Empowering Visible-Infrared Person Re-Identification with Large Foundation Models

Zhangyi Hu[1*] **Bin Yang**[1*] **Mang Ye**[1†]
[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, China.
`{zhangyi_hu,yangbin_cv,yemang}@whu.edu.cn`
`https://github.com/WHU-HZY/TVI-LFM`

Abstract

Visible-Infrared Person Re-identification (VI-ReID) is a challenging cross-modal retrieval task due to significant modality differences, primarily caused by the absence of detailed color information in the infrared modality. The development of large foundation models like Large Language Models (LLMs) and Vision Language Models (VLMs) motivates us to investigate a feasible solution to empower VI-ReID performance with off-the-shelf large foundation models. To this end, we propose a novel Text-enhanced VI-ReID framework driven by Large Foundation Models (TVI-LFM). The basic idea is to enrich the representation of the infrared modality with textual descriptions automatically generated by VLMs. Specifically, we incorporate a pre-trained VLM to extract textual features from texts generated by VLM and augmented by LLM, and incrementally fine-tune the text encoder to minimize the domain gap between generated texts and original visual modalities. Meanwhile, to enhance the infrared modality with extracted textual representations, we leverage modality alignment capabilities of VLMs and VLM-generated feature-level filters. This allows the text model to learn complementary features from the infrared modality, ensuring the 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 retrieval strategy is proposed to leverage complementary strengths of each query modality 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 large foundation models in downstream data-demanding multi-modal retrieval tasks.

1 Introduction

Person Re-Identification (ReID) aims to retrieve images of the same identity across different cameras, which is an important task for intelligent surveillance and urban security [14, 62]. Although RGB-based methods [22, 16, 8, 23, 33, 25, 69, 55, 47, 63, 68, 13] have shown promising results during the daytime, their performance greatly diminishes at night, as RGB cameras fail to capture adequate information about a person in low-light conditions. Infrared cameras can obtain pedestrian appearance at dark environments. Therefore, Visible-Infrared Person Re-Identification (VI-ReID) is proposed to match images of individuals captured by visible and infrared cameras, enabling 24-hour surveillance. However, the absence of crucial information, such as color, in infrared images creates significant differences between infrared and visible modalities, posing the major challenge for VI-ReID.

*Equal Contribution.
†Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

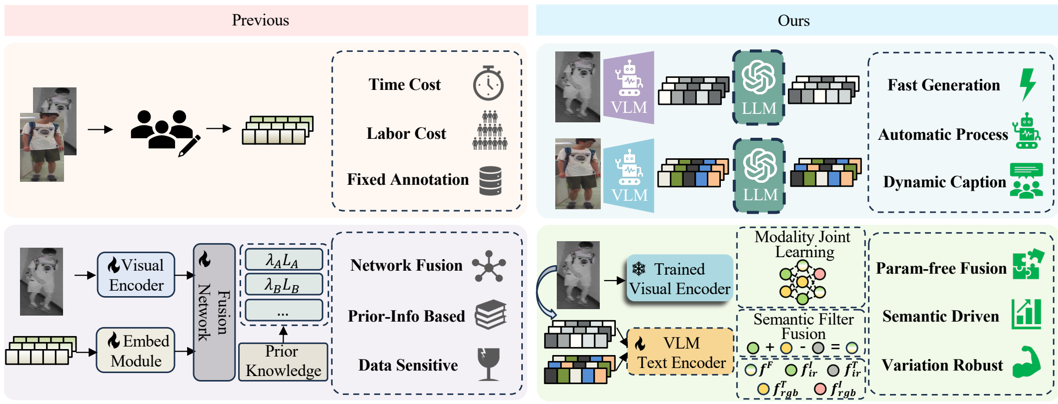


Figure 1: Illustration of our idea. Existing methods rely on fixed manual annotation, complex architecture and prior-knowledge-based optimization to enrich infrared modality. Leading to significant time and labor cost, additional parameters and data sensitivity. In contrast, our method employs VLM and LLM to automatically generate dynamic text, improving the robustness against text variation; fine-tunes a pre-trained VLM through aligning features across all modalities, enabling the framework to create fusion features semantically consistent with visible modality in a parameter-free manner.

Most existing VI-ReID methods, such as supervised methods [59, 57, 20, 56, 28, 58, 11, 53, 60], semi-supervised [39, 45] and unsupervised methods [51, 38, 37, 49, 50, 52, 48, 61], mainly focus on mining modality-shared features, while paying less attention to complementing information absence in the infrared modality, which limits further improvements in cross-modal retrieval performance.

In real scenarios, human descriptions are based on visible modality, providing rich detailed information, such as color, which can serve as vital auxiliary clues. However, existing methods that utilize auxiliary text descriptions [9] to enhance the infrared modality, as shown in Fig. 1, heavily rely on human annotating to collect fixed text descriptions, leading to significant time and labor costs. Moreover, they depend on prior knowledge, such as pre-defined color vocabularies or hyper-parameters, to design complex loss functions and modules with additional parameters for modality alignment and fusion. This reliance leads to sensitivity to data variations and diminishes the effectiveness of the fusion process.

Recent advancements in large foundation models [2], particularly LLMs and VLMs, show potential for data-demanding multi-modal retrieval tasks. This motivates us to explore a feasible solution to complement missing vital information with off-the-shelf large foundation models. To this end, we propose a text-enhanced VI-ReID framework driven by large foundation models (TVI-LFM) comprising Modality-Specific Caption (MSC), Incremental Fine-tuning Strategy (IFS), and Modality Ensemble Retrieval (MER). The basic idea is to enrich infrared representations with generated text, which is a cross-modality retrieval approach bolstered by heterogeneous text descriptions. Specifically, MSC employs LLM to augment VLM-generated texts for dynamic descriptions of visible and infrared images, reducing labor and time cost, and improving the model's robustness against text variation. Then, IFS incorporates a pre-trained VLM to extract features from text generated by MSC, and incrementally fine-tunes the text encoder to minimize the domain gap between the generated texts and the original visual modalities. To enhance the infrared modality with extracted textual representations, IFS leverages modality alignment capabilities of VLMs and VLM-generated feature-level filters to create a fusion modality. 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, IFS introduces modality joint learning to align features of all modalities, ensuring that textual features maintain a stable semantic representation of overall pedestrian appearance during complementary information learning. Additionally, MER is introduced to leverage complementary strengths of each query modality to form ensemble queries, further improving retrieval performance. Finally, by employing the above three modules for dynamic text generation, semantic alignment, and the integration of complementary queries, our method effectively addresses the information absence in the infrared modality, improving cross-modal retrieval performance.

The main contributions can be summarized as follows:

- We design a Text-enhanced VI-ReID framework driven by Large Foundation Models (TVI-LFM). It enriches infrared representations with generated textual descriptions, effectively mitigating the absence of critical information, e.g. color, in the infrared modality and significantly improving the performance of cross-modal retrieval.
- We propose IFS that fine-tunes a pre-trained VLM to align generated texts with original images. It creates a fusion modality to learn complementary information from the infrared modality and jointly align features across all modalities. It ensures stable semantic consistency of text and fusion features with the visible modality during complementary information learning.
- We propose Modality Ensemble Retrieval that leverages the complementary strengths of all query modalities to form ensemble queries, further improving the performance of cross-modality retrieval bolstered by heterogeneous text descriptions.
- We introduce three extended VI-ReID datasets with VLM-generated textual descriptions for every image. Extensive experiments on these expanded datasets demonstrate the competitive performance of our TVI-LFM framework, paving the way for utilizing large foundation models in downstream data-demanding multi-modal retrieval tasks.

2 Related Work

2.1 Visible-Infrared Person Re-Identification

VI-ReID aims to retrieve images across visible and infrared modalities, suffering from the absence of critical information, e.g. color, in infrared modality. Previous methods [59, 4, 64, 27, 57] focus on mining modality-shared information and optimizing features extracted by CNNs or Transformers [54] but pay less attention to complementing the missing vital information in the infrared modality, limiting the further improvements of the retrieval performance. Some methods [9, 5, 1] explore auxiliary information compensation. Specifically, [9] uses coarse descriptions as textual identity labels, while [5] and [1] integrate attribute embeddings with visual features. These methods heavily rely on handcrafted annotations for each identity. Furthermore, they require prior knowledge, such as hyper-parameters and pre-defined color vocabulary, to design complicated modules and loss. This results in additional parameters and increases sensitivity to auxiliary data variations. In contrast, we propose a framework that automatically expands dynamic textual descriptions for VI-ReID datasets and fine-tunes a pre-trained VLM to align the generated texts with original images. By leveraging VLM's modality alignment capabilities and feature-level filters, it creates a fusion modality that enables the text model to learn complementary features from the infrared modality, while maintaining semantic consistency with the visible modality.

2.2 Large Foundation Model

Large foundation models [2], pre-trained on extensive datasets, have shown great potential in downstream tasks. Recent advancements in VLMs like [44, 24, 34, 7, 26] and LLMs such as [35, 3, 67, 41, 18], demonstrate remarkable data generation and text-visual alignment capabilities. For instance, BLIP [24] excels at generating textual captions from images, and can be fine-tuned to accommodate various image styles such as infrared images. Vicuna [67], pre-trained on extensive text data, is great at customized text generation and understanding with prompts. CLIP [34]'s pre-training on large-scale image-text pairs enables its basic capability to align text-image semantics. It can also be fine-tuned on downstream cross-modal retrieval tasks [19]. Our approach integrates generative VLMs and LLMs for automatic text generation and dynamic augmentation. And it incorporates a pre-trained VLM into the VI-ReID system to extract features from texts and utilize them to enrich infrared representations.

2.3 Multi-modal Analogical Reasoning in VI-ReID

As demonstrated in [30], language analogical reasoning in the language embedding space can be represented using vector arithmetic. For example, the analogy "*man is to woman as king is to* ?" can be solved by finding the word whose embedding vector is closest to:

$$\vec{v}_{\text{``king''}} - \vec{v}_{\text{``man''}} + \vec{v}_{\text{``woman''}}. \tag{1}$$

Subsequently, [21] investigates an identical regularity manifested in the multi-modal vector space. For instance, the relationship can be expressed as:

$$\vec{v}_{\text{image(blue car)}} - \vec{v}_{\text{``blue''}} + \vec{v}_{\text{``red''}} \approx \vec{v}_{\text{image(red car)}}. \tag{2}$$

We establish this property by aligning features across all modalities, subsequently mapping them into a unified embedding space. This enables us to create fusion features in a parameter-free manner.

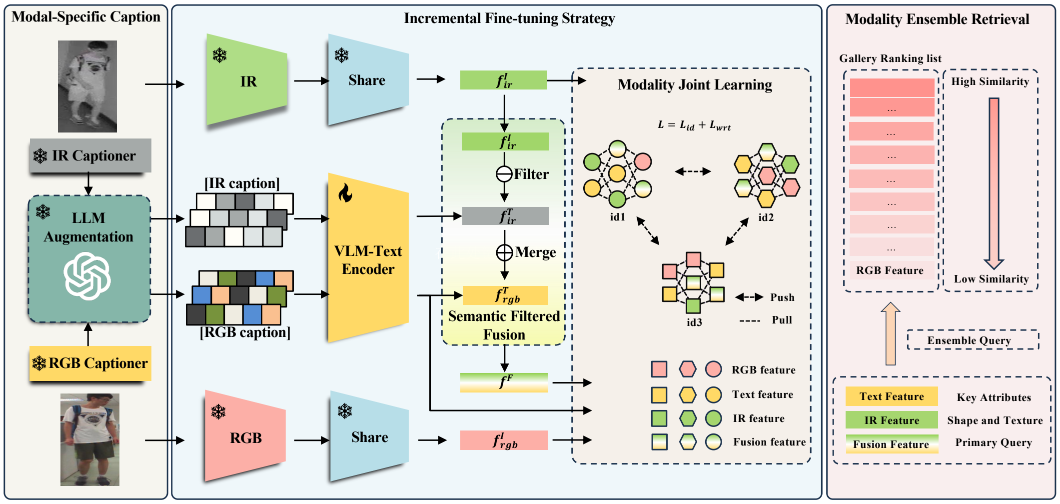


Figure 2: Illustration of our TVI-LFM, including Modality-Specific Caption (MSC), Incremental Fine-tuning Strategy (IFS), and Modality Ensemble Retrieval (MER). MSC utilizes fine-tuned VLMs as modal-specific captioners and employs an LLM for augmentation. IFS fine-tunes a pre-trained VLM to create fusion features semantically consistent with visible features. MER leverages the strengths of all query modalities to form ensemble queries, thereby improving the retrieval performance.

3 Proposed Method

Task Setting. Humans can provide textual descriptions based on visible modality. Containing critical information, such as colors, they can serve as auxiliary clues for identifying individuals. Therefore, we propose the cross-modality retrieval bolstered by heterogeneous text descriptions. During inference, each query sample consists of an infrared image V_{ir}, a text description T_{ir} generated from V_{ir}, and a randomly selected textual description T_{rgb} of visible images for the same person. These elements combine to form the query $q = \{V_{ir}, T_{ir}, T_{rgb}\}$. The gallery, meanwhile, contains only visible images represented as $g = \{V_{rgb}\}$. Then, compute similarity ranking lists based on each query and all gallery representations as results.

Overview. Detailed in Fig. 2, TVI-LFM contains MSC, IFS and MER. MSC employs two fine-tuned Blips [24] to automatically generate textual descriptions from visible/infrared images and utilizes LLM for augmentation. IFS trains a VI-ReID backbone to extract vision features, and incrementally fine-tunes a pre-trained CLIP [34] to align generated texts with original images. Then, it creates a fusion modality to learn complementary features from the infrared modality and jointly align features across all modalities. It ensures a stable semantic consistency of text and fusion features with the visible modality during complementary information learning. Additionally, MER leverages the strengths of each query modality, forming ensemble queries for more accurate retrieval.

3.1 Modal-Specific Caption (MSC)

MSC utilizes fine-tuned VLMs to automatically generate text from visible and infrared images and employs an off-the-shelf LLM for textual augmentation, consequently creating dynamic descriptions for VI-ReID datasets. This module reduces the time and labor cost of manual annotations and increases the system's robustness against auxiliary text variations.

VLM based Textual Generation. Currently, there are no publicly available large-scale VI-ReID datasets with image-level text annotations. Thus, to reduce labor and time costs, we fine-tune two VLMs to automatically generate text descriptions with critical information, such as color, for every visible and infrared image. Consequently, we construct three expanded datasets: Tri-SYSU-MM01, Tri-LLCM, and Tri-RegDB, each derived from the original datasets [46, 65, 31] respectively.

1) RGB Captioner: At the beginning, train a Blip [24] on a large-scale pedestrian image-text dataset [40] as the RGB captioner, which is able to generate text descriptions for visible images.

2) IR Captioner: Then, randomly select visible and infrared images pairs in SYSU-MM01's training split for every identity, then apply the RGB captioner in step 1 to generate textual descriptions for every visible image in these pairs. Then, remove color-related terms from these generated texts by regular expression filters, and build Infrared-Text (filtered) pairs dataset with filtered text descriptions and corresponding infrared images in the same expanded visible-infrared image pairs. Finally we fine-tune the Blip [24] in step 1 again on the Infrared-Text (filtered) dataset as the IR Captioner, which is able to generate text descriptions without color for infrared images.

3) Text Expanding: Finally, utilize the two refined modality-specific captioners in former steps to generate text descriptions for visible and infrared images.

The statistics and samples visualization of the expanded datasets Tri-LLCM, Tri-RegDB and Tri-SYSU-MM01 are shown in Appendix A. Through this expansion process, the framework can automatically generate text descriptions for VI-ReID datasets without large-scale manual annotations.

LLM based Textual Augmentation. To ensure that the framework can extract robust representations from generated descriptions against text variations while preserving original semantics of sentences, we propose the LLM-based textual augmentation module applied during the training stage. This module regenerates diverse descriptions by rephrasing the original text for the same target. In detail, given an original description T, the module employs an LLM to rephrase it, producing an augmented textual description T^*. The LLM is guided by the prompt *"Rephrase the person's description using similar words, without changing the original semantics"*. The augmentation is applied as follows:

$$T^* = \begin{cases} llm(T \mid Prompt), & \textit{with probability } p \\ T, & \textit{with probability } 1-p \end{cases}, \tag{3}$$

where $p = 0.5$ reflects 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 original meanings. This forces the model to focus on the core information of person appearance, thus enhancing the robustness of our system against text variation. Moreover, we can also apply this augmentation method directly on existing frameworks involving text data processing, without changing the original structure.

3.2 Incremental Fine-tuning Strategy (IFS)

IFS incrementally fine-tunes a CLIP [34] based on the frozen visual features extracted by a trained VI-ReID backbone, to minimize the domain gap between the generated texts and original visual modalities, namely, to align the complementary feature across vision and text, thereby creating fusion features semantically consistent with visible modality. The detailed steps are as follow:

Features Extraction. IFS first utilizes the Channel Augmentation (CA) [57] strategy to train a dual-stream ResNet-50 [15] as the VI-ReID backbone, detailed in Appendix B. Then, fix its parameters and use this visual backbone to extract infrared features f_{ir}^I and visible features f_{rgb}^I. Meanwhile, IFS incorporates a CLIP [34] to extract the features of visible images descriptions f_{rgb}^T as "text features" and the features of infrared images descriptions f_{ir}^T as "filter features".

Semantic Filtered Fusion (SFF). Meanwhile, to enhance the infrared modality with the generated texts, we propose the SFF module that leverages the text-visual alignment capability of VLM and VLM-generated filter features to create fusion features.

Benefiting from the large-scale pre-training on image-text pairs, the VLM possesses powerful text-visual alignment capability, ensuring that the features extracted from the generated texts contain the same information as the features of original images. Therefore, we regard the text feature f_{rgb}^T as an alternative for the visible feature f_{rgb}^I. Similarly, we use the filter feature f_{ir}^T as an alternative for the infrared feature f_{ir}^I. Next, we formulate the complementary features for the infrared modality by decomposing the visible feature f_{rgb}^I and the text feature f_{rgb}^T as:

$$f_{rgb}^I = f_{ir}^I + f_{comp}^I, \tag{4}$$

$$f_{rgb}^T = f_{ir}^T + f_{comp}^T, \tag{5}$$

where f_{comp}^I denotes the visual complementary feature for the infrared modality. Similarly, f_{comp}^T denotes the textual complementary feature for the infrared modality. Finally, with Eq. (4), Eq. (5), and the alignment of generated texts and original images benefiting from the pre-trained VLM, the representation of fusion features f^F with the same semantic structure as the visible modality can be

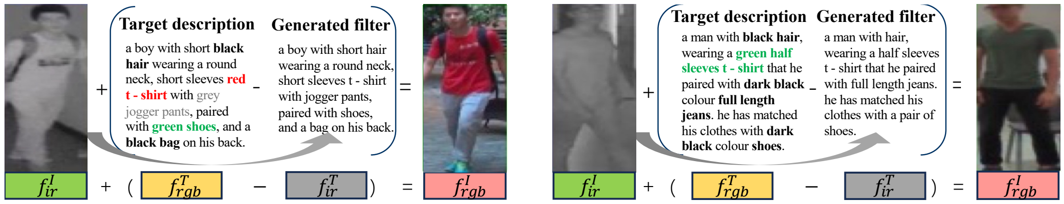


Figure 3: The Visualization of SFF. With the aligned features of generated texts and original images, SFF creates fusion features semantically consistent with visible modality by arithmetically adding the textual complementary information for infrared modality to the infrared features.

derived, represented as:

$$
\begin{aligned}
f^I_{rgb} &= f^I_{ir} + f^I_{comp} \\
&= f^I_{ir} + (f^I_{rgb} - f^I_{ir}) \\
&\approx f^I_{ir} + (f^T_{rgb} - f^T_{ir}). \\
&= f^I_{ir} + f^T_{comp} \\
&\triangleq f^F
\end{aligned}
\tag{6}
$$

By leveraging the powerful text-visual alignment capability of VLM and VLM-generated filter features, SFF approximates visual complementary features for infrared modality $f^I_{rgb} - f^I_{ir}$ with the textual complementary features for the infrared modality $f^T_{rgb} - f^T_{ir}$. Thus, the framework can create fusion features by adding this textual complementary features to the infrared features, as shown in Fig.3. This allows the text model to learn complementary features from the infrared modality, while ensuring semantic structural consistency between the fusion modality and the visible modality.

Modality Joint Learning (MJL). Furthermore, MJL is proposed to optimize the pre-trained VLM. It incrementally fine-tunes the CLIP based on the infrared and visible features extracted from the frozen VI-ReID backbone trained before, thereby refining fusion features to be further aligned with visible modality. This training strategy and its effectiveness on avoiding the conflicts during the representations learning of visual and textual part are discussed in Appendix E.

This method utilizes a classic ReID loss for fine-tuning, thereby eliminating the prior knowledge reliance, such as hyper-parameters and pre-defined vocabulary, during optimization. The loss consists of cross-entropy loss L_{id} and weighted regularized triplet loss L_{wrt} [59], represented as:

$$
L_{total} = L_{id}(y, f^*) + L_{wrt}(y, f^*), f^* \in \{f^T_{rgb}, f^I_{rgb}, f^I_{ir}, f^F\}, \tag{7}
$$

where f^I_{rgb} denotes the **frozen** visible feature, f^I_{ir} denotes the **frozen** infrared feature, f^T_{rgb} denotes the **trainable** text feature, f^F denotes the **trainable** fusion feature. During this process, MJL pulls all these features of the same identity y together and pushes them away from features of different identities, aligning the semantics across all modalities. Since we solely adjust the parameters of the text encoder, this alignment can be regarded as refining both the text features f^T_{rgb} and fusion features f^F, enabling them to learn representations equivalent to visible features f^I_{rgb}, under the supervision of the identity label y and with a regularization from f^T_{ir}.

By aligning the textual complementary features for infrared modality f^T_{comp} integrated in the fusion features with the corresponding part f^I_{comp} in the visible features, MJL effectively reduces the discrepancy between fusion and visible modality. Additionally, it ensures the overall semantic consistency of text features f^T_{rgb} with visible features f^I_{rgb} during complementary information learning, which enables the following MER to form effective ensemble queries.

Thereby, considering the established two alignments above while taking into account the feature decomposition in Eq. (4) and Eq. (5), the alignment between infrared features f^I_{ir} and filter features f^T_{ir} can be derived and represented as:

$$
\begin{aligned}
f^I_{ir} &= f^I_{rgb} - f^I_{comp} \\
&\approx f^T_{rgb} - f^T_{comp}. \\
&= f^T_{ir}
\end{aligned}
\tag{8}
$$

Consequently, by establishing the alignment between f^I_{rgb} and f^T_{rgb}, and ensuring the alignment between f^I_{ir} and f^T_{ir}, we successfully minimize the domain gap between the generated texts and original visual modalities, namely, align each part of the visual and textual complementary features, thereby enabling the alignment between fusion features f^F and the visible features, leading to improved retrieval accuracy.

According to Eq. (4) and Eq. (5), f^T_{rgb} can be decomposed as trainable f^T_{ir} and f^T_{comp}, while f^F can be decomposed as frozen f^I_{ir} and trainable f^T_{comp}. Considering that frozen infrared features f^I_{ir} also participate in aligning with the text features f^T_{rgb} and the fusion features f^F, it can be regarded as a regularization to constrain the complexity of textual complementary features f^T_{comp} while aligning with filter features f^T_{ir}, enabling the framework to construct better fusion features.

3.3 Modality Ensemble Retrieval (MER)

MER aims to comprehensively utilize the complementary advantages and rich semantics in query modalities mined from IFS to form ensemble query features f_q for more accurate retrieval. The ensemble query feature f_q is represented as:

$$f_q = (f^I_{ir} + f^T_{rgb} + f^F)/3, \tag{9}$$

where ***fusion features*** f^F aim to provide a combined feature with semantic structure the same as the visible features, serving as the primary matching modality. ***Infrared features*** f^I_{ir} provide contiguous visual semantics. Their similarity with visible images can serve as a supplementary reference for texture and shape information. ***Text features*** f^T_{rgb}, aligned with visible features in MJL, provide descriptive information such as the color and the pattern of the clothes. The similarity between text features and visible features serves as a reference for these key matching features. Therefore, when encountering challenging scenarios that are hard to distinguish a person's clothing or shape but distinguishing color is feasible, or vice versa, the ***ensemble query features*** f_q are able to leverage the features of two modalities in addition to the primary matching contribution from fusion modality to explore their similarities in visual texture and key attributes respectively. Thus, the similarity score s is defined as the dot product of f_q and f_{rgb}. Thus, the strengths of all query modalities can be integrated into the final similarity score, consequently enhancing the accuracy against hard cases of retrieval. In fact, by plugging Eq. (9) into the definition of the similarity score $f_q \cdot f_{rgb}$, an equivalent definition as the similarity between two high-dimension features is derived, represented as:

$$s = ([f^I_{ir}, f^T_{rgb}, f^F] \cdot [f^I_{rgb}, f^I_{rgb}, f^I_{rgb}])/3. \tag{10}$$

It is equivalent to increasing the feature dimension for retrieval but use ensemble features with less dimensions and computational cost. Due to the larger distances between classes in higher dimensional feature space, the models can more easily distinguish features of different identities in the ensemble feature space, therefore utilizing ensemble queries can further improve the retrieval performance.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We evaluate our framework on Tri-SYSU-MM01, Tri-RegDB, and Tri-LLCM. These datasets with text descriptions for each visible and infrared image are expanded from the original VI-ReID datasets [46, 65, 31], utilizing fine-tuned Blip [24] as captioner. The splits of the training set and testing set for each dataset are available in Appendix D.

Evaluation Protocols. In line with established VI-ReID settings [59, 57], we assess the performance of the infrared query and the fusion query using Rank-k matching accuracy, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [59] within our TVI-LFM framework. To get stable performance on Tri-SYSU-MM01 and Tri-LLCM, we evaluate our model 10 times with random splits of the gallery set; as for Tri-RegDB, we evaluate our model on 10 trials with different train/test splits and report the average performance on each dataset.

Implementation Overview. We utilize a dual-stream ResNet-50 [58] pre-trained on ImageNet [36] as the visual backbone and a transformer in CLIP [34] as the textual encoder. Training involves visible and infrared images alongside text descriptions generated by two modality-specialized fine-tuned Blip [24] models. All text descriptions are augmented by vicuna-7b [67] with a random rephrasing strategy. Incremental fine-tuning is applied by fixing the visual parameters while tuning the textual part of the framework. All details are described in Appendix B.

Table 1: Ablation study on fusion query ($\boldsymbol{I} + \boldsymbol{T} \rightarrow \boldsymbol{R}$) 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	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**

4.2 Ablation Study

To thoroughly evaluate the effect of each component of our proposed method, we conduct comprehensive ablation studies on Tri-LLCM and Tri-SYSU-MM01. 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. By leveraging text-visual modality alignment capability of VLM and VLM-generated filter features, SFF fuses textual complementary information with infrared features to create fusion features semantically consistent with visible modality. Compared to the baseline, the method obtains a 4.48% Rank-1 improvement in Tri-SYSU-MM01 and a 2.10% Rank-1 improvement in Tri-LLCM, as shown in Table 1. The results demonstrate that the module effectively integrates information from different modalities.

Effect of Modality Joint Learning. In cooperation with SFF, MJL aligns features across all modalities to minimize the domain gap between the generated texts and original visual modalities, thereby mitigating the fusion-visible modality difference. Based on the experimental results in Table 1 and compared to the baseline with SFF, adding MJL 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. This result demonstrates the effectiveness of MJL, which greatly minimizes the modality gap.

Effect of Modality Ensemble Retrieval. MER leverages the complementary advantages of different query modalities to construct ensemble query features thereby improving the retrieval accuracy. The results demonstrate its effectiveness, in Table 1, incorporating MER provides an additional improvement of 0.91% in Rank-1, 0.92% in mAP, and 1.11% in mINP in the Tri-SYSU-MM01 dataset over the baseline with MJL+SFF. Similarly, on the Tri-LLCM dataset, MER achieves 0.33% Rank-1 improvement, 0.29% mAP improvement, and 0.27% mINP improvement.

Effect of LLM based Textual Augmentation. To extract robust text representations against text variation, we implement LLM based augmentation module by randomly rephrasing the generated descriptions for dynamic expression. As shown in Table 1, incorporating it further improves the overall performance and robustness against text variation. It also works well with other modules, achieving 84.90% Rank-1 and 58.19% Rank-1 in Tri-SYSU-MM01 and Tri-LLCM respectively.

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

We present a comprehensive comparison of TVI-LFM against state-of-the-art methods 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) [59]. For fair comparison, we re-run YYDS on the proposed expanded 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-LFM greatly improves the performance of the VI-ReID backbone and outperforms all previous methods under 'All Search' and 'Indoor Search' conditions. Specifically, TVI-LFM achieves significant improvements in Rank-1, reaching 84.90% and 89.06% respectively, compared to the next best result of 77.78% by PartMix [20] in All Search and 83.20% by SAAI [10] in Indoor Search. Furthermore, in terms of mAP, TVI-LFM posts scores of 81.47% and 90.78%, which are substantial increases from the previous high scores of 77.03% and 88.01%, respectively.

Table 2: Comparison 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 [46]	ICCV-17	$I \to R$	14.80	15.95	-	20.58	26.92	-
HCML [56]	AAAI-18		14.32	16.16	-	24.52	30.08	-
cmGAN [6]	IJCAI-18		26.97	27.80	-	31.63	42.19	-
AlignGAN [43]	ICCV-19		42.40	40.70	-	45.90	54.30	-
AGW [59]	TPAMI-21		47.50	47.65	35.30	54.17	62.97	59.23
DDAG [58]	ECCV-20		54.75	53.02	39.62	61.02	67.98	62.61
CM-NAS [12]	ICCV-21		61.99	60.02	-	67.01	72.95	-
DART [53]	CVPR-22		68.7	66.3	-	82.0	73.8	-
CAJ [57]	ICCV-21		69.88	66.89	53.61	76.26	80.37	76.79
DEEN [65]	CVPR-23		74.70	71.80	-	80.30	83.30	-
SAAI [10]	ICCV-23		75.90	77.03	-	83.20	88.01	-
MSCLNet [64]	ECCV-22		76.99	71.64	-	78.49	81.17	-
SGIEL [11]	CVPR-23		77.12	72.33	-	82.07	82.95	-
PartMix [20]	CVPR-23		77.78	74.62	-	81.52	84.38	-
YYDS [9]	Arxiv-24	$I + T \to R$	74.60	70.35	56.01	81.35	83.64	79.56
VI-ReID Backbone	-	$I \to R$	69.89	66.74	53.34	76.91	80.64	76.70
TVI-LFM	-	$I + T \to R$	**84.90**	**81.47**	**70.85**	**89.06**	**90.78**	**88.39**

Table 3: Comparison with the state-of-the-art methods on the proposed Tri-RegDB and Tri-LLCM.

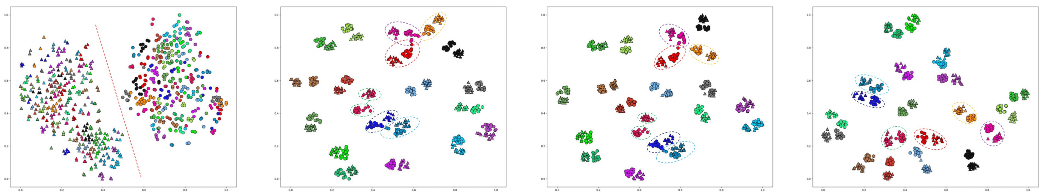
Methods	Venue	Type	Tri-RegDB			Tri-LLCM		
			R-1	mAP	mINP	R-1	mAP	mINP
DDAG [58]	ECCV-20	$I \to R$	68.06	61.80	48.62	40.3	48.4	-
AGW [59]	TPAMI-21		70.49	65.90	51.24	43.6	51.8	-
CAJ [57]	ICCV-21		84.8	77.8	61.56	48.8	56.6	-
DART [53]	CVPR-22		82.0	73.8	-	52.2	59.8	-
MMN [66]	MM-21		87.5	80.5	-	52.5	58.9	-
DEEN [65]	CVPR-23		89.5	83.4	-	54.9	62.9	-
YYDS [9]	Arxiv-24	$I + T \to R$	90.95	84.22	70.12	58.13	64.91	61.77
VI-ReID Backbone	-	$I \to R$	89.51	83.51	69.65	53.53	59.77	56.40
TVI-LFM	-	$I + T \to R$	**91.38**	**85.92**	**72.73**	**58.19**	**65.08**	**61.83**

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-LFM 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 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.

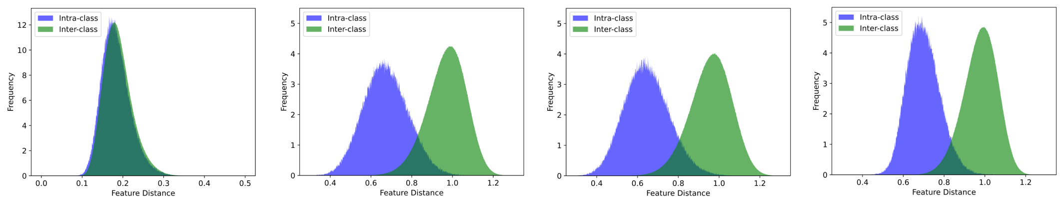
4.4 Visualization

Feature Distribution Visualization. To explore the reason why our method is effective, we utilize t-SNE [42] 2D feature space and visualize cosine distances of the intra-class and inter-class features on Tri-SYSU-MM01 dataset. From the (a) to (d) in Fig. 4, 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. For the feature distance distribution shown in Fig. 4 (e-h), which corresponds to the 2D t-SNE [42] feature distribution, the inter-class and intra-class distance distributions are increasingly well separated, particularly noting that the excessive intra-class distance is also significantly 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 C. For the same query image, our method significantly enhances retrieval performance utilizing generated descriptions compared to baseline and VI-ReID backbone.



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



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

Figure 4: First row (a-d) shows the t-SNE feature distribution of the 20 randomly selected identities, triangles means infrared features (w/wo textual enhancement), circles means visible features. Different colors indicate different identities. Figures in the second row (e-h) represent the intra-class (blue) and inter-class (green) distances of infrared features (w/wo textual fusion) and visible features.

5 Conclusion

To alleviate the absence of detailed color information in the infrared modality, this paper presents a VI-ReID framework driven by Large Foundation Models (TVI-LFM) to enrich the infrared representation with VLM-generated textual descriptions, which is a cross-modality retrieval approach bolstered by heterogeneous text descriptions. To enhance the infrared modality with text, MSC utilizes one off-the-shelf LLM to augment VLM-generated text descriptions. Then, IFS incorporates a pre-trained VLM to extract features from generated texts, and incrementally fine-tunes the text encoder to align generated texts and original visual modalities. To enhance the infrared modality with extracted textual representations, IFS leverages modality alignment capabilities of VLMs and VLM-generated feature-level filters to create fusion modality. 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, IFS introduces modality joint learning to align features of all modalities, maintaining a stable semantic representation of overall pedestrian appearance for text features, during complementary information learning. Additionally, MER leverages complementary strengths of query modalities to form ensemble queries, further improving retrieval performance. Extensive experiments on three expanded VI-ReID datasets demonstrate that our method achieves a competitive performance, paving the way for utilizing large foundation models in downstream data-demanding multi-modal retrieval tasks.

limitations and future research

While the proposed TVI-LFM shows promising performance, the retrieval accuracy hinges on text quality, and the performance on hard datasets, such as LLCM[65], still have rooms for improvement. High-quality text enhances retrieval accuracy by improving text-vision correspondences during training and providing precise information for infrared compensation during inference. Therefore, for future improvements, *1) more advanced generative models*; *2) image augmentations during generator fine-tuning*; *3) progressive generation strategies focusing on fine-grained attributes*, could be introduced to enhance text quality, thereby improving the accuracy of cross-modality retrieval bolstered by heterogeneous textual descriptions.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China under Grant (62176188, 62361166629, 62225113), Key Research and Development Project of Hubei Province (2022BAD175), and Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20241268). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Zheng Aihua, Pan Peng, Li Hongchao, Li Chenglong, Luo Bin, Tan Chang, and Jia Ruoran. 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, and et. al. On the opportunities and risks of foundation models. *ArXiv*, 2022.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, and et. al. Language models are few-shot learners. In *NeurIPS*, 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] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, 2023.

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

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

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

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

[12] 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.

[13] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *CVPR*, 2021.

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

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

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

[17] 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.

[18] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, and et. al. Mixtral of experts. *ArXiv*, 2024.

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

[20] 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.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *ArXiv*, 2014.

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

[23] 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.

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

[25] Mingkun Li, Chun-Guang Li, and Jun Guo. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. *IEEE TIP*, 2022.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[27] 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.

[28] 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.

[29] 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.

[30] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *ACL*, 2013.

[31] 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.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, and et. al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] Jiangming Shi, Xiangbo Yin, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *ArXiv*, 2024.

[39] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, 2023.

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

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, and et. al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, 2023.

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

[43] 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.

[44] 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.

[45] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Semi-supervised learning with heterogeneous distribution consistency for visible infrared person re-identification. *IEEE TIP*, 2024.

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

[47] Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE TIFS*, 2022.

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

[49] Bin Yang, Jun Chen, Cuiqun Chen, and Mang Ye. Dual consistency-constrained learning for unsupervised visible-infrared person re-identification. *IEEE TIFS*, 2024.

[50] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *ICCV*, 2023.

[51] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *CVPR*, 2024.

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

[53] Mouxing Yang, Peng Huang Zhenyu, Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, 2022.

[54] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *ArXiv*, 2024.

[55] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *ACM MM*, 2019.

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

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

[58] 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.

[59] 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.

[60] Mang Ye, Zesen Wu, Cuiqun Chen, and Bo Du. Channel augmentation for visible-infrared re-identification. *IEEE TPAMI*, 2024.

[61] Xiangbo Yin, Jiangming Shi, Yachao Zhang, Yang Lu, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. *ArXiv*, 2024.

[62] Zhiming Luo Yansong Qu Rongrong Ji Min Jiang Yunpeng Gong, Zhun Zhong. Cross-modality perturbation synergy attack for person re-identification. *ArXiv*, 2024.

[63] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiaoping Tan, Errui Ding, and Yi Yang. Understanding image retrieval re-ranking: A graph neural network perspective. *ArXiv*, 2020.

[64] 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.

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

[66] 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.

[67] 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 *NeurIPS*, 2023.

[68] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Weishi Zheng, and Xing Sun. Viewpoint-aware loss with angular regularization for person re-identification. *ArXiv*, 2019.

[69] Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, and Chi Zhang. Discrepant and multi-instance proxies for unsupervised person re-identification. In *ICCV*, 2023.

A Details of Expanded Datasets

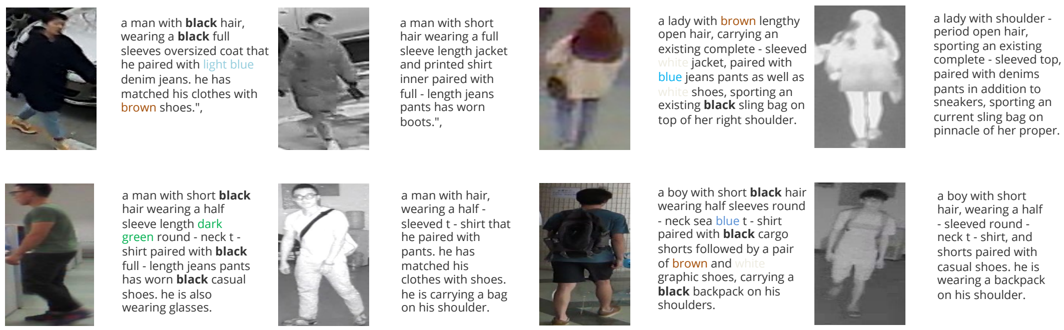


Figure 5: Visualization of the data samples selected from the expanded three datasets.

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

All the fine-tuning process of VLMs can be found in documentations from huggingface: `https://huggingface.co/docs/transformers/main/en/tasks/image_captioning`.

The generator model we use refers to the official implementation released in huggingface: `https://huggingface.co/Salesforce/blip-image-captioning-large`.

B Implementation Details

We implement our framework in PyTorch [32] 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 the same as CAJ [57]. All text descriptions are generated by two modality-specialized fine-tuned VLMs and augmented by the proposed LLM rephrasing augmentation with a probability of 0.5, here we use vicuna-7b [67] as our LLM model, use Blip [24] as our VLM model, whose tuning process can be found in Sec. 3.1. We employ a dual-stream resnet50 model [58] pre-trained on ImageNet [36] as the visual backbone and a transformer model with parameters derived from CLIP [34] as the textual backbone. For incrementally fine-tuning our TVI-LFM, at first, we should get an available well-trained visual backbone. Here we utilize the augmentation method [57] 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 text encoder from CLIP [34] and a simple ReID bottleneck [29] applied for each feature for 20 epochs. We use the Adam [17] 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 Retrieve Result Examples w/wo Text

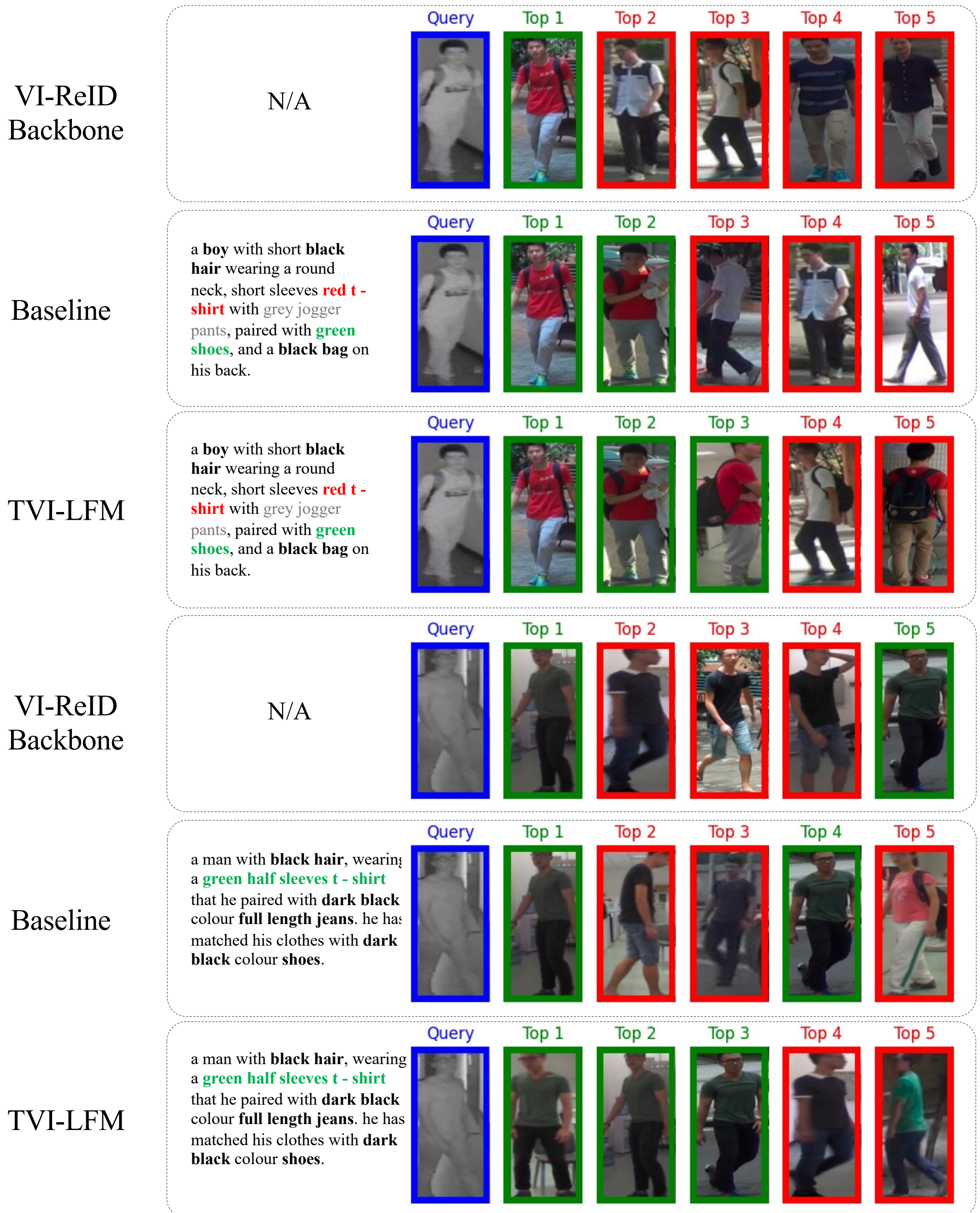


Figure 6: Visualization of the rank-5 retrieval results obtained by the VI-ReID backbone, the baseline, and our method on the proposed Tri-SYSU-MM01.

The VI-ReID backbone and the baseline still includes misidentifications. But our method fully leverages complementary information from textual data, significantly enhancing retrieval performance through semantic filtered fusion.

D Assets Details

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

- **SYSU-MM01** [46]
  - *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** [65]
  - *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** [31]
  - *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

E Discussion of Incremental Fine-tuning Strategy

To optimize the whole framework, we first train a simple VI-ReID backbone, then incrementally fine-tune the VLM textual encoder based on the well-trained frozen backbone to inherit its visual perception capability and integrate text information for infrared modality compensation. If we train the whole framework from scratch, as shown in the Table 5 above, the performance of the VI-ReID backbone suddenly declines by 5.43% and 4.24% in Rank-1 in the two datasets respectively, indicating the loss of visual perception capability, thereby the performance textually enhanced task ($I + T \rightarrow R$) 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 incrementally fine-tuning strategy, which avoids the potential performance influence caused by conflicts of modeling visual features and textual enhanced infrared features optimization.

Table 5: The impact of incrementally fine-tuning the framework based on a frozen, well-trained visual backbone versus training from scratch is evaluated for two scenarios: 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**. To specifically analyze the influence on visual perception capability and fusion feature modeling, we **exclude** the **MER** strategy from the fusion query to eliminate the effect of combining original features from both the infrared modality and text 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
From Scratch	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
TVI-LFM	**84.17**	**80.72**	**70.02**	**57.13**	**64.06**	**60.72**
From Scratch	84.03↓**0.14**	79.85↓**0.87**	68.06↓**1.97**	55.47↓**1.66**	62.23↓**1.83**	58.86↓**1.86**

F Broader Impacts

Our TVI-LFM 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.

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 large foundation models to generate dynamic textual modalities effectively addresses the lack of critical information, such as color, in infrared modality. The experimental results across all three expanded VI-ReID datasets show great improvements on the accuracy of our proposed cross-modality retrieval bolstered by heterogeneous text descriptions.

   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 Sec. 5.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

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

   Answer: [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-LFM architecture in Sec. 3 and Fig. 2, the method of expanding the existing open-source dataset in Sec. 3.1 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 Sec. 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: [Yes]

Justification: We will release our data and code soon in `https://github.com/WHU-HZY/TVI-LFM`.

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 in Sec. 4.1, while the analysis of experiments results can be found in Ablation Study in Sec. 4.2 and Comparison with state-of-the-art methods in Sec. 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, 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 F

    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 [67] 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 D.

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: [Yes]

Justification: All the documentations of original datasets can be viewed at the github urls in Appendix D. Documentations and statistics of the expanded data will also be provided at this paper's github url.

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.