Empowering Visible-Infrared Person Re-Identification with Foundation Models

Zhangyi Hu[1*] **Bin Yang**[1] **Mang Ye**[1†]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Hubei Luojia Laboratory, Wuhan University, Wuhan, China.
{zhangyi_hu,yangbin_cv,yemang}@whu.edu.cn

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. The development of foundation models like Large Language Models (LLMs) and Language Vision Models (LVMs) motivates us to investigate a feasible solution to empower existing VI-ReID backbones 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 textual descriptions automatically generated by LVMs. Specifically, we incorporate a pre-trained LVM to extract textual features from descriptions generated by two modal-specialized fine-tuned LVM captioners and augmented by LLM. To enrich infrared features with generated textual information, we use modality alignment capabilities of LVMs and LVM-Generated feature-level filters to create a preliminary fusion modality. This enables the text model to learn complementary features according to the infrared modality, ensuring semantic consistency between the fusion and visible modalities. Then, modality joint learning aligns features of all modalities, incrementally fine-tunes text encoder to adapt to frozen VI-ReID backbone, maintaining stability of overall semantic of text representations while refining text-enriched infrared representations, thus minimizing the domain gap between enriched infrared and visible modalities. Additionally, Modality Ensemble Retrieving is proposed to leverage complementary strengths of each query modality to improve retrieval performance 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 [16, 13, 7, 17, 24], their effectiveness diminishes in low-light conditions at night. Infrared images provide a valuable visual alternative to visible images in low-light scenarios. Various techniques [**?** 39] have been proposed for visible-infrared person re-identification (VI-ReID), ensuring 24-hour surveillance. However substantial differences exist between infrared and visible modalities. This disparity mainly caused by the lack of detail in infrared images, poses significant challenges for current VI-ReID methods [33, 36, 39].

Most existing VI-ReID methods [39, 37, 15, 20, 38, 10, 35] do not adequately address the issue of information absence in the infrared modality. Instead, they primarily focus on extracting discriminative information shared between the infrared and visible modalities. However, the lack of information in

Preprint. Under review.

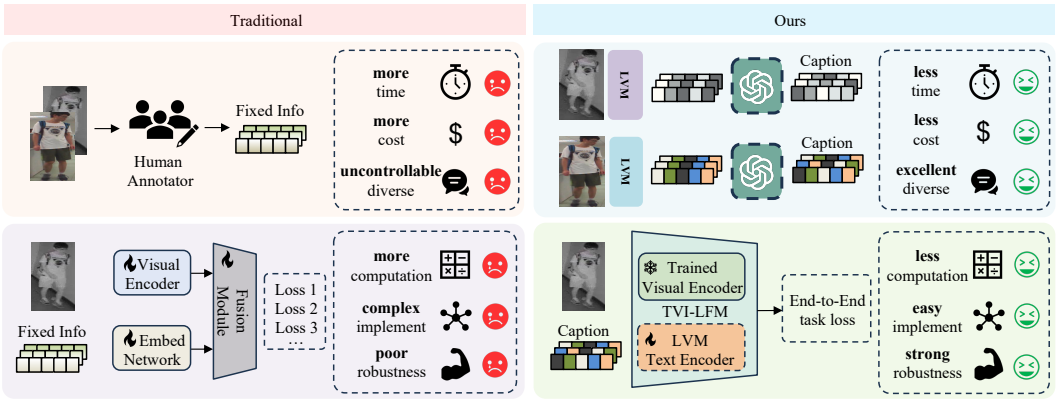


Figure 1: text-enhanced VI-ReID driven by foundation models compared to traditional methods

the infrared modality limits the performance of these methods. There also exist methods [8, 43, 5] that utilize auxiliary information like text descriptions [8] or attributes [43, 5] to enhance the infrared modality. However, as shown in Fig. 1, these methods heavily rely on human-annotated data, leading to significant time, labor costs and uncontrollable data quality. Moreover, due to neglecting the natural semantic connection between the auxiliary and visual modalities, they struggle to leverage the auxiliary data to enhance the VI-ReID task. Additionally, they rely on prior knowledge [8], complex model structures [8, 5], or manually designed loss functions [43] to extract auxiliary information, requiring training the whole framework from scratch to adapt the new information. These approaches cannot deal with the model's sensitivity to auxiliary data variations that may appear in real scenarios and result in considerable computational costs and limited performance.

Recent advancements in foundation models [2], particularly LLMs and LVMs, demonstrate strong inherent capabilities for text-visual modality alignment and personalized data generation, built on the extensive prior knowledge and powerful representation ability acquired during pre-training on vast datasets. Leveraging these capabilities, foundation models can automatically generate text from images and semantically link language with vision, providing vital complementary information for the infrared modality. This can effectively compensate for informational absence in infrared modality compared to the visible modality, thus bridging the gap between them. Motivated by this potential, to handle the challenges mentioned before, we introduce the Text-enhanced VI-ReID framework driven by Foundation Models (TVI-FM). This framework aims to seamlessly integrate auxiliary information into existing VI-ReID systems using off-the-shelf models, enhancing their overall performance by enriching the representation of the infrared modality with automatically generated textual descriptions. It comprises Modal-Specific Caption (MSC), Incremental Fine-tuning Strategy (IFS), and Modality Ensemble Retrieving (MER). Specifically, MSC is proposed to automate the generation of diverse textual descriptions from original infrared and visible image datasets. Initially, we pre-train a generative LVM on a massive pedestrian image-text dataset [28]. Using this model, we expand the text modality for randomly selected visible-infrared image pairs by generating descriptions from visible images and subsequently removing color-specific terms to create infrared-text pairs. These adapted pairs are then used to fine-tune a second model, resulting in two specialized models capable of producing accurate textual descriptions for both infrared and visible images. Additionally, to further enhance the textual quality and diversity, we employ an LLM to perform random paraphrasing augmentation on the generated text during the training process. This approach significantly reduces manual annotation efforts and improves the architecture's robustness against text variations. To enhance the utilization of auxiliary information generated by MSC, we introduce the Incremental Fine-tuning Strategy (IFS), which comprises Semantic Filtered Fusion (SFF) and Modality Joint Learning (MJL). SFF extracts features from text generated based on visible-infrared image pairs. Visible-derived textual features are then semantically filtered using infrared-derived textual features and combined with infrared visual features, enabling us to selectively merge the infrared modality and its complementary information. This process leverages the intrinsic language-vision alignment capabilities of foundational models to create preliminary fusion features sharing a similar semantic structure with visible modalities, setting the stage for further feature refinement. Built on this, MJL further refines these enhanced infrared features and text features in

an incremental manner. We freeze the VI-ReID backbone and focus solely on fine-tuning the text encoder of the foundation model through end-to-end task-oriented training for aligning textual and fusion representations with all other modalities. This enables the text model to effectively learn complementary information from infrared features while maintaining the overall semantic stability of the text, thus creating fusion features more similar to visible features. Utilizing the powerful capability of text-visual alignment in foundation model, IFS incrementally fine-tunes LVM textual encoder to achieve text-visual alignment on expanded VI-ReID datasets, thereby seamlessly integrating complementary textual information into the existing VI-ReID framework to enrich infrared modality, effectively bridging the gap between infrared and visible modalities. Additionally, to maximize the utilization of semantically-rich query representations derived from the Incremental Fine-tuning Strategy, we introduce the Modality Ensemble Retrieving (MER) strategy. This approach averages features from the infrared, textual, and fusion modalities to create a robust composite query feature, enhancing retrieval accuracy. MER capitalizes on the unique strengths of each modality: infrared features provide continuous visual semantics; textual features offer descriptive information; fusion features ensure sharing the same semantic structure with visible features. This strategy fully utilizes information from all modalities, improves the robustness and accuracy of the retrieval.

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 text, reducing the cost of text annotations and enhancing the performance of cross-modality retrieval.
- We develop novel modules including Modal-Specific Caption (MSC), Incremental Fine-tuning Strategy (IFS) and Modality Ensemble Retrieving (MER), utilizing off-the-shelf foundation models to seamlessly empower the performance of the existing VI-ReID framework from data, optimization and inference level
- Extensive experiments demonstrate that our method significantly improves retrieval performance on three expanded cross-modality re-identification datasets, paving the way for applying foundation models in downstream cross-modality 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 faces significant modality gaps cased by limited information in IR images. Previous works [39, 4, 40, 19, 37] attempt to bridge this gap by mining shared discriminative information, but the absence of information in IR images hampers performance. [8] introduces manually designed coarse descriptions and relies on costly metric learning and independent modules for auxiliary information integration. It neglects the semantic connection between auxiliary information and vision, depends heavily on prior knowledge and is sensitive to variations in auxiliary data. In contrast, our framework can automatically generate diverse textual descriptions to enhance infrared data. By integrating a text encoder capable of text-visual alignment with existing VI-ReID backbone, and incrementally fine-tuning the textual encoder using classic ReID loss, this approach seamlessly enhances existing VI-ReID backbone without the need for additional complex implementations, with better robustness against variations in auxiliary data.

2.2 Foundation Model

Foundation models, pre-trained on extensive and diverse datasets[2], have shown great potential across various domains. Recent advancements in Language-Vision Models (LVMs) like GIT[32], BLIP[18], and CLIP[25], alongside Large Language Models (LLMs) such as GPT-2[26], GPT-3[3], Vicuna[44], and LLaMa2[29], have demonstrated remarkable data generation and semantic understanding capabilities. For instance, BLIP[18] excels at generating relevant textual descriptions from images, which can be fine-tuned on diverse image styles, thus can handle different visual modalities. Vicuna[44], a leading LLM, leverages its extensive pre-training on textual data for sophisticated text manipulation without losing semantic integrity, ideal for personalized text enhancements. Similarly, CLIP[25]'s pre-training on large-scale image-text pairs has enabled its ability to align text-image modalities and embed features into the same semantic space, streamlining modality alignment. Building on these capabilities, our approach integrates generative LVMs and LLMs for automatic textual data generation and augmentation. We also incorporate a text encoder pre-trained on vision-language pairs into the traditional VI-ReID system, enhancing cross-modality performance with textual information.

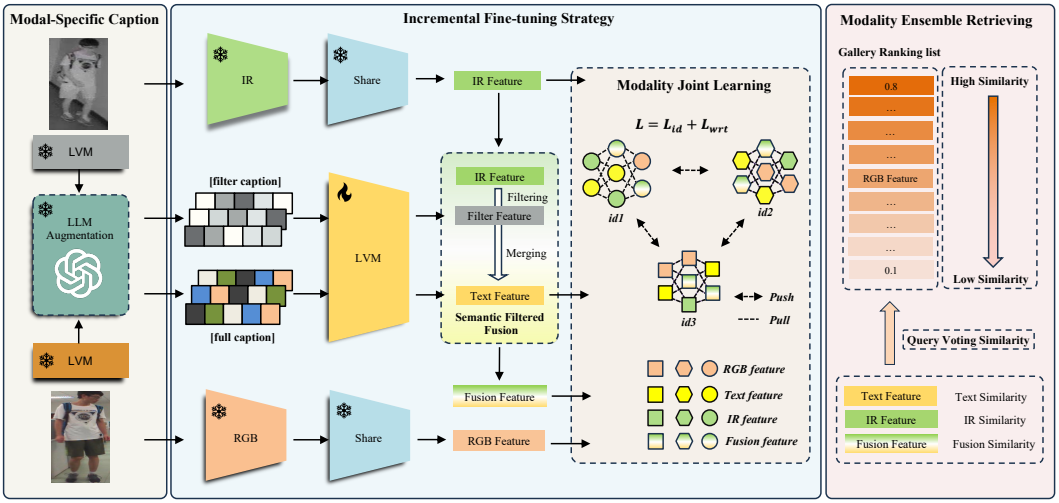


Figure 2: Illustration of TVI-FM: text-enhanced VI-ReID framework integrating a pre-trained LVM textual encoder[25] with a well-trained dual-stream visual backbone[12, 38].

3 Proposed Method

Task Setting. We utilize generated descriptions to enrich query representations for retrieval. Closely aligning with real-world application scenarios where witnesses provide varied descriptions to enhance query images for identifying individuals, each query q consists of an infrared sample V_i from the query set of original cross-modal datasets and a randomly selected textual description T_t generated from visible images for the same person, combining as $q = \{V_i, T_t\}$, while the gallery is composed by visible samples V_r. Then, we extract query and gallery representations through the model and compute ranking lists based on each query representation and all gallery representations as the retrieval results. Unlike traditional methods that solely utilize infrared queries, this query mode compensates for the infrared modality by integrating auxiliary information from descriptive sentences, having potential to enhance retrieval accuracy and robustness.

Overview. Our TVI-FM system, as depicted in Fig. 2, contains three parts. Modal-Specific Caption (MSC) employs LVMs to generate textual descriptions from visible and infrared images, subsequently utilizing LLM-based random rephrasing for their augmentation, thus creating diverse descriptions from cross-modal datasets without manual annotation efforts. Following this, the Incremental Fine-tuning Strategy (IFS) leverages the foundation model's intrinsic text-visual alignment capabilities to fine-tune the textual encoder on expanded datasets with representations of all modalities, seamlessly integrating complementary information from filtered textual features into the infrared modality to get fusion modality sharing the same semantic structure with visible modality. This effectively bridges the gap between infrared and visible modalities. To fully utilize all modalities' information, Modality Ensemble Retrieving (MER) aggregates features from all query modalities, thereby leveraging the combined strengths of each modality to improve retrieval accuracy and robustness.

3.1 Modal-Specific Caption

The Modal-Specific Caption (MSC) aims to automate the generation of high quality auxiliary descriptions for both infrared and visible images. It utilizes fine-tuned LVMs to generate text data from cross-modal datasets, then employs a Large Language Model (LLM) for textual augmentation, consequently creating diverse descriptions for visible and infrared images. This module reduces manual annotations and increases the system's robustness against auxiliary text variations.

LVM based Textual Generation. Due to the scarcity of large-scale Text-Visible-Infrared person re-identification datasets, we utilize Language Vision Models (LVMs) to generate textual descriptions for both visible and infrared images. Initially, we pre-train a Blip model on a vast pedestrian image-text dataset. We then randomly select visible-infrared image pairs from the SYSU-MM01's[34] training split and use this model to generate descriptions for visible images. After removing color-specific terms of the generated descriptions, we pair these adapted texts with corresponding infrared images in the same pairs to form an infrared-text sub-dataset. This sub-dataset was used to fine-

tune the Blip model, creating an infrared-specific captioner. This process enabled the autonomous generation of textual descriptions for datasets containing both modalities, eliminating the need for manual annotation. Through this method, we successfully constructed three cross-modal datasets: Tri-SYSU-MM01, Tri-LLCM, and Tri-RegDB. Further details are available in Appendix A.

LLM based Textual Augmentation. To ensure our framework extracting robust representations from generated textual descriptions against data variations, while preserving semantic integrity to enhance VI-ReID model, we implement an augmentation module based on LLM. This module regenerates more diverse descriptions for the same target, forcing the encoder to extract features with core semantics of person appearance. In detail, given an original description T, the module employs 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:

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

where $p = 0.5$ reflects the assumption that each description variant is equally probable. Utilizing the powerful prompt-driven text generation capability of LLM, this approach diversifies the textual descriptions while maintaining their core meanings. This forces the model to focus on extracting the core semantic of person appearance, thus enhancing the robustness of our system against text variation. Moreover, we can also apply this augmentation method directly on existing framework related with text data, without any change of the original structure.

3.2 Incremental Fine-tuning Strategy

The Incremental Fine-tuning Strategy (IFS) leverages the intrinsic capability of text-visual alignment contained in foundation models, aiming to integrate the complementary information of generated text into infrared modalities through incrementally tuning the textual encoder of LVM[25], thereby mitigating the gap between infrared and visible modalities. In detail, Semantic Filtered Fusion (SFF) refines textual features with filter features generated from infrared images, primarily creating enriched infrared features that share the same semantic structure as visible features. Modality Joint Learning (MJL) then optimizes the global association by aligning representations from all modalities, maximizing the effectiveness of information fusion in capturing complementary information for the infrared modality, while preserving the overall semantic stability of the text.

Semantic Filtered Fusion. Through LLM-based textual augmentation, the diverse textual descriptions can provide complementary information to infrared images. To leverage the intrinsic language-vision alignment capability for compensating infrared modality with complementary information from auxiliary textual descriptions, we introduce Semantic Filtered Fusion (SFF). First, we get features of infrared images by VI-ReID backbone and employ the textual encoder of a pre-trained LVM extract representations from descriptions generated for visible and infrared images. Moreover, leveraging the powerful text-visual capability, the textual encoder from the LVM maps textual features into a semantic space preliminarily aligned with vision, setting the stage for further refinement during the Modality Joint Learning (MJL) phase. Consequently, we can preliminarily form the fusion features jointly with text features and infrared features by summation and subtraction. However, while fusing the auxiliary information in text features f_t with infrared features f_i, they may contain redundant semantic details about the person's appearance, thus disrupting the semantic structural consistency with features from other modalities, so we selectively filter out redundant semantics in textual representations f_t that overlap with those found in infrared images by subtracting representations f_{filter} of descriptions generated for infrared images. Then we get the fusion features f_{sum} by composing the refined text features retaining rich complementary information with the infrared features:

$$f_{sum} = f_i + f_t - f_{filter} \tag{2}$$

This strategy leverages LVM's text-visual alignment capability to create fusion modalities. These preliminary fusion features composed of infrared feature and complementary information share the similar semantic structure with visible features, in following MJL, we refine the fusion modalities to further mitigate the difference between fusion features and visible features.

Modality Joint Learning. To further refine the fusion and textual features, thereby seamlessly integrating auxiliary information into existing VI-ReID frameworks, we propose Modality Joint Learning (MJL) for optimization. This strategy inherits the basic visual capabilities of the VI-ReID

system by freezing the backbone and focuses on incrementally fine-tuning the text encoder of the foundation model through end-to-end task-oriented training. This approach enables the text to align with vision and mine complementary information from text according to infrared modality. MJL optimizes the overall framework by adjusting the associations of representations from all modalities, including visible, infrared, textual, and fused modalities, using a task-oriented loss function. The total loss is composed of cross-entropy loss L_{id} and weighted regularized triplet loss L_{wrt}[39]:

$$L = L_{id} + L_{wrt} \tag{3}$$

Unlike other related works[8, 43, 5], leveraging the powerful intrinsic text-visual alignment capability of foundation models, MJL fine-tunes the whole framework with the classic ReID loss to integrate auxiliary textual information into infrared modalities and simultaneously mines discriminative information related to person identities, eliminating the need for manually designed processes or hyper-parameters. This strategy aligns representations across all modalities, enabling the text model to effectively learn complementary information according to infrared features while preserving the stability of the overall semantics of textual features. Consequently, it enhances infrared features to more closely resemble the visible modality, significantly bridging the gap between enriched infrared and visible modalities.

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 and robust retrieval, the Modality Ensemble Retrieving (MER) strategy is employed to comprehensively synthesize the unique and complementary advantages of different modalities. This involves averaging the features from the infrared modality f_i, textual modality f_t, and fusion modality f_{sum} to form a comprehensive query feature:

$$f_{agg} = mean(f_i, f_t, f_{sum}) \tag{4}$$

- ***Fusion features*** f_{sum} provide a comprehensive and enriched description of the target and aim to learn features with the same semantic structure as the visible modality, serving as the primary matching modality.
- ***Infrared features*** f_i provide valuable and contiguous visual semantics. Their similarity with visible images can serve as a supplementary reference for visual information.
- ***Textual features*** f_t provide descriptive details that may not be visually apparent or recognizable in infrared images. The similarity between textual features and visible features serves as an explicit reference for the missing or blurred appearance information in the infrared modality.
- ***The comprehensive features*** The retrieval results of f_{agg} and visible features f_r integrate the similarity scores of multiple query modalities with the visible modality to obtain a voting score, effectively harnessing the complementary strengths of each modality, reducing the potential impact of abnormal scores in sole-modal query-based retrieval lists through averaging with other modalities, and 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 cross-modal datasets with text description for each image are expanded from the original visible-infrared images datasets SYSU-MM01[34], RegDB[22], and LLCM[41] by the fine-tuned generative LVMs named Blip[18] in three stages (Details in Appendix A). The splits of the training set and testing set for each dataset are available in Appendix F.

Evaluation Protocols. In line with established VI-ReID settings [39, 37], we assess the performance of the infrared query mode and the textual enhanced infrared query mode using Rank-k matching accuracy, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP[39]) within our TVI-FM framework. To get stable performance on SYSU-MM01 and LLCM, we evaluate our model 10 times with random splits of the gallery set; as for RegDB, we evaluate our model on 10 trials with different training/testing splits. Finally, we report our model's average performance on each dataset.

Implementation Overview. We utilize a dual-stream ResNet-50[38] pretrained on ImageNet[27] as the visual backbone and a transformer in CLIP[25] as the textual encoder. Training involves visible and infrared images alongside text descriptions generated by two modality-specialized fine-tuned Blip[18] models, as detailed in Appendix A. All text descriptions are augmented by vicuna-7b[44]

Table 1: Ablation study on Text-enhanced Infrared query ($\boldsymbol{I + T \rightarrow R}$) 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	SFF	MJL	LLM	MER	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**

with a random rephrasing strategy. Incremental fine-tuning is applied by fixing the visual parameters and only tuning the textual part of the framework. All details are described in Appendix C.

4.2 Ablation Study

To thoroughly evaluate the effect of each component of our proposed method, we conduct comprehensive ablation studies on the Tri-LLCM and Tri-SYSU-MM01 datasets. These studies involve gradually adding the proposed modules to our baseline, systematically removing specific modules from our framework and assessing their impact on performance. The overall experimental setup remained consistent, with only the module under evaluation being modified.

Effect of Semantic Filtered Fusion. In order to form fusion queries sharing the same semantic structure with the visible modality, we implement a feature-level filtering mechanism utilizing LVM-generated filter features from IR images to compensate infrared features with filtered textual features. Compared with the baseline, the filter module enhances the framework's ability to comprehend the complementary semantics from text, while the baseline cannot effectively extract sufficient features from text. The method obtains a 4.48% Rank-1 improvement in Tri-SYSU-MM01 and a 1.90% Rank-1 improvement in Tri-LLCM, as shown in Table 1.

Effect of Modality Joint Learning. Incorporating SFF, and to further integrate auxiliary textual information into existing VI-ReID backbones, we propose Modality Joint Learning (MJL) to optimize the whole framework by aligning representations from all modalities. Based on the experimental results in Table 1, compared to the baseline only with filter mechanisms, adding this method gains a significant enhancement of 6.97% Rank-1 improvement, 6.67% mAP improvement, and 7.96% mINP improvement in Tri-SYSU-MM01, and 2.03% Rank-1 improvement, 2.63% mAP improvement, and 2.71% mINP improvement in Tri-LLCM.

Effect of Modality Ensemble Retrieving. The Modality Ensemble Retrieving strategy synthesizes the unique advantages of all query modalities, minimizing the potential impact of abnormal scores from single-modal queries 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. This demonstrates that the aggregation of different query modalities leads to more accurate and robust overall performance.

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

Discussion of Freezing Operation in IFS To seamlessly enhance the existing VI-ReID system with foundation models, we choose to fine-tune the textual LVM encoder and freeze the parameters of the existing VI-ReID model to inherit its capability of processing visual information and apply textual enhancement based on it. When we allow the visual backbone to update its parameters, as shown in the table from Appendix B, the performance of the integrated VI-ReID backbone suddenly declines by 5.43% and 4.24% in Rank-1 in the two datasets respectively. The performance of our

Table 2: 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 [33]	ICCV-17	$I \rightarrow R$	14.80	15.95	-	20.58	26.92	-
HCML [36]	AAAI-18		14.32	16.16	-	24.52	30.08	-
cmGAN [6]	IJCAI-18		26.97	27.80	-	31.63	42.19	-
AlignGAN [31]	ICCV-19		42.40	40.70	-	45.90	54.30	-
AGW [39]	TPAMI-21		47.50	47.65	35.30	54.17	62.97	59.23
DDAG [38]	ECCV-20		54.75	53.02	39.62	61.02	67.98	62.61
CM-NAS [11]	ICCV-21		61.99	60.02	-	67.01	72.95	-
DART [35]	CVPR-22		68.7	66.3	-	82.0	73.8	-
CAJ [37]	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 [41]	CVPR-23		74.70	71.80	-	80.30	83.30	-
SAAI [9]	ICCV-23		75.90	77.03	-	83.20	88.01	-
MSCLNet [40]	ECCV-22		76.99	71.64	-	78.49	81.17	-
SGIEL [10]	CVPR-23		77.12	72.33	-	82.07	82.95	-
PartMix [15]	CVPR-23		77.78	74.62	-	81.52	84.38	-
YYDS[8]	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 3: 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 [38]	ECCV-20	$I \rightarrow R$	68.06	61.80	48.62	40.3	48.4	-
AGW [39]	TPAMI-21		70.49	65.90	51.24	43.6	51.8	-
CAJ [37]	ICCV-21		84.8	77.8	61.56	48.8	56.6	-
DART [35]	CVPR-22		82.0	73.8	-	52.2	59.8	-
MMN [42]	MM-21		87.5	80.5	-	52.5	58.9	-
DEEN [41]	CVPR-23		89.5	83.4	-	54.9	62.9	-
YYDS[8]	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**

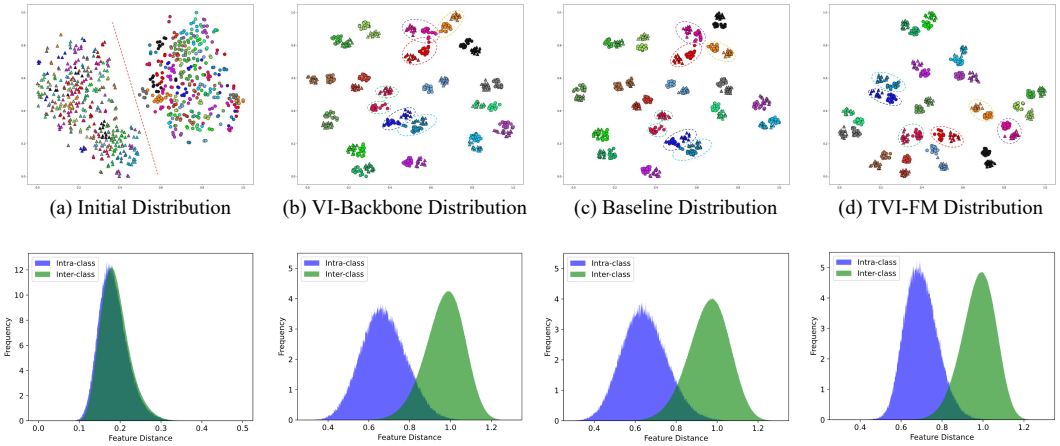
textually enhanced framework ($I + T \rightarrow R$) is also affected, with a decline of 0.14% Rank-1 in Tri-SYSU-MM01 and 1.66% Rank-1 in Tri-LLCM. This demonstrates the importance of freezing the integrated backbone to avoid the potential performance influence caused by conflicts of infrared feature learning and fusion feature learning during training.

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

In this section, we present comprehensive comparison of the proposed TVI-FM against state-of-the-art models on different datasets as outlined in Table 2 and Table 3. Our evaluation includes a variety of metrics: Rank-1 (R-1), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). For fair comparison, we re-run YYDS on our proposed Tri-modal datasets with the same image size: 288×144.

Performance on Tri-SYSU-MM01 Dataset As shown in Table 2, with the enhancement of generated text, TVI-FM greatly improves the performance of the VI-ReID backbone and outperforms all previous methods under 'All Search' and 'Indoor Search' conditions. Specifically, TVI-FM achieves significant improvements 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 are substantial increases from the previous high scores of 77.03% and 88.01%, respectively.

Performance on Tri-RegDB and Tri-LLCM Dataset Table 3 outlines our method's performance on the two datasets. In the Tri-RegDB dataset, TVI-FM obtains a Rank-1 of 91.38% and an 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



(a) Initial Distribution (b) VI-Backbone Distribution (c) Baseline Distribution (d) TVI-FM Distribution

(e) Initial Distance (f) VI-Backbone Distance (g) Baseline Distance (h) TVI-FM Distance

Figure 3: First row (a-d) show the t-SNE feature distribution of the 20 randomly selected identities, triangle means infrared features(w/wo textual enhancement), 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.

Tri-LLCM dataset, our method leads with a Rank-1 of 58.19% and an 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[30] 2D feature space and visualize cosine distances of the intra-class and inter-class features on Tri-SYSU-MM01 dataset in Fig. 3. From the 'Initial' to 'TVI-FM' in Fig. 3(a-d), the t-SNE feature distribution shows that our method greatly enhances the ability of distinguishing features from different identities with text and reduces extreme outliers of the same identity and samples with too large cross modal discrepancy. While for feature distance distribution in Fig. 3(e-h), corresponding to 2D t-SNE[30] feature distribution, the inter/intra-class distance distributions are increasingly separated well, especially, the 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 VI-ReID backbone, baseline and our method on the Tri-SYSU-MM01 dataset in Appendix E. For the same query image, our method significantly enhances retrieval performance utilizing generated descriptions compared to baseline and VI-ReID backbone.

5 Conclusion

This paper introduces a novel framework for text-enhanced Visible-Infrared Person Re-identification (VI-ReID) driven by Foundation Models (TVI-FM). Traditional VI-ReID often struggles compared to RGB-based ReID due to significant modality differences, notably the absence of information of infrared modality. Our approach addresses this by enriching the infrared modality with automatically generated textual descriptions. We incorporate a pre-trained LVM to extract textual features from descriptions generated by fine-tuned LVMs and augmented by LLM. To enrich infrared features with generated text, we use modality alignment capabilities of LVMs and LVM-Generated feature-level Filters to create preliminary fusion modality. This enables text model to learn complementary information according to infrared modality, ensuring semantic consistency between the fusion and visible modalities. Then, modality joint learning aligns features of all modalities by fine-tuning the text model to adapt to the frozen VI-ReID backbone, maintaining the stability of the overall semantics of text representations while refining text-enriched infrared representations, thus minimizing the domain gap between enriched infrared and visible modalities. Additionally, the Modality Ensemble Retrieving strategy enhances retrieval performance by leveraging the strengths of each query modality. Extensive experiments on three expanded cross-modal datasets demonstrate significant improvement in retrieval performance, paving the way for applying 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 *ACM MM*, 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, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy 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 Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark 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, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan 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 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*, 2022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NIPS*, 2020.

[4] 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 TIP*, 2022.

[5] 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*, 2023.

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

[7] Neng Dong, Shuanglin Yan, Hao Tang, Jinhui Tang, and Liyan Zhang. Multi-view information integration and propagation for occluded person re-identification. *IF*, 2024.

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

[9] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *ICCV*, 2023.

[10] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *CVPR*, 2023.

[11] 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 *ICCV*, 2021.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[14] 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*, 2017.

[15] 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 *CVPR*, 2023.

[16] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal transformer with conv-patchify for person re-identification. In *ACM MM*, 2022.

[17] 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 TIFS*, 2021.

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

[19] 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*, 2022.

[20] 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 TIM*, 2024.

[21] 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*, 2019.

[22] 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*, 2017.

[23] 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 *NIPS*, 2019.

[24] 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 *ECCV*, 2017.

[25] 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 *ICML*, 2021.

[26] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[27] 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. *IJCV*, 2014.

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

[29] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta,

Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, 2023.

[30] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[31] 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 *ICCV*, 2019.

[32] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *ArXiv*, 2022.

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

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

[35] 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 *CVPR*, 2022.

[36] Mang Ye, Xiangyuan Lan, Jiawei Li, and P C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[37] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, 2021.

[38] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, 2020.

[39] 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 TPAMI*, 2022.

[40] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *ECCV*, 2022.

[41] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, 2023.

[42] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *ACM MM*, 2021.

[43] 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 *ACM MM*, 2022.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NIPS*, 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[8] 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 fine-tune LVMs to generate textual descriptions for both visible and infrared images, thereby constructing three large scale tri-modality datasets: Tri-SYSU-MM01, Tri-LLCM and Tri-RegDB from the original datasets SYSU-MM01[34], LLCM[41], RegDB[22] separately. The details are introduced in following steps below:

1) *Getting the LVM ables to Generate Textual description from RGB images*: We pre-train Blip[18] on a large-scale pedestrian image-text dataset [28] to get the captioner for visible images.

2) *Getting the LVM able to Generate Textual description from IR images*: Firstly randomly select visible and infrared images pairs in SYSU-MM01's training split for every identity, then apply the captioner we got in **step 1** to generate textual descriptions for every visible images in these pairs. Then we remove color-related terms from these generated text by regular expression filter, build Infrared-Text(filtered) pairs dataset with filtered text descriptions and corresponding infrared images in the same expanded visible-infrared pairs. Finally we fine-tune the Blip[18] got from **step 1** on the IR-Text(filtered) dataset, get the captioner for infrared 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 4. And the visualization on samples of our datasets are shown in 4. All the fine-tuning process of LVMs are from documentations from huggingface `https://huggingface.co/docs/transformers/main/en/tasks/image_captioning`, the generator model we use refers to `https://huggingface.co/Salesforce/blip-image-captioning-large`.

Table 4: 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

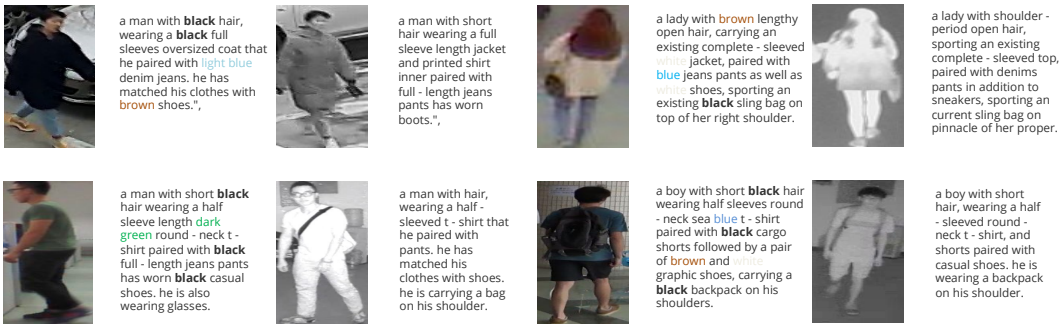


Figure 4: Visualization of the data samples selected from the expanded three datasets.

B Ablation Study of Freezing Operation in IFS

Table 5: The influence of whether to freeze visual backbone on case of infrared query ($\boldsymbol{I} \rightarrow \boldsymbol{R}$) and text-enhanced 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 **MER** 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**

C Implementation Details

We implement our framework in PyTorch [23] 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 [37]. All text descriptions are generated by two modality-specialized fine-tuned LVMs and augmented by the proposed LLM rephrasing augmentation with a probability of 0.5, here we use vicuna-7b [44] as our LLM model, use Blip[18] as our LVM model, whose tuning process can be found in AppendixA. We employ a dual-stream resnet50 model [38] pre-trained on ImageNet [27] as the visual backbone and a transformer model with parameters derived from CLIP [25] 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 [37] 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[21] applied for each feature for 20 epochs. We use the Adam[14] 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.

D 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.

E Retrieve Result Examples w/wo Text

Figure 5: Visualization of the rank-5 retrieval results obtained by the VI-ReID backbone on the proposed Tri-SYSU-MM01.

The VI-ReID backbone and baseline still includes misidentifications. But our method optimally leverages rich complementary information from textual data, significantly enhancing retrieval performance through modality fusion at a semantic level. It can be found that even the hard query samples that fail to retrieve correct targets still exhibit high appearance semantic similarity with the target identity.

F Assets Details

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

- **SYSU-MM01**[34]
  - *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**[41]
  - *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**[22]
  - *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

G 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)Improving room still exists in handling challenging datasets such as LLCM[41]. 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.

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 G.

   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 C, 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, hyperparameters, 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 C.

   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 D

    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[44] 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 F.

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 F.

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.