Research Article

When Large Vision-Language Models Meet Person Re-Identification

Qizao Wang[1], Bin Li[1], Xiangyang Xue[1]

1. Fudan University, China

Large Vision-Language Models (LVLMs) that incorporate visual models and Large Language Models (LLMs) have achieved impressive results across various cross-modal understanding and reasoning tasks. In recent years, person re-identification (ReID) has also started to explore cross-modal semantics to improve the accuracy of identity recognition. However, effectively utilizing LVLMs for ReID remains an open challenge. While LVLMs operate under a generative paradigm by predicting the next output word, ReID requires the extraction of discriminative identity features to match pedestrians across cameras. In this paper, we propose LVLM-ReID, a novel framework that harnesses the strengths of LVLMs to promote ReID. Specifically, we employ instructions to guide the LVLM in generating one pedestrian semantic token that encapsulates key appearance semantics from the person image. This token is further refined through our Semantic-Guided Interaction (SGI) module, establishing a reciprocal interaction between the semantic token and visual tokens. Ultimately, the reinforced semantic token serves as the pedestrian identity representation. Our framework integrates the semantic understanding and generation capabilities of LVLMs into end-to-end ReID training, allowing LVLMs to capture rich semantic cues from pedestrian images during both training and inference. Our method achieves competitive results on multiple benchmarks without additional image-text annotations, demonstrating the potential of LVLM-generated semantics to advance person ReID and offering a promising direction for future research.

Corresponding authors: Qizao Wang, qzwang22@m.fudan.edu.cn; Bin Li, libin@fudan.edu.cn; Xiangyang Xue, xyxue@fudan.edu.cn

1. Introduction

Person re-identification (ReID) is a crucial task in computer vision, aimed at accurately matching pedestrians across different camera views[1]. With the continuous advancements in deep learning techniques, person ReID methods have evolved significantly[2][3]. In the past decade, a large body of research has significantly improved ReID accuracy by optimizing the distances between features[4][5] and designing refined modules[6][7][8][9], following the paradigm shown in Fig. 1 (a). However, challenges such as lighting variations, occlusions, and changes in appearance still persist, prompting researchers to explore more robust feature extraction models.

 doi.org/10.32388/AIoCSN

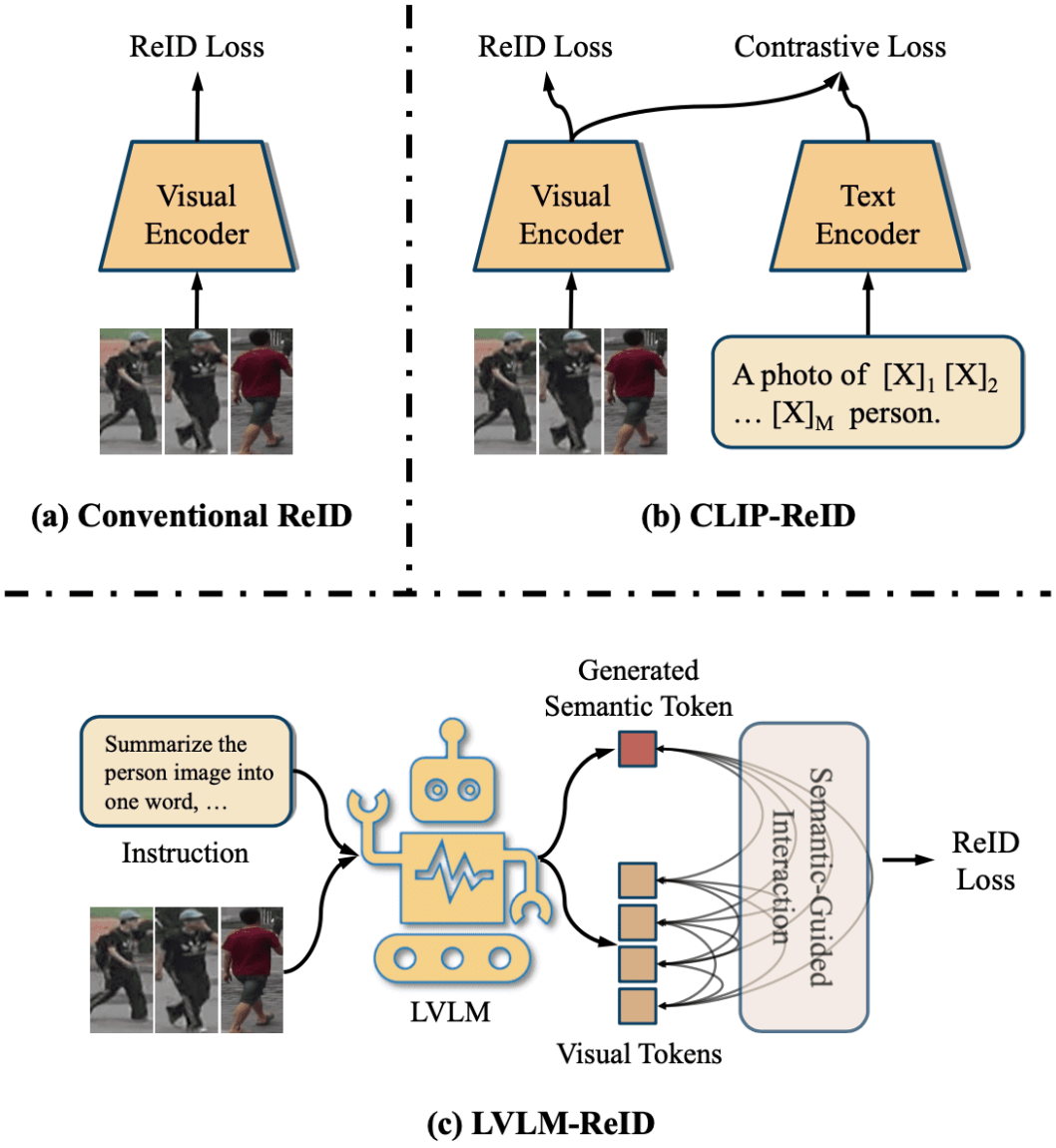

Figure 1. Comparison of different person ReID frameworks. (a) Conventionally, a visual encoder is applied to extract pedestrian identity representations, overlooking the supplemented semantics from other modalities. (b) CLIP-ReID uses the text encoder of CLIP to introduce text semantics based on the contrastive learning paradigm. (c) Our proposed LVLM-ReID incorporates LVLM in the ReID pipeline. Through instruction, LVLM generates one pedestrian semantic token to enhance visual representations.

Due to the difficulty of learning rich pedestrian semantic information from a single modality, cross-modal learning has received close attention in recent years. For example, in the context of the development of pre-trained Vision-Language Models (VLMs), CLIP-ReID[10] based on the representative VLM model CLIP[11] to leverage the semantic information in text. As shown in Fig. 1 (b), it enhances visual features through cross-modal contrastive learning with image-text pairs. Meanwhile, Large Language Models (LLMs)[12][13][14] have attracted widespread attention due to their powerful

capabilities in text generation and comprehension. Large Vision-Language Models (LVLMs)[15][16][17][18][19] enhance LLMs by incorporating visual perception and understanding, demonstrating considerable potential in multi-modal learning tasks. However, integrating LVLMs with person re-identification remains an underexplored challenge.

LVLMs typically operate on a generative paradigm, training and functioning by predicting the next word in a sequence. Thanks to pre-training and instruction tuning, LVLMs can follow instructions and converse with humans. As a result, a direct approach might be to have the model to identify the input person images. However, ReID gallery databases are usually very large (comprising tens of thousands of pedestrian images) [20][21]. For each query image, the time and cost of comparing identities one by one with LVLMs are substantial. Processing multiple images simultaneously would also lead to an unacceptable increase in visual tokens. Therefore, we consider whether it's possible to leverage the reasoning and understanding capabilities of LVLMs while adhering to the mainstream ReID paradigm of feature extraction combined with feature similarity-based retrieval[1]. A potential solution involves using LVLMs to create textual descriptions of pedestrian images and fine-tuning the visual encoder via tasks such as image-text matching or image caption prediction. However, this approach presents several limitations: (1) High-quality and diverse text annotations are expensive to obtain. (2) The goals of image-text matching or image caption prediction tasks may not align well with those of image-based ReID. (3) During the inference phase, the potential of LVLMs is often underutilized, as they are not effectively integrated with the visual features.

To address these issues, we propose a new framework called LVLM-ReID. We propose to leverage the superior semantic understanding and generation ability of LVLMs to assist ReID. Specifically, as shown in Fig. 1 (c), we use instruction to guide the LVLM to focus on specific visual semantics in pedestrian images, generating a semantic token representing the pedestrian's appearance information. We then design an effective interaction module between the generated token and visual tokens, refining the visual representations of pedestrians while reinforcing the semantic token as a discriminative identity representation. Ultimately, the reinforced semantic token is optimized and used during inference to achieve person retrieval. Our framework integrates the generative process of LVLMs into the ReID model, eliminating the need for additional image caption annotations and enabling end-to-end effective learning. More importantly, during the inference phase, we continue to leverage the generative power of LVLMs to adaptively enhance visual features. Our experiments show that one generated semantic token can effectively facilitate the learning of pedestrian representations. Our contributions are summarized as follows:

- We propose a novel framework that incorporates LVLMs into the person ReID task, offering a new perspective on using generative language models to assist discriminative visual models.
- We propose to utilize the generative capability of LVLMs to produce a semantic token for pedestrians and design a semantic-guided interaction module leveraging the generated semantic token to enhance identity representations.
- Experimental results show that, without requiring additional annotations, our method effectively improves the discriminability of identity features and achieves competitive results across multiple datasets.

2. Related Work

2.1. Person Re-Identification

With the development of deep learning techniques, Convolutional Neural Network (CNN)-based approaches have seen widespread adoption in person ReID[1]. In addition to extracting global feature representations from pedestrian images, part-level and multi-granularity features[22][23][24] play an important role in fine-grained pedestrian identity recognition. Moreover, DG-Net[25] proposes a joint learning framework that couples ReID learning and data generation end-to-end. SAN[6] incorporates semantics-aligned feature representation learning through delicate supervision designs. Many methods also attempt to learn better pedestrian representations and relationships through well-designed modules[7][26][8][9]. With the popularity of the Transformer architecture[27], methods like TransReID[28] explore to leverage Vision Transformer (ViT)[29] to enhance the model's ability in learning rich structural patterns. Based on the Transformer baseline, DCAL[30] extends self-attention modules to better learn subtle feature embeddings, and AAformer[31] integrates part features for retrieval. Recently, visual language pre-training significantly improves the performance of many downstream tasks by training to match images and language[11][32]. CLIP-ReID[10] utilizes the contrastive cross-modal alignment in the CLIP paradigm[11] and adopts a two-stage strategy to facilitate a better visual representation.

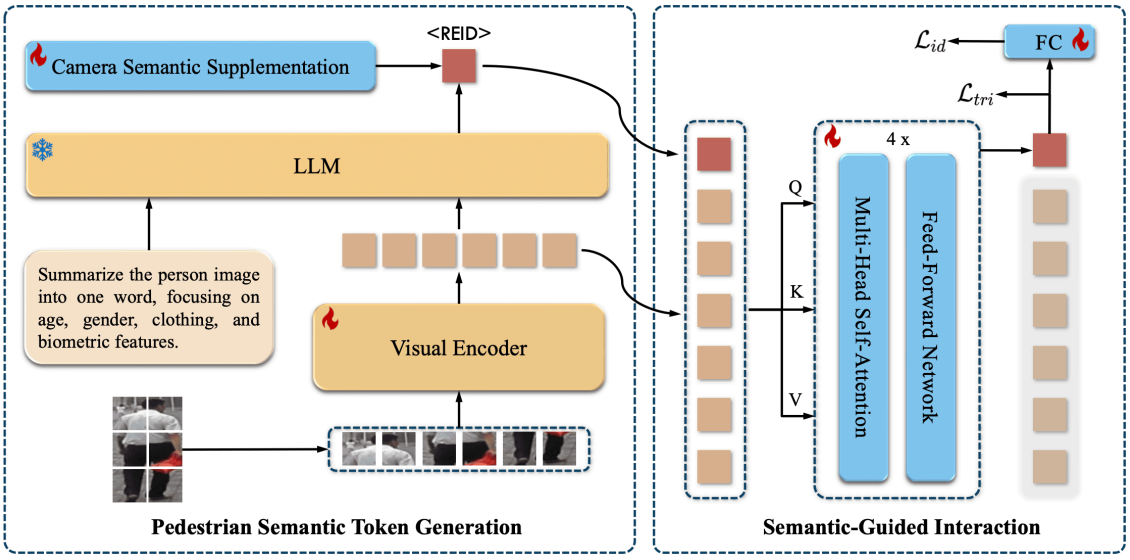

Figure 2. Framework of our LVLM-ReID. It leverages clear instructions to guide the frozen LLM towards focusing on particular visual semantics within pedestrian images, resulting in the generation of one semantic token that encapsulates the pedestrian's appearance information. Subsequently, an efficient interaction module is designed to facilitate refinement between the generated token and the visual tokens. Finally, the reinforced token as a distinctive identity descriptor is optimized and employed for person retrieval.

2.2. Large Vision-Language Model

Building on the impressive reasoning and understanding capabilities of LLMs[12][12][14], researchers have been working to adapt these strengths to the visual domain, leading to the development of Large Vision-Language Models (LVLMs). LVLMs have become a key technology in multimodal learning, enabling the processing and generation of complex visual and textual information. For instance, Flamingo[33] introduces a cross-attention mechanism that enables the model to attend to visual contexts, supporting visual in-context learning. Other models, such as BLIP-2[18] and mPLUG-OWL[34], use visual encoders to process image features, which are then combined with text embeddings and input into the LLM. Additionally, LLaVA[16] and MiniGPT-4[35] align the image and text features as a preliminary step, followed by instruction tuning to refine the model's instruction following ability. Recently, Qwen2-VL[19] employs a unified paradigm for processing both images and videos and support varying resolutions, achieving highly competitive performance across various multimodal benchmarks. LVLMs can effectively facilitate cross-modal understanding of both image and text inputs, while how to leverage their advantages in ReID tasks remains an underexplored issue. Based on one of the representative LVLMs, Qwen2-VL, we explore the possibility of using the semantic understanding and generation capabilities of LVLM to enhance pedestrians' semantic representation in person ReID.

3. Methodology

In this section, we first introduce the overall framework of LVLM in Sec. 3.1. Then, we elaborate on our proposed Pedestrian Semantic Token Generation (PSTG) in Sec. 3.2. PSTG aims to generate one semantic token that encapsulates instructive appearance information of the pedestrian, and the generated semantic token is then used for Semantic-Guided Interaction (SGI) with visual tokens (see Sec. 3.3). Finally, we introduce our end-to-end optimization and inference scheme in Sec. 3.4. The framework of our proposed LVLM-ReID is shown in Fig. 2.

3.1. Overview of LVLM

Overall framework.

A typical LVLM consists of three key components: a visual encoder, a vision-language connector, and an LLM. The visual encoder extracts rich visual representations from images, which are then processed by the vision-language connector that converts visual features into the word embedding space. The LLM, trained for next-word prediction, generates text based on the encoded visual content. This generative structure enables LVLM to handle multimodal inputs, allowing for efficient image-text interaction and the generation of new textual information. In this work, we leverage Qwen2-VL[19], one of the most advanced LVLMs, known for its superior capabilities in instruction-following, semantic understanding, and text generation across diverse tasks. Qwen2-VL combines a Vision Transformer (ViT)[29] as the visual encoder and the Qwen2[14] as the LLM. The vision-language connector between the two components is a simple MLP layer that also compresses the extracted visual tokens.

Visual token extraction.

Before inputting a pedestrian image into the LLM, the image is first encoded and compressed by the visual encoder. Specifically, each input RGB image $x \in \mathbb{R}^{H\times W\times 3}$, where H and W are its height and width, is first divided into patches of size $P \times P$. These patches are then embedded and flattened into a feature vector $x^p \in \mathbb{R}^{N\times d}$, where $N = H \times W/P^2$ represents the number of patches, and d is the embedding dimension. The resulting patch embeddings are processed through multiple layers of Transformer self-attention blocks[27], producing visual representations $f \in \mathbb{R}^{N\times d}$. To enhance the model's ability to capture spatial dependencies, Multimodal Rotary Position Embedding (M-RoPE)[19] is used in the process. Afterward, a simple MLP layer compresses adjacent 2×2 tokens into a single token, producing the final visual tokens V, which is formulated as:

$$V = \mathrm{MLP}(f) = [v_1; v_2; \cdots; v_n] \in \mathbb{R}^{n\times d} \tag{1}$$

where $n = \frac{N}{4}$. Notably, instead of using the traditional [class] token[29], the image is transformed into a set of visual tokens. These visual tokens will then be passed to the LLM for further processing and interaction.

3.2. Pedestrian Semantic Token Generation

We aim to integrate the advanced visual semantic understanding and generation capabilities of LVLM into the feature extraction pipeline, by guiding the ReID model to generate one semantic token that encapsulates instructive appearance information of the pedestrian. To achieve this, we propose the Pedestrian Semantic Token Generation (PSTG) strategy, where we use instructions to direct the LVLM to generate a semantic token that summarizes the pedestrian's visual appearance. Considering that representative attributes, such as age, gender, and clothing, are crucial for identifying pedestrians, the instruction is carefully formulated as follows:

<|vision_start|> *V* <|vision_end|> Summarize the person image into one word, focusing on age, gender, clothing, and biometric features.

where V represents the extracted visual tokens, while the special tokens <|vision_start|> and <|vision_end|> are used to mark the beginning and end of the visual token sequence. With this instruction, the LVLM is guided to focus on the appearance-related semantics in the image, and then generate a semantic token that summarizes the relevant identity features. We denote this generated token as <REID>, which serves as a compact representation of the pedestrian's visual appearance. The generated semantic token is then used in the following stages of our framework to guide identity feature learning.

The quality of the instruction is crucial for obtaining a useful token. Through empirical evaluation, we find that simple, clear instructions work effectively in guiding the LVLM. Future work could explore more sophisticated instruction designs to improve the semantic token generation process and, by extension, ReID performance.

Camera semantic supplementation.

The semantic token generation process overlooks the influence of camera variations. To improve pedestrian semantic consistency across cameras, we explicitly model and account for these camera-induced feature variations. Specifically, we assign a unique learnable embedding vector to each camera, which allows the model to learn the inherent feature shifts caused by cameras. These camera embeddings are used to adjust the pedestrian's semantic representation by incorporating camera-specific information. We denote the set of learnable camera embeddings as $V_{cam} = \{v_{cam}^i | i = 1, 2, \cdots, N^c\}$, where N^c is the total number of cameras. One direct implementation is to supplement the generated pedestrian semantic token with the camera semantics as follows:

$$\bar{v}_{reid} = v_{reid} + v_{cam}^{y^c} \tag{2}$$

where v_{reid} is the encoding of the <REID> token, y^c is the camera ID corresponding to the image x. However, this late supplementation strategy may affect the visual model weakly. We thus try to transfer the usage of camera embeddings to the input of visual model, where the camera embeddings are added to the patch embeddings x^p. We evaluate the two variants and discuss their influences in Sec. 4.3.

3.3. Semantic-Guided Interaction

We design the Semantic-Guided Interaction (SGI) module to facilitate bidirectional interaction between the generated semantic token and the visual tokens. Specifically, the generated semantic token is first concatenated with the visual tokens. Formally,

$$z = [v_{reid}; v_1; v_2; \cdots; v_n] \in \mathbb{R}^{(n+1)\times d} \tag{3}$$

This concatenated token sequence is then passed through 4 layers of Transformer blocks, each consisting of a multi-head self-attention layer and a feed-forward network. The module refines the visual features to capture identity-relevant information under the guidance of the semantic token. Meanwhile, the semantic token, serving as the pivot for information aggregation, distills more discriminative features from the visual representations, enhancing the overall understanding of the pedestrian's identity. Through the semantic-guided interaction module, the model produces the reinforced representation as:

$$\hat{z} = [\hat{v}_{reid}; \hat{v}_1; \hat{v}_2; \cdots; \hat{v}_n] = \mathrm{SGI}(z) \tag{4}$$

Then, the reinforced semantic token representation $\hat{v}_{reid}$ is used to compute the Re-ID losses, i.e., identity classification loss[2] and triplet loss[36]. Specifically, identity classification loss ensures that the reinforced semantic token correctly maps to the pedestrian's identity category. A Fully Connected (FC) layer is employed as the identity classifier, and p_i represents the predicted logits for the i-th identity category. The identity classification loss is computed as:

$$\mathcal{L}_{id} = -\sum_{i=1}^{N^p} q_i \log p_i \tag{5}$$

$$q_i = \begin{cases} 1 - \epsilon + \frac{\epsilon}{N^p} & , i = y \\ \frac{\epsilon}{N^p} & , \text{otherwise} \end{cases} \tag{6}$$

where N^p is the total number of training identities, y is the corresponding identity label for the image x, and ϵ is a small constant for label smoothing regularization, which is typically set to 0.1.

Additionally, to further improve the identity discrimination of the learned features, triplet loss is used. It ensures that the identity representations of different pedestrians maintain the correct relative distances in the feature space. The triplet loss is defined as:

$$\mathcal{L}_{tri} = \max\left(m + d_p - d_n, 0\right) \tag{7}$$

where d_p and d_n are the feature distances between a positive pair and a negative pair mined in the training batch, respectively, and m is the margin. A positive pair consists of images from the same pedestrian, while a negative pair contains images from different pedestrians. The triplet loss encourages the model to minimize the distance between images of the same identity and maximize the distance between images of different identities.

3.4. Optimization and Inference

During training, we optimize the parameters of both the visual model and the SGI module while keeping the LLM parameters frozen, though we retain its gradients. The LLM's role in generating the semantic token, guided by the instruction, enables the model to focus on identity-relevant regions and characteristics within the pedestrian image. By leveraging the generated <REID> token in conjunction with the SGI module, we achieve joint end-to-end training that harnesses the strengths of LVLM in instruction-following and visual semantic understanding. This process allows for the integration of rich semantic cues into the visual representations, improving pedestrian identity recognition accuracy. The overall training loss is a weighted combination of the identity classification loss $\mathcal{L}_{id}$ and the triplet loss $\mathcal{L}_{tri}$, which is expressed as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{id} + \alpha_2 \mathcal{L}_{tri} \tag{8}$$

where α_1 and α_2 are balancing factors that control the contribution of each loss term.

During inference, the LVLM is also used to generate the <REID> token for each input image. Then, the reinforced semantic token representation, $\hat{v}_{reid}$, is used to compute the cosine similarity between different person images. These similarity scores are employed for identity matching, allowing the model to accurately identify pedestrians. Note that the identity representations of persons in the large gallery databases need to be extracted only once in applications.

Backbone	Methods	DukeMTMC-reID		Market-1501		CUHK03	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
CNN	MGN[23]	78.4	88.7	86.9	95.7	67.4	68.0
	DG-Net[25]	74.8	86.6	86.0	94.8	-	-
	SAN[6]	75.5	87.9	88.0	**96.1**	76.4	80.1
	Pyramid[24]	79.0	89.0	88.2	95.7	76.9	78.9
	Relation-Net[7]	78.6	89.7	88.9	95.2	75.6	77.9
	RGA-SC[26]	-	-	88.4	**96.1**	77.4	81.1
	CDNet[8]	76.8	88.6	86.0	95.1	-	-
	CAL[9]	76.4	87.2	87.0	94.5	-	-
ViT	TransReID[28]	80.6	89.6	88.2	95.0	-	-
	DCAL[30]	80.1	89.0	87.5	94.7	-	-
	AAformer[31]	80.0	90.1	88.0	95.4	79.0	80.3
	PFD[37]	82.2	**90.6**	**89.6**	95.5	-	-
	CLIP-ReID[10]	**82.5**	90.0	**89.6**	95.5	**80.3**	**81.6**
LVLM	LVLM-ReID	**82.8**	**92.2**	**89.2**	**95.6**	**82.3**	**84.6**

Table 1. Comparison with the state-of-the-art methods on DukeMTMC-reID, Market-1501, and CUHK03. The results of our proposed method and the best results of comparison methods are shown in bold.

4. Experiments

4.1. Experimental Settings

4.1.1. Dataset

We evaluate our methods on three person Re-ID datasets:

- **DukeMTMC-reID**[20] consists of 36,411 images of 1,404 identities, captured from 8 different cameras. The dataset includes 16,522 images for training and 19,889 images for testing.
- **Market-1501**[21] is captured by 6 cameras at Tsinghua University, containing 12,936 images of 751 identities for training and 19,281 images of 750 identities for testing.

- **CUHK03**[38] consists of 1,467 pedestrians. Following[39], 767 identities are used for training and 700 identities for testing. The labeled version is used with manually labeled bounding boxes from 14,096 images.

4.1.2. Evaluation Metrics

We follow the common practices to adopt Cumulative Matching Characteristics (CMC) at Rank-1 and mean Average Precision (mAP) for performance evaluation.

4.1.3. Implementation Details

Our method is implemented on PyTorch. We employ Qwen2-VL-2B[19] considering its efficiency with limited resources, while larger model sizes such as 7B and 72B have better LLM capabilities. The model adopts BFloat16 mixed precision. H, W, P are set to 280, 140, 14, respectively, resulting in $n = 50$. In other words, 50 visual tokens are included in the input of LLM and our SGI module. Following[2], random horizontal flipping, padding, random cropping, and random erasing[40] are used for data augmentation. 16 identities and 4 images per person are randomly sampled to constitute a training batch. Adam optimizer with weight decay of 3×10^{-4} is adopted, with the warmup strategy that linearly increases the learning rate from 3×10^{-5} to 3×10^{-4} in the first 10 epochs. We train the model for 60 epochs, with a learning rate decay factor of 0.1 at the 30th epoch. α_1 and α_2 are set to 0.25 and 1 following[10]. The margin m of triplet loss is set to 0.3.

4.2. Comparison with State-of-the-Art Methods

We compare our method with the state-of-the-art methods on three widely used person ReID benchmarks in Tab. 1. Methods based on CNNs achieve solid performance by designing elaborate modules for person ReID. TransReID[28], on the other hand, explores the potential of Transformers[27][29] in ReID, establishing itself as a strong baseline with superior capability. As shown in Tab. 1, ViT-based methods achieve consistent performance across different datasets due to the effectiveness of pre-training and Transformer architecture. Our LVLM-ReID adopts LVLM as the backbone, leveraging the advantages of Transformer and capabilities of large language models. Rather than designing elaborate modules for interactions between image pairs[30], or leveraging part-level features[31] or pose semantics[37] based on ViT, we introduce LVLM's advanced understanding and generative processes into the ReID framework. Our method achieves consistently better results across the three datasets.

More concretely, on the DukeMTMC-reID dataset, which is known for occlusions and variations in appearance, LVLM-ReID achieves an mAP of 82.8% and a Rank-1 accuracy of 92.2%, surpassing previous advanced methods, such as PFD[37] (mAP: 82.2%, Rank-1: 90.6%) and CLIP-ReID[10] (mAP: 82.5%, Rank-1: 90.0%). The results indicate that LVLM-ReID is effective in handling the challenging variations from varying cameras and complex environmental conditions. On the CUHK03 dataset, LVLM-ReID achieves an mAP of 82.3% and a Rank-1 accuracy of 84.6%, significantly outperforming other methods like CLIP-ReID[10] (mAP: 80.3%, Rank-1: 81.6%). LVLM-ReID also achieves competitive results on the Market-1501 dataset. The strong performance of LVLM-ReID across datasets demonstrates its capability of leveraging LVLM. Note that CLIP-ReID[10] leverages a VLM pre-trained through contrastive learning on large-scale image-text pairs, and it discards the text encoder during inference. Differently, our proposed LVLM-ReID integrates LVLM into ReID

training and inference stages in a novel paradigm, moving beyond traditional model designs. The comparison results demonstrate its effectiveness in advancing person ReID performance.

4.3. Ablation Studies

To incorporate LVLM to promote ReID, we introduce two key components, i.e., Pedestrian Semantic Token Generation (PSTG) and Semantic-Guided Interaction (SGI). We validate their importance and necessity in Tab. 2. We also discuss the camera semantic supplementation design in Tab. 3 and our semantic-guided interaction design in Tab. 4. To demonstrate the effectiveness of our end-to-end training process, we further ablate one variant that eliminates the gradient from LLM in Tab. 5.

Methods	DukeMTMC-reID		Market-1501	
	mAP	Rank-1	mAP	Rank-1
Baseline	79.0	90.2	87.3	94.7
Ours w/o PSTG	80.9	91.0	88.3	95.0
Ours w/o SGI	79.0	90.0	87.3	94.5
Ours	**82.8**	**92.2**	**89.2**	**95.6**

Table 2. Ablation studies of our key two components on DukeMTMC-reID and Market-1501.

Effectiveness of the generated pedestrian semantic token.

(1) Our baseline is based on the visual model of the LVLM, and the visual tokens are averaged to compute loss and feature similarity during training and inference. The baseline only uses the visual model, overlooking the role of LVLM in visual semantic understanding and achieving inferior performance. (2) In the variant "Ours w/o PSTG", we replace the LVLM-generated semantic token with a learnable token, similar to the design of the [class] token[29], to integrate visual information. As shown in Tab. 2, this substitution leads to a substantial performance drop since the randomly initialized learnable token lacks rich semantic cues. This result underscores the importance of our PSTG mechanism, which contributes to a more comprehensive understanding of pedestrian images.

Effectiveness of the SGI module.

In the "Ours w/o SGI" variant, we remove the SGI module and rely solely on the LVLM-generated semantic token for ReID. As shown in Tab. 2, this configuration still achieves reasonably good performance, suggesting that our PSTG mechanism effectively captures essential pedestrian semantic information. However, the variant struggles to outperform the baseline, emphasizing the importance of the SGI module in leveraging the generated semantic token. The SGI module not only refines the visual tokens by allowing them to interact with the semantic token but also reinforces its identity-specific

information, resulting in a more comprehensive representation. The performance improvement of introducing SGI highlights its role in obtaining a more robust and discriminative pedestrian representation for person ReID.

Methods	DukeMTMC-reID		Market-1501	
	mAP	Rank-1	mAP	Rank-1
w/o CSS	81.6	91.4	89.1	95.2
CSS-v_{reid}	82.3	92.1	88.4	95.3
CSS-x^p	**82.8**	**92.2**	**89.2**	**95.6**

Table 3. Ablation of the camera semantic supplementation (CSS) strategy. CSS-v_{reid} and CSS-x^p denote adding the camera embedding to v_{reid} and x^p, respectively.

Ablation of the camera semantic supplementation strategy.

As shown in Tab. 3, we compare two variants that supplement camera semantics for the generated tokens and visual inputs. The result of "CSS-v_{reid}" shows that camera semantics can improve the representation ability of the generated tokens for pedestrians. However, since it indirectly enhances the robustness of the visual model to camera changes through our semantic-guided interaction module, the late supplementation strategy may affect the visual model weakly. When transferring the usage of camera embeddings to the input of the visual model (denoted by CSS-x^p), we observe a better performance. Interestingly, the observation is consistent with the work only using ViT[28]. In our LVLM-ReID framework, this design helps to improve the robustness of the generated semantic token and the extracted visual features, further improving the model's ability to match pedestrians across cameras.

Methods	DukeMTMC-reID		Market-1501	
	mAP	Rank-1	mAP	Rank-1
v_{reid} **as Query**	80.5	89.4	88.6	95.2
Ours	**82.8**	**92.2**	**89.2**	**95.6**

Table 4. Ablation of the SGI module design. "v_{reid} as Query" treats the generated semantic token as query, with image tokens serving as keys and values in a cross-attention mechanism[27], and uses the resulting output as the pedestrian representation.

doi.org/10.32388/AIoCSN

Ablation of the SGI design.

In the SGI module, we adopt bidirectional interaction between the generated semantic token and the visual tokens. To assess the effectiveness of this design, we evaluate an alternative configuration, "v_{reid} as Query", in Tab. 4. However, this variant results in a noticeable performance decrease, validating the effectiveness and rationale of our SGI design. Its inferior performance suggests that limiting interaction to a single directional flow from visual tokens to the semantic token does not leverage the mutual enhancement potential between them. In contrast, our SGI allows the semantic token to guide and refine the visual features while being dynamically influenced by the visual content. This reciprocal exchange strengthens the visual representations with relevant semantic context.

Methods	DukeMTMC-reID		Market-1501	
	mAP	Rank-1	mAP	Rank-1
Stop Gradient	73.1	86.8	84.7	93.5
Ours	**82.8**	**92.2**	**89.2**	**95.6**

Table 5. Ablation studies of our end-to-end training design. "Stop Gradient" prevents gradient flow from the LLM to the visual model.

Effectiveness of our end-to-end design.

We evaluate a variant of our proposed LVLM-ReID, denoted as "Stop Gradient", in Tab. 5. In the variant, while the LLM generates semantic tokens, they do not impact the visual model's learning process. The variant fails to harness the full benefit of joint training, as it restricts the cross-modal optimization loop that allows the generated semantics to iteratively enhance visual feature learning. Therefore, it shows unsatisfactory performance. The results highlight that our integrated design not only fosters tighter synergy between the visual model and the semantic token but also enables our model to capture more nuanced identity-relevant details, ultimately driving stronger ReID performance.

4.4. Qualitative Analysis

To understand the identity-related information in the semantic token and demonstrate the effectiveness of LVLM in enriching pedestrian semantics, we analyze the attention maps using[41] in Fig. 3, and retrieval results in Fig. 4.

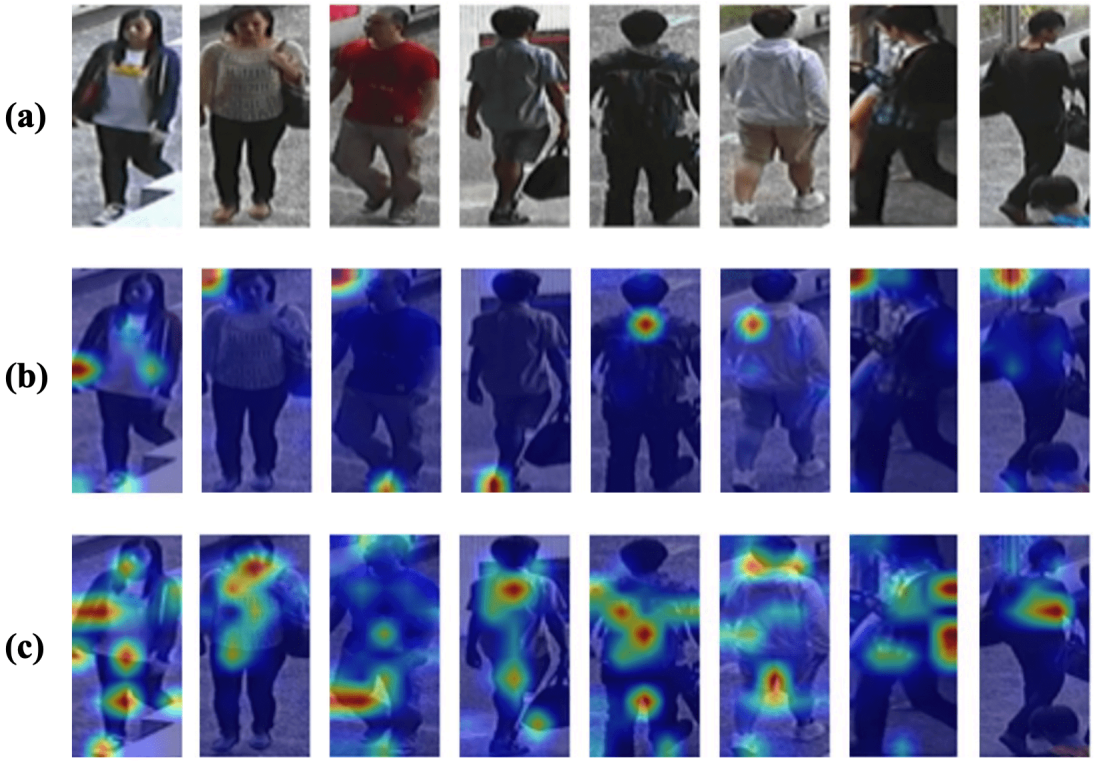

Figure 3. Visualization of attention maps. We show (a) the original images, and compare the attentions of (b) the "Ours w/o PSTG" variant, and (c) our LVLM-ReID model, on CUHK03.

Visualization of attention maps.

As shown in Fig. 3, the attention map visualizations reveal a clear advantage of our method in enhancing semantic understanding. Since a learnable `[class]` token lacks pedestrian semantics, the "Ours w/oPSTG" variant tends to shift attention toward background regions, relying on dataset-specific biases rather than intrinsic pedestrian attributes. This results in an inability to learn robust identity representations under such complex conditions. In contrast, our method, guided by the semantic token, concentrates attention on key identity-specific regions, such as unique clothing patterns and distinctive body parts. In challenging scenarios, such as those involving occlusions or the presence of other pedestrians in the background (as shown in the last two columns), our method demonstrates superior robustness. It focuses on the primary body of pedestrians, ensuring that meaningful identity-specific features are prioritized. This focused attention demonstrates the effectiveness of our semantic guidance in directing the model toward meaningful visual cues, thereby enhancing identity recognition accuracy and robustness across varied scenes and backgrounds.

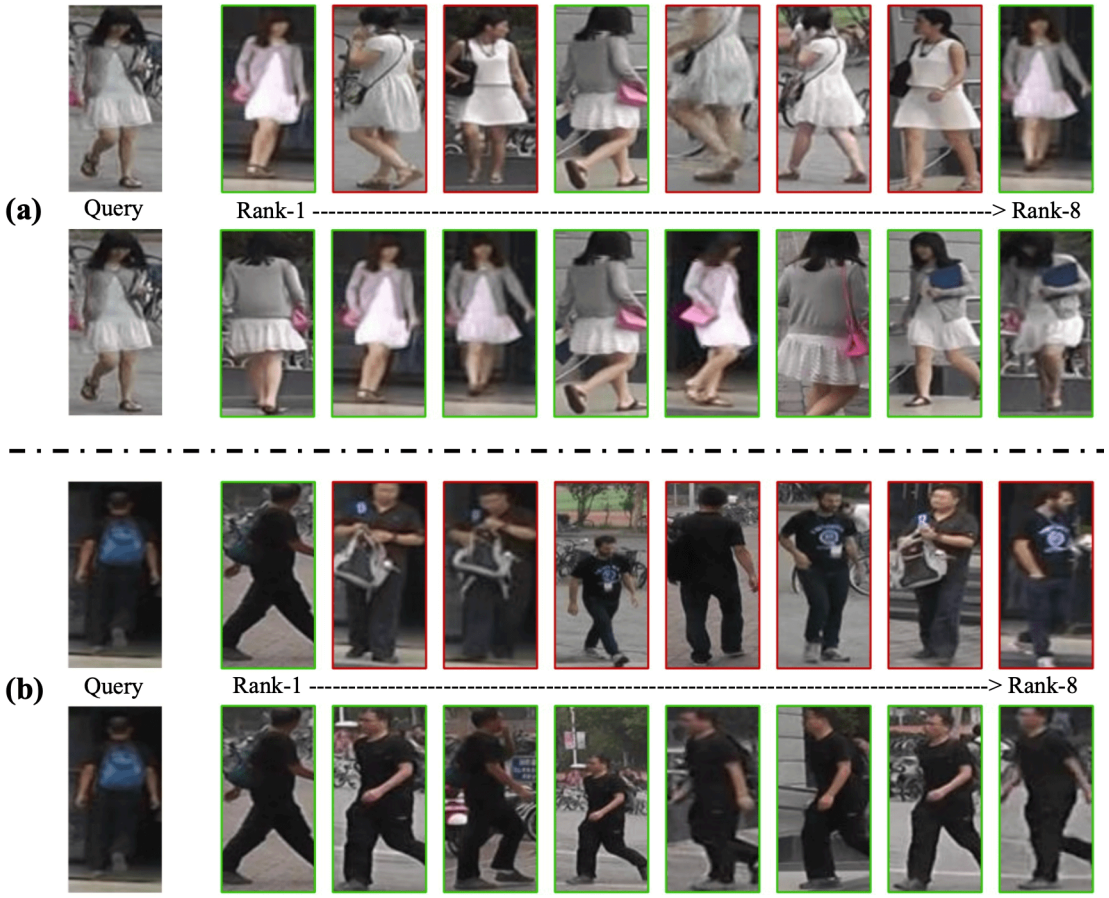

Figure 4. Visualization of retrieval results. For each query, the first and the second rows show the top-8 retrieval results of the baseline and our method on Market-1501, respectively. Retrieved images with green and red boxes are correct and incorrect results, respectively. Best viewed in color and zoomed in.

Visualization of retrieval results.

As displayed in Fig. 4, the baseline model often returns false positives, particularly when individuals in the images share similar attributes with the query image, such as clothing color or style. In contrast, our method effectively captures nuanced identity-specific features, accurately identifying the correct individuals. For example, our method demonstrates robustness to variations in image resolution and human pose (as shown in Fig. 4 (a)), and handles well scale changes (as shown in Fig. 4 (b)), achieving consistently higher precision in ReID compared to the baseline. With the help of LVLM, our method can refine visual representations and enhance the discriminative power of the identity features. This leads to more reliable and accurate ReID in complex scenarios.

5. Conclusion

In this paper, we introduce LVLM-ReID, a novel framework that leverages the semantic understanding and generation capabilities of LVLMs to enhance the performance of person ReID. We design two key components: Pedestrian Semantic

Token Generation (PSTG) and Semantic-Guided Interaction (SGI). We certify that LVLM can be integrated into the ReID process by generating one pedestrian semantic token, which can be used to improve the visual identity representations via an efficient interaction module. In our framework, LVLM effectively helps capture and utilize the rich semantics of pedestrians. Our experimental findings underscore the importance of semantic guidance in strengthening visual representations, and highlight the advantages of our end-to-end design. Our work sets a new direction for integrating LVLMs in the area of person ReID.

Limitations and future work.

We validate the effectiveness of our framework on a 2B parameter model, while the performance gains from more advanced LVLMs or larger model series still need to be explored. While a larger model will bring greater computational overhead, exploring more lightweight LVLMs or optimization techniques is also important.

References

1. [a], [b], [c]*Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SCH (2021). "Deep learning for person re-identification: A survey and outlook". IEEE Transactions on Pattern Analysis and Machine Intelligence. **44** (6): 2872–2893.*
2. [a], [b], [c]*Luo H, Jiang W, Gu Y, Liu F, Liao X, Lai S, Gu J (2019). "A strong baseline and batch normalization neck for deep person re-identification". IEEE Transactions on Multimedia. **22** (10): 2597–2609.*
3. [^]*Wang Q, Qian X, Li B, Fu Y, Xue X (2023). "Rethinking person re-identification from a projection-on-prototypes perspective". arXiv preprint arXiv:2308.10717. Available from: https://arxiv.org/abs/2308.10717.*
4. [^]*Zheng Z, Zheng L, Yang Y (2017). "A discriminatively learned CNN embedding for person reidentification". ACM Transactions on Multimedia Computing, Communications, and Applications. **14** (1): 1–20.*
5. [^]*Zhai Y, Guo X, Lu Y, Li H (2019). "In defense of the classification loss for person re-identification". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 0–0.*
6. [a], [b], [c]*Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, Zhibo Chen. "Semantics-aligned representation learning for person re-identification." In Proceedings of the AAAI Conference on Artificial Intelligence, pages 11173–11180, 2020.*
7. [a], [b], [c]*Park H, Ham B. Relation network for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020:11839-11847.*
8. [a], [b], [c]*Li H, Wu G, Zheng WS (2021). "Combined depth space based architecture search for person re-identification". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 6729–6738.*
9. [a], [b], [c]*Rao Y, Chen G, Lu J, Zhou J (2021). "Counterfactual attention learning for fine-grained visual categorization and re-identification". Proceedings of the IEEE/CVF International Conference on Computer Vision. 1025–1034.*
10. [a], [b], [c], [d], [e], [f], [g]*Li S, Sun L, Li Q (2023). "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels". Proceedings of the AAAI Conference on Artificial Intelligence. pages 1405–1413.*
11. [a], [b], [c]*Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR; 2021. p. 8748-8763.*

12. ^a ^b ^cTouvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023. Available from: https://arxiv.org/abs/2302.13971.
13. ^Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
14. ^a ^b ^cYang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F, et al. Qwen2 technical report. arXiv preprint arXiv:2407.10671. 2024.
15. ^Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774. 2023. Available from: https://arxiv.org/abs/2303.08774.
16. ^a ^bLiu H, Li C, Wu Q, Lee YJ (2024). "Visual instruction tuning". Advances in Neural Information Processing Systems. 36.
17. ^Liu H, Li C, Li Y, Lee YJ. "Improved baselines with visual instruction tuning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 26296–26306.
18. ^a ^bLi J, Li D, Savarese S, Hoi S. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In: International Conference on Machine Learning. PMLR; 2023. p. 19730-19742.
19. ^a ^b ^c ^d ^eWang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191. 2024.
20. ^a ^bRistani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016). "Performance measures and a data set for multi-target, multi-camera tracking". Proceedings of the European Conference on Computer Vision Workshops. pp. 17–35.
21. ^a ^bZheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015). "Scalable person re-identification: A benchmark." Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1116–1124.
22. ^Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018). "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)". In Proceedings of the European Conference on Computer Vision, pages 480–496.
23. ^a ^bWang G, Yuan Y, Chen X, Li J, Zhou X (2018). "Learning discriminative features with multiple granularities for person re-identification." In Proceedings of the 26th ACM International Conference on Multimedia, pages 274–282.
24. ^a ^bZheng F, Deng C, Sun X, Jiang X, Guo X, Yu Z, Huang F, Ji R (2019). "Pyramidal person re-identification via multi-loss dynamic training". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8514–8522.
25. ^a ^bZheng Z, Yang X, Yu Z, Zheng L, Yang Y, Kautz J (2019). "Joint discriminative and generative learning for person re-identification". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2138–2147.
26. ^a ^bZhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, Zhibo Chen. "Relation-aware global attention for person re-identification." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3186–3195, 2020.
27. ^a ^b ^c ^dVaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017). "Attention is all you need". In: Advances in Neural Information Processing Systems, pp. 5998–6008.
28. ^a ^b ^c ^dHe S, Luo H, Wang P, Wang F, Li H, Jiang W (2021). "Transreid: Transformer-based object re-identification". Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15013–15022.
29. ^a ^b ^c ^d ^eDosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations; 2021.

30. ^a ^b ^cZhu H, Ke W, Li D, Liu J, Tian L, Shan Y (2022). "Dual cross-attention learning for fine-grained visual categorization an d object re-identification". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4692–4 702.
31. ^a ^b ^cZhu K, Guo H, Zhang S, Wang Y, Liu J, Wang J, Tang M. "Aaformer: Auto-aligned transformer for person re-identificatio n". IEEE Transactions on Neural Networks and Learning Systems. 2023.
32. ^Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T (2021). "Scaling up visual and vision-languag e representation learning with noisy text supervision". International Conference on Machine Learning. PMLR: 4904–4916.
33. ^Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, et al. Flamingo: a visual l anguage model for few-shot learning. Advances in Neural Information Processing Systems. **35**: 23716–23736, 2022.
34. ^Ye Q, Xu H, Xu G, Ye J, Yan M, Zhou Y, Wang J, Hu A, Shi P, Shi Y, et al. mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178. 2023. Available from: https://arxiv.org/abs/2304.14178.
35. ^Zhu D, Chen J, Shen X, Li X, Elhoseiny M (2023). "Minigpt-4: Enhancing vision-language understanding with advanced larg e language models". arXiv preprint arXiv:2304.10592. Available from: https://arxiv.org/abs/2304.10592.
36. ^Hermans A, Beyer L, Leibe B (2017). "In defense of the triplet loss for person re-identification". arXiv preprint arXiv:1703.077 37. Available from: https://arxiv.org/abs/1703.07737.
37. ^a ^b ^cWang T, Liu H, Song P, Guo T, Shi W (2022). "Pose-guided feature disentangling for occluded person re-identification b ased on transformer". Proceedings of the AAAI conference on artificial intelligence. pages 2540–2549.
38. ^Li W, Zhao R, Xiao T, Wang X. Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2014. p. 152-159.
39. ^Zhong Z, Zheng L, Cao D, Li S (2017). "Re-ranking person re-identification with k-reciprocal encoding". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 3652–3661.
40. ^Zhong Z, Zheng L, Kang G, Li S, Yang Y (2020). "Random erasing data augmentation". Proceedings of the AAAI Conference o n Artificial Intelligence. pages 13001–13008.
41. ^Chefer H, Gur S, Wolf L (2021). "Generic attention-model explainability for interpreting bi-modal and encoder-decoder tran sformers". Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.