MDReID: Modality-Decoupled Learning for Any-to-Any Multi-Modal Object Re-Identification

Yingying Feng[1]*, **Jie Li**[2]*, **Jie Hu**[3], **Yukang Zhang**[2],
Lei Tan[3]†, **Jiayi Ji**[2,3]‡
[1]Northeastern University [2]Xiamen University [3]National University of Singapore
fengyingying@stumail.neu.edu.cn, tanlei@nus.edu.sg

Abstract

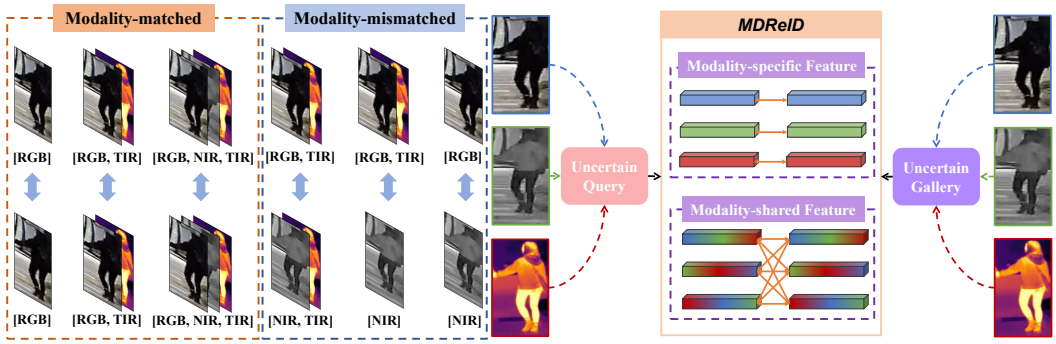
Real-world object re-identification (ReID) systems often face modality inconsistencies, where query and gallery images come from different sensors (e.g., RGB, NIR, TIR). However, most existing methods assume modality-matched conditions, which limits their robustness and scalability in practical applications. To address this challenge, we propose MDReID, a flexible any-to-any image-level ReID framework designed to operate under both modality-matched and modality-mismatched scenarios. MDReID builds on the insight that modality information can be decomposed into two components: modality-shared features that are predictable and transferable, and modality-specific features that capture unique, modality-dependent characteristics. To effectively leverage this, MDReID introduces two key components: the Modality Decoupling Learning (MDL) and Modality-aware Metric Learning (MML). Specifically, MDL explicitly decomposes modality features into modality-shared and modality-specific representations, enabling effective retrieval in both modality-aligned and mismatched scenarios. MML, a tailored metric learning strategy, further enforces orthogonality and complementarity between the two components to enhance discriminative power across modalities. Extensive experiments conducted on three challenging multi-modality ReID benchmarks (RGBNT201, RGBNT100, MSVR310) consistently demonstrate the superiority of MDReID. Notably, MDReID achieves significant mAP improvements of 9.8%, 3.0%, and 11.5% in general modality-matched scenarios, and average gains of 3.4%, 11.8%, and 10.9% in modality-mismatched scenarios, respectively. The code is available at: https://github.com/stone96123/MDReID.

1 Introduction

Object Re-Identification (ReID) focuses on identifying and retrieving specific objects across non-overlapping camera views. Conventional object re-identification (ReID) methods [1, 2, 3, 4, 5, 6, 7] primarily rely on RGB images. However, RGB-based approaches often face significant limitations under challenging environmental conditions, such as poor illumination, shadows, and low image resolution. These adverse conditions frequently lead to the extraction of misleading features, thereby diminishing discriminative capability. To overcome these limitations, multi-modal ReID [8, 9, 10, 11, 12, 13], which integrates complementary information from multiple spectrums, has emerged as a promising approach. By leveraging the inherent strengths and complementary features of different modalities, multi-modal ReID significantly enhances feature representations, enabling more robust and accurate identification in complex real-world scenarios.

*These authors contributed equally to this work.
†Corresponding Author: Lei Tan
‡Project Leader

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Examples of modality-(mis)matched ReID scenarios. (b) The motivation of proposed MDReID.

Figure 1: **Illustration of the motivation of MDReID. (a)** Though the availability of spectral modalities (e.g., RGB, NIR, TIR) varies across queries and galleries, recent methods only focus on the modality-matched scenarios, which limits their practical applicability. **(b)** MDReID overcomes the rigidity of modality constraints by disentangling modality-shared and modality-specific features, enabling effective matching between queries and galleries from arbitrary modalities.

The introduction of multi-modal data has significantly enhanced the performance of ReID in challenging scenarios, thereby stimulating extensive research efforts in this field [8, 9, 10, 11, 12]. EDITOR [11] selects diverse tokens from Vision Transformer (ViT) to mitigate the impact of irrelevant backgrounds and narrow the gap between modalities, achieving strong detection performance. TOP-ReID [12] incorporates a cyclic token permutation module to aggregate multi-spectral features and a complementary reconstruction module to minimize the distribution gap across different image spectra. These innovations enable TOP-ReID to achieve robust performance under both modality-complete and modality-missing scenarios. Although these methods effectively leverage the characteristics of multi-modal data, they are developed based on a fundamental assumption that all modalities within the dataset are strictly aligned (as illustrated in Figure 1 (a) Modality-matched). Ideally, every camera would simultaneously provide RGB, NIR, and TIR modalities. However, due to practical constraints such as device differences and deployment diversity, achieving fully matched retrieval conditions is challenging in real-world scenarios. Therefore, it is crucial to develop a flexible image-level any-to-any ReID framework capable of effectively operating under both modality-matched and modality-mismatched scenarios (as illustrated in Figure 1 (a)).

To address this limitation, we propose MDReID, a flexible any-to-any ReID framework designed to support retrieval tasks involving arbitrary combinations of query and gallery modalities. The core challenge of this problem lies in learning effective representations that remain robust under both modality-matched and modality-mismatched conditions. A straightforward solution, as adopted by TOP-ReID [12], attempts to predict the missing modality representation from the available one, thereby enabling modality-aligned retrieval. However, as highlighted by RLE [14], predicting cross-spectrum features solely from visual inputs constitutes an ill-posed problem. The modality-specific characteristics that are inherently unpredictable often lead to suboptimal learning. To this end, MDReID introduces a Modality-Decoupled Learning (MDL), which introduces a modality-shared and modality-specific token into the ViT architecture, leveraging multi-layer attention within the transformer to extract shared and specific features from multi-modal inputs. MDL explicitly disentangles the modality representations into two complementary components: a modality-shared component that captures the predictable cross-modal representation and a modality-specific component that preserves the unpredictable modality characteristics. This decoupling enables the model to better generalize across diverse modality combinations while maintaining the modality-specific cues.

Furthermore, to enhance the disentanglement between the two components, we propose a Modality-aware Metric Learning (MML) strategy. MML comprises two complementary objectives: a representation orthogonality loss (ROL) and a knowledge discrepancy loss (KDL). ROL is applied at the channel level to both promote the aggregation of modality-shared features across modalities and enforce orthogonality between shared and specific components, ensuring they capture distinct and non-overlapping information. KDL, on the other hand, encourages representational complementarity between shared and specific components by ensuring that the combined representation is more discriminative than either component alone.

To sum up, the main contributions of this paper are as follows:

- We propose MDReID, a flexible any-to-any object re-identification framework that supports retrieval across arbitrary query-gallery modality combinations, addressing the practical limitations of strictly aligned multi-modal datasets.
- We introduce a Modality-Decoupled Learning (MDL) strategy, coupled with a Modality-aware Metric Learning (MML) strategy. MDL explicitly disentangles modality-shared and modality-specific representations, while MML further strengthens this disentanglement by introducing a representation orthogonality loss and a knowledge discrepancy loss, encouraging the two components to encode distinct and complementary information.
- Extensive experiments on multi-spectral object ReID datasets (RGBNT201, RGBNT100, MSVR310) demonstrate the superior adaptability and performance of the MDReID across diverse scenarios. MDReID achieves significant mAP improvements of 9.8%, 3.0%, and 11.5% in general modality-matched scenarios, and average gains of 3.4%, 11.8%, and 10.9% in modality-mismatched scenarios, respectively.

2 Related Work

Compared to the general RGB-to-RGB object re-identification [15, 16, 17, 18, 19, 20, 21], multi-modal object re-identification (ReID) has demonstrated significant practical value, achieving widespread applications in cross-scenario systems [22, 23] (e.g., security surveillance, intelligent transportation, etc.). Current methods [24, 9, 25] primarily enhance re-identification accuracy by effectively leveraging the complementarity and integration of multi-modal features. For example, Zheng et al. [26] proposed the cross-directional consistency network (CCNet), which overcomes discrepancies in both modality and sample aspects, thereby achieving more effective multi-modal data collaboration. To address the modal-missing problem, Zheng et al. [27] proposed the DENet model, which incorporates a feature transformation module to recover information from missing modalities and a dynamic enhancement module to improve multi-modality representation. Li et al. [8] proposed HAMNet, which automatically fuses different spectral features using a specially designed heterogeneous score coherence loss. He et al. [28] proposed the GPFNet model, which utilizes graph learning to fuse multi-modal features.

Notably, models based on Vision Transformer (ViT) [29, 30, 10, 31, 32, 33] for object re-identification have achieved excellent results in recent years and gained widespread attention. Pan et al. [34] proposed the H-ViT model, which alleviates challenges caused by heterogeneous multi-modalities and reduces feature deviations arising from modal variations. Additionally, they introduced a progressively hybrid transformer (PHT [35]) that effectively fuses multi-modal complementary information through random hybrid augmentation and a feature hybrid mechanism. Crawford et al. [36] addressed the issue of modality laziness in multi-modal fusion by proposing the UniCat model based on a ViT architecture. By selecting diverse tokens from ViT to mitigate the impact of irrelevant backgrounds and narrow the gap between modalities, Zhang et al. [11] developed the EDITOR model. Wang et al. [12] designed the Top-ReID model, which achieves state-of-the-art accuracy by incorporating a cyclic token permutation module to aggregate multi-spectral features and a complementary reconstruction module to minimize the distribution gap across different image spectra.

Despite advancements in multi-modal object ReID, existing methods rely on modality-aligned conditions, limiting their adaptability to real-world scenarios. To address this, we introduce MDReID, which disentangles modality-shared and modality-specific features for effective retrieval across both matched and mismatched conditions. It incorporates Modality-Decoupled Learning (MDL) to refine feature separation and Modality-aware Metric Learning (MML) to enhance discrimination through orthogonality and knowledge discrepancy losses. As a result, MDReID enables robust object re-identification across diverse modality configurations.

3 Methodology

3.1 MDReID: Any-to-any Object ReID

In ideal multispectral person re-identification (ReID) settings, complete RGB, NIR, and TIR modalities are available for both query and gallery (RNT-to-RNT). However, real-world deployments often suffer from different modalities due to heterogeneous sensor availability and deployment

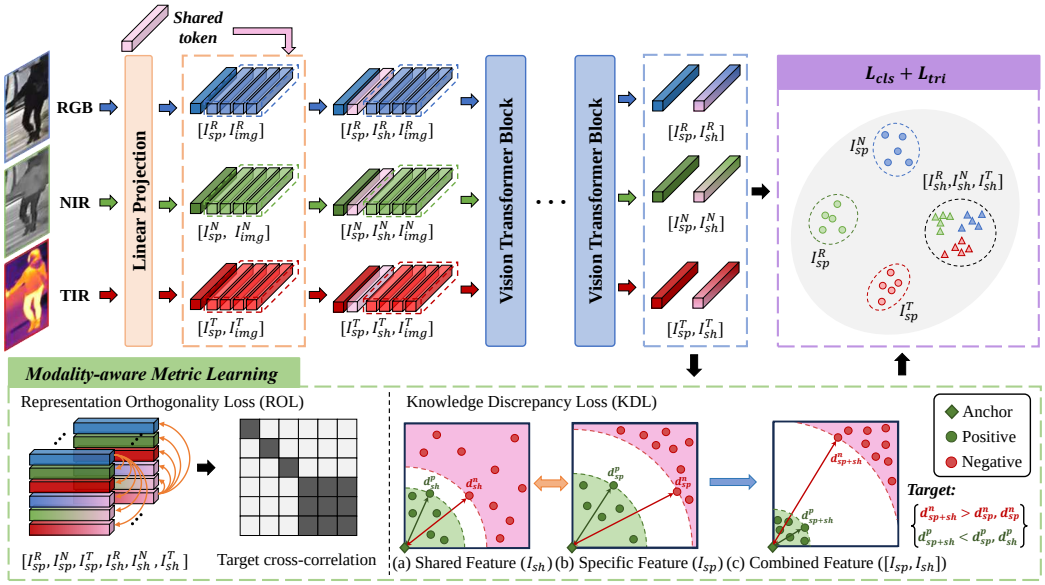


Figure 2: **Overall framework of MDReID.** MDReID is designed to support retrieval across arbitrary modality combinations. It disentangles features into shared and specific components to boost performance in both matched and mismatched scenarios. Additionally, by leveraging representation orthogonality loss (ROL) and knowledge discrepancy loss (KDL), MDReID refines feature separation and enhances retrieval robustness.

constraints. To address this fundamental challenge, we propose MDReID, a flexible, any-to-any ReID framework that supports arbitrary combinations of input and output modalities. As illustrated in Fig. 2, MDReID adopts a Vision Transformer (ViT) backbone and integrates two core components: Modality-Decoupled Learning (MDL) and Modality-Aware Metric Learning (MML). MDL explicitly disentangles each modality's representation into modality-shared features (crucial for cross-modality retrieval) and modality-specific features (essential for preserving discriminative cues in matched scenarios). MML further enhances this decoupling by enforcing metric-level consistency across modalities. This design enables robust and flexible object re-identification in diverse and modality-incomplete environments, significantly advancing the applicability of ReID in practical settings.

3.2 Modality Decoupled Learning

Generally, with a ViT as the backbone, we process input images from different modalities ($M \in \{R, N, T\}$, corresponding to RGB, NIR, and TIR). Standard ViT procedure involves segmenting the image into non-overlapping patches, which are then linearly projected into patch embeddings as:

$$I^M = \{[I^M_{\text{cls}}], I^M_1, I^M_2, \ldots, I^M_n\}, \quad \text{where } M \in \{R, N, T\}. \tag{1}$$

To explicitly model and decouple modality characteristics crucial for cross-modal ReID, we deviate from the standard use of a single [CLS] token. Instead, we prepend two distinct learnable tokens to the patch embedding sequence for each modality: a modality-specific token and a modality-shared token. This augmented sequence is then processed by the ViT encoder to get:

$$I^M_e = \{[I^M_{\text{sp}}, I^M_{\text{sh}}], I^M_1, I^M_2, \ldots, I^M_n\}, \quad \text{where } M \in \{R, N, T\}. \tag{2}$$

where I^M_{sp} denotes the modality-specific features, I^M_{sh} denotes the modality-shared features, and I^M_e represents the encoded features.

Leveraging these decoupled features from different modalities (I^R_{sp},I^N_{sp},I^T_{sp},I^R_{sh},I^N_{sh},I^T_{sh}), we construct a unified representation for each sample using a fixed-dimensional structure and an associated availability mask. This allows consistent handling across different modality presence scenarios. Specifically, we define a potential full feature vector, v_{full}, by concatenating all possible specific and shared components, assuming all three modalities (RGB, NIR, TIR) are present:

$$v_{full} = \left[I^R_{\text{sp}}, I^N_{\text{sp}}, I^T_{\text{sp}}, I^R_{\text{sh}}, I^N_{\text{sh}}, I^T_{\text{sh}}\right]. \tag{3}$$

Correspondingly, a binary availability mask, $mask_{full}$, indicates the presence of each component:

$$mask_{full} = [1, 1, 1, 1, 1, 1], \tag{4}$$

where '1' signifies that the feature component at that position is present and valid, and '0' (as introduced below) signifies absence or invalidity.

The actual feature vector v and mask $mask$ used for any given input sample are derived from this full structure based on the modalities available for that specific sample. If a modality $M \in \{R, N, T\}$ is missing for a sample, the corresponding modality-specific feature I_{sp}^{M} and the modality-shared feature I_{sh}^{M} in the vector v are replaced with a zero vector $\mathbf{0}$. Also, the entries in the mask $mask$ corresponding to these zeroed-out features (I_{sp}^{M} and I_{sh}^{M}) are set to 0.

For example, if the TIR modality (T) is missing for a sample, its feature vector v and mask $mask$ would be:

$$v = \left[I_{\text{sp}}^{R}, I_{\text{sp}}^{N}, \mathbf{0}, I_{\text{sh}}^{R}, I_{\text{sh}}^{N}, \mathbf{0}\right], mask = [1, 1, 0, 1, 1, 0]. \tag{5}$$

This structured representation (v and $mask$) is used consistently across all scenarios, including both modality-matched and modality-mismatched retrieval. In modality-matched scenarios (*e.g.*, RN-to-RN), both the query and gallery samples will have the same modalities present (or absent), resulting in identical structures for their masks (*e.g.*, [1, 1, 0, 1, 1, 0] for both if T is missing). In modality-mismatched scenarios (*e.g.*, R-to-N), the query and gallery samples will inherently have different available modalities. Their respective v and $mask$ representations are generated according to the rules above based on their own available data. For instance, the RGB query sample would have $v_q = [I_{\text{sp}}^{R}, \mathbf{0}, \mathbf{0}, I_{\text{sh}}^{R}, \mathbf{0}, \mathbf{0}]$ and $mask_q = [1, 0, 0, 1, 0, 0]$. And the NIR gallery sample would have $v_g = [\mathbf{0}, I_{\text{sp}}^{N}, \mathbf{0}, \mathbf{0}, I_{\text{sh}}^{N}, \mathbf{0}]$ and $mask_g = [0, 1, 0, 0, 1, 0]$.

The subsequent retrieval process utilizes both the feature vectors (v_q, v_g) and their corresponding availability masks ($mask_q$, $mask_g$) to compute a similarity score that respects the decoupled nature of the features and handles missing modalities robustly. We define the total similarity, $Sim(v_q, v_g)$, as the combine of two components, *i.e.*, the modality-specific similarity Sim_{sp} and the modality-shared similarity Sim_{sh}. The modality-specific similarity compares modality-specific features between identical modalities only. The similarity is calculated as the sum of dot products between corresponding specific features, masked for joint availability:

$$Sim_{sp} = \frac{1}{sum(mask_{sp,q}) * sum(mask_{sp,g})} \sum_{M \in \{R,N,T\}} \left((I_{sp,q}^{M})^{T}(I_{sp,g}^{M})\right) \cdot mask_{sp,q}^{M} \cdot mask_{sp,g}^{M}. \tag{6}$$

The modality-shared similarity computes the similarity between all pairs of available shared features across modalities. Define $V_{sh,q}$ be a matrix with shared features $I_{sh,q}^{R}$, $I_{sh,q}^{N}$, $I_{sh,q}^{T}$ (or zero vectors if unavailable) as columns, and similarly for $V_{sh,g}$, $mask_{sh,q}$ and $mask_{sh,g}$. We first compute the matrix of all pairwise shared feature similarities:

$$S_{sh} = V_{sh,q}^{T} V_{sh,g}. \tag{7}$$

This results in a 3×3 matrix where $(S_{sh})_{ij}$ is the dot product $(I_{sh,q}^{i})^{T}(I_{sh,g}^{j})$ for $i, j \in \{R, N, T\}$. To account for missing modalities, we define a pairwise availability mask matrix with the outer product of the shared masks:

$$M_{sh\_pair} = mask_{sh,q}^{T} \cdot mask_{sh,g}, \tag{8}$$

where $(M_{sh\_pair})_{ij}$ is 1 if and only if both the i-th shared feature of the query and the j-th shared feature of the gallery are present. The total shared similarity is the sum of all valid pairwise similarities, obtained by Hadamard product of S_{sh} and M_{sh_pair}:

$$Sim_{sh} = \frac{1}{sum(M_{sh\_pair})} \sum_{i,j} (S_{sh} \odot M_{sh\_pair})ij. \tag{9}$$

Finally, the total similarity score is:

$$Sim_{total}(v_q, v_g) = (Sim_{sp} + Sim_{sh})/2. \tag{10}$$

This approach, based on a fixed-size zero-padded feature vector and an explicit availability mask, provides a flexible and robust mechanism for handling arbitrary combinations of modality presence and absence, effectively adapting to all modality-matched and mismatched object ReID scenarios addressed by our method.

3.3 Modality-aware Metric Learning

After MDL, the features are separated into modality-specific and modality-shared components. To achieve more complete and thorough feature decoupling, we propose a modality-aware metric learning approach. Note that, during the training phase, we assume all the modalities for the samples are available. The core objective of this metric learning is twofold: 1) **Enhance Shared Feature Consistency**: Promote high similarity between modality-shared features irrespective of their originating modality, which is crucial to bridge the modality gap in mismatched retrieval scenarios. 2) **Reinforce Specific Feature Purity**: Ensure modality-specific features are distinct from each other and also orthogonal to all modality-shared features. This preserves unique modality characteristics and prevents leakage between specific and shared representations.

To achieve these goals simultaneously, we define a representation orthogonality loss, L_{ROL} , that minimizes the discrepancy between the computed pairwise similarities of the decoupled features and a predefined target similarity structure. Within our framework, we operate on the feature vector v and its availability mask $mask$ generated for each sample, as defined previously. We compute the 6×6 pairwise similarity matrix $Vsim$ based on the L_2-normalized vector v:

$$V_{sim}(i,j) = v_i^T v_j, \quad \text{for } i,j \in \{1,\dots,6\}. \tag{11}$$

We then define the ideal target similarity matrix as:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{12}$$

Finally, the loss L_{ROL} is then formulated as the sum of squared errors between the computed similarities V_{sim} and the target similarities A, masked by M_{pair}:

$$L_{ROL} = \sum_{i=1}^{6}\sum_{j=1}^{6} \left((V_{sim}(i,j) - A(i,j))^2 \right). \tag{13}$$

Additionally, the concept of triplet loss is incorporated to further enhance the model's feature decoupling capability. In other words, the goal of feature decoupling is to achieve a synergistic effect and prevent information from collapsing into modality-specific representations. To this end, as shown in Fig. 2, we introduce a knowledge discrepancy loss (KDL) to enforce the complementarity between shared and modality-specific features, ensuring that their combination yields better retrieval performance than using either alone. Specifically, within a batch of samples, given an anchor a, the positive samples P_a denotes the samples with the same label as a while the negative samples N_a denotes the opposite samples. The distances using specific feature from the anchor a to the batch of positive or negative samples are computed as follows:

$$\begin{aligned} d_S(a,T) &= \{\|I_S(a) - I_S(t)\|_2 \mid t \in T\}, \\ &\text{where } \; T \in \{P_a, N_a\}, S \in \{sp+sh, sp, sh\}, \end{aligned} \tag{14}$$

where S indicates the features used (modality-specific one, modality-shared one, or the combined one). Note that we omit the superscript representing the modality for simplicity, considering the modalities are completed, and the features of RGB, NIR, TIR modalities are concatenated to calculate the distance. Then, for the combination of shared and modality-specific features, the maximum positive sample distance should be as small as possible compared to using a single type of feature, and the minimum negative sample distance should be as large as possible compared to using a single type of feature. As shown in Fig. 2, this formulation can be described as: $\max(d_{sp+sh}(a,p)) > \max(d_{sp}(a,p)), \max(d_{sh}(a,p))$ and $\min(d_{sp+sh}(a,n)) < \min(d_{sp}(a,n)), \min(d_{sh}(a,n))$. Therefore, we define the knowledge discrepancy loss L_{KDL} as:

$$\begin{aligned} L_{KDL} &= \|D_p - 0\|_1 + \|D_n - 1\|_1, \\ \text{where } D_p &= \frac{\max(d_{sp+sh}(a,p))}{\max\left(d_{sp+sh}(a,p)\right) + \max\left(d_{sp}\left(a,p\right)\right) + \max(d_{sh}(a,p))}, \\ D_n &= \frac{\min(d_{sp+sh}(a,n))}{\min\left(d_{sp+sh}(a,n)\right) + \min\left(d_{sp}(a,n)\right) + \min(d_{sh}(a,n))}. \end{aligned} \tag{15}$$

It is worth noting that detach operations are applied to $d_{sp}(a,n)$ and $d_{sh}^{RNT}(a,n)$ to avoid gradient computation. Thus, minimizing D_p reduces the farthest positive sample distance for the combined shared and modality-specific features, while minimizing $\|D_n - 1\|_1$ increases the nearest negative sample distance. Consequently, improving the retrieval capability of the combined features enhances the performance of model. In summary, the loss function for modality-aware metric learning is defined as follows:

$$L_{MML} = w_1 \times L_{ROL} + w_2 \times L_{KDL}. \tag{16}$$

3.4 Objective Function

As illustrated in Fig. 2, our model incorporates two primary loss functions: one for the ViT backbone and another for modality-aware metric learning. For the ViT backbone, we adopt label smoothing cross-entropy loss L_{ce} and triplet loss L_{tri} following previous research to optimize the representation. In summary, the total loss function is as follows:

$$L = L_{ce} + L_{tri} + L_{MML}. \tag{17}$$

4 Experiments

4.1 Datasets and Evaluation Protocols

We evaluate the proposed **MDReID** framework on three publicly available datasets spanning both person (RGBNT201 [24]) and vehicle re-identification (RGBNT100 [8] and MSVR310 [26]) tasks. For evaluating modality mismatch scenarios, we focus on representative cross-modal configurations, including RT-to-NT, RT-to-N, R-to-N, and R-to-NT. Here, NIR is the second most widely used modality in surveillance systems after RGB, and RGB-to-NIR research continues to attract significant interest. Therefore, we evaluate our method in four R-to-N-based application scenarios. Performance is measured using mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at ranks 1, 5, and 10 (R-1, R-5, R-10). We also report key complexity indicators, including the number of trainable parameters and floating point operations (FLOPs), to quantify the efficiency of the model.

RGBNT201 [24] is a multi-modality person re-identification dataset designed to overcome the limitations of single-modal imaging in challenging surveillance scenarios. It was captured on a university campus using four non-overlapping camera views with a synchronized triple-camera system that simultaneously records RGB, near-infrared (NIR), and thermal-infrared (TIR) images. The dataset contains 201 identities, each represented by at least 20 non-adjacent image triplets, resulting in 4,787 images per modality. The collection covers a range of real-world conditions, including severe illumination changes, occlusion, viewpoint variations, and background clutter. The dataset is split into 141 identities for training, 30 for validation, and 30 for testing, with the entire test set serving as the gallery and 10 records per identity sampled as probes. RGBNT201 thus provides a robust benchmark for evaluating multi-modal fusion strategies and addressing missing-modality issues in person re-identification.

RGBNT100 [8] contains 17,250 spatially aligned image triples from 100 vehicles captured under uniform conditions. Each triple includes RGB, NIR, and TIR images. Fifty vehicles (8,675 triples) are used for training, and the remaining 50 vehicles (8,575 triples) form the testing/gallery set, with 1,715 triples randomly selected as queries. The RGB and NIR images are recorded at a resolution of 1920×1080, while the TIR images are captured at 640×480, all at a consistent frame rate of 25 fps. TIR images provide thermal information that is robust to illumination changes, further enhancing the multi-spectral data for challenging vehicle re-identification scenarios.

MSVR310 [26] is a high-quality multi-spectral vehicle re-identification benchmark captured under diverse, challenging conditions. It contains 2,087 triplet samples (6,261 images in total) from 310 vehicles. Each sample includes spatially aligned images from three modalities: RGB, NIR, and TIR. RGB images are captured by a 360 D866 camera during the day and by a Mi8 mobile phone at night, NIR images are obtained with the 360 D866 in near-infrared mode, and TIR images are recorded with a FLIR SC620 camera at 640×480 resolution. The number of samples per vehicle ranges from 2 to 20, and time labels are provided for cross-time matching. The dataset is divided into 1,032 samples from 155 vehicles for training and 1,055 samples from 155 vehicles for testing/gallery, with 591 samples

Table 1: **Performance comparison on RGBNT201, RGBNT100, and MSVR310.** The best and second results are in bold and underlined, respectively.

Method	RGBNT201				Method	RGBNT100		MSVR310	
	*m*AP	R-1	R-5	R-10		*m*AP	R-1	*m*AP	R-1
Single									
MUDeep [39]	23.8	19.7	33.1	44.3	DMML [40]	58.5	82.0	19.1	31.1
HACNN [41]	21.3	19.0	34.1	42.8	BoT [42]	78.0	95.1	23.5	38.4
MLFN [43]	26.1	24.2	35.9	44.1	Circle Loss [44]	59.4	81.7	22.7	34.2
PCB [45]	32.8	28.1	37.4	46.9	HRCN [46]	67.1	91.8	23.4	44.2
OSNet [47]	25.4	22.3	35.1	44.7	TransReID [1]	75.6	92.9	18.4	29.6
CAL [48]	27.6	24.3	36.5	45.7	AGW [16]	73.1	92.7	28.9	46.9
Multi									
HAMNet [8]	27.7	26.3	41.5	51.7	GAFNet [25]	74.4	93.4	-	-
PFNet [24]	38.5	38.9	52.0	58.4	GraFT [49]	76.6	94.3	-	-
IEEE [9]	47.5	44.4	57.1	63.6	GPFNet [28]	75.0	94.5	-	-
DENet [27]	42.4	42.2	55.3	64.5	PHT [35]	79.9	92.7	-	-
UniCat [36]	57.0	55.7	-	-	UniCat [36]	79.4	96.2	-	-
HTT [10]	71.1	73.4	83.1	87.3	CCNet [26]	77.2	96.3	36.4	55.2
EDITOR [11]	66.5	68.3	81.1	88.2	EDITOR [11]	82.1	96.4	39.0	49.3
RSCNet [31]	68.2	72.5	-	-	RSCNet [31]	82.3	**96.6**	39.5	49.6
TOP-ReID [12]	72.3	76.6	84.7	89.4	TOP-ReID [12]	81.2	96.4	35.9	44.6
MDReID (Ours)	**82.1**	**85.2**	**90.3**	**92.6**	**MDReID (Ours)**	**85.3**	95.6	**51.0**	**68.9**

(from 52 vehicles) used as queries. MSVR310 provides a comprehensive platform for addressing intra-class appearance variations and modality differences to support robust vehicle re-identification.

4.2 Implementation Details

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU with CUDA 12.5 and Python 3.8. We adopt the CLIP-Base [37] visual encoder as the backbone. Input images are resized to 256×128 for RGBNT201 and 128×256 for RGBNT100 and MSVR310. We employ random horizontal flipping, cropping, and erasing [38] for data augmentation. The model is optimized using Adam with a batch size of 64. The base learning rate is initialized at 3.5×10^{-4}, while the visual encoder is fine-tuned with a reduced rate of 5×10^{-6}. Training is performed for 50 epochs.

4.3 Comparison with State-of-the-Art Methods

Multi-spectral Object ReID (RNT-to-RNT) We compare our MDReID against a range of state-of-the-art single-spectral and multi-spectral approaches on RGBNT201, RGBNT100, and MSVR310. For person ReID on RGBNT201, as shown in Table 1, the TOP-ReID [12] achieves 72.3% mAP, 76.6% R-1, 84.7% R-5, and 89.4% R-10. In comparison, our MDReID outperforms these results by 9.8%, 8.6%, 5.6%, and 3.2%, respectively, highlighting the effectiveness of our proposed MDReID. For vehicle re-ID, MDReID achieves the highest mAP on RGBNT100 (85.3%) and surpasses the next-best approach on MSVR310 by a significant margin of 11.5% in mAP and 13.7% in R-1. These results clearly demonstrate the superior discriminative power and robustness of our modality-decoupled design in multi-spectral re-ID scenarios where all modalities are available.

Evaluation on Missing-spectral Scenarios. In practical applications, sensor limitations or environmental factors often lead to incomplete modality inputs. Following the setting of TOP-ReID [12], we evaluate the robustness of our method under all six possible missing-modality configurations across the RGB, NIR, and TIR channels using the RGBNT201 dataset. As shown in Table 2, our MDReID consistently outperforms the TOP-ReID, achieving an average improvement of 10.0% in mAP and 9.4% in Rank-1 accuracy across all missing-modality scenarios. These results demonstrate that our modality-decoupled framework remains highly effective even when partial modality information is absent, highlighting its strong generalization ability in incomplete multispectral settings.

Evaluation on Modality-mismatched Scenarios. To evaluate the generalizability of our approach under realistic deployment conditions, we benchmark MDReID across four representative modality-mismatched scenarios on three public datasets. We compare against two strong baselines, TOP-ReID [12] and EDITOR [11], by reproducing their results using official open-source implementations. As shown in Table 3, EDITOR is designed exclusively for modality-matched settings, which exhibits limited effectiveness when modalities differ across query and gallery. Although TOP-ReID attempts

Table 2: **Performance of missing-modality settings on RGBNT201.** "M (X)" means missing the X image modality. The best and second results are in bold and underlined, respectively.

	Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
		*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1
Single	MUDeep [39]	19.2	16.4	20.0	17.2	18.4	14.2	13.7	11.8	11.5	6.5	12.7	8.5	15.9	12.9
	HACNN [41]	12.5	11.1	20.5	19.4	16.7	13.3	9.2	6.2	6.3	2.2	14.8	12.0	13.3	10.7
	MLFN [43]	20.2	18.9	21.1	19.7	17.6	11.1	13.2	12.1	8.3	3.5	13.1	9.1	15.6	12.4
	PCB [45]	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4	19.7	18.1
	OSNet [47]	19.8	17.3	21.0	19.0	18.7	14.6	12.3	10.9	9.4	5.4	13.0	10.2	15.7	12.9
Multi	PFNet [24]	-	-	31.9	29.8	25.5	25.8	-	-	-	-	26.4	23.4	-	-
	DENet [27]	-	-	35.4	36.8	33.0	35.4	-	-	-	-	32.4	29.2	-	-
	TOP-ReID [12]	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	26.0	34.1	31.7	44.4	45.4
	MDReID (Ours)	**67.0**	**67.9**	**75.8**	**80.9**	**61.7**	**59.8**	**48.6**	**51.1**	**31.4**	**29.2**	**42.1**	**40.0**	**54.4**	**54.8**

Table 3: **Performance of modality-mismatched Scenarios.** We show the best average score in bold.

	Methods	RT-to-NT		RT-to-N		R-to-N		R-to-NT		Average	
		*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1
RGBNT201	EDITOR [11]	27.3	27.9	2.80	0.0	4.0	2.3	4.3	4.1	8.5	7.5
	TOP-ReID [12]	43.0	44.6	14.4	13.3	15.4	**14.0**	11.9	8.7	18.2	18.0
	MDReID (Ours)	**53.1**	**51.7**	**16.7**	**13.8**	**16.6**	11.1	**15.1**	**10.9**	**21.6**	**19.1**
RGBNT100	EDITOR [11]	42.1	59.9	2.6	0.8	2.8	1.5	2.7	1.5	11.9	15.5
	TOP-ReID [12]	59.0	81.8	21.9	28.5	26.2	34.0	21.7	25.4	26.8	36.1
	MDReID (Ours)	**69.4**	**85.5**	**39.6**	**47.9**	**45.4**	**56.4**	**37.6**	**41.8**	**38.6**	**47.4**
MSVR310	EDITOR [11]	6.4	11.0	2.2	2.5	1.6	0.2	1.7	1.5	2.5	3.4
	TOP-ReID [12]	18.4	30.5	12.9	19.5	13.7	21.3	12.6	18.8	11.2	17.8
	MDReID (Ours)	**35.1**	**52.5**	**24.7**	**34.5**	**28.6**	**39.8**	**27.0**	**36.0**	**22.1**	**31.7**

to address missing modalities via reconstruction, its performance remains constrained due to the ill-posed nature of the reconstruction problem. In contrast, our MDReID framework consistently improves the average mAP by 3.4%, 11.8%, and 10.9% on the RGBNT201, RGBNT100, and MSVR310 datasets, respectively. In addition, we train four expert models on the RGBNT201 dataset for four specific scenarios. Results show that our method outperforms these experts by 9.2% on average in mAP and by 8.9% in R-1. Importantly, our single-model approach adapts to all four scenarios and offers greater flexibility. In summary, these results highlight the robustness and adaptability of MDReID in handling modality-mismatched ReID, affirming its effectiveness capable of supporting arbitrary modality combinations in both query and gallery inputs.

4.4 Ablation Study

To evaluate the effectiveness of each component in our proposed MDReID framework, we conduct comprehensive ablation studies on the RGBNT201 dataset. To ensure a thorough evaluation across both modality-matched and modality-mismatched scenarios, we report the average performance over 8 evaluation settings, including RNT-to-RNT, RT-to-RT, RT-to-NT, RT-to-N, R-to-N, and R-to-NT. The detailed results are presented in Table 4 (a). Note that configuration "1" corresponds to using a single classifier for all modalities, while configuration "3" corresponds to assigning a separate classifier to each modality. Ablation results show that, compared with one classifier, modality-specific classifiers effectively improve performance, increasing mean mAP by 13.4% and mean R-1 by 13.7%.

Modality Decoupled Learning Our base framework integrates a shared vision transformer, utilizing label smoothing cross-entropy along with triplet loss for optimization. To better address modality-mismatched object re-identification, we design a modality decoupling learning (MDL) that separates features into shared and specific representations. Experimental results demonstrate that this module boosts performance, with an average increase of 11.5% in mAP and 11.1% in Rank-1. This demonstrates that modality decoupling not only benefits re-identification in mismatched settings but also enhances retrieval performance in matched scenarios.

Modality-aware Metric Learning To further decouple shared and specific features, we design a modality-aware metric learning (MML) loss function. This loss consists of two components: the first aligns shared features across multiple modalities, while the second enhances the decoupling of features based on the aligned representation. Experimental results show that introducing ROL improves average mAP and Rank-1 by 1.8% and 2.7%, respectively. Similarly, introducing KDL increases mAP by 0.5% and Rank-1 by 1.7%. When both components are combined, the average mAP and Rank-1 improve by 3.8% and 4.1%, respectively. These results indicate that ROL significantly

Table 4: **Ablation study and hyper-parameter settings of MDReID.** We show the best average score in bold.

(a) **Ablation study.** MDL, L_{ROL}, and L_{KDL} indicate the Modality Decoupled Learning, Representation Orthogonality Loss (ROL), and Knowledge Discrepancy Loss (KDL), respectively.

Index	MDL	L_{ROL}	L_{KDL}	*m*AP	R-1
1	×	×	×	27.8	27.1
2	✓	×	×	39.4	38.2
3	✓	✓	×	41.2	40.8
4	✓	×	✓	39.9	40.9
5	✓	✓	✓	**43.2**	**42.3**

(b) **Performance under different** w_1. The optimal performance achieves when w_1 is set to 1.5.

w_1	*m*AP	R-1
0.5	39.6	38.1
1.0	40.3	40.0
1.5	**41.2**	**40.8**
2.0	38.9	38.1
3.0	38.2	37.0

(c) **Performance under different** w_2. The optimal performance achieves when w_2 is set to 5.25.

w_2	*m*AP	R-1
4.5	41.7	41.1
5.0	40.2	39.6
5.25	**43.2**	**42.3**
5.5	41.6	41.3
6.0	41.9	40.4

enhances re-identification accuracy in mismatch scenarios. Moreover, while KDL alone provides limited improvement, its combination with ROL effectively boosts overall model performance.

Modality Decoupled Learning We design the KDL objective to enforce a knowledge gap between specific and shared features. Without KDL, the model may degrade into relying only on shared features. Therefore, we ensure that combining specific and shared features yields higher retrieval performance than using either feature alone. To achieve this, we apply formulation 15 to impose distance constraints in the feature space based on the described relations. As shown in Table 4 (a), index 3 and 5, adding KDL improves mAP by 2.0% and Rank-1 by 1.5% across multiple scenarios. In summary, the MDReID framework, comprising the modality decoupled learning and modality-aware metric learning, achieves an average performance of 43.2% mAP, 42.3% Rank-1, 50.2% Rank-5, and 54.9% Rank-10 across various scenarios, including both modality-matched and modality-mismatched cases. Its modular design effectively decouples shared and specific features, enhancing object re-identification accuracy under multi-spectral conditions while maintaining robust adaptability across diverse modalities.

4.5 Discussions[4]

Hyper-parameter settings of MDReID In Eq. (16), we introduce two hyperparameters w_1 and w_2 to balance the importance of ROL and KDL. Therefore, this part evaluates the performance under different hyperparameter settings and shows the results in Table 4. We first vary w_1 without KDL, observing that the model achieves optimal performance at $w_1 = 1.5$, highlighting the importance of enforcing orthogonality between shared and specific components. Subsequently, with $w_1 = 1.5$, we tune w_2 and find that the best results are obtained at $w_2 = 5.25$, confirming the benefit of encouraging representational synergy and complementarity. These observations validate that a contribution from both ROL and KDL is crucial for maximizing retrieval accuracy.

5 Conclusion

In this paper, we propose MDReID, an image-level any-to-any object re-identification framework that supports retrieval across arbitrary query-gallery modality combinations. Our approach leverages Modality-Decoupled Learning (MDL) to explicitly disentangle modality-shared and modality-specific features, while Modality-aware Metric Learning (MML) refines the feature space to enhance discrimination. Extensive evaluations on the RGBNT201, RGBNT100, and MSVR310 datasets validate MDReID's superior performance in both modality-matched and modality-mismatched scenarios, demonstrating its practical potential for any-to-any ReID tasks.

Acknowledgements. This research was supported in part by the China Postdoctoral Science Foundation under Grant Number 2025M771584 and 2025T180439, and the National Natural Science Foundation of China under Grant 6250071985

[4]A comprehensive analysis of the model's computational complexity (Sec. A.1), along with additional visualization results (Sec. A.4), is provided in the supplementary material.

References

[1] 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 (ICCV)*, pages 15013–15022, October 2021.

[2] Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1415–1423, 2023.

[3] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14133–14142, 2023.

[4] 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*, volume 37, pages 1405–1413, 2023.

[5] Guan'an Wang, Xiaowen Huang, Jitao Sang, et al. Denoiserep: Denoising model for representation learning. *Advances in Neural Information Processing Systems*, 37:40032–40056, 2024.

[6] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 531–540, 2022.

[7] Lei Tan, Pingyang Dai, Jie Chen, Liujuan Cao, Yongjian Wu, and Rongrong Ji. Partformer: Awakening latent diverse representation from vision transformer for object re-identification. *arXiv preprint arXiv:2408.16684*, 2024.

[8] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11345–11353, Apr. 2020.

[9] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2633–2641, Jun. 2022.

[10] Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5850–5858, Mar. 2024.

[11] 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 (CVPR)*, pages 17117–17126, June 2024.

[12] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5758–5766, Mar. 2024.

[13] 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*.

[14] Lei Tan, Yukang Zhang, Keke Han, Pingyang Dai, Yan Zhang, Yongjian Wu, and Rongrong Ji. Rle: A unified perspective of data augmentation for cross-spectral re-identification. *Advances in Neural Information Processing Systems*, 37:126977–126996, 2024.

[15] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2907, 2021.

[16] 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.

[17] Chuanming Wang, Yuxin Yang, Mengshi Qi, Huanhuan Zhang, and Huadong Ma. Towards efficient object re-identification with a novel cloud-edge collaborative framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7600–7608, 2025.

[18] Lei Tan, Jiaer Xia, Wenfeng Liu, Pingyang Dai, Yongjian Wu, and Liujuan Cao. Occluded person re-identification via saliency-guided patch transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5070–5078, 2024.

[19] Yizhen Jia, Rong Quan, Yue Feng, Haiyan Chen, and Jie Qin. Doubly contrastive learning for source-free domain adaptive person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3949–3957, 2025.

[20] Ziniu Yin, Yanglin Feng, Ming Yan, Xiaomin Song, Dezhong Peng, and Xu Wang. Roda: Robust domain alignment for cross-domain retrieval against label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9535–9543, 2025.

[21] Haoran Liu, Ying Ma, Ming Yan, Yingke Chen, Dezhong Peng, and Xu Wang. Dida: Disambiguated domain alignment for cross-domain retrieval with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3612–3620, 2024.

[22] Chao Su, Huiming Zheng, Dezhong Peng, and Xu Wang. Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20610–20618, 2025.

[23] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, 2023.

[24] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3529–3537, May 2021.

[25] Jinbo Guo, Xiaojing Zhang, Zhengyi Liu, and Yuan Wang. Generative and attentive fusion for multi-spectral vehicle re-identification. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1565–1572, 2022.

[26] Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark, 2023.

[27] Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification, 2023.

[28] 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.

[29] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, pages 1–31, 2024.

[30] Yingquan Wang, Pingping Zhang, Dong Wang, and Huchuan Lu. Other tokens matter: Exploring global and local features of vision transformers for object re-identification. *Computer Vision and Image Understanding*, 244:104030, 2024.

[31] 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*, 2024.

[32] Yanping Li, Yizhang Liu, Hongyun Zhang, Cairong Zhao, Zhihua Wei, and Duoqian Miao. Occlusion-aware transformer with second-order attention for person re-identification. *IEEE Transactions on Image Processing*, 2024.

[33] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[34] Wenjie Pan, Hanxiao Wu, Jianqing Zhu, Huanqiang Zeng, and Xiaobin Zhu. H-vit: Hybrid vision transformer for multi-modal vehicle re-identification. In Lu Fang, Daniel Povey, Guangtao Zhai, Tao Mei, and Ruiping Wang, editors, *Artificial Intelligence*, pages 255–267, Cham, 2022. Springer Nature Switzerland.

[35] Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multi-modal vehicle re-identification. *Sensors*, 23(9), 2023.

[36] Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. Unicat: Crafting a stronger fusion baseline for multimodal re-identification, 2023.

[37] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[38] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13001–13008, Apr. 2020.

[39] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. Deep meta metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[41] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[42] 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 (CVPR) Workshops*, June 2019.

[43] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[45] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[46] Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 205–214, October 2021.

[47] 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 (ICCV)*, October 2019.

[48] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1025–1034, October 2021.

[49] Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. Graft: Gradual fusion transformer for multimodal re-identification, 2023.

[50] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2153–2162, June 2023.

A Supplemental Material

A.1 Analysis of model computational complexity

Table 5 compares the computational cost of our method with recent state-of-the-art approaches. MDReID achieves the most efficient design, requiring only **57.2M** parameters and **22.3G** FLOPs, which is significantly lower than HTT (85.6M / 33.1G), EDITOR (117.5M / 38.6G), and TOP-ReID (278.2M / 34.5G). Despite its efficient architecture, MDReID consistently outperforms these heavier models in accuracy, demonstrating a superior trade-off between performance and efficiency. This highlights the effectiveness of our MDL and MML designs, which provide robust alignment in all kinds of query-gallery combinations without incurring large computational overhead. Furthermore, we evaluate the total feature generation and retrieval times of different models on the RGBNT201 test set under the RNT-to-RNT, RT-to-NT, and R-to-N scenarios. Theoretically, our method reduces retrieval cost compared to Top-ReID. For example, in the RT-to-NT scenario, we only perform T-to-T (specific features) and R-to-T (shared features) matching, whereas Top-ReID first generates missing-modality features before conducting a full RNT-to-RNT matching. Experimental results confirm this, showing that our method incurs no additional time and generates features faster.

Table 5: **Comparison of computational cost with recent methods.** We show the best result in bold.

Methods	Params(M)	Flops(G)	RNT-to-RNT(s)		RT-to-NT(s)		R-to-N(s)	
			Extract	Retrieval	Extract	Retrieval	Extract	Retrieval
HTT [10]	85.6	33.1	8.32	**0.12**	-	-	-	-
EDITOR [11]	117.5	38.6	31.12	**0.12**	30.99	0.12	30.98	**0.12**
TOP-ReID [12]	278.2	34.5	71.78	**0.12**	73.32	**0.11**	73.71	**0.12**
MDReID (Ours)	**57.2**	**22.3**	**3.27**	0.14	**2.90**	**0.11**	**3.14**	**0.12**

A.2 Analysis of specific and shared features

In MDReID, both shared and modality-specific features are essential. For the RT-to-NT task, using only shared features overlooks the unique characteristics of the T modality, while using only modality-specific features prevents valid R-to-N matching. To quantify this, we evaluate RT-to-NT performance under three settings: shared features only, modality-specific features only, and both combined. As shown in Table 6, combining shared and modality-specific features yields the best performance.

Table 6: **Performance analysis under different features.** We show the best result in bold.

Feature	***m*AP**	**R-1**
Specific feature	52.0	50.5
Shared feature	52.7	51.2
Specific and shared feature	**53.1**	**51.7**

A.3 Comparison with cross-modality method

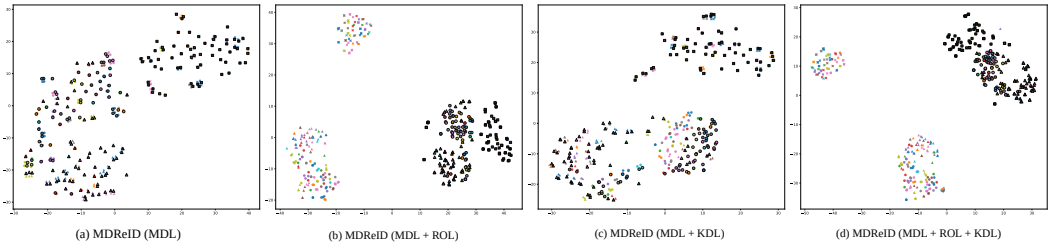
We compare our method with the cross-modality object re-identification model DEEN [50] on the RGBNT201 dataset under the R-to-N scenario. Table 7 shows that even when DEEN is specially trained for R-to-N, its mean precision is 3.5% lower than ours. Notably, our model demonstrates stronger generalization across diverse application scenarios.

Table 7: **Comparison with cross-modality method.** We show the best result in bold.

Feature	***m*AP**	**R-1**	**Average**
DEEN [50]	11.5	9.2	10.35
Ours	16.6	11.1	13.85

A.4 Visualization

To better illustrate the impact of different components on feature disentanglement, we visualize the learned modality-specific and modality-shared features across RGB, NIR, and TIR using t-SNE, as shown in Figure 3. Notably, the baseline method does not use MDL, so it does not separate specific and shared features and only outputs fused features, hence no corresponding visualization.



(a) MDReID (MDL) (b) MDReID (MDL + ROL) (c) MDReID (MDL + KDL) (d) MDReID (MDL + ROL + KDL)

Figure 3: The visualization of the ablation study. Triangles, circles, and X symbols correspond to the RGB, NIR, and TIR modalities, respectively. Black borders highlight shared features for clarity. Without ROL, distinguishing between shared and modality-specific features is challenging, but its introduction effectively clusters the shared features.

Specifically, in (a), with only MDL applied, the features exhibit noticeable overlap, and the separation between modality-specific and shared components remains unclear. In (c), the addition of KDL slightly improves the clustering and separation of shared and specific features. In contrast, (b) shows that introducing ROL significantly enhances feature orthogonality, forming clearer boundaries between shared and modality-specific clusters. Finally, (d) demonstrates that the combination of both ROL and KDL yields the most structured and disentangled feature space, with shared features tightly clustered and well-separated from their modality-specific counterparts. These results confirm the complementary roles of ROL and KDL in refining the representation space.

A.5 Limitations and Broader Impact

Although our work is the first to explore object re-identification under the uncertain query-gallery modality setting and achieves significant performance improvements, the current accuracy is still insufficient for deployment in real-world applications. Furthermore, existing datasets are limited in both scale and modality diversity, making it unclear how well MDReID would generalize to larger and more complex datasets involving a broader range of modality types. Nevertheless, MDReID represents the first effort to address the *any-to-any* object re-identification task, marking a significant step toward this emerging direction. We believe our work will inspire new perspectives and challenges in the re-identification community.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of this paper accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of this work in Section A.5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have introduced all the details in our theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper has provided detailed information for reproducing the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at: https://github.com/stone96123/MDReID.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper has specified all the training and test details, such as dataset and hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not provide error bars due to the limited resources. Also, previous methods do not provide error bars either.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper has provided sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper meets the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper has no societal impact to the best of our knowledge.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks since it does not release data or models that have risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper has properly cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper only uses LLMs for writing and editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.