u, X^m, m\right)\right]. \tag{11}$$

The three factors correspond to source visual support, target text insufficiency, and target visual insufficiency, respectively. Only when all three conditions hold does $\mu_{m \leftarrow u,k} = 1$, allowing the token to contribute to identity semantic completion for the target modality.

Modality-specific Semantic Compensation. The content token features from the two source modalities satisfying $\mu_{m \leftarrow u,k} = 1$ form the candidate sequence $\mathcal{U}^m$. Using the target modality's global visual feature c^m as the query and $\mathcal{U}^m$ as the keys and values, cross-attention aggregates complementary identity cues. These cues are added to the target visual feature together with the shared identity semantic increment, yielding the output $\widetilde{c}^m$

$$\widetilde{c}^m = c^m + g^m + \beta\, \mathrm{CA}\left(Q \leftarrow c^m,\ K, V \leftarrow \mathcal{U}^m\right). \tag{12}$$

Here, CA denotes cross-attention, and β controls the strength of modality-specific supplementation.

3.4 Objective Function

Following prior work, we optimize the model using an identity classification loss and a triplet loss (Hermans et al., 2017). The overall objective is

$$\mathcal{L}_{\mathrm{final}} = \mathcal{L}_{\mathrm{id}} + \mathcal{L}_{\mathrm{tri}}. \tag{13}$$

4 Experiments

4.1 Comparison with State-of-the-Art Methods

We evaluate CTRL on three public multimodal vehicle ReID datasets, MSVR310, RGBNT100, and WMVeID863, in terms of mAP and Rank-K accuracy. Dataset statistics, evaluation protocols,

Table 1. Performance comparison on RGBNT100, WMVeID863, and MSVR310. The highest and second highest results among all methods are shown in **bold** and in underline respectively. † denotes CLIP-based methods, ∗ denotes ViT-based methods, and the others are CNN-based methods.

	Methods	Venue	RGBNT100		WMVeID863		MSVR310			
			mAP	R-1	mAP	R-1	mAP	R-1	R-5	R-10
Single	BoT (Luo et al., 2019)	CVPRW'19	78.0	95.1	51.1	55.7	23.5	38.4	56.8	64.8
	OSNet (Zhou et al., 2019)	ICCV'19	75.0	95.6	42.9	46.8	28.7	44.8	66.2	73.1
	AGW (Ye et al., 2022)	TPAMI'22	73.1	92.7	30.3	35.3	28.9	46.9	64.3	70.7
	TransReID* (He et al., 2021)	ICCV'21	75.6	92.9	67.0	74.7	26.9	43.5	62.4	70.7
Multimodal	HAMNet (Li et al., 2020)	AAAI'20	74.5	93.3	45.6	48.5	27.1	42.3	61.6	69.5
	PFNet (Zheng et al., 2021)	AAAI'21	68.1	94.1	50.1	55.9	23.5	37.4	57.0	67.3
	IEEE (Wang et al., 2022)	AAAI'22	61.3	87.8	45.9	48.6	21.0	41.0	57.7	65.0
	TOP-ReID* (Wang et al., 2024)	AAAI'24	81.2	96.4	67.7	75.3	35.9	44.6	–	–
	DeMo† (Wang et al., 2025a)	AAAI'25	86.2	97.6	68.8	77.2	49.2	59.8	–	–
	IDEA† (Wang et al., 2025b)	CVPR'25	87.2	96.5	–	–	47.0	62.4	–	–
	ICPL-ReID† (Li et al., 2026c)	TMM'26	87.0	**98.6**	67.2	74.0	56.9	77.7	87.6	91.5
	Signal† (Liu et al., 2026)	AAAI'26	86.3	97.6	67.8	76.2	53.6	71.9	–	–
	PEFT-BoA† (Li et al., 2026a)	AAAI'26	85.1	97.8	68.8	75.4	57.4	74.5	–	–
	CCL† (Su et al., 2026)	ICML'26	88.0	98.0	–	–	53.8	69.5	–	–
	DCG† (Zheng et al., 2026)	TCSVT'26	87.6	97.6	71.4	80.1	62.9	77.5	91.7	94.2
	PRISM† (Zhou et al., 2026)	TIP'26	86.1	97.8	70.7	78.1	47.6	64.8	–	–
	NEXT† (Li et al., 2026b)	TIFS'26	88.2	97.7	70.9	77.8	60.8	79.0	89.2	92.2
	CTRL†	**Ours**	**89.2**	98.1	**73.1**	**81.0**	**67.6**	**81.2**	**92.6**	**95.4**

and implementation details are provided in Appendix B. Table 1 compares CTRL with existing single-modal and multimodal vehicle ReID methods on three datasets. CTRL achieves the best mAP on all three datasets and the best Rank-1 accuracy on MSVR310 and WMVeID863. On MSVR310, CTRL ranks first across all four metrics, achieving 67.6% mAP and 81.2% Rank-1 accuracy. It surpasses DCG by 4.7 points in mAP and NEXT by 2.2 points in Rank-1 accuracy. On RGBNT100, CTRL achieves 89.2% mAP and 98.1% Rank-1 accuracy, improving the best reported mAP by 1.0 points while remaining competitive in Rank-1 accuracy. On WMVeID863, it reaches 73.1% mAP and 81.0% Rank-1 accuracy, exceeding the previous best results by 1.7 and 0.9 points, respectively.

4.2 Ablation Studies

To evaluate ITAP and CTSC, we add each module separately to a CLIP-based baseline on MSVR310 and WMVeID863. Table 2 reports the performance and complexity of all variants.

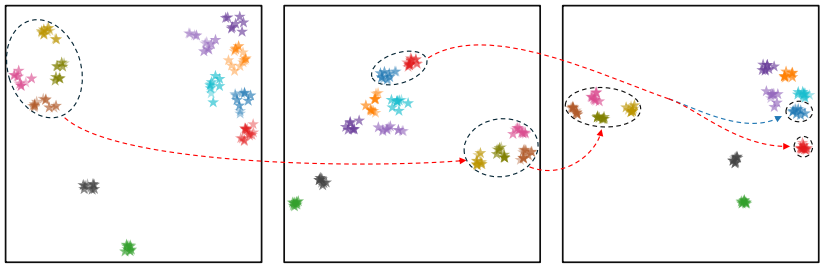
Effects of Key Modules. On MSVR310, ITAP and CTSC individually improve the baseline mAP and Rank-1 accuracy from 53.2% and 70.4% to 58.8% and 73.6%, and to 63.9% and 77.7%, respectively. Their combination reaches 67.6% mAP and 81.2% Rank-1 accuracy, improving the baseline by 14.4 and 10.8 points. On WMVeID863, ITAP and CTSC individually achieve 70.5% and 78.1%, and 71.4% and 79.0%, while their combination reaches 73.1% and 81.0%, improving the baseline by 3.4 and 3.6 points. These results demonstrate the effectiveness and complementarity of ITAP and CTSC.

Model Complexity. The full model increases the parameter count from 86.38 M to 89.86 M and FLOPs from 33.29 G to 35.55 G, corresponding to increases of only 4.0% and 6.8%, respectively. These results show that CTRL achieves substantial performance gains while introducing only modest increases in model size and computational cost relative to the baseline.

Effect of ITAP Routing Components. Table 4 analyzes the effect of text guidance and visual scoring on ITAP routing. Using visual scoring or text guidance alone yields 66.9% and 64.6% mAP, respectively; combining the two raises mAP and Rank-1 accuracy to 67.6% and 81.2%. These results show that visual scoring and text guidance provide complementary cues for ITAP routing, jointly improving the identification of identity-discriminative patches and the resulting retrieval performance.

Table 2. Ablation studies of different modules on vehicle datasets MSVR310 and WMVeID863.

Index	Modules		MSVR310				WMVeID863				Params M	FLOPs G
	ITAP	CTSC	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10		
(a)	×	×	53.2	70.4	84.3	89.2	69.7	77.4	82.2	84.9	86.38	33.29
(b)	✓	×	58.8	73.6	86.6	91.7	70.5	78.1	84.2	86.5	86.97	34.36
(c)	×	✓	63.9	77.7	91.4	94.6	71.4	79.0	84.5	86.7	89.27	34.48
(d)	✓	✓	**67.6**	**81.2**	**92.6**	**95.4**	**73.1**	**81.0**	**85.2**	**88.3**	89.86	35.55



(a) Baseline **(b) Baseline+ITAP** **(c) Baseline+ITAP+CTSC**

Figure 4. t-SNE visualization of feature distributions, with identities coded by distinct colors.

4.3 More Analysis

Feature Distribution. Figure 4 presents a t-SNE (van der Maaten & Hinton, 2008) visualization of the identity features learned by the baseline, Baseline+ITAP, and the full model. Compared with Figure 4(a), introducing ITAP makes features of the same identity more compact in Figure 4(b). Further incorporating CTSC yields clearer separation between different identities in Figure 4(c). As highlighted by the dashed circles, the full model produces more compact and distinct identity clusters, demonstrating the effectiveness of the two modules in learning discriminative representations.

Effectiveness of Description Quality. Table 5 compares different texts in IDEA and CTRL. Compared with IDEA-Text and NEXT-Text, our generated text attains higher mAP and Rank-1 accuracy on both models. CTRL with our text reaches 67.6% mAP and 81.2% Rank-1, exceeding CTRL with NEXT-Text by 5.5 and 5.7 points, respectively. Our generated text has an average length of 56.99 words, essentially the same as IDEA-Text (56.00) and clearly shorter than NEXT-Text (71.36). These results show that our consistency-aware text generation pipeline provides more effective identity semantics while maintaining compact descriptions. The gains on both IDEA and CTRL indicate that this benefit is not tied to a specific model.

Parameter Analysis. Figure 5 shows the joint sensitivity of λ and β on MSVR310 and WMVeID863. On MSVR310, performance remains strong and stable when λ lies in $[0.8, 2]$ and β lies in $[1, 8]$; very small or large λ values and very small β values reduce performance. On WMVeID863, both mAP and Rank-1 remain high for λ in $[0.4, 1]$ and β in $[0.1, 1]$, while large β values reduce performance.

ITAP Activation Layer and retention ratio Analysis. Figure 6 analyzes the effect of the enabled ITAP layer and the retention ratio κ. When the layer is shallow, both mAP and Rank-1 remain low, as the features have not yet formed stable identity semantics and routing decisions lack a reliable basis; performance improves steadily as the layer deepens and peaks at the tenth layer. When κ is small, only a few patches are assigned to the identity-discriminative set, so some patches that carry identity cues are also assigned to the weak-identity set, which hurts performance. As κ increases, performance rises and peaks between 0.75 and 0.85; higher κ degrades results, indicating that an overly wide protected set also admits background, occlusion, and other noisy regions, which weakens the routing constraint.

Effect of CTSC Gating Factors. Table 3 evaluates CTSC's three semantic gating factors on MSVR310 and WMVeID863. Starting from Source Visual Support, adding Target Text Insufficiency and then Target Visual Insufficiency improves mAP and Rank-1 on both datasets. These consistent improvements demonstrate the effectiveness of CTSC's three-factor semantic gating design.

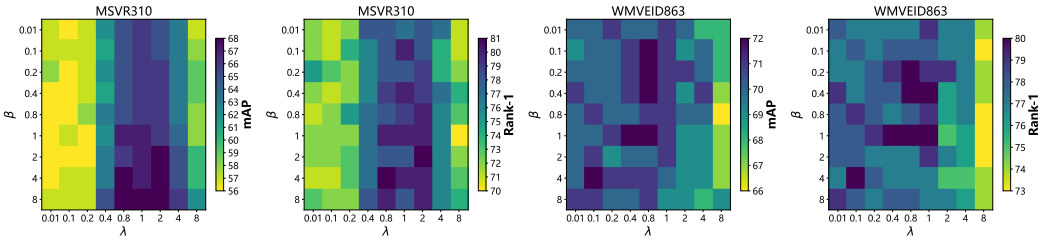


Figure 5. Parameter sensitivity of λ and β on MSVR310 and WMVeID863.

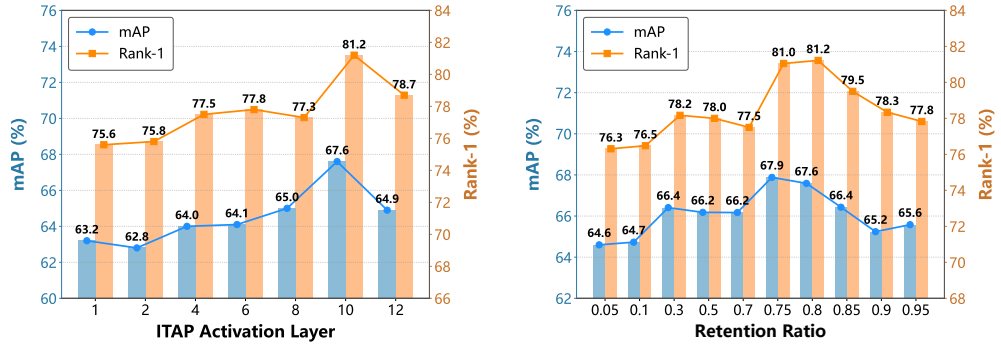


Figure 6. Effects of the ITAP activation layer and retention ratio on MSVR310.

Table 3. Effects of CTSC semantic gating factors on MSVR310 and WMVeID863.

Source Visual Support	Target Text Insufficiency	Target Visual Insufficiency	MSVR310		WMVeID863	
			mAP	R-1	mAP	R-1
✓	×	×	66.6	78.5	70.0	77.6
✓	✓	×	67.3	80.5	71.2	78.6
✓	✓	✓	**67.6**	**81.2**	**73.1**	**81.0**

Table 4. Ablation of ITAP routing components on MSVR310.

Text Scoring	Visual Scoring	mAP	R-1
✓	×	64.6	76.0
×	✓	66.9	78.2
✓	✓	**67.6**	**81.2**

Table 5. Description-type comparison across IDEA and CTRL on MSVR310. Mean words denotes average description length.

Description Type	IDEA		CTRL (Ours)		Mean words
	mAP	R-1	mAP	R-1	
IDEA-Text	47.7	62.7	51.0	68.0	56.00
NEXT-Text	52.0	66.6	62.1	75.5	71.36
Ours-Text	**56.3**	**68.8**	**67.6**	**81.2**	56.99

5 Conclusion

This paper revisits how generated text is used in multimodal vehicle ReID and categorizes its use into two non-exclusive roles, namely guiding visual feature learning and providing identity information. Accordingly, we propose CTRL, a consistency-aware text-driven representation learning framework. CTRL jointly examines images from three modalities through a consistency-aware text generation pipeline, unifies attribute values, and organizes descriptions according to modality visibility to improve intra-modal image-text consistency while preserving cross-modal attribute-value consistency. ITAP uses the corresponding descriptions during visual encoding to prevent regions with weak identity information from interfering with identity-discriminative features, while refining those regions through controlled updates. CTSC selectively supplements each target visual representation with cues estimated to have source visual support but insufficiently represented in the target text and visual features, limiting redundant supplementation and strengthening identity information underrepresented in the target modality. Extensive experiments on three public multimodal vehicle ReID benchmarks demonstrate reliable performance improvements over existing methods.

AI USE STATEMENT

Generative AI tools were used to assist with selected parts of this manuscript. Specifically, they helped translate and polish portions of the text and prepare preliminary drafts for a small number of specific sections. We then revised and polished these drafts iteratively. The tools supported literature retrieval and discovery by helping us search for potentially relevant references and related work. We independently located, read, and verified candidate references before citing or relying on them. They also supported parts of the experimental implementation and code execution under our supervision. Additionally, we applied these tools to the three existing vehicle ReID benchmark datasets. We used MLLMs to generate candidate attribute values and modality-specific visibility judgments from the original images, and used these outputs to produce synthetic textual descriptions. The descriptions were paired with original images and used in our experiments; the source images and identity labels themselves were not generated by AI. We reviewed and revised all AI-assisted drafts, translations, and polished manuscript text, checked the experimental code and settings, examined the generated descriptions, and verified the reported results against the original experimental outputs. We take full responsibility for the methods, results, and final content of this paper.

ETHICS STATEMENT

Our work aims to advance multimodal vehicle re-identification. We acknowledge that vehicle re-identification could also be used for unauthorized cross-camera tracking and may raise privacy concerns. Our experiments are limited to existing public research benchmarks and do not involve real-world deployment. We encourage responsible use of the method in accordance with dataset usage terms and applicable privacy safeguards.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 describes the CTRL architecture and training objectives. Appendix A details the text generation procedure, including attribute value aggregation, modality visibility estimation, and description construction. Appendix B reports the datasets, evaluation protocols, implementation details, and training settings used in our experiments.

REFERENCES

Qianru Han, Xinwei He, Zhi Liu, Sannyuya Liu, Ying Zhang, and Jinhai Xiang. CLIP-SCGI: Synthesized caption-guided inversion for person re-identification. arXiv preprint arXiv:2410.09382, 2024.

Qiaolin He, Zefeng Lu, Zihan Wang, and Haifeng Hu. Graph-based progressive fusion network for multi-modality vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):12431–12447, 2023.

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*, pp. 15013–15022, 2021.

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

Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. In *Advances in Neural Information Processing Systems*, 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

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*, pp. 11345–11353, 2020.

Hongchao Li, Guangxing Liu, Xixi Wang, Baihe Liang, and Yonglong Luo. PEFT-BoA: Parameter-efficient fine-tuning with bag-of-adapters for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6127–6135, 2026a.

Shihao Li, Huaibo Huang, Junxian Duan, Aihua Zheng, Jin Tang, and Jixin Ma. NEXT: Multi-grained mixture of experts via text-modulation for multi-modal object re-identification. *IEEE Transactions on Information Forensics and Security*, 21:8371–8384, 2026b.

Shihao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. ICPL-ReID: Identity-conditional prompt learning for multi-spectral object re-identification. *IEEE Transactions on Multimedia*, 28: 599–613, 2026c.

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*, pp. 1405–1413, 2023.

Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yangyang Liu, Yuhao Wang, and Pingping Zhang. Signal: Selective interaction and global-local alignment for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7359–7367, 2026.

Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Chao Su, Shuying Li, Ruitao Pu, Dezhong Peng, Zhenwen Ren, and Yuan Sun. Correspondence cognitive learning for multi-modal object re-identification. In *International Conference on Machine Learning*, 2026.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

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*, pp. 5758–5766, 2024.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. DeMo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8141–8149, 2025a.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29701–29710, 2025b.

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*, pp. 2633–2641, 2022.

Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17343–17353, 2024.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

Aihua Zheng, Zi Wang, Zi-Han Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3529–3537, 2021.

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.

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.

Aihua Zheng, Ya Gao, Shihao Li, Chenglong Li, and Jin Tang. DCG ReID: Disentangling collaboration and guidance fusion representations for multi-modal vehicle re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13001–13008, 2020.

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*, 2019.

Weixiang Zhou, Jiabei Zuo, Yuhao Wang, Cong Wang, Huchuan Lu, and Zhixun Su. Multi-modal object re-identification with Prompt-S6 and semantic-aware knowledge guidance. *IEEE Transactions on Image Processing*, 35:8306–8321, 2026.

APPENDIX

A TEXT GENERATION PIPELINE DETAILS

This section describes the text generation process in detail. As shown in Figure 7, panel (a) presents the overall pipeline from attribute generation to the three modality descriptions, while panel (b) provides the vehicle attribute template, listing the attributes to be described. We provide the RGB, NIR, and TIR images of the same vehicle, together with this template, to multiple multimodal large language models (MLLMs). Each model examines all three images together and outputs results for the same attribute slots, allowing the outputs of different models to be aligned and aggregated attribute by attribute. We then determine unified attribute values, form a subset for each modality according to attribute visibility, and generate the corresponding descriptions.

A.1 ATTRIBUTE OUTPUTS AND CONFIDENCE

Let M be the number of models and J the number of predefined attributes. Let $e \in \{1, \ldots, M\}$ denote the model index, $j \in \{1, \ldots, J\}$ the attribute index, and $m \in \{R, N, T\}$ the modality, where R, N, and T correspond to RGB, NIR, and TIR. For attribute j, model e provides a candidate value $a_j^{(e)}$ and its confidence $w_j^{(e)}$. For this candidate value, the model also provides a visibility judgment $o_{m,j}^{(e)} \in \{0, 1\}$ and its confidence $w_{m,j}^{(e)}$ for each modality m. The judgment equals one when the image of modality m can independently verify the candidate value and zero when it cannot.

A visibility judgment concerns the complete semantics formed by an attribute and its candidate value. Although each model examines all three images together, its judgment of whether a given modality

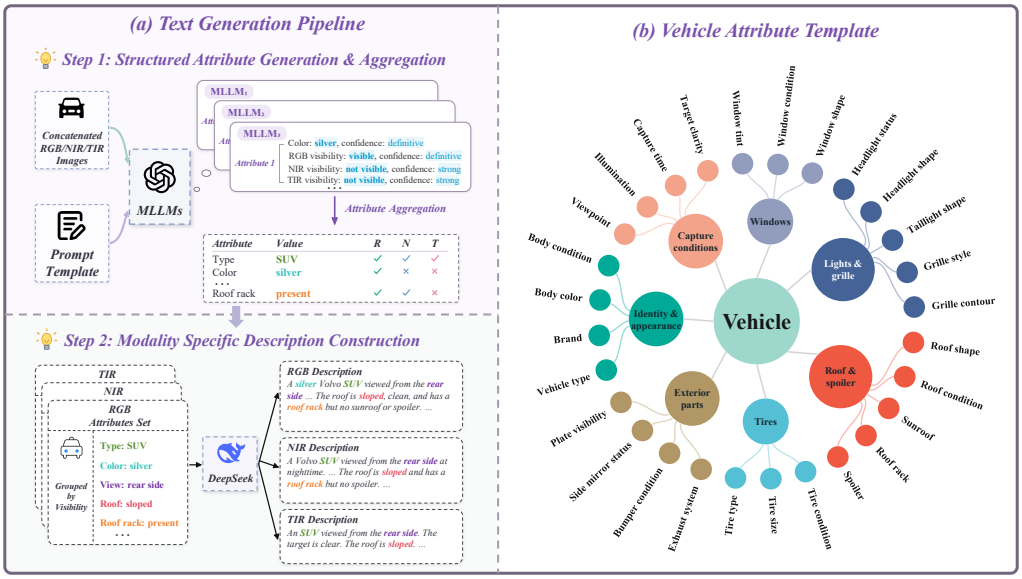


Figure 7. Overview of the consistency-aware text generation pipeline and predefined vehicle attribute template. (a) The figure illustrates a two-stage pipeline in which structured attribute outputs are generated and aggregated before modality-specific descriptions are constructed. (b) The figure depicts the vehicle attribute template, which organizes the attributes to be described into seven groups covering identity and appearance, capture conditions, windows, lights and grille, roof and spoiler, tires, and exterior parts.

can verify the value must rely only on visual information from that modality, without borrowing image evidence from the other modalities. Thus, the same attribute value may be visible in one modality but not in another. Visibility judgments may also differ across samples of the same identity.

Both types of confidence use four levels, `weak`, `moderate`, `strong`, and `definitive`. During aggregation, the mapping ϕ assigns these levels weights of 0.25, 0.50, 0.75, and 1.00, respectively. These weights are used only to combine the judgments of multiple models and do not appear in the final attribute sets or natural-language descriptions.

A.2 Unified Attribute Values and Visibility Aggregation

Different models may provide different candidate values for the same attribute. Let a be any candidate value proposed by the models for attribute j, and let $\mathbf{1}[\cdot]$ denote the indicator function, which equals one when the enclosed condition holds and zero otherwise. The mapping ϕ converts the models' confidence levels into the numerical weights defined above. We denote the confidence-weighted support score for candidate a by $S_j(a)$ and compute it as

$$S_j(a) = \frac{1}{M} \sum_{e=1}^{M} \mathbf{1}[a_j^{(e)} = a]\, \phi(w_j^{(e)}). \tag{14}$$

A model that does not support a contributes zero, while the denominator remains the total number of models, M. Thus, $S_j(a)$ reflects both the number of supporting models and the confidence of their judgments. We select the highest-scoring value among the candidates actually proposed by the models and denote it by a_j, the unified value of attribute j.

$$a_j = \arg\max_a S_j(a). \tag{15}$$

If multiple candidate values share the highest score, we select the lexicographically smallest one. All three modalities use the same a_j rather than determining attribute values independently. This prevents the descriptions of the same vehicle from assigning conflicting values to a shared attribute.

After determining a_j, we aggregate visibility judgments using only models whose candidate value for attribute j equals a_j. We denote the set of their model indices by $\mathcal{I}_j$.

$$\mathcal{I}_j = \{e \mid 1 \le e \le M,\ a_j^{(e)} = a_j\}. \tag{16}$$

For modality m and attribute j, let $o \in \{0, 1\}$ denote the visibility judgment whose support is being evaluated. Here, $o = 1$ means that the modality image can independently verify the unified value a_j, whereas $o = 0$ means that it cannot. We denote by $S_{m,j}(o)$ the confidence-weighted support score contributed by models in $\mathcal{I}_j$ whose visibility judgment is o.

$$S_{m,j}(o) = \frac{1}{M} \sum_{e \in \mathcal{I}_j} \mathbf{1}[o_{m,j}^{(e)} = o]\, \phi(w_{m,j}^{(e)}), \qquad o \in \{0, 1\}. \tag{17}$$

Here, $o_{m,j}^{(e)}$ is model e's judgment, while o is the outcome being evaluated. The indicator includes only models whose judgments equal o, and $\phi(w_{m,j}^{(e)})$ weights them by their confidence in that judgment. The denominator remains M, but only models in $\mathcal{I}_j$ contribute. We compare the two scores to obtain the final visibility $o_{m,j}$ of a_j in modality m.

$$o_{m,j} = \arg \max_{o \in \{0,1\}} S_{m,j}(o). \tag{18}$$

If the two scores are equal, we set $o_{m,j} = 0$. The final visibility always concerns the selected attribute value. A model whose candidate value was not selected contributes no visibility judgment. Applying the calculation to RGB, NIR, and TIR separately yields the selected value's visibility in each modality.

A.3 Modality-Specific Attribute Sets and Description Generation

We denote by $\mathcal{E}$ the single attribute set containing every attribute index paired with its unified value.

$$\mathcal{E} = \{(j, a_j) \mid 1 \le j \le J\}. \tag{19}$$

For each modality m, let $\mathcal{E}_m$ denote the subset containing only attribute-value pairs that can be verified from that modality's image. This subset is determined by the final visibility judgment $o_{m,j}$.

$$\mathcal{E}_m = \{(j, a_j) \in \mathcal{E} \mid o_{m,j} = 1\}, \qquad m \in \{R, N, T\}. \tag{20}$$

The three subsets may share attributes or retain different attributes. Shared attributes can be verified in multiple modalities, whereas the remaining attributes reflect the different visual information available in each modality's image. Regardless of which subsets contain an attribute, its value comes from the unified set $\mathcal{E}$ and does not change across modalities.

Let G be the text generation model, P_m the prompt for modality m, and D_m the resulting description. The model uses $\mathcal{E}_m$ and P_m to generate the description for modality m.

$$D_m = G(\mathcal{E}_m, P_m), \qquad m \in \{R, N, T\}. \tag{21}$$

Generation is instructed to express only attributes in $\mathcal{E}_m$, without changing their selected values or adding information outside the subset. Each modality-specific subset inherits unified attribute values and is filtered according to visibility judgments. Thus, the three descriptions are constructed to agree on shared attributes and to contain only information judged to be verifiable from their corresponding modality images.

B Experimental Settings

B.1 Datasets and Evaluation Protocols

Datasets. We evaluate our method on three public multimodal vehicle ReID datasets, namely MSVR310 (Zheng et al., 2023), RGBNT100 (Li et al., 2020), and WMVeID863 (Zheng et al., 2025). We use Qwen-3.5-Plus and GPT-5 to generate confidence-graded vehicle attributes and modality visibility judgments, and DeepSeek-V4-Flash to produce final modality-specific descriptions. MSVR310 contains 2,087 image triplets from 310 identities, including 1,032 for training, 1,055 for the gallery, and 591 gallery triplets selected as queries. RGBNT100 contains 17,250 image triplets from 100 identities, including 8,675 for training, 8,575 for the gallery, and 1,715 gallery triplets selected as queries. WMVeID863 contains image triplets from 863 vehicle identities captured across eight camera views under diverse illumination conditions. We use 603 identities for training and the remaining 260 for testing to evaluate cross-spectral generalization.

Evaluation Protocols. We use mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at Rank-K, where $K = 1, 5, 10$, to evaluate retrieval performance.

B.2 Implementation Details

The model is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We use a pretrained CLIP (Radford et al., 2021) ViT-B/16 as the visual encoder and its frozen text encoder to extract description features. All input images are resized to 128×256. Random horizontal flipping, padding and cropping, and random erasing (Zhong et al., 2020) are used for data augmentation. The batch size is set to 64 for all three datasets. We sample four image triplets per identity on MSVR310 and WMVeID863 and 16 image triplets per identity on RGBNT100. The model is trained for 50 epochs using the Adam optimizer (Kingma & Ba, 2015). The base learning rate is set to 3.5×10^{-4} for RGBNT100 and WMVeID863 and 5×10^{-6} for MSVR310. The visual encoder is fine-tuned with a learning rate of 5×10^{-6} on all three datasets, while the learning rate of the ITAP router is set to 3.5×10^{-4}. The Adam optimizer uses a weight decay of 1×10^{-4} on all three datasets.

C Additional Experimental Results

C.1 Component Ablations of ITAP and CTSC

Table 6 evaluates Text Scoring and Visual Scoring for ITAP routing on MSVR310 and WMVeID863. Text Scoring measures each patch's relevance to the corresponding modality description, while Visual Scoring assesses evidence from the patch's visual feature. On MSVR310, Text Scoring alone obtains 64.6% mAP and 76.0% Rank-1, Visual Scoring alone obtains 66.9% and 78.2%, and combining them reaches 67.6% and 81.2%. On WMVeID863, the corresponding results are 70.9% and 78.1%, 71.9% and 79.5%, and 73.1% and 81.0%. These results demonstrate that combining Text Scoring and Visual Scoring yields more effective ITAP routing than either component alone on both datasets.

Table 6. Ablation of ITAP routing components on MSVR310 and WMVeID863.

Text Scoring	Visual Scoring	MSVR310		WMVeID863	
		mAP	R-1	mAP	R-1
✓	×	64.6	76.0	70.9	78.1
×	✓	66.9	78.2	71.9	79.5
✓	✓	**67.6**	**81.2**	**73.1**	**81.0**

Table 7. Ablation of CTSC semantic gating factors on MSVR310 and WMVeID863.

Source Visual Support	Target Text Insufficiency	Target Visual Insufficiency	MSVR310		WMVeID863	
			mAP	R-1	mAP	R-1
✓	×	×	66.6	78.5	70.0	77.6
✓	✓	×	67.3	80.5	71.2	78.6
✓	✓	✓	**67.6**	**81.2**	**73.1**	**81.0**

Table 7 evaluates three gating factors for CTSC on both datasets. Source Visual Support assesses whether a candidate source-text token receives sufficient support from source image patches. Target Text Insufficiency and Target Visual Insufficiency assess whether its semantics are insufficiently represented in the target description and target image patches, respectively. On MSVR310, Source Visual Support alone obtains 66.6% mAP and 78.5% Rank-1 accuracy; adding Target Text Insufficiency raises these results to 67.3% and 80.5%, and adding Target Visual Insufficiency further raises them to 67.6% and 81.2%. On WMVeID863, the corresponding results are 70.0% and 77.6%, 71.2% and 78.6%, and 73.1% and 81.0%. The consistent improvements on both datasets demonstrate the effectiveness of CTSC's three-factor semantic gating design, which jointly considers Source Visual Support, Target Text Insufficiency, and Target Visual Insufficiency.

C.2 Effect of Modality-invisible Attribute Injection

We investigate the effect of intra-modal image-text consistency on vehicle ReID by adding attributes judged not to be independently verifiable from the corresponding modality image to the modality

descriptions. Following the definitions in Appendix A, $\mathcal{E}$ is the unified attribute set for a vehicle, $\mathcal{E}_m$ is the subset visible in modality $m \in \{R, N, T\}$, a_j is the unified value of attribute j, and $o_{m,j} \in \{0, 1\}$ indicates whether the image of modality m can independently verify that value. For each sample and target modality m, we form the set of target-modality-invisible attributes for injection from the unified attribute set as

$$\mathcal{H}_m = \{(j, a_j) \in \mathcal{E} \mid o_{m,j} = 0,\ \exists n \in \{R, N, T\} \setminus \{m\},\ o_{n,j} = 1\}. \tag{22}$$

The attributes in $\mathcal{H}_m$ are invisible in the target modality but can be verified from the image of at least one other modality. We retain their unified values a_j when injecting them, rather than predicting new values for the target modality. The intervention thus changes the correspondence between the added attributes and the target image without introducing cross-modal attribute-value conflicts.

For each sample and target modality m, we randomly shuffle the invisible attributes in $\mathcal{H}_m$ once and assign consecutive indices according to the shuffled order. For injection ratios $\eta = 0, 0.5$, and 1, the selected subset $\mathcal{H}_m^{(\eta)}$ contains no attributes, the first 50% of attributes by index, and all attributes, respectively. We add the selected subset to the original visible attribute set, giving

$$\mathcal{E}_m^{(\eta)} = \mathcal{E}_m \cup \mathcal{H}_m^{(\eta)}. \tag{23}$$

The same indexing is used for all three ratios, so the attributes selected at 50% are also included at 100%. We then keep the text generation model G and modality prompt P_m unchanged when generating a description from each attribute set as

$$D_m^{(\eta)} = G\left(\mathcal{E}_m^{(\eta)}, P_m\right). \tag{24}$$

The generation model still expresses only the attributes in its input set and preserves their unified values. We then use the three groups of descriptions to train CTRL separately and compare retrieval results under the same MSVR310 training and evaluation settings. In Table 8, Mean attributes is the average number of input attributes per modality description, and Mean words is the average number of English words in the generated descriptions. Both statistics are computed over the full set of generated modality descriptions.

Table 8. Effects of modality-invisible attribute injection on MSVR310 retrieval and descriptions.

Attribute composition	**mAP**	**R-1**	Mean words	Mean attributes
Visible attributes only (Ours)	**67.6**	**81.2**	56.99	17.84
Visible + 50% invisible attributes	66.9	79.2	64.70	20.80
Visible + 100% invisible attributes	66.0	79.4	70.53	23.42

The three rows of Table 8 correspond to using only visible attributes, adding 50% of the invisible attributes, and adding all invisible attributes, respectively. With visible attributes alone, each description contains an average of 17.84 attributes and 56.99 words, and CTRL achieves 67.6% mAP and 81.2% Rank-1 accuracy. Injecting 50% of the invisible attributes increases the mean numbers of attributes and words to 20.80 and 64.70, while mAP and Rank-1 decrease to 66.9% and 79.2%. Injecting all invisible attributes further increases these means to 23.42 attributes and 70.53 words, while mAP decreases to 66.0% and Rank-1 is 79.4%. Although injecting more invisible attributes produces descriptions with more attributes and words, retrieval performance under both injection settings remains below the visible-only baseline, and mAP declines as the injection ratio increases. This indicates that injecting attributes invisible in the target modality weakens the consistency between its description and image and may lead to poorer retrieval performance, even when attribute values remain consistent across modalities.

C.3 Cosine similarity distribution visualization

Figure 8 shows the cosine-similarity distributions of same-identity (positive) and different-identity (negative) pairs on MSVR310. Dashed lines mark the mean similarities, and Δ denotes their difference. In the first row, the baseline has a mean gap of 0.4312. Adding ITAP or CTSC individually increases it to 0.4381 and 0.4520, respectively, showing that each module improves the mean separation of positive and negative pairs. In the second row, CTRL is trained with different generated

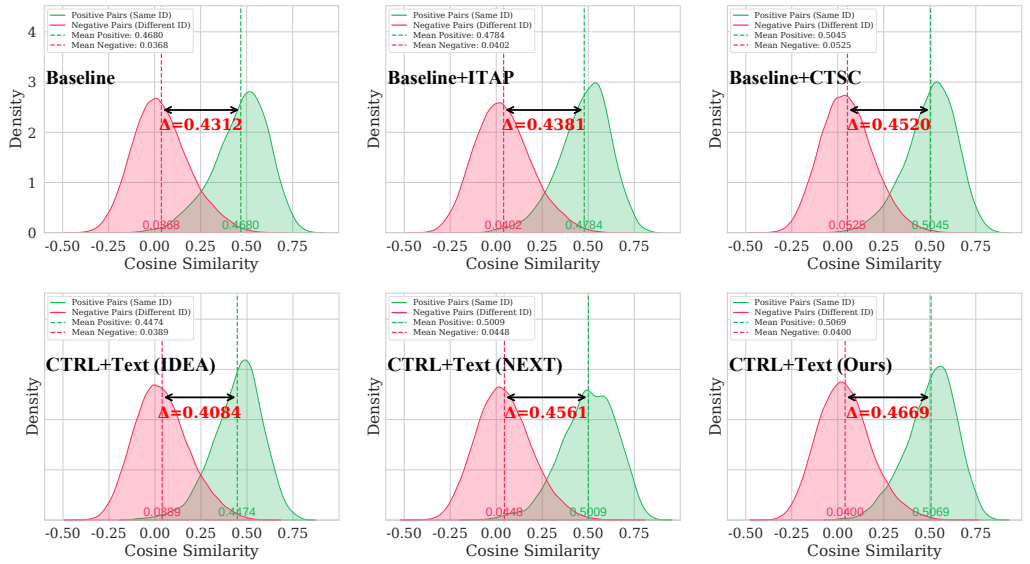


Figure 8. Cosine-similarity distributions on MSVR310. The first row shows the baseline and the ITAP/CTSC ablations; the second row compares CTRL trained with IDEA-Text, NEXT-Text and Our-Text.

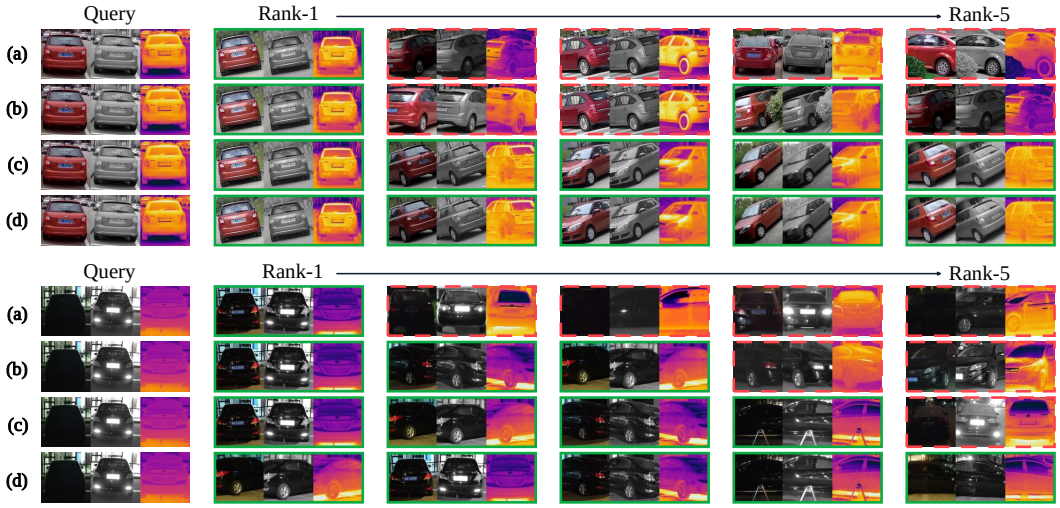


Figure 9. Qualitative results of multimodal vehicle retrieval.

descriptions while its architecture remains fixed. IDEA-Text and NEXT-Text yield gaps of 0.4084 and 0.4561, whereas our text achieves the largest gap of 0.4669. Its positive-pair mean reaches 0.5069 while its negative-pair mean remains at 0.0400. Together, these comparisons show that the proposed modules and our descriptions both contribute to more distinguishable identity representations.

C.4 Retrieval Results for ITAP and CTSC Ablations

Figure 9 presents the top-5 retrieval results for daytime and nighttime queries on MSVR310. Rows (a)–(d) correspond to the baseline, the baseline with ITAP alone, the baseline with CTSC alone, and the model with both modules, respectively, as defined in Table 2. Green and red boxes indicate correct and incorrect matches. The baseline retrieves only one correct top-5 match for each query, while other high-ranked results include vehicles with similar appearances. ITAP alone raises the correct-match counts to two and three for the daytime and nighttime queries, respectively, whereas CTSC alone raises them to five and four. With both modules, all five results are correct for each query. These examples show the individual and combined improvements brought by ITAP and CTSC.

C.5 Retrieval Results with Different Generated Texts

Figures 10 and 11 compare the top-5 retrieval results of IDEA and CTRL using different generated texts. Within each model, the three text settings use the same queries. Green and red boxes indicate correct and incorrect matches, respectively. IDEA-Text produces many incorrect matches for both models. NEXT-Text improves the results, but incorrect matches remain in the top five. In contrast, with our text, all top-5 matches are correct for every query shown. Together with the quantitative results in Table 5, these examples suggest that maintaining cross-modal attribute-value consistency and intra-modal image-text consistency helps produce more discriminative descriptions. These retrieval results further reflect the high quality of our generated text and indicate that it is effective across different models.

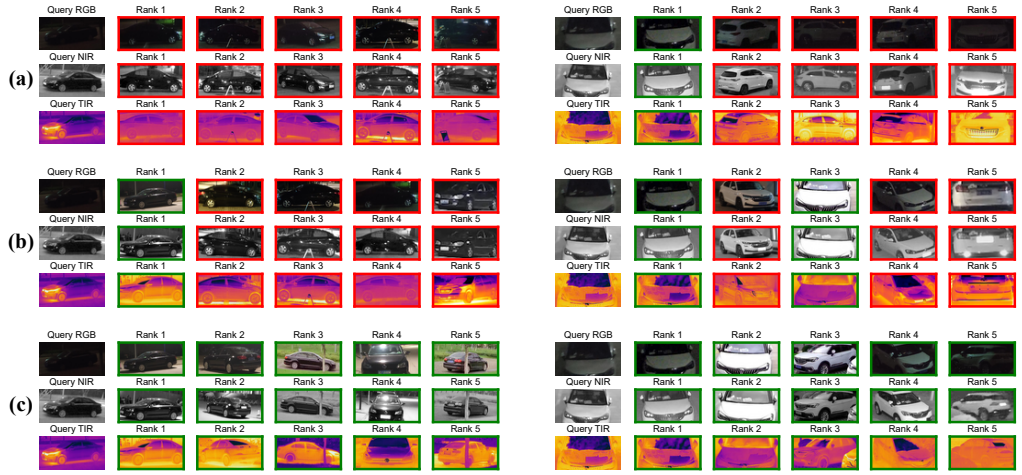


Figure 10. Top-5 retrieval results of IDEA on MSVR310 using (a) IDEA-Text, (b) NEXT-Text, and (c) our text.

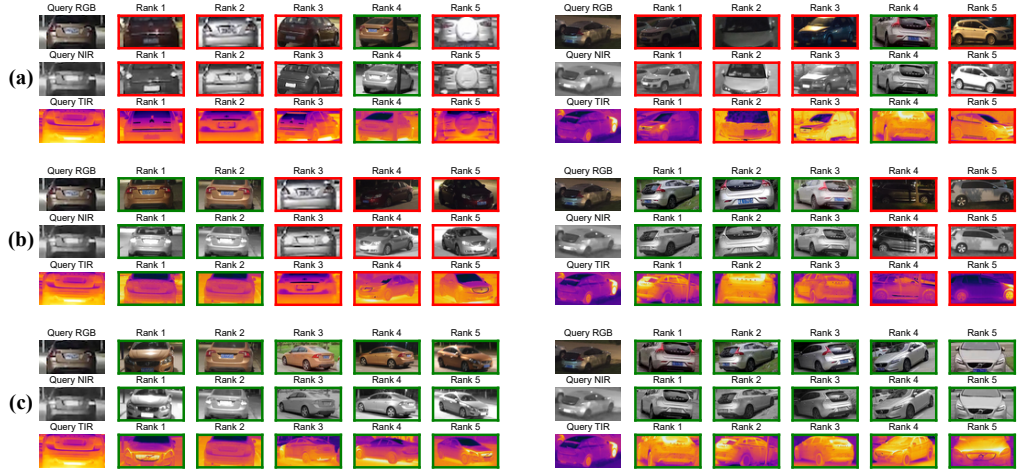


Figure 11. Top-5 retrieval results of CTRL on MSVR310 using (a) IDEA-Text, (b) NEXT-Text, and (c) our text.