Empowering Visible-Infrared Person Re-Identification with Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Visible-Infrared Person Re-identification (VI-ReID) often underperforms compared to RGB-based ReID due to significant modality differences, primarily caused by the absence of detailed information in the infrared modality. With the development of Large Language Models (LLMs) and Language Vision Models (LVMs), this motivates us to investigate a feasible solution to empower VI-ReID performance with off-the-shelf foundation models. To this end, we propose a novel text-enhanced VI-ReID framework driven by Foundation Models (TVI-FM). The basic idea is to enrich the representation of the infrared modality with automatically generated textual descriptions. Specifically, we incorporate a pretrained multimodal language vision model (LVM) to extract textual features and incrementally fine-tune the text encoder to minimize the domain gap between generated texts and original visual images. Meanwhile, to enhance the infrared modality with text, we employ LLM to augment textual descriptions, leveraging modality alignment capabilities of LVMs and LVM-generated feature-level filters. This allows the text model to learn complementary features from the infrared modality, ensuring semantic structural consistency between the fusion modality and the visible modality. Furthermore, we introduce modality joint learning to align features of all modalities, ensuring that textual features maintain stable semantic representation of overall pedestrian appearance during complementary information learning. Additionally, a modality ensemble retrieving strategy is proposed to consider each query modality for leveraging their complementary strengths to improve retrieval effectiveness and robustness. Extensive experiments demonstrate that our method significantly improves retrieval performance on three expanded cross-modal re-identification datasets, paving the way for utilizing foundation models in downstream data-demanding tasks. The code will be released.

1 Introduction

Person Re-Identification (ReID) aims to retrieve images of the same identity across different cameras, which is crucial for urban security. While RGB-based methods have shown promising results [14, 11, 5, 15, 23], their effectiveness diminishes in low-light conditions at night. To address this issue, Visible-Infrared Person Re-Identification (VI-ReID) [29] is proposed to enable cross-modality retrieval using visible and infrared images, ensuring 24-hour surveillance. Thus this area gains increasing interest among researchers. Infrared images offer an alternative source of visual information to visible images, particularly in low-light conditions. However, there remains a significant disparity between the infrared and visible modalities. Additionally, the absence of detailed information in the infrared modality also presents considerable challenges for existing VI-ReID method [29, 32, 35].

Most existing VI-ReID methods [35, 33, 13, 18, 34, 8, 31] don't take into account the problem of information absence in infrared modality, mostly aim to force the model to focus on mining

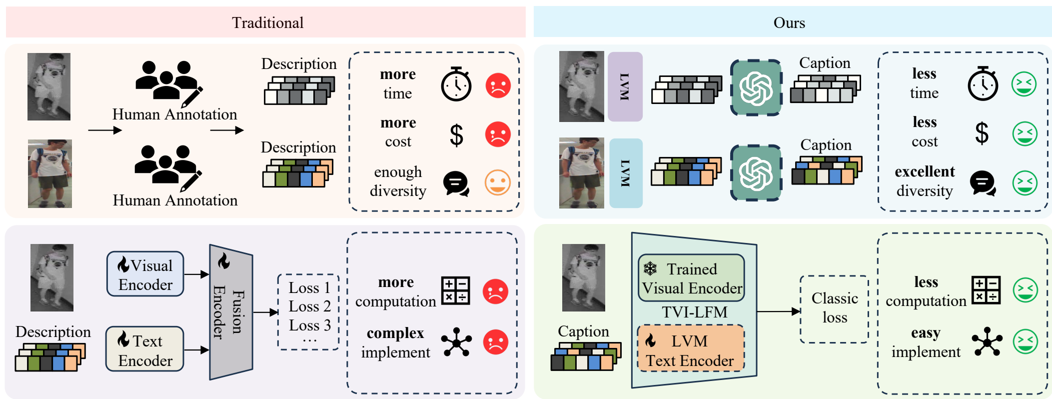


Figure 1: FM improved text-enhanced VI-ReID compared to traditional methods

discriminative information from infrared and visible modalities. Due to the absence of information in infrared modalities, the performance of these methods is limited. There exists methods [6, 40, 4] consider about utilizing auxiliary information like text descriptions or attributes to enhance infrared modality. However, As shown in Figure 1, these methods heavily rely on specially manual annotated auxiliary data, leading to great time and labor cost. Moreover, they struggle to use these auxiliary data with structure with massive learnable parameters and complex metric learning methods, which demands a lot of computation cost. For example, YYDS[6] processes manual annotated coarse descriptions with multiple encoders and complex regularizaton for text-image modality alignment.

Recent advancements in foundation models[2], especially LLM, LVM pre-trained on vast datasets as detailed in 2.2, facilitate customized data generation and augmentation. This further enables effective alignment of textual and visual modality, having great potential to enhance VI-ReID performance with text. This motivates us to investigate a feasible solution to empower the VI-ReID performance with off-the-shelf foundations models. Thus to handle these problems, we introduce a novel Text-enhanced VI-ReID framework driven by Foundation Models (TVI-FM), including Incremental Fine-tuning Strategy (IFS) and Modality Ensemble Retrieving (MER) modules. The basic idea is to enrich the representation of infrared modality with the automatically generated textual descriptions. This expansion method saves considerable time and labor cost for data annotation, which is introduced in Appendix 4.

To address these challenges, we propose a novel Text-enhanced VI-ReID framework, TVI-FM, which incorporates Incremental Fine-tuning Strategy (IFS) and Modality Ensemble Retrieving (MER) modules. This approach significantly reduces the time and labor required for data annotation, as explained in Appendix 4. Specifically, to seamlessly integrate text into existing VI-ReID framework, the Incremental Fine-tuning Strategy (IFS) is proposed to optimize the whole framework. IFS utilizes a pretrained multimodal language vision model (LVM) to extract textual features and incrementally fine-tune the text encoder to minimize the domain gap between generated texts and original visual images. Meanwhile, to enhance the infrared modality with text, we employ LLM to augment textual descriptions, leveraging modality alignment capabilities of LVMs and LVM-generated feature-level filters. This allows the text model to learn complementary features from the infrared modality, ensuring semantic structural consistency between the fusion modality and the visible modality. Furthermore, we introduce modality joint learning to align representations of all modalities, this ensures that textual features uphold stable semantic representation of holistic pedestrian appearance while mining complementary information. Additionally, Modality Ensemble Retrieving (MER) is proposed to enhance retrieval robustness and accuracy by aggregating multiple modalities similarity.

The main contributions can be summarized as follows:

- We propose a novel text-enhanced VI-ReID framework driven by Foundation Models (TVI-FM), which enriches the representation of infrared modality with the automatically generated textual descriptions, reducing the cost of text annotations and enhancing the performance of cross-modality retrieval.
- We develop an Incremental Fine-tuning Strategy (IFS) to employ LLM to augment textual descriptions and incorporate a pre-trained LVM to extract textual features, leveraging modality alignment

capabilities of LVMs and feature-level filters generated by LVMs to enhance infrared modality with information fusion and modality joint learning.

- We introduce Modality Ensemble Retrieving (MER) strategy to comprehensively take into account the queries' similarity with gall features for leveraging their complementary advantages to improve retrieval effectiveness and robustness.
- Extensive experiments demonstrate that our method improves retrieval performance on three expanded cross-modality re-identification datasets, paving the way for utilizing LLMs in downstream data-demanding tasks.

2 Related Work

2.1 Visible-Infrared Person Re-Identification

Visible-Infrared Person Re-Identification (VI-ReID) aims to match identities across visible and infrared images, but facing challenges of significant modality gap and the absence of information in IR modality. Previous works [35, 3, 37, 17, 33] attempt to bridge modality gap by mining discriminative information shared by modalities, but the limited information in blurred IR images results in poor performance. To address these issues, [6] introduces coarse texual descriptions as auxiliary information to enhance cross-modality retrieval. However, it heavily relies on manual annotations, structures with massive learnable parameters and complex metric learning, increasing computational costs and limiting flexibility. Different from existing works, our approach introduces a novel text-enhanced VI-ReID driven by foundation models, which automatically generates semantically rich text to complement visual data. It integrates a textual encoder with excellent capability of text-image alignment and a well-trained visual backbone, incrementally fine-tunes the texual encoder with classical ReID losses. This ensures effective enhancement for existing VI-ReID without additional complex implementations.

2.2 Foundation Model

Foundation models, pre-trained on large and varied datasets, show promise in many fields. Advances in Language-Vision Models (LVMs) like GIT, BLIP, and CLIP, as well as Large Language Models (LLMs) such as GPT-2, GPT-3, Vicuna, and LLaMa2, highlight their data generation and semantic understanding skills. BLIP is adept at creating image-based text descriptions, adaptable to various visual styles. Vicuna, a top LLM, uses its extensive training for complex text adjustments while preserving semantics, suited for custom text enhancements. CLIP's training on vast image-text datasets lets it align these modalities and embed them in a shared semantic space, aiding modality alignment. Our approach leverages these models for automatic text generation and augmentation, and incorporates a pre-trained text encoder into the VI-ReID system to improve performance with textual data.

3 Proposed Method

Our TVI-FM system, as depicted in Fig. 2, leverages Language Vision Models (LVMs) to automatically generate textual modality, which enriches the representations of the infrared modality. This integration leads to significant performance improvements on existing VI-ReID backbones through our Incremental Fine-tuning Strategy (IFS) and Modality Ensemble Retrieving (MER). The IFS is a comprehensive approach that improves the robustness and accuracy of existing VI-ReID systems by complementing the infrared modality with information from generated text. It employs LLM-based Textual Augmentation to diversify textual descriptions while maintaining core semantics, which strengthens the model's resilience to text variability and improves the robustness of extracted textual features. Enhance IR with LVM Generated Filter is used to refine these textual features, focusing on extracting their complementary information according to infrared modality. Modality Joint Learning (MJL) then optimizes the global association of all modalities, aligning semantic representations and preserving the textual semantic of overall pedestrian appearance during complementary information learning. The MER strategy aligns query features from different modalities, capitalizing on their unique strengths to achieve more accurate retrieval. This integrated approach ensures semantic consistency across modalities, maximizing the effectiveness of Information Fusion in capturing complementary information for the infrared modality. As a result, our system advances the capabilities of VI-ReID systems, delivering enhanced performance and semantic understanding.

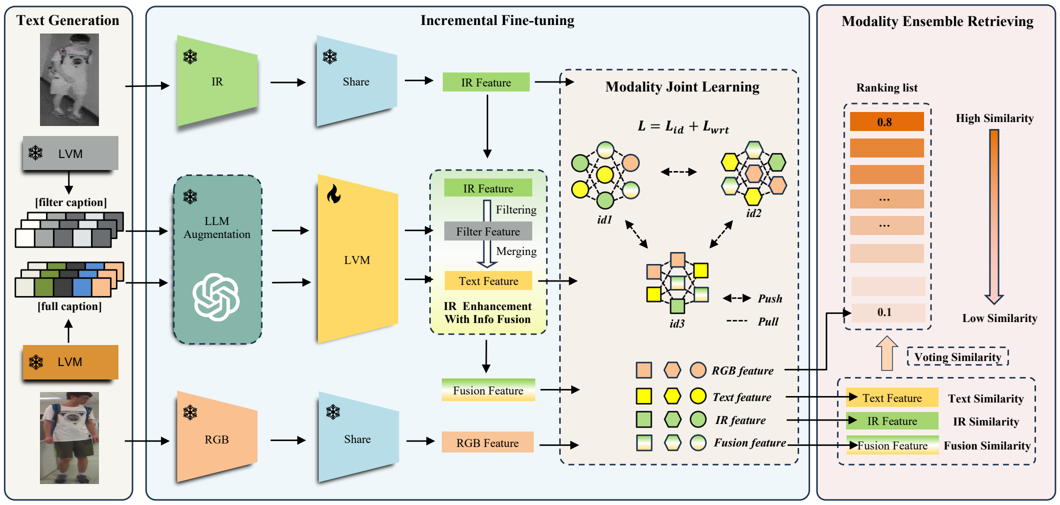


Figure 2: Illustration of TVI-FM: text-enhanced VI-ReID framework integrating a pre-trained LVM textual encoder[24] with a well-trained dual-stream visual backbone[10, 34]. Trained by Incremental Fine-tuning strategy and utilize Modality Ensemble Retreving for robust results.

3.1 Baseline

To seamlessly enhance the performance of existing VI-ReID backbone utilizing textwe employ a frozen well-trained VI-ReID backbone as visual encoder E_v to extract the visual features $f_m = E_v(V_m), m \in \{i, v\}$ from visible images V_r and infrared images V_i respectively, while the text features $f_{t_r} = E_t(T_r)$ are extracted by a textual encoder E_t of LVM from textual descriptions T_r. Benefiting from the image-text pre-training parameters of LVM textual encoder, the text-visual modality gap can be significantly mitigated during training, so we combine text and infrared features using a simple summation-fusion strategy to enhance infrared modality with the complementary information from text, defined as $f_{sum} = f_i + f_{t_r}$. Then we apply the sum of an identity loss and a weighted regularized triplet loss as the total framework:

$$L = L_{id}(f) + L_{wrt}(f) \tag{1}$$

$\boldsymbol{f} \in \{f_r^n\}_{n=1}^{N_r} \cup \{f_{sum}^n\}_{n=1}^{N_{sum}}$ are the extracted features from visible images and the fusion feature composed by infrared feature and corresponding text feature, where $\boldsymbol{N}_r$ denotes the number of visible images, $\boldsymbol{N}_{sum}$ denotes the number of fusion samples.

3.2 Incremental Fine-tuning Strategy

Our Incremental Fine-tuning Strategy (IFS) is a multi-faceted approach designed to enhance the robustness and accuracy of VI-ReID systems through tailored optimization and integration of generated textual and infrared modalities. Through Language Layout Models (LLM) Based Textual Augmentation, we diversify textual descriptions while preserving core semantics, fortifying the model against input variability and enabling the model to extract more robust textual features. Enhance IR with LVM Generated Filter refines textual features, focusing on extracting complementary information for the infrared modality. Modality Joint Learning (MJL) optimizes the global association of all modalities, aligning semantic representations and preventing distortion during fusion feature learning. This comprehensive strategy ensures semantic consistency across modalities, maximizing the effectiveness of Information Fusion in capturing complementary information for infrared modality, thus advancing the capabilities of VI-ReID systems.

3.2.1 LLM based Textual Augmentation

To ensure getting robust representations from diverse generated textual descriptions to enhance VI-ReID model, while preserving semantic integrity, we implement a probabilistic augmentation module based on LLM. This module regenerates more diverse descriptions for the same target and prevents overfitting to incidental information in sentences apart from the core features of person appearance, such as sentence style. In detail, given an original description $\boldsymbol{T}$, the module employs an LLM to augment the textual descriptions, controlled by the prompt ***"Rephrase the person's description using similar words, without changing the original semantics."*** The transformation is applied as follows:

$$T^* = \begin{cases} LLM(T \mid \text{Prompt}), & \text{with probability } p \\ T, & \text{with probability } (1-p) \end{cases} \tag{2}$$

where $p = 0.5$ reflects the assumption that each description variant is equally probable. Utilizing the powerful customized text understanding and generation capability of LLM, this approach not only diversifies the textual descriptions but also maintains their essential meanings. This forces the model to focus on extracting the core semantic of person appearance, thus enhancing the robustness of our system against text variability. Moreover, we can apply this augmentation method directly on existing framework related with text data, without any change of the original structure.

3.2.2 Enhance IR with LVM Generated Filter.

Through the LLM based textual augmentation, our textual encoder can extract more robust representations from the diverse textual description, containing rich complementary semantic compared to the infrared images. In order to fully mine the complementary semantic for infrared modality from auxiliary text data, which has been aligned with image modality in pre-train stage, we develop a novel feature-level information filtering mechanism to filter the redundant semantics in textual representations that are as same as infrared images. To get the filter features with semantic contained by infrared, we employ the fine-tuned LVM in Section A to generate textual description T_i from infrared images V_i as filter features, the generation of the filter features $\boldsymbol{f}_{t_i}$ can be represented as $\boldsymbol{f}_{t_i} = LVM(V_i)$. Considering that the fusion features $\boldsymbol{f}_{sum}$ are conbined with text features $\boldsymbol{f}_{t_r}$ and infrared features $\boldsymbol{f}_i$, which contains the same redundant semantic of the person appearance and may affect the semantic structural consistency with features of other modalities. With the proposed filter text features $\boldsymbol{f}_{t_i}$, we can obtain the refined text features by directly subtracting filter features, containing rich complementary information for infrared modality. So we form the fusion features jointly with the refined text features and infrared features. In detail, the refined fusion features can be represented as:

$$\boldsymbol{f}_{sum} = \boldsymbol{f}_i + \boldsymbol{f}_{t_r} - \boldsymbol{f}_{t_i} \tag{3}$$

Then we can force our model to focus on learning complementary information by solely fine-tuning the filtered text features without other redundant information.

3.2.3 Modality Joint Learning

In corperating with the learning of complementary information in Section 3.2.2, we introduce the Modality Joint Learning (MJL) strategy. The filter mechanism focuses on refining textual representations to capture complementary information for the infrared modality, while MJL optimizes the global association of all modalities. By jointly optimizing the total framework with visible, infrared, textual, and fusion modalities, MJL aligns corresponding semantics across all modalities, which enhances semantic consistency in textual representations and provides more informative hard samples for multi-modal representation learning. This maximizes the effectiveness of Information Fusion in capturing complementary information for the infrared modality while preventing potential semantic distortion during fusion feature learning. In detail, we jointly optimize the total framework with visible features $\boldsymbol{f}_r$, infrared features $\boldsymbol{f}_i$, textual features $\boldsymbol{f}_{t_r}$, fusion modality $\boldsymbol{f}_{sum}$ by the combination of cross-entropy loss L_{id} and weighted regularized triplet loss L_{wrt}:

$$L^* = L_{id}(f^*) + L_{wrt}(f^*) \tag{4}$$

In our training loss L^*, all the feature $\boldsymbol{f^*} \in \{f_r^n\}_{n=1}^{N_r} \cup \{f_i^n\}_{n=1}^{N_i} \cup \{f_{t_r}^n\}_{n=1}^{N_{t_r}} \cup \{f_{sum}^n\}_{n=1}^{N_{sum}}$ from different modalities share the same classifier and the global distance association with any other feature is also optimized, where $\boldsymbol{N}_r$ denotes the number of visible images, $\boldsymbol{N}_{sum}$ denotes the number of fusion samples, $\boldsymbol{N}_i$ denotes the number of infrared samples, $\boldsymbol{N}_t$ denotes the number of textual samples.

3.3 Modality Ensemble Retrieving

To maximize utilization of query representations with rich semantics mined from Incremental Fine-tuning Strategy in Section 3.2 for more accurate retrieval, the Modality Ensemble Retrieving (MER) strategy is employed to comprehensively take into account the unique and complementary advantages

Table 1: Ablation study on **Infrared** query enhanced by **auxiliary textual descriptions** ($\boldsymbol{I}+\boldsymbol{T} \rightarrow \boldsymbol{R}$) about each component on the performance of **Tri-SYSU-MM01** and **Tri-LLCM** datasets. **Rank** (R) at first accuracy (%), **mAP**(%), and **mINP**(%) are reported.

$I+T \rightarrow R$					Tri-SYSU-MM01			Tri-LLCM		
B	Filter	MJL	LLM	MES	R1	mAP	mINP	R1	mAP	mINP
✓					72.52	69.15	55.93	52.63	58.82	55.43
✓	✓				77.00	73.73	61.50	54.73	60.95	57.64
✓	✓	✓			83.97	80.40	69.46	56.76	63.58	60.35
✓	✓	✓	✓		84.17	80.72	70.02	57.13	64.06	60.72
✓	✓	✓		✓	84.88	81.32	70.57	57.09	63.87	60.62
✓	✓	✓	✓	✓	**84.90**	**81.47**	**70.85**	**58.19**	**65.08**	**61.83**

of different modalities. This involves averaging the features from infrared modality $\boldsymbol{f}_i$, textual modality $\boldsymbol{f}_{t_r}$, and fusion modality $\boldsymbol{f}_{sum}$ to form a comprehensive query feature:

$$\boldsymbol{f}_{agg} = mean(\boldsymbol{f}_i, \boldsymbol{f}_{t_r}, \boldsymbol{f}_{sum}) \tag{5}$$

Fusion features $\boldsymbol{f}_{sum}$ provide a comprehensive and enriched description of the target and aims to learn features with the same semantic structure of RGB modality, serving as the primary matching modality. ***Infrared features*** $\boldsymbol{f}_i$ provide valuable and contiguous visual semantics. Their similarity with RGB images can serve as a supplementary reference for visual information. ***Textual features*** $\boldsymbol{f}_{t_r}$ provide descriptive details that may not be visually apparent or recognizable in infrared images. The degree of similarity between textual features and RGB images serves as a explicit reference for the missing or blurred appearance information in the infrared modality. ***The comprehensive features*** $\boldsymbol{f}_{agg}$ used to retrieve RGB features $\boldsymbol{f}_r$ integrate the similarity scores of multiple query modalities with RGB to obtain a voting score, effectively harness the complementary strengths of each modality and reduce the potential impact of extreme scores in the fusion query retrieval list, enhancing the overall effectiveness and robustness of the retrieval system.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We evaluate our framework on the expanded datasets, including Tri-SYSU-MM01, Tri-RegDB, and Tri-LLCM. The proposed three multi-modal datasets with text description for each image are expanded from original visible-infrared images datasets SYSU-MM01[30], RegDB[20], and LLCM[38] by the fine-tuned generative LVMs named Blip[16] in three stages (Detail in Appendix A). The splits of the training set and testing set for each dataset are available in Appendix G.

Evaluation Protocols. In line with established VI-ReID settings [36, 33], we assess performance of infrared query mode and textual enhanced infrared query mode using Rank-k matching accuracy, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP[36]) within our TVI-FM framework. To get stable performance on SYSU-MM01 and LLCM, we evaluate our model for 10 times with random split of the gallery set, as for RegDB we evaluate our model on the 10 trials with different training/testing splits. Finally we report our model's average performance on each dataset. The task settings with detail of different query modes are shown in Appendix C.

Implementation Overview. We utilize a dual-stream resnet-50[34] pretrained on ImageNet[25] as the visual backbone and a transformer in CLIP[24] for the textual backbone. Training involves of visible and infrared images alongside text descriptions generated from these images, which are augmented by vicuna-7b[41] with a probabilistic rephrasing strategy. Incremental fine-tuning is applied by fixing the visual parameters and only tuning the texual part of the framework. All Details are described in Appendix B.

4.2 Ablation Study

To thoroughly evaluate the effect of each component to our proposed method, we conduct comprehensive ablation studies on the Tri-LLCM and Tri-SYSU-MM01 datasets. These studies involved gradually adding the proposed modules on our baseline, systematically removing specific modules

Table 2: The influence of whether to froze visual backbone on case of infrared query ($\boldsymbol{I} \rightarrow \boldsymbol{R}$) and fusion query($\boldsymbol{I} + \boldsymbol{T} \rightarrow \boldsymbol{R}$) on the performance of **Tri-SYSU-MM01** and **Tri-LLCM**. In order to focus on the impact of IFS on the learning of infrared features and fusion features separately, we **remove** the **MES** strategy for fusion query to avoid the effect of aggregating original information from infrared modality and text modality together with fusion modality.

$I \rightarrow R$	Tri-SYSU-MM01			Tri-LLCM		
	R1	mAP	mINP	R1	mAP	mINP
VI-ReID Backbone	69.89	66.74	53.34	53.53	59.77	56.40
Ours - Frozen	64.46↓**5.43**	61.31↓**5.43**	46.94↓**6.40**	49.29↓**4.24**	55.78↓**3.99**	52.12↓**4.28**

$I + T \rightarrow R$	Tri-SYSU-MM01			Tri-LLCM		
	R1	mAP	mINP	R1	mAP	mINP
Ours	**84.17**	**80.72**	**70.02**	**57.13**	**64.06**	**60.72**
Ours - Frozen	84.03↓**0.14**	79.85↓**0.87**	68.06↓**1.97**	55.47↓**1.66**	62.23↓**1.83**	58.86↓**1.86**

from our framework and assessing the impact on its performance. The overall experimental setup remained consistent, with only the module under evaluation being modified.

Effect of Enhancing IR with LVM Generated Filter. In order to enhance the semantic uniformity of fusion queries and other modalities while filtering out redundant information, we implement a feature-level filtering mechanism utilizing a Language Vision Model [16] to generate the filter features from IR images. Compared with the baseline, the filter module achieves enhancement of the comprehension of the textual complementary semantic, while the baseline cannot extract enough effective feature from text very well. The method obtains 4.48% Rank-1 improvement in Tri-SYSU-MM01 and 1.90% Rank-1 improvement in Tri-LLCM respectively in Table 1.

Effect of Modality Joint Learning. For fully making use of the multi-modal representations and learn robust, deeply aligned and semantic-consistent features for each identity, we propose a global modality joint learning method to incorporating with the filter mechanisms. Based on the experiment result in Table 1, compared to baseline only with filter mechanisms, adding this method gains a great enhancement of 6.97% Rank-1 improvement, 6.67% mAP improvement, 7.96% mINP% improvement in Tri-SYSU-MM01 and 2.03% Rank-1 improvement, 2.63% mAP improvement, 2.71% mINP improvement in Tri-LLCM.

Effect of Modality Ensemble Retrieving. The Modality Ensemble Searching strategy fully take account into all query modalities, minimizing the potential impact of extreme scores with a comprehensive query representation. From Table 1, it can be observed that incorporating MES provides an additional improvement of 0.71% in Rank-1, 0.60% in mAP, and 0.55% in mINP in the Tri-SYSU-MM01 dataset over the joint learning method with filter mechanisms. Similarly, on the Tri-LLCM dataset, MES achieves a 1.10% Rank-1 improvement, 1.21% mAP improvement, and 1.21% mINP improvement. These results demonstrate that the aggregation of different query modality leads to better overall performance, enabling more accurate retrieval.

Effect of LLM based Textual Augmentation. To extract more robust representations from diverse textual descriptions for the same person against the potential over-fitting while maintaining semantic integrity. We implement a probabilistic augmentation module based on Large Language Model (LLM). With LLM based augmentation, as the result shown in Table 1, it further improves our model's performance assisted with auxiliary text, and it can works well with other modules, achieving 84.90% Rank-1 and 58.19% Rank-1 in Tri-SYSU-MM01 and Tri-LLCM respectively.

Discussion of Frozen Operation It's important for VI-ReID system to handle cases with auxiliary text descriptions and without them in uncertain real scenarios application. When we allow the visual backbone to update parameters, as shown in Table 2, performance with infrared query suddenly declines by 5.43% and 4.24% of Rank-1 in the two datasets respectively. The performance on fusion query ($I + T \rightarrow R$) is also affected, with a decline of 0.14% Rank-1 in Tri-SYSU-MM01 and 1.66% Rank-1 in Tri-LLCM. This demonstrates the importance of freezing the integrated backbone to avoid the potential performance influence caused by conflict of infrared feature learning and fusion feature learning during training. With Frozen Operation we can seamlessly enhancing existing VI-ReID framework by only fine-tuning textual encoder.

Table 3: Compare with the state-of-the-art methods on the proposed Tri-SYSU-MM01

Methods	Venue	Type	All Search			Indoor Search		
			R-1	mAP	mINP	R-1	mAP	mINP
Zero-Padding [29]	ICCV-17		14.80	15.95	-	20.58	26.92	-
HCML [32]	AAAI-18		14.32	16.16	-	24.52	30.08	-
cmGAN [22]	IJCAI-18		26.97	27.80	-	31.63	42.19	-
AlignGAN [28]	ICCV-19		42.40	40.70	-	45.90	54.30	-
AGW [35]	TPAMI-21		47.50	47.65	35.30	54.17	62.97	59.23
DDAG [34]	ECCV-20		54.75	53.02	39.62	61.02	67.98	62.61
CM-NAS [9]	ICCV-21		61.99	60.02	-	67.01	72.95	-
DART [31]	CVPR-22	$I \rightarrow R$	68.7	66.3	-	82.0	73.8	-
CAJ [33]	ICCV-21		69.88	66.89	53.61	76.26	80.37	76.79
PAENet [1]	MM-22		74.22	73.90	-	78.04	83.54	-
DEEN [38]	CVPR-23		74.70	71.80	-	80.30	83.30	-
SAAI [7]	ICCV-23		75.90	77.03	-	83.20	88.01	-
MSCLNet [37]	ECCV-22		76.99	71.64	-	78.49	81.17	-
SGIEL [8]	CVPR-23		77.12	72.33	-	82.07	82.95	-
PartMix [13]	CVPR-23		77.78	74.62	-	81.52	84.38	-
YYDS[6]	Arxiv-24	$I + T \rightarrow R$	74.60	70.35	56.01	81.35	83.64	79.56
VI-ReID Backbone	-	$I \rightarrow R$	69.89	66.74	53.34	76.91	80.64	76.70
TVI-FM	-	$I + T \rightarrow R$	**84.90**	**81.47**	**70.85**	**89.06**	**90.78**	**88.39**

Table 4: Compare with the state-of-the-art methods on the proposed Tri-RegDB and Tri-LLCM

Methods	Venue	Type	Tri-RegDB			Tri-LLCM		
			R-1	mAP	mINP	R-1	mAP	mINP
DDAG [34]	ECCV-20		68.06	61.80	48.62	40.3	48.4	-
AGW [35]	TPAMI-21		70.49	65.90	51.24	43.6	51.8	-
CAJ [33]	ICCV-21	$I \rightarrow R$	84.8	77.8	61.56	48.8	56.6	-
DART [31]	CVPR-22		82.0	73.8	-	52.2	59.8	-
MMN [39]	MM-21		87.5	80.5	-	52.5	58.9	-
DEEN [38]	CVPR-23		89.5	83.4	-	54.9	62.9	-
YYDS[6]	Arxiv-24	$I \rightarrow R$	90.95	84.22	70.12	58.13	64.91	61.77
VI-ReID Backbone	-	$I \rightarrow R$	89.51	83.51	69.65	53.53	59.77	56.40
TVI-FM	-	$I + T \rightarrow R$	**91.38**	**85.92**	**72.73**	**58.19**	**65.08**	**61.83**

4.3 Comparison with the State-of-the-art Methods

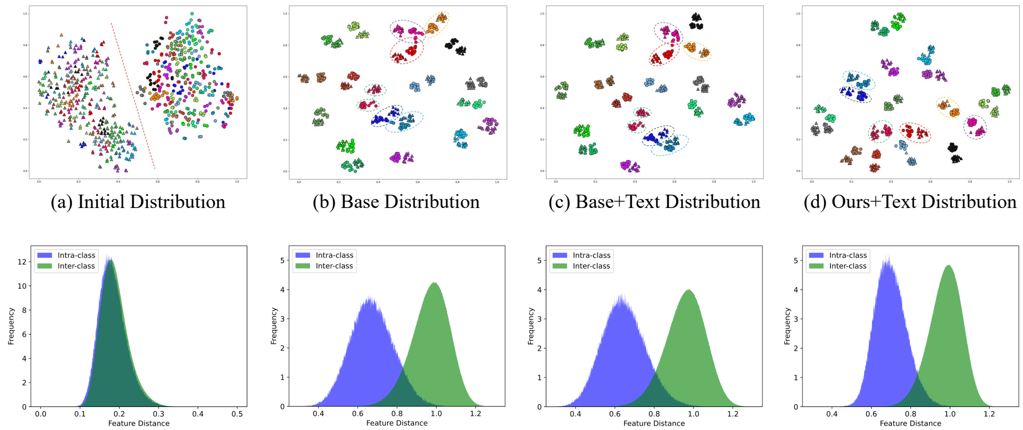
In this section, we present a comprehensive comparison of the proposed TVI-FM, against state-of-the-art models across different datasets as outlined in Table 3 and Table 4. Our evaluation includes a variety of metrics: Rank-1 (R-1), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP).

Performance on Tri-SYSU-MM01 Dataset As shown in Table3, with auxiliary textual descriptions, TVI-FM outperforms all previous methods under 'All Search' and 'Indoor Search' conditions. Specifically, TVI-FM achieves a significant improvement in Rank-1, reaching 84.90% and 89.06% respectively, compared to the next best result of 77.78% by PartMix in All Search and 82.07% by SGIEL in Indoor Search. Furthermore, in terms of mAP, TVI-FM posts scores of 81.47% and 90.78% which is a substantial increase from the previous high scores of 77.03% and 88.01%, respectively.

Performance on Tri-RegDB and Tri-LLCM Dataset Table 4 outlines our method's performance on the two datasets. In the Tri-RegDB dataset, TVI-FM obtains an Rank-1 of 91.38% and mAP of 85.92%, higher than the prior top scores of 90.95% in Rank-1 and 84.22% in mAP by YYDS. In the Tri-LLCM dataset, our method leads with an Rank-1 of 58.19% and mAP of 65.08%, surpassing the prior top scores of 58.13% in Rank-1 and 64.91% in mAP, both held by YYDS.

4.4 Visualization

Feature Distribution Visualization. To explore the reason why our method is effective, we utilize t-SNE[27] 2D feature space and visualize cosine distances of the intra-class and inter-class on the proposed Tri-SYSU-MM01 dataset in Fig.3. From the 'Initial' to 'Ours+Text' in 3(a-d), the t-SNE



(a) Initial Distribution (b) Base Distribution (c) Base+Text Distribution (d) Ours+Text Distribution

(e) Initial Distance (f) Base Distance (g) Base+Text Distance (h) Ours+Text Distance

Figure 3: Figures in the first row (a-d) show the t-SHE feature distribution of the 20 randomly selected identities, triangle means infrared features(w/wo textual fusion) and circle means visible features. Different colors indicate different identities. Figures in the second row (e-h) represent the intra-class(blue) and inter-class(green) distance of infrared features(w/wo textual fusion) and visible features.

feature distribution shows that with textual assistance, our method greatly enhances the ability of distinguishing features from different identities and reduces extreme outliers of the same identity and samples with too large cross modal discrepancy. While for feature distance distribution in 3(e-h), corresponding to 2D t-SHE[27] feature distribution, the inter/intra-class distance distributions are increasingly sparated well, especially, the situation of excessive intra-class distance has also been greatly reduced.

Retrieval Result. To intuitively present the performance of our method, we visualize some retrieval results of the Base VI-ReID model, Text-assisted base model and our method with text on the Tri-SYSU-MM01 dataset in Appendix F. For the same query image, with assistance of text description, the base model does a slightly better job of identifying samples that more closely correspond to the original RGB image, but there are still other identities in the results. Our method can mine the rich complementary information contained in text data to the maximum extent, the modality fusion greatly enhances the retrieval performance at fine-grained semantic level, even the failed retrieval samples still have high similarity with the target identity.

5 Conclusion

This paper proposes a novel framework of text-enhanced VI-ReID driven by Foundation Models (TVI-FM). VI-ReID often lags behind RGB-based ReID due to the inherent differences between modalities, particularly the absence of information in the infrared modality. Our method enriches the representation of the infrared modality by integrating automatically generated textual descriptions and augments them with LLM. We utilize a pretrained multimodal LVM to extract textual features and fine-tune the text encoder to minimize the domain gap between generated texts and original visual images. Leveraging LVMs' modality alignment capabilities and feature-level filters, this approach enables the text model to learn complementary features from the infrared modality, ensuring semantic structural consistency between the fusion modality and the visible modality. We further introduce modality joint learning to align features of all modalities, ensuring stable semantic representation of overall pedestrian appearance during complementary information learning. Moreover, a modality ensemble retrieving strategy is proposed to leverage the complementary strengths of each query modality, enhancing retrieval effectiveness and robustness. Extensive experiments on three expanded cross-modal re-identification datasets demonstrate significant improvements in retrieval performance, paving the way for utilizing foundation models in downstream data-demanding tasks.

References

[1] Hongchao Li Chenglong Li Bin Luo Chang Tan Ruoran Jia Aihua Zheng, Peng Pan. Progressive attribute embedding for accurate cross-modality person re-id. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4309–4317. ACM, 2022.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

[3] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022.

[4] Zhuxuan Cheng, Huijie Fan, Qiang Wang, Shiben Liu, and Yandong Tang. Dual-stage attribute embedding and modality consistency learning-based visible–infrared person re-identification. *Electronics*, 12(24), 2023.

[5] Neng Dong, Shuanglin Yan, Hao Tang, Jinhui Tang, and Liyan Zhang. Multi-view information integration and propagation for occluded person re-identification. *Information Fusion*, 104:102201, 2024.

[6] Yunhao Du, Zhicheng Zhao, and Fei Su. Yyds: Visible-infrared person re-identification with coarse descriptions, 2024.

[7] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11270–11279, October 2023.

[8] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22752–22761, June 2023.

[9] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11803–11812, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[12] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

[13] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18621–18632, June 2023.

[14] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal transformer with conv-patchify for person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 7317–7326, New York, NY, USA, 2022. Association for Computing Machinery.

[15] Huafeng Li, Yiwen Chen, Dapeng Tao, Zhengtao Yu, and Guanqiu Qi. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:1480–1494, 2021.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19366–19375, June 2022.

[18] Min Liu, Yeqing Sun, Xueping Wang, Yuan Bian, Zhu Zhang, and Yaonan Wang. Pose-guided modality-invariant feature alignment for visible–infrared object re-identification. *IEEE Transactions on Instrumentation and Measurement*, 73:1–10, 2024.

[19] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019.

[20] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17:605, 03 2017.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[22] Dai Pingyang, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. pages 677–683, 07 2018.

[23] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *European Conference on Computer Vision*, 2017.

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

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.

[26] A V Subramanyam, Niranjan Sundararajan, Vibhu Dubey, and Brejesh Lall. Iiitd-20k: Dense captioning for text-image reid, 2023.

[27] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[28] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[29] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5390–5399, 2017.

[30] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14308–14317, 2022.

[32] Mang Ye, Xiangyuan Lan, Jiawei Li, and P C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 04 2018.

[33] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13567–13576, October 2021.

[34] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 2020.

[35] 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*, 2022.

[36] 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.

[37] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. *Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification*, pages 462–479. 10 2022.

[38] 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.

[39] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021.

[40] Aihua Zheng, Peng Pan, Hongchao Li, Chenglong Li, Bin Luo, Chang Tan, and Ruoran Jia. Progressive attribute embedding for accurate cross-modality person re-id. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4309–4317, New York, NY, USA, 2022. Association for Computing Machinery.

[41] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Datasets Expansion

Given that there are almost no publicly available large-scale RGB-Text-Infrared person re-identification datasets up to now. The only existing VI-ReID dataset with text is labeled manually, YYDS[6] using only one Coarse description for all images with the same identity, which probably causes serious overfitting and cannot deal with the complex and various description in the real-world application. In order to get text data with various styles and rich semantic detail like for every RGB and IR image without any manual annotation, We construct three multi-modal dataset called Tri-SYSU-MM01, Tri-LLCM and Tri-RedDB from the original datasets SYSU-MM01[30], LLCM[38], RegDB[20] separately, following steps below:

1) Getting the LVM able to Generate Textual description from RGB images: We pre-trained Blip[16] on a large-scale pedestrian image-text dataset [26] to get the captioner for RGB modality.

2) Getting the LVM able to Generate Textual description from IR images: Firstly utilize the captioner for RGB modality we got before to generate textual descriptions from visible images in SYSU-MM01's training split, which contains various visible and infrared images for every identity. Then we remove rgb modal-related terms from these generated text by regular expression filter, build an IR-Text(filtered) dataset according to the same identity label shared by filtered text descriptions and infrared images. Finally we fine-tune the Blip[16] got from **step 1** on the IR-Text(filtered) dataset, get the captioner for IR modality

3) Getting Textual description from any dataset contains visible-infrared images: Utilize the refined LVM respectively we get in former steps as captioners for RGB modality and IR modality, to zero-shot generate text descriptions for datasets containing visible-infrared images.

The statisctics of our expanded dataset Tri-LLCM, Tri-RegDB and Tri-SYSU-MM01 are shown in Table 5. And the visualization on samples of our datasets are shown in 4

Table 5: Dataset statistics

Datasets	#ID	#RGB	#IR	#Text
Tri-LLCM	1064	25626	21141	46767
Tri-RegDB	412	4120	4120	8240
Tri-SYSU-MM01	491	30071	15792	45863

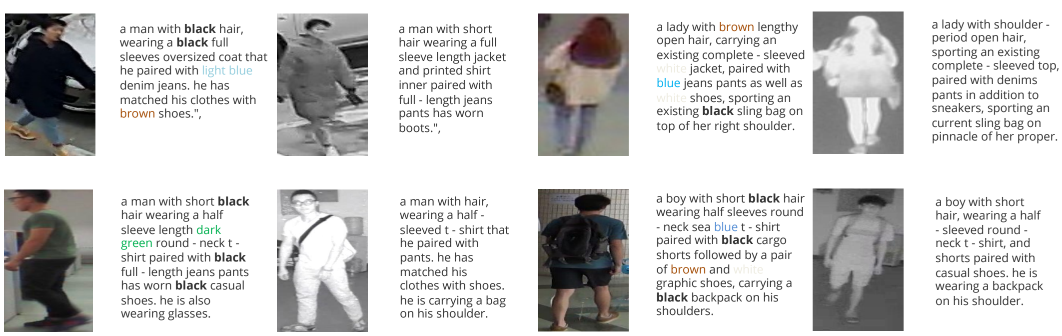


Figure 4: Visualization of the data samples selected from the expanded three datasets.

B Implementation Details

We implement our framework in PyTorch [21] utilizing a single NVIDIA RTX 3090 GPU for training. For visual backbone training, it takes about 9GB memory for training and about 3GB memory for testing, about 9 hours are needed for training on Tri-SYSU-MM01 and Tri-LLCM, about 1 hour for smaller Tri-RegDB. For incremental fine-tuing, it takes about 5GB memory for training and about 3GB memory for testing, about 1 hour are needed for fine-tuning on Tri-SYSU-MM01 and Tri-LLCM, about 10 miniutes for smaller Tri-RegDB. Each batch consists of 8 identities, with each identity containing 4 visible images, 4 infrared images, 4 text descriptions generated from visible images, and 4 text descriptions generated from infrared images. All input images are resized to 3 $\times$ 288 $\times$ 144, with full augmentation strategy as the same as CAJ [33]. All text descriptions are augmented by the proposed LLM rephrasing augmentation with a probability of 0.5, here we use vicuna-7b [41] as our LLM model. We employ a dual-stream resnet50 model [34] pre-trained on ImageNet [25] as the visual backbone and a transformer model with parameters derived from CLIP [24] as the textual backbone. For incrementally fine-tuning our TVI-FM, firstly we should get an available well-trained visual backbone. Here we utilize the augmentation method [33] to train the visual backbone for 120 epochs by cross-entropy loss and weighted regularized triplet loss, finally get the well-trained visual backbone. Then we integrate the well-trained VI-ReID model and fine-tune the textual backbone and a simple ReID bottleneck[19] applied for each feature for 20 epochs. We use the Adam[12] for optimization. For the Tri-SYSU-MM01 and Tri-LLCM datasets, in both visual and textual parts, the learning rate is set to 3.5e-4 and the weight decay to 5e-4. For the Tri-RegDB dataset, the learning rate for the visual part is 2e-3 with weight decay of 5e-4, and for the textual part, the learning rate is 1e-5 with weight decay of 4e-5. The learning rate rises up to the initial value by a linear warm-up scheme for the first 10 epochs, then decays by a linear scheme with a decay-factor of 0.1 at the milestones of 40, 60, and 100 epochs.

C Task Settings

C.1 Setting Details

We define $\mathcal{R} = \{V_r^n\}_{n=1}^{N^r}$ and $\mathcal{I} = \{V_i^n\}_{n=1}^{N^i}$ as the samples for visible and infrared images respectively, where each V_r^n and V_i^n represents an individual image from the visible and infrared modality. Textual descriptions $\mathcal{T}_r = \{T_r^n\}_{n=1}^{N^r}$ are autonomously generated for the visible images using a Language Vision Model (LVM). These descriptions provide detailed semantic insights about the identities, depicted across various textual styles. $N_v = N^r + N^i$ indicate the total numbers of samples in each modality. The identity labels for these images are given by $\mathcal{Y}^r = \{y_r^n\}_{n=1}^{N^r}$ for visible images and $\mathcal{Y}^i = \{y_i^n\}_{n=1}^{N^i}$ for infrared images. The task aims to retrieve visible representations $G^r = \{f_r^n\}_{n=1}^{N^r}$ through the infrared representations $Q^i = \{f_i^n\}_{n=1}^{N^i}$ or infrared representations enhanced by text representations $Q^t = \{f_{t_r}^n\}_{n=1}^{N^i}$. Inference is performed in two modes: one involves using a **fusion query**, as the same as the real-world conditions that descriptions provided by witnesses for the same target are diverse, each query $q_{sum} = \{f_i, f_{t_r}\}, f_i \in Q^i, f_{t_r} \in Q^t$ is composed by a infrared feature f_i and a corresponding randomly selected feature f_{t_r} of textual description. The other mode uses solely an **infrared query** $q_i = \{f_i\}, f_i \in Q^i$. The system then computes a ranking list based on the similarity of each query feature to gallery features $g = \{f_r\}, f_r \in G^r$ in visible modality, in order to find out all the person with the same identity.

D limitations and future research

While the TVI-FM framework has shown promising outcomes, two limitations still remain: 1)Its performance is linked to the quality of textual descriptions. High-quality textual descriptions will improve the accuracy of retrieval, which plays a crucial role in driving performance improvements in our framework. 2)Challenges persist in effectively handling challenging datasets such as LLCM[38]. Future researches on LLM and LVM is expected to generate higher-quality textual descriptions. Leveraging these advancements could lead to more robust and accurate retrieval results.

E Broader Impacts

Our TVI-FM framework offers significant advancements in urban security by enhancing person re-identification in low-light conditions, boosting surveillance effectiveness. It automates text generation from IR and RGB images, reducing annotation workload and improving text robustness, aiding multi-modal research and smart security system development. However, it's crucial to address environmental impact concerns related to large models' energy consumption and the privacy risks associated with re-identification technology. Governments and regulatory bodies must enact stringent regulations to prevent misuse and ensure identification accuracy to avoid societal disruptions.

F Retrieve Result Examples w/wo Text

Figure 5: Visualization of the rank-5 retrieval results obtained by the base model and our model on the proposed Tri-SYSU-MM01 dataset.

G Assets Details

This section provides the necessary details for the data assets utilized in our research: SYSU-MM01, LLCM, and RegDB.

- **SYSU-MM01**[30]
  - *Source and Citation*: The SYSU-MM01 dataset was created by researchers at Sun Yat-sen University (SYSU). Ancong Wu, et al. “RGB-IR Person Re-Identification by Cross-Modality Similarity Preservation” (2020) is the seminal paper associated with this dataset.
  - *data splits*: The training set contains 22,258 visible images and 11,909 infrared images of 395 identities. The testing set contains 96 identities, with 3,803 infrared images for query and 301 (single-shot) randomly selected visible images as the gallery set.
  - *URL*: The dataset can be accessed through a GitHub repository: `https://github.com/wuancong/SYSU-MM01` , where users must agree to the data release agreement.
  - *License*: We cannot find out the license SYSU-MM01 uses, but the author requires signing the usage agreement notice and contact him through e-mail to get the dataset. The detailed usage agreement refers to the github url mentioned above.
- **LLCM**[38]
  - *Source and Citation*: The LLCM dataset was introduced by researchers from Xiamen University. Yukang Zhang and Hanzi Wang’s paper “Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification” (2023) discusses this dataset.
  - *data splits*: The training set contains 30,921 images of 713 identities, and the test set contains 13,909 images of 351 identities.
  - *URL*: The dataset is available on GitHub `https://github.com/ZYK100/LLCM`.
  - *License*: CC-BY 4.0
  - *Code*: We use its code for feature visualization.
- **RegDB**[20]
  - *Source and Citation*: The RegDB dataset was developed at Dongguk University from the paper named "Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras".
  - *data splits*: The training set contains 206 identities and the testing set contains 206 identities. There are 10 visible images and 10 infrared images for each person.
  - *URL*: We can only find the paper’s doi `https://doi.org/10.3390/s17030605`
  - *License*: CC-BY 4.0

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: We demonstrate clearly our main claim that leveraging foundation models to generate enhanced and encoded textual modalities effectively addresses the challenges faced in VI-ReID and enhances retrieval performance. Our experimental results show significant improvements on retrieval accuracy across all three proposed datasets.

   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 discuss the limitation in the Appendix D.

   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: We don't have proofs.

   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 provides a clear and comprehensive description of the proposed TVI-FM architecture in section 3 with a figure 2, the method of expanding the existing open-source dataset in Appendix A as well as the complete implementation details of constructing the whole framework in Appendix B, training and testing, along with the detailed steps of the experiments in section 4. This ensures the replicability of our 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: [No]

Justification: We will release our code soon in the future.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All of experimental settings are shown insection 4.1, while the analysis of experiments results can be found in Ablation Study in section 4.2 and Comparison with state-of-the-art methods in section 4.3.

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: We don't have error bars, but in order to get stable performance, we evaluate our model for 10 times with random split of the gallery set in all experiments on Tri-SYSU-MM01 and Tri-LLCM datasets; for RegDB we evaluate our model on the 10 trials with different training/testing splits, and finally we report our model's average performance on each dataset, as the same as existing related works did.

   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: All the details can be found in Appendix B.

   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: Yes, the research conducted in the paper conform, in every respect, with 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: [Yes]

    Justification: We discuss the Broader Impacts in Appendix E

    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 pretrained language model[41] mentioned above we used are safe and come from open source community, and we don't post any new pre-trained language model.

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: Yes of course, we cite the author and owners for all used assets. And we also respect and follow all the license and terms of use explicitly mentioned. The detail of data and code assets we used are shown in Appendix G.

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: We introduced a method of expanding text modality from existing VI-ReID datasets. All the documentations of datasets we used can be viewed at the github urls of original datasets in Appendix G.

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: Our work didn't relate to any crowdsourcing and 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 human participants.

    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.