Unbiased Prototype Consistency Learning for Multi-Modal and Multi-Task Object Re-Identification

Zhongao Zhou[1*] **Bin Yang**[1*] **Wenke Huang**[1] **Jun Chen**[1†] **Mang Ye**[1†]
[1] School of Computer Science, Wuhan University

Abstract

In object re-identification (ReID) task, both cross-modal and multi-modal retrieval methods have achieved notable progress. However, existing approaches are designed for specific modality and category (person or vehicle) retrieval task, lacking generalizability to others. Acquiring multiple task-specific models would result in wasteful allocation of both training and deployment resources. To address the practical requirements for unified retrieval, we introduce Multi-Modal and Multi-Task object ReID (M^3T-ReID). The M^3T-ReID task aims to utilize a unified model to simultaneously achieve retrieval tasks across different modalities and different categories. Specifically, to tackle the challenges of modality distibution divergence and category semantics discrepancy posed in M^3T-ReID, we design a novel Unbiased Prototype Consistency Learning (UPCL) framework, which consists of two main modules: Unbiased Prototypes-guided Modality Enhancement (UPME) and Cluster Prototype Consistency Regularization (CPCR). UPME leverages modality-unbiased prototypes to simultaneously enhance cross-modal shared features and multi-modal fused features. Additionally, CPCR regulates discriminative semantics learning with category-consistent information through prototypes clustering. Under the collaborative operation of these two modules, our model can simultaneously learn robust cross-modal shared feature and multi-modal fused feature spaces, while also exhibiting strong category-discriminative capabilities. Extensive experiments on multi-modal datasets RGBNT201 and RGBNT100 demonstrates our UPCL framework showcasing exceptional performance for M^3T-ReID. The code is available at https://github.com/ZhouZhongao/UPCL.

1 Introduction

Object re-identification (ReID) [75, 49, 65, 45, 18, 60, 3, 9, 75, 73, 69, 30, 4] leverages computer vision techniques to identify specific objects (such as persons or vehicles) in videos and still images. ReID technology has been widely applied in intelligent video surveillance, public security, and other related fields. Traditional ReID predominantly focuses on single-modal scenario, where both the query and gallery consist of RGB images. However, RGB cameras are highly sensitive to illumination variations, making it difficult to accurately capture target information under low-light or overexposed conditions. To address the above challenges, Near-Infrared (NI) and Thermal Infrared (TI) modalities have been introduced into ReID tasks, enabling robust imaging in challenging environments [72, 15, 29, 55, 17, 71, 6, 63]. Depending on the retrieval scenarios, the existing ReID methods can be broadly categorized into cross-modal ReID [58, 53, 36, 70, 41, 19, 2] and multi-modal ReID [47, 50, 68, 67]. Specifically, cross-modal ReID focuses on retrieval between two different modalities (e.g.,NI-RGB, TI-RGB), whereas multi-modal ReID utilizes RGB, NI, and TI fusion to achieve feature matching.

[*]Equal contribution.
[†]Corresponding Author.

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

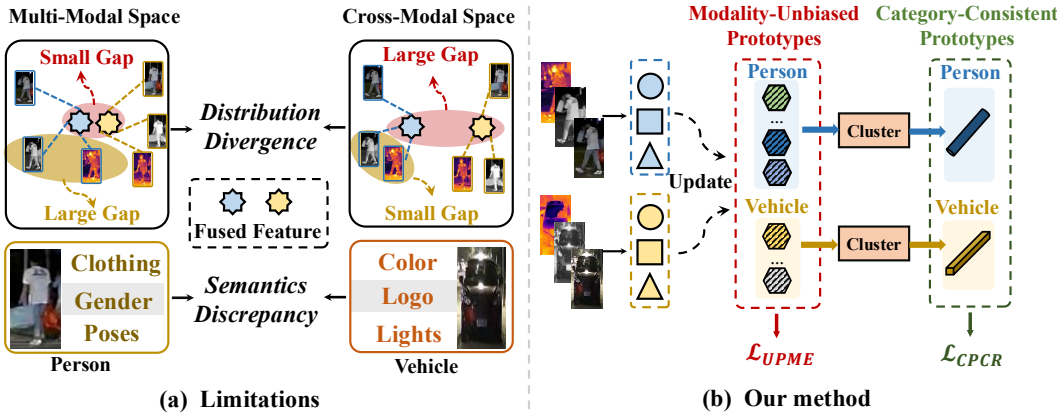

Figure 1: **Illustration of limitations in M^{3}T-ReID and our method.** (a) *Distribution divergence*: In multi-modal feature space, the multi-modal fused features exhibit smaller gap while cross-modal gap remain larger. Conversely, the opposite characteristic holds in cross-modal feature space. *Semantics discrepancy*: Different categories of objects possess distinct discriminative semantics. (b) We introduce modality-unbiased prototypes and cluster-derived category-consistent prototypes to enhance the model's comprehensive retrieval capability from both modality and category perspectives.

As illustrated in Figure 1, although existing cross-modal and multi-modal approaches have achieved remarkable results, they still suffer from two significant limitations: **1)** Real-world surveillance environments present extreme complexity, where targets may appear in scenarios captured by either single-modality cameras or aligned multi-modal imaging systems. While cross-modal ReID primarily focuses on learning a shared cross-modal feature space, multi-modal ReID emphasizes the effective fusion of different modalities to obtain more robust fused features. The cross-modal and multi-modal approaches follow different optimization directions, thereby resulting in distribution divergence. Therefore, existing ReID models cannot simultaneously handle both cross-modal and multi-modal retrieval. **2)** The retrieval tasks are confined to a specific category, necessitating separate model training for either person or vehicle ReID tasks. In practical scenarios such as criminal investigations, surveillance systems require the capability to simultaneously retrieve both suspects and vehicles. Due to the discrepancy in semantics between diverse categories, existing ReID methods lack the generalization capability to perform unified retrieval across categories. While training dedicated retrieval models for distinct modalities and categories may serve as a feasible solution, this approach inevitably leads to substantial redundancy in both training and deployment resources.

To meet the real-world demands for retrieval across diverse modalities and categories, we propose the Multi-Modal and Multi-Task object ReID (M^{3}T-ReID). The M^{3}T-ReID task aims to achieve a unified model for simultaneous retrieval across multiple modalities and diverse categories. However, achieving high-performance of M^{3}T-ReID introduces several challenges. Firstly, since cross-modal and multi-modal retrieval models optimize fundamentally different objectives, this leads to **challenge I** : *How to jointly learn both a robust cross-modal shared feature space and an effective multi-modal fusion feature space.* Secondly, discriminative features vary significantly across object categories. For instance, person ReID primarily focuses on attributes like pose and clothing while vehicle ReID emphasizes vehicle type and color, which raises the **challenge II** : *How to enable a model to simultaneously learn category-specific discriminative representations for heterogeneous objects.*

To address the aforementioned chanllenges in M^{3}T-ReID, we propose Unbiased Prototype Consistency Learning framework (UPCL) which comprises two key modules: Unbiased Prototypes-guided Modality Enhancement (UPME) and Cluster Prototype Consistency Regularization (CPCR). For **challenge I**, UPME enhances both cross-modal shared features and multi-modal fused features through modality-unbiased prototypes, thereby bridging the discrepancy across heterogeneous modalities and simultaneously improving robustness of the overall feature space. For **challenge II**, CPCR derives category-consistent features through prototypes clustering, thereby regulating the model to stably acquire category-specific discriminative semantics from diverse categories.

The main contributions of this paper can be summarized as follows:

- To address the practical demands for retrieval of diverse modalities and categories, we propose a novel Multi-Modal and Multi-Task object ReID (M^3T-ReID).
- To address the challenges in M^3T-ReID, we propose UPCL which comprises two main components: UPME and CPCR. UPME leverages modality-unbiased prototypes to simultaneously enhance cross-modal shared features and multi-modal fused features, and CPCR regulates the learning of discriminative semantics across diverse object categories through prototypes clustering.
- Extensive experiments on the public multi-modal ReID benchmarks RGBNT201 and RGBNT100 have verified the advantage of our methods, achieving significantly higher accuracy compared to existing counterparts in both cross-modal and multi-modal retrieval scenarios.

2 Related Work

2.1 Cross-modal and Multi-modal Re-identification

Cross-modal re-identification [23, 14, 20, 5, 8, 62, 31, 59, 11] aims to retrieve target RGB images across heterogeneous modalities. The cross-modal retrieval capability of a model primarily depends on the robustness of itscross-modal shared feature space. Wu *et al.*[52] utilize a zero-padding one-stream network with grayscale inputs to learn the shared feature between RGB andNI images. Ye *et al.*[61] introduce a Channel Augmentation (CA) mechanism that mitigates the modality gap by generating color-irrelevant person representations. Liu *et al.*[23] propose the Memory-Augmented Unidirectional Metric (MAUM) method to enhance the cross-modality correlation by utilizing two unidirectional metrics. Liang *et al.*[20] make an early attempt at unsupervised cross-modal ReID with a two-stage framework. Yang *et al.*[57] design an Augmented Dual-Contrastive Aggregation (ADCA) learning framework for Unsupervised Learning Visible-Infrared Person ReID.

Multi-modal re-identification [47, 15] jointly leverages complementary information from multiple modalities to extract more robust fused features, thereby improving retrieval accuracy. Zheng *et al.*[72] propose PFNet which hierarchically fuses RGB,NI, and TI features to obtain more robust representations. Wang *et al.*[49] design a Cross-Modal Interacting Module to enhance modality-specific information during feature fusion. Wang *et al.*[50] utilize the relationship among heterogeneous modalities to fine-tune the network prior to inference, thereby improving generalization to unseen data. Zhang *et al.*[68] propose a general PromptMA framework, which employs learnable prompts to aggregate modalities and bridge the modality distribution gap. Wang *et al.*[46] introduce the token permutation to enhance inter-modal interaction and facilitate multi-spectral feature alignment. Zhang *et al.*[67] propose EDITOR framework, which obtains spatial–frequency masks to refine multi-modal features. Wang *et al.*[48] adaptively balances decoupled features using a mixture-of-experts mechanism to produce more robust multi-modal representations.

Due to the divergent optimization objectives between cross-modal and multi-modal, a single model cannot be effectively applied to both retrieval scenarios simultaneously. Furthermore, significant semantic discrepancies exist among objects of different categories, models trained on one category cannot achieve generalized to retrieval of other categories. To address diverse retrieval requirements in real-world scenarios, we propose the UPCL framework, which trains a unified model capable of performing retrieval across multiple modalities and object categories.

2.2 Prototypes Learning

Prototype is calculated as the mean feature of the instances belonging to the same ID [38]. Due to its simplicity and scalability, prototype plays an indispensable role across various domains, such as few-shot learning [56, 38, 13, 27, 42] , unsupervised learning [54, 7, 16, 53],incremental learning [74, 76, 66, 43, 37], and federated learning [21, 12, 26, 22, 28, 25, 40, 64, 34, 35]. In object ReID research [24, 57, 58, 53], prototypes and other feature-centric concepts have also demonstrated significant effectiveness. Luo *et al.*[24] introduce Center Loss to enhance the feature similarity of same-ID samples in the model. Yang *et al.*[57] assigns pseudo-labels to unannotated data through clustering and dynamically update the centers of samples corresponding to each pseudo-ID as supervisory signals for network optimization.

Current applications of prototypes or similar concepts primarily leverage their statistical properties at ID-level to enhance feature compactness within the same category, thereby improving network robustness. In this work, we innovatively utilize modality-unbiased prototypes from a modality-

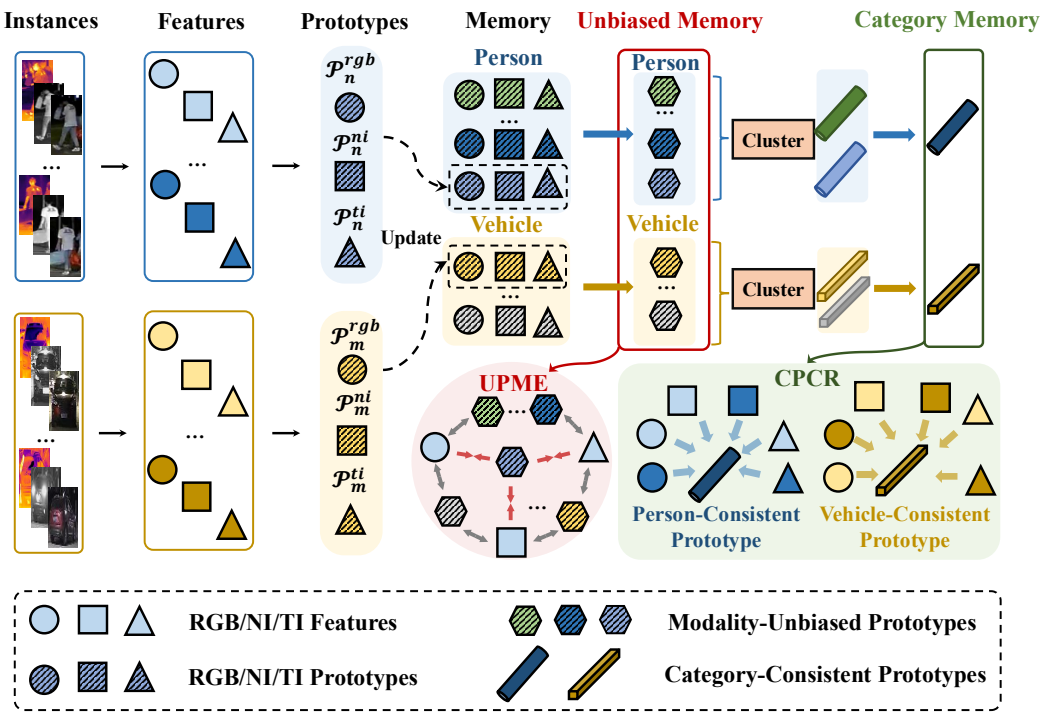

Figure 2: **Overview of of the proposed UPCL framework.** It consists of two core components: Unbiased Prototypes-guided Modality Enhancement (UPME) and Cluster Prototype Consistency Regularization (CPCR). We dynamically update the multi-modal prototype memory for each ID during every training iteration. UPME aggregates modality-unbiased prototypes via Equation (9), leveraging their identity-consistent information to strengthen both cross-modal and multi-modal representation learning. CPCR further utilizes category-consistent prototypes clustered from modality prototypes, exploiting their category-consistent semantics to regularize the model and achieve more discriminative category-wise decision boundaries.

consistency perspective and employ clustering strategies to obtain category-level (rather than identity-level) discriminative features. This approach significantly enhances the unified ReID model's capability to perform robust retrieval across multiple categories and modalities.

3 Method

In this section, we present the Unbiased Prototype Consistency Learning framework (UPCL) which consists of two main modules. The Unbiased Prototypes-guided Modality Enhancement (UPME) module leverages modality-unbiased prototypes to simultaneously enhance both cross-modal shared features and multi-modal fused features, thereby improving the model's performance across different retrieval modes. The Cluster Prototype Consistency Regularization (CPCR) module utilizes modality-unbiased prototypes via clustering to derive category-consistent prototypes, which are then utilized to regulate the model's discriminative semantic learning process for different categories. The overview of UPCL is illustrated in Figure 2 and the details are discussed in the following subsections.

3.1 Overall Architecture

Our method utilizes the pretrained CLIP [32] model as the visual encoder which is shared with RGB, NI and TI modalities. Specifically, for the i-th multi-modal instance $V_i = \{V_i^{rgb}, V_i^{ni}, V_i^{ti}\}$, the images of three modalities are cropped into equal-sized patches and then mapped to embedding vectors with fixed dimensions. Next, we feed these embedding vectors into the visual encoder to obtain their corresponding class token $f_i^m \in \mathbb{R}^D$ and patch tokens $p_i^m \in \mathbb{R}^{N_p \times D}$, where $m \in \{rgb, ni, ti\}$, N_p denotes the number of patch tokens and D is the embedding dimension.

Consistent with existing multi-modal methods [72, 49, 50, 46], we employ the label smoothing cross-entropy loss $\mathcal{L}_{ce}$ [39] and triplet loss $\mathcal{L}_{tri}$ [10] to supervise the learning of visual encoder:

$$\mathcal{L}_g = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{1}$$

To further encourage cross-modal feature alignment, we introduce an additional cross-modal alignment loss function:

$$\mathcal{L}^{(m\to n)}(i) = -\log \frac{\exp\left(\langle f_i^m, f_i^n \rangle / \tau\right)}{\sum_{j=1}^{B} \exp\left(\langle f_i^m, f_j^n \rangle / \tau\right)}, \tag{2}$$

$$\mathcal{L}^{(n\to m)}(i) = -\log \frac{\exp\left(\langle f_i^n, f_i^m \rangle / \tau\right)}{\sum_{j=1}^{B} \exp\left(\langle f_i^n, f_j^m \rangle / \tau\right)}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ represents the cosine similarity function. f_i^m and f_i^n denote the features of i-th instance with different modalities $m, n \in \{rgb, ni, ti\}$. B is the batch size, and τ is the temperature parameter. Then the cross-modal loss function between modalities m and n can be formulated as:

$$\mathcal{L}_{m\leftrightarrow n} = \frac{1}{2B} \sum_{i=1}^{B} [\mathcal{L}^{(m\to n)}(i) + \mathcal{L}^{(n\to m)}(i)]. \tag{4}$$

Building upon this, we derive the cross-modal loss function $\mathcal{L}_{cross}$:

$$\mathcal{L}_{cross} = \mathcal{L}_{rgb\leftrightarrow ni} + \mathcal{L}_{rgb\leftrightarrow ti} + \mathcal{L}_{ni\leftrightarrow ti}. \tag{5}$$

Finally, we obtain the base objective $\mathcal{L}_b$ of the framework :

$$\mathcal{L}_b = \mathcal{L}_g + \mathcal{L}_{cross}. \tag{6}$$

3.2 Unbiased Prototypes-guided Modality Enhancement

To enable the model to achieve high performance in both cross-modal and multi-modal retrieval, it is essential to design optimization strategies that maintain directional consistency between these two paradigms. Inspired by the successful application of prototypes in other domains, we leverage the identity-consistent information encapsulated in prototypes to guide the feature learning.

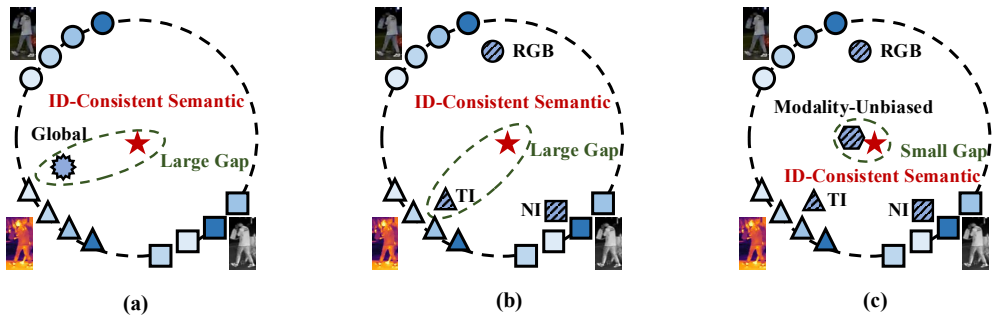

Figure 3: Comparison of different prototypes. (a) Global prototypes; (b) Modality-specific prototypes; (C) Modality-unbiased Prototypes (Ours). It is obviously observed that our proposed Modality-Unbiased Prototypes exhibit superior ID-consistent semantic representation in the feature space.

We compare two commonly used prototypes (global prototypes and modality-specific prototypes) in multi-modal learning alongside our newly proposed unbiased prototype in Figure 3.

Global Prototypes. The features of all samples across three modalities under the same ID are fused to obtain the global prototype:

$$\mathcal{P}_c^{global} = \frac{1}{\|I(c)\|} \sum_{i\in I(c)} L_{fused}(f_i^{rgb}, f_i^{ni}, f_i^{ti}), \tag{7}$$

where $I(c)$ denotes the indices of all instances with identity c. L_{fused} is an MLP designed to fuse the features from the three modalities, producing a global feature in $\mathbb{R}^D$.

Modality-Specific Prototypes. We compute modality-specific prototypes for identity c :

$$\mathcal{P}_c^m = \frac{1}{\|I(c)\|} \sum_{i \in I(c)} f_i^m, m \in \{rgb, ni, ti\}. \tag{8}$$

Then this yields a set of prototypes $\mathcal{P}_c = \{\mathcal{P}_c^{rgb}, \mathcal{P}_c^{ni}, \mathcal{P}_c^{ti}\}$ for three modalities .

Modality-Unbiased Prototypes. Global prototypes tend to incorporate semantic features biased toward the dominant modality, thereby deviating from identity-consistent semantics. Conversely, modality-specific prototypes primarily focus on intra-modal identity information, which inherently carries modality-specific bias relative to identity consistency. Therefore, in order to optimize the features towards identity-consistent semantics, we comprehensively derive a modality-unbiased prototype from $\mathcal{P}_c = \{\mathcal{P}_c^{rgb}, \mathcal{P}_c^{ni}, \mathcal{P}_c^{ti}\}$:

$$\mathcal{U}_c = (\mathcal{P}_c^{rgb} + \mathcal{P}_c^{ni} + \mathcal{P}_c^{ti})/3. \tag{9}$$

As illustrated in Figure 3, compared with global prototypes and modality-specific prototypes, the modality-unbiased prototypes mitigate both inter-modal and intra-modal distribution discrepancies, thereby thereby capturing more robust identity representations.

Under the guidance of modality-unbiased prototypes, we enhance features of modality m by:

$$\mathcal{L}_{UPME}^m = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\langle f_i^m, \mathcal{U}_{c_i} \rangle / \tau\right)}{\sum_{j=1}^{B} \exp\left(\langle f_i^m, \mathcal{U}_{c_j} \rangle / \tau\right)}, \tag{10}$$

where c_i represents the identity of the i-th sample. Finally, we formulate the Unbiased Prototypes-guided Modality Enhancement loss through summation:

$$\mathcal{L}_{UPME} = \mathcal{L}_{UPME}^{rgb} + \mathcal{L}_{UPME}^{ni} + \mathcal{L}_{UPME}^{ti}. \tag{11}$$

By enhancing semantic consistency across multi-modal features, the discriminative capability for identities in any modality is improved, thereby boosting the cross-modal retrieval performance. Simultaneously, the approach reduces inter-modal divergence and improves fusion efficiency, yielding more robust multi-modal representations.

3.3 Cluster Prototype Consistency Regularization

Current ReID methods typically train dedicated models for specific object categories (e.g., persons or vehicles). However, since discriminative semantic features vary significantly across different object categories, models trained on specific category lack generalization capability. To develop a unified model capable of retrieving diverse object categories, we need to enable the model to accurately capture category-consistent semantic information. As detailed in Section 3.2, the modality-unbiased prototypes inherently encapsulate identity-consistent semantic information that is strongly correlated with sample categories. This intrinsic correlation motivates us to explore a novel approach for extracting category-consistent semantics directly from these prototypes.

Assuming that there are N_p and N_v identities belonging to persons and vehicles, respectively, we utilize $\mathcal{U}^p = \{\mathcal{U}_c\}_{c=1}^{N_p}$ and $\mathcal{U}^v = \{\mathcal{U}_c\}_{c=N_p+1}^{N_p+N_v}$ to denote the sets of modality-unbiased prototypes. To obtain category-consistent information, we aim to fully integrate prototypes from all identities within a specific category. While the identities of any two prototypes in $\mathcal{U}^p$ or $\mathcal{U}^h$ are distinct, they may share highly similar semantic characteristics (e.g. males wearing short sleeves). Such semantic similarity naturally forms clusters among these identities, where the dominant clusters contribute more representative information within $\mathcal{U}^p$ or $\mathcal{U}^v$.

To better capture discriminative semantic features of specific categories, we propose to utilize cluster-level statistics instead of identity-level information. The acquisition of reliable cluster-level statistics fundamentally depends on achieving proper clustering of $\mathcal{U}^p$ and $\mathcal{U}^v$. Unlike conventional clustering algorithms K-means [1] and HAC [51], FINCH [33] operates in a completely parameter-free manner, automatically determining the optimal number of clusters based on the inherent similarity relationships among all prototypes. Therefore, we employ FINCH to cluster $\mathcal{P}^{u,p}$ and $\mathcal{P}^{u,v}$, obtaining the sets of multiple clustered prototypes $\mathcal{C}^p$ and $\mathcal{C}^v$:

$$\mathcal{U}^p = \{\mathcal{U}_c\}_{c=1}^{N_p} \xrightarrow{Cluster} \mathcal{C}^p = \{\mathcal{C}_l^p\}_{l=1}^{L_p}, \tag{12}$$

$$\mathcal{U}^v = \{\mathcal{U}_c\}_{c=N_p+1}^{N_v} \xrightarrow{Cluster} \mathcal{C}^v = \{\mathcal{C}_l^v\}_{l=1}^{L_v}, \tag{13}$$

where L_p and L_v denote the number of elements in $\mathcal{C}^p$ and $\mathcal{C}^v$. The l-th cluster prototypes for person and vehicle are denoted as $\mathcal{C}_l^u$ and $\mathcal{C}_l^u$, which are calculated by averaging all modality-unbiased prototypes in l-th cluster. Then we obtain the category-consistent prototype through:

$$\mathcal{C}ate^p = \frac{1}{L_p}\sum_{l=1}^{L_p}\mathcal{C}_l^p, \tag{14}$$

$$\mathcal{C}ate^v = \frac{1}{L_v}\sum_{l=1}^{L_v}\mathcal{C}_l^v. \tag{15}$$

Then, we introduce the CPCR loss function with $\mathcal{C}ate^p$ and $\mathcal{C}ate^v$ for specific category:

$$\mathcal{L}_{CPCR}^p = \frac{1}{\|I^p\|}\sum_m\sum_{i\in I^p}^{B^p}\log\frac{\exp\left(\langle f_i^m, \mathcal{C}ate^p\rangle/\tau\right)}{\exp\left(\langle f_i^m, \mathcal{C}ate^p\rangle/\tau\right)+\exp\left(\langle f_i^m, \mathcal{C}ate^v\rangle/\tau\right)}, \tag{16}$$

$$\mathcal{L}_{CPCR}^v = \frac{1}{\|I^v\|}\sum_m\sum_{i\in I^v}^{B^v}\log\frac{\exp\left(\langle f_i^m, \mathcal{C}ate^v\rangle/\tau\right)}{\exp\left(\langle f_i^m, \mathcal{C}ate^p\rangle/\tau\right)+\exp\left(\langle f_i^m, \mathcal{C}ate^v\rangle/\tau\right)}, \tag{17}$$

where I^p and I^v denote the index sets of all person-class and vehicle-class instances within a batch, respectively. It is natural to derive the total CPCR loss:

$$\mathcal{L}_{CPCR} = \mathcal{L}_{CPCR}^p + \mathcal{L}_{CPCR}^v. \tag{18}$$

Under the regularization of category-consistent prototypes, the model effectively learns discriminative semantic features that represent object categories, thereby achieving robust performance across diverse retrieval tasks involving different object categories.

Finally, we integrate all components to formulate a comprehensive optimization objective:

$$\mathcal{L}_{total} = \mathcal{L}_b + \alpha\mathcal{L}_{UPME} + \beta\mathcal{L}_{CPCR}, \tag{19}$$

where α and β are the the hyper-parameters to balance the contributions of each loss function.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Protocols. To evaluate the performance of our UPCL, we combined RGBNT201 [72] and RGBNT100 [15] to construct a multi-modal dataset with diverse object categories. Specifically, RGBNT201 is the first multi-modal person ReID dataset, where each pedestrian ID contains RGB, NI and TI modalities. RGBNT100 contains the same modalities as RGBNT201, but collects vehicle images instead. We evaluate the performance of our proposed UPCL with Rank-k matching accuracy, mean Average Precision (mAP) which are the commonly utilized metrics in object ReID tasks. The evaluation protocol encompasses six cross-modal testing scenarios (*R* $\rightarrow$ *N*, *N* $\rightarrow$ *R*, *R* $\rightarrow$ *T*, *T* $\rightarrow$ *R*, *N* $\rightarrow$ *T*, and *T* $\rightarrow$ *N*) along with one multi-modal testing configuration (*RNT* $\rightarrow$ *RNT*).

Implementation Details. The implementation platform is Pytorch with a NVIDIA 3090 GPU. We utilize the pre-trained CLIP as the visual encoder. Images of all modalities are resized to 256×128. For data augmentation, we apply random horizontal flipping, cropping and erasing. The batch size is set to 64, sampling 8 images per identity. The training process is conducted with the Adam optimizer for 50 epochs, and the initial learning rate is set to 3.5e-4. We select 0.03 as the temperature parameter τ. The hyper-parameters α and β are set as 2.0 and 0.5 respectively. During the testing phase, cross-modal retrieval directly computes similarity using features from two modalities, while multi-modal retrieval concatenates features from three modalities for matching.

Table 1: Comparison with the state-of-the-art methods. Each model is trained on a combined dataset consisting of RGBNT201 and RGBNT100, and evaluated separately.

Methods	*RGBNT201*															
	$R \rightarrow N$		$N \rightarrow R$		$R \rightarrow T$		$T \rightarrow R$		$N \rightarrow T$		$T \rightarrow N$		*RNT* $\rightarrow$ *RNT*		*Harm_Mean*	
	mAP	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*
HTT [50]	4.43	3.31	3.59	3.26	2.73	0.48	3.26	0.36	3.26	2.27	4.25	4.31	9.16	4.67	3.83	1.05
TOP-ReID [46]	10.50	8.61	10.71	9.33	3.55	1.32	3.54	1.20	5.40	4.31	6.10	4.67	63.74	64.95	6.27	3.07
PromptMA [68]	20.37	17.70	19.39	14.35	11.80	7.06	13.77	12.32	11.56	8.49	10.28	6.34	**65.71**	**68.18**	15.32	10.95
EDITOR [67]	3.70	2.03	3.38	2.27	3.83	2.39	3.72	0.60	4.37	2.51	3.32	0.96	56.03	56.70	4.26	1.56
DeMo [48]	3.98	2.27	4.33	2.63	3.30	1.44	4.09	3.59	3.10	0.60	3.44	2.75	64.35	63.76	4.22	1.82
UPCL (Ours)	**22.33**	**23.21**	**20.09**	**18.54**	**16.77**	**14.59**	**18.19**	**18.54**	**17.93**	**14.47**	**17.56**	**21.17**	64.91	67.12	**20.75**	**19.97**
Methods	*RGBNT100*															
	$R \rightarrow N$		$N \rightarrow R$		$R \rightarrow T$		$T \rightarrow R$		$N \rightarrow T$		$T \rightarrow N$		*RNT* $\rightarrow$ *RNT*		*Harm_Mean*	
	mAP	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*
HTT [50]	5.94	5.89	4.64	2.33	3.00	0.64	3.59	1.05	3.57	2.33	3.90	2.33	32.21	53.00	4.48	1.76
TOP-ReID [46]	15.47	20.06	12.24	12.01	3.50	2.86	3.17	1.40	3.96	4.31	3.80	1.92	71.47	89.15	5.48	3.57
PromptMA [68]	41.73	53.64	42.64	51.78	8.54	7.46	9.96	8.28	7.87	4.90	10.37	8.92	71.07	86.53	13.93	11.28
EDITOR [67]	2.59	1.40	2.99	2.04	3.45	2.92	3.20	2.39	2.61	1.57	3.19	2.16	74.25	93.00	3.44	2.28
DeMo [48]	3.12	1.69	4.14	2.39	2.42	0.76	3.07	2.33	4.31	2.51	4.55	3.27	79.12	93.47	3.97	2.02
UPCL (Ours)	**48.41**	**64.08**	**49.83**	**64.61**	**16.50**	**19.42**	**17.20**	**18.31**	**17.26**	**21.40**	**17.25**	**18.08**	**79.41**	**94.87**	**22.63**	**27.50**

4.2 Comparison with State-of-the-art Methods

We present a comprehensive comparison of UPCL with state-of-the-art methods as outlined in Table 1. To thoroughly evaluate the model's holistic performance on six cross-modal and one multi-modal testing scenarios, we employ the harmonic mean of task-specific metrics as the aggregated performance measure, given that the harmonic mean places greater emphasis on smaller values compared to the arithmetic mean.

As demonstrated in Table 1, on the person-category dataset (RGBNT201) , UPCL significantly outperforms all competing methods across all six cross-modal scenarios, which strongly validates its cross-modal matching capability. For multi-modal retrieval results, compared with PromptMA which contains module specifically designed for multi-modal fusion matching, our approach trails by merely 0.8 percentage points in mAP. Regarding the arithmetic mean of all seven metrics, UPCL achieves substantial superiority over other methods, surpassing the second-best approach by 5.43 and 9.02 percentage points on these two metrics, respectively. For the vehicle-category dataset (RGBNT100) , our UPCL demonstrates consistent superiority across every scenario. Across both category-specific datasets (person and vehicle) under all seven evaluation protocols, our method achieves remarkable performance, which conclusively demonstrates its effectiveness for the M^3T-ReID task.

4.3 Ablation Studies

To thoroughly validate the effectiveness of our proposed UPME and CPCR modules, we conducted extensive ablation experiments on both RGBNT201 and RGBNT100 datasets. The experimental protocol involves incrementally integrating our proposed modules into the baseline model without introducing any extra modifications, enabling direct performance comparison through controlled ablation studies. As illustrated in Table 2, all reported metrics are presented as the harmonic mean of performance across seven evaluation scenarios.

Effect of UPME. By leveraging the identity-consistent information captured in modality-unbiased prototypes, the UPME module enhances feature representations across different modalities, thereby boosting cross-modal and multi-modal representation capabilities. The experimental results demonstrate that our UPME module achieves mAP improvements of 3.20% and 2.19% on RGBNT201 and RGBNT100 respectively, compared to the baseline. When aggregated across both datasets, the module delivers average gains of 2.41% mAP and 3.37% rank-1 accuracy.

Effect of CPCR. The CPCR module clusters modality-unbiased prototypes from UPME to derive category-consistent prototypes, then exploits their category-discriminative semantics to guide model optimization. After incorporating CPCR, the model achieves a 4.10% mAP and 5.99% rank-1 improvement on RGBNT201, while attaining 0.73% mAP and 2.91% rank-1 gains on RGBNT100.

Table 2: Ablation study of each component on RGBNT201 and RGBNT100.

Components		*RGBNT201*				*RGBNT100*				*Average*			
UPME	*CPCR*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*	*mAP*	*Rank-1*	*Rank-5*	*Rank-10*
		13.85	10.06	20.62	27.93	19.89	21.78	24.39	26.39	16.87	15.92	22.51	27.16
✔		16.65	13.98	25.59	33.56	21.90	24.59	28.06	30.61	19.28	19.29	26.83	32.09
✔	✔	**20.75**	**19.97**	**36.03**	**46.13**	**22.63**	**27.50**	**31.13**	**32.79**	**21.69**	**23.74**	**33.58**	**39.46**

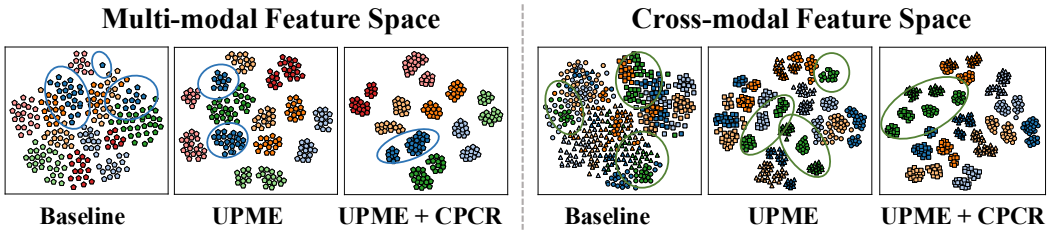

Figure 4: The t-SNE visualization of several randomly selected identities. Color indicate identities.

Visualization. To intuitively demonstrate the effectiveness of UPME and CPCR, we visualize the t-SNE[44] feature distribution of several identities in Figure 4. Specifically, the three subplots on the left visualize the embedding space of multi-modal fused features. Under the effects of UPME and CPCR, the fused features of the same ID progressively converge, while the discriminability between different identities becomes increasingly pronounced. The cross-modal features of identical IDs in the right figure exhibit consistent variation trends. Therefore, the t-SNE visualization clearly indicates that integrating UPME and CPCR enhances the intra-identity consistency and inter-identity discriminability of both fused and cross-modal features.

Effects of Diverse Categories . As discussed earlier, different categories of retrieval targets contain distinct discriminative semantics, making it challenging for a model to maintain high performance across retrieval tasks involving diverse categories. To validate this claim, Table 3 presents a comparison between category-specific model and a unified model. The category-specific model is trained on a single-category dataset, either RGBNT201 or RGBNT100, while the unified model is trained on a multi-category dataset formed by combining RGBNT201 and RGBNT100. It can be clearly observed that, for the same method, the unified model exhibits a significant performance drop compared to its corresponding category-specific model. This suggests that mixing multiple categories in training introduces substantial interference, hindering the model's ability to simultaneously learn discriminative semantics for all categories. These findings strongly support our hypothesis that multi-category training can negatively impact model optimization.

Table 3: Comparison between specifc and unified model. Each model is trained on a combined dataset consisting of RGBNT201 and RGBNT100, and evaluated separately. mAP(%) is reported.

Methods	*RGBNT201*		*RGBNT100*	
	Specific	*Unified*	*Specific*	*Unified*
HTT [50]	69.0	9.16	75.7	32.21
TOP-ReID [46]	72.3	63.74	81.2	71.47
PromptMA [68]	78.4	65.71	85.3	71.07
EDITOR [67]	66.5	56.03	79.8	74.25
DeMo [48]	79.0	64.35	86.2	79.12

5 Cross-domain Retrieval Evaluation

In conventional Re-ID evaluations, both training and testing sets are typically collected under the same scene conditions. As a result, the learned feature space may be constrained to a specific environment, making the evaluation results insufficient to reflect the generalization capability required in real-world applications. To assess the cross-domain generalization ability of our model, we train it on a combined dataset of RGBNT201 and RGBNT100, and evaluate it on the unseen MSVR310 dataset.

The comparison results are presented in Table 4. All models show significantly lower detection accuracy on the MSVR310 dataset compared to RGBNT201 and RGBNT100, indicating that cross-domain generalization remains highly challenging for existing Re-ID models. In particular, methods

Table 4: Performance analysis of cross-domain generalization on MSVR310 dataset. All models are trained on the combined dataset of RGBNT201 and RGBNT100, and evaluated on the MSVR310 dataset. Rank-k accuracy (%) and mAP(%) are reported.

Methods	*MSVR310*															
	$R \rightarrow N$		$N \rightarrow R$		$R \rightarrow T$		$T \rightarrow R$		$N \rightarrow T$		$T \rightarrow N$		$RNT \rightarrow RNT$		*Harm_Mean*	
	mAP	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*	*mAP*	*Rank-1*
HTT [50]	1.99	1.18	2.00	1.52	1.88	1.18	2.33	1.34	1.91	1.02	1.88	2.20	5.90	8.97	2.20	1.02
TOP-ReID [46]	4.31	6.43	3.87	3.21	2.23	0.85	1.84	1.52	2.30	1.02	2.54	2.71	14.91	25.55	2.94	1.89
PromptMA [68]	11.92	14.72	10.71	13.37	2.43	1.02	2.64	1.52	2.82	2.53	3.70	3.21	23.14	31.98	4.28	2.78
EDITOR [67]	2.16	1.18	1.53	0.52	1.59	0.51	1.49	0.63	1.45	0.85	1.35	1.02	17.05	30.96	1.70	0.82
DeMo [48]	2.12	0.34	1.73	0.17	1.87	1.35	1.77	1.02	1.66	1.02	1.76	0.51	15.89	26.90	2.07	0.52
UPCL (Ours)	**13.89**	**20.30**	**14.09**	**20.14**	**9.43**	**10.15**	**9.03**	**9.14**	**9.57**	**11.68**	**9.35**	**11.17**	**24.60**	**38.07**	**11.44**	**13.77**

such as HTT, TOP-ReID, EDITOR, and DeMo exhibit severe performance degradation under both cross-modal and multi-modal testing scenarios on the unseen MSVR310 dataset. In contrast, the UPCL and PromptMA methods achieve considerably better results, with UPCL consistently outperforming PromptMA across various detection settings. These experimental results provide strong evidence that our model possesses excellent cross-domain generalization capability. The superior cross-domain retrieval performance of UPCL further validates our method from another perspective—not only does it enhance the robustness of the fused feature space across modalities, but it also strengthens the model's ability to discriminate semantic features across different categories.

6 Limitations

Although our method has demonstrated promising performance on existing multi-modal ReID datasets, it still exhibits several notable limitations. The current multi-modal ReID datasets are relatively scarce, mainly restricted to the pedestrian and vehicle domains. Considering the practical demands of real-world applications, it is crucial to explore how incorporating a broader range of categories affects retrieval performance. In addition, our unified model may slightly underperform task-specific models in a very limited number of test scenarios. Although this is reasonable, this also indicates that our approach still has potential for further enhancement. Therefore, our future work will focus on extending UPCL to more generalized multi-modal and multi-task retrieval scenarios.

7 Conclusion

In this paper, we introduce a novel M^{3}T-ReID task to address the practical demands of retrieval across diverse modalities and categories. To tackle the challenges in M^{3}T-ReID, we propose the Unbiased Prototype Consistency Learning framework (UPCL) which consists of two main modules UPME and CPCR. Specifically, UPME mitigates the divergence between cross-modal shared spaces and multi-modal fusion distributions by leveraging identity-consistent information from modality-unbiased prototypes, thereby enhancing both cross-modal and multi-modal representations. Meanwhile, CPCR reduces semantic discrepancies across categories by clustering modality-unbiased prototypes to obtain category-consistent prototypes with discriminative semantics. Extensive experiments on multiple datasets validate the superiority and effectiveness of our method, demonstrating its robustness and generalization ability in diverse retrieval scenarios.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under Grants (62501428, 62176188), the Innovative Research Group Project of Hubei Province under Grants (2024AFA017), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003, 2025BEA002), Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20241268, 2024M762479), Hubei Postdoctoral Talent Introduction Program (2024HBB-HJD070), Hubei Provincial Natural Science Foundation of China (2025AFB219) and WHU-Kingsoft Joint Lab. The numerical calculations in this paper had been supported by the super-computing system in the Supercomputing Center of Wuhan University.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017.

[2] Cuiqun Chen, Mang Ye, and Ding Jiang. Towards modality-agnostic person re-identification with descriptive query. In *CVPR*, pages 15128–15137, 2023.

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *CVPR*, pages 3425–3435, 2021.

[4] Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *CVPR*, pages 16614–16623, 2024.

[5] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 6, 2018.

[6] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *CVPR*, pages 4313–4322, 2022.

[7] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *CVPR*, pages 9706–9715, 2022.

[8] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *ICCV*, pages 16403–16412, 2021.

[9] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021.

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

[11] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. In *NeurIPS*, pages 117363–117387, 2024.

[12] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322. IEEE, 2023.

[13] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr. Prototypical priors: From improving classification to zero-shot learning. *arXiv preprint arXiv:1512.01192*, 2015.

[14] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797, 2023.

[15] Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *AAAI*, volume 34, pages 11345–11353, 2020.

[16] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[17] Qiwei Li, Kunlun Xu, Yuxin Peng, and Jiahuan Zhou. Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation. *IJCV*, 132(11):4850–4865, 2024.

[18] 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 *AAAI*, volume 37, pages 1415–1423, 2023.

[19] Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, Xi Peng, and Peng Hu. Robust duality learning for unsupervised visible-infrared person re-identification. *IEEE TIFS*, 2025.

[20] Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE TIP*, 30: 6392–6407, 2021.

[21] Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Huabin Zhu, Yanchao Tan, Jun Wang, and Yue Qi. Hyperfed: hyperbolic prototypes exploration with consistent aggregation for non-iid data in federated learning. *arXiv preprint arXiv:2307.14384*, 2023.

[22] Jiale Liu, Yu-Wei Zhan, Xin Luo, Zhen-Duo Chen, Yongxin Wang, and Xin-Shun Xu. Prototype-based layered federated cross-modal hashing. In *ICASSP*, pages 1–2, 2023.

[23] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *CVPR*, pages 19366–19375, 2022.

[24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR*, pages 0–0, 2019.

[25] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *NeurIPS*, pages 5972–5984, 2021.

[26] Feng Lyu, Cheng Tang, Yongheng Deng, Tong Liu, Yongmin Zhang, and Yaoxue Zhang. A prototype-based knowledge distillation framework for heterogeneous federated learning. In *ICDCSW*, pages 1–11, 2023.

[27] Pascal Mettes, Elise Van der Pol, and Cees Snoek. Hyperspherical prototype networks. In *NeurIPS*, 2019.

[28] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *FGCS*, 143:93–104, 2023.

[29] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[30] Zhiqi Pang, Junjie Wang, Lingling Zhao, and Chunyu Wang. Identity-clothing similarity modeling for unsupervised clothing change person re-identification. In *CVPR*, 2025.

[31] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, pages 27197–27206, 2024.

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

[33] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, pages 8934–8943, 2019.

[34] Wei Shen, Wenke Huang, Guancheng Wan, and Mang Ye. Label-free backdoor attacks in vertical federated learning. In *AAAI*, volume 39, pages 20389–20397, 2025.

[35] Wei Shen, Mang Ye, Wei Yu, and Pong C Yuen. Build yourself before collaboration: Vertical federated learning with limited aligned samples. *IEEE TMC*, 2025.

[36] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *ECCV*, pages 456–474. Springer, 2024.

[37] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *ICCV*, pages 1772–1781, 2023.

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[40] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, volume 36, pages 8432–8440, 2022.

[41] Xiao Teng, Long Lan, Dingyao Chen, Kele Xu, and Nan Yin. Relieving universal label noise for unsupervised visible-infrared person re-identification by inferring from neighbors. In *AAAI*, volume 39, pages 7356–7364, 2025.

[42] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, pages 266–282, 2020.

[43] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *CVPR*, pages 16732–16741, 2022.

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[45] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *AAAI*, volume 36, pages 2540–2549, 2022.

[46] Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *AAAI*, volume 38, pages 5758–5766, 2024.

[47] Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. Mambapro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *AAAI*, volume 39, pages 8150–8158, 2025.

[48] Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *AAAI*, volume 39, pages 8141–8149, 2025.

[49] 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 *AAAI*, volume 36, pages 2633–2641, 2022.

[50] Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. In *AAAI*, volume 38, pages 5850–5858, 2024.

[51] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *JASA*, 58(301):236–244, 1963.

[52] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[53] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *CVPR*, pages 9548–9558, 2023.

[54] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[55] Kunlun Xu, Chenghao Jiang, Peixi Xiong, Yuxin Peng, and Jiahuan Zhou. Dask: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification. In *AAAI*, volume 39, pages 8915–8923, 2025.

[56] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, pages 21969–21980, 2020.

[57] Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *ACM MM*, page 2843–2851, 2022.

[58] Bin Yang, Jun Chen, Xianzheng Ma, and Mang Ye. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE TIP*, 2023.

[59] Fan Yang, Yang Wu, Zheng Wang, Xiang Li, Sakriani Sakti, and Satoshi Nakamura. Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval. *IEEE TMM*, 23:2347–2360, 2020.

[60] Yue Yao, Tom Gedeon, and Liang Zheng. Large-scale training data search for object re-identification. In *CVPR*, pages 15568–15578, 2023.

[61] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, October 2021.

[62] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 44(6):2872–2893, 2021.

[63] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *IJCV*, 133(5):2410–2440, 2025.

[64] Mang Ye, Wei Shen, Bo Du, Eduard Snezhko, Vassili Kovalev, and Pong C Yuen. Vertical federated learning for effectiveness, security, applicability: A survey. *ACM Computing Surveys*, 57(9):1–32, 2025.

[65] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *ICCV*, pages 11185–11195, 2023.

[66] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *CVPR*, pages 6982–6991, 2020.

[67] Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, pages 17117–17126, 2024.

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

[69] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *CVPR*, pages 3436–3445, 2021.

[70] Yukang Zhang, Yang Lu, Yan Yan, Hanzi Wang, and Xuelong Li. Frequency domain nuances mining for visible-infrared person re-identification. *IEEE TIFS*, 20:5411–5424, 2025.

[71] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Adaptive middle modality alignment learning for visible-infrared person re-identification. *IJCV*, 133(4):2176–2196, 2025.

[72] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *AAAI*, volume 35, pages 3529–3537, 2021.

[73] Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification. In *CVPR*, pages 19691–19701, 2023.

[74] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880, 2021.

[75] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, pages 4692–4702, 2022.

[76] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *CVPR*, pages 9296–9305, 2022.

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 claims are clearly stated in the abstract and the introduction.

   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 included the discussion of the limitations of our work.

   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: [NA]

Justification: Our work focuses on computer vision applications, not including theoretical results.

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: We provide all the details of our proposed methods to reproduce the 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: [No]

Justification: The datasets used in this paper are fused directy by two public datasets which can be downloaed from https://github.com/924973292/DeMo. We will release the relevant algorithm code after the paper is accepted.

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: The training and testing details are available in the paper.

   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: Due to the limited resources, our paper does not provide error bars. Also, previous relative 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: The information of computer resources is provided.

   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: The research conforms to 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: Our research has no negative societal impacts.

    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: The paper poses no such 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: The creators and original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

    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: The 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: The 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: The 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: The core method development in this research does not involve LLMs.

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.