Beyond Pedestrians: Caption-Guided CLIP Framework for High-Difficulty Video-based Person Re-Identification

Supplementary Material

A. Dataset

A.1. Source MOT dataset

To evaluate video-based person ReID methods in challenging scenarios with high visual similarity between individuals, we construct two new datasets. We focus on Multi-Object Tracking (MOT) datasets, as they provide person ID labels for each individual along with their bounding boxes in each frame. We adopt the SportsMOT dataset [2] and the DanceTrack dataset [5]. The SportsMOT dataset contains 240 video clips collected from three sports: basketball, soccer, and volleyball. The DanceTrack dataset comprises 100 video clips capturing group dance scenes. Both datasets feature multiple individuals wearing similar clothing, making it extremely challenging to distinguish each person.

A.2. Dataset creation process

MOT datasets typically include videos, per-frame person ID labels, and bounding boxes. First, using the person ID labels and bounding boxes, we crop the regions corresponding to each individual from every frame of each video. Since the labels in MOT datasets are assigned independently for each video and do not correspond across videos, it is not possible to automatically detect and match the same individual appearing in different videos. Therefore, we employ two annotators to describe the characteristics of each individual (*e.g.*, gender, hairstyle, clothing, socks, shoes, and jersey number) for every video. Table S.1 shows examples of the annotations provided in our SportsVReID dataset. Note that the actual annotations are originally created in Japanese and translated into English for this paper. Based on these descriptions, we manually perform person matching across videos and reassign new labels to the entire dataset.

Table S.1. Examples of manual annotations for SportsVReID.

ID	Sports	Gender	Uniform	Number	Others
1	basketball	woman	yellow	5	bun hair
2	basketball	woman	blue	10	red shoes
91	soccer	man	gray, blue	32	black socks
92	soccer	man	white	1	red shoes

Subsequently, we filter out images where the target individual is barely visible or the image size is too small. We then divide the videos into tracklets (short video clips) for each person, ensuring that the maximum frame length is 50. Through this process, we create the SportsVReID and DanceVReID datasets from the SportsMOT and DanceTrack datasets, respectively. It is worth noting that each MOT dataset is divided into three subsets: train, val, and test. However, since the test set does not include ground truth (*i.e.*, no labels or bounding box information), we use only the train and val sets, with the train set for training and the val set for evaluation.

A.3. Caption generation

In this paper, we utilize Multi-modal Large Language Models (MLLMs), such as Phi-4-Multimodal (phi-4-mm) [1], to perform image captioning, caption augmentation, and translation for generating captions. For each task, we first input a few examples into GPT-4o [3] to generate response examples, which are then included in the prompt to conduct caption generation in a few-shot manner.

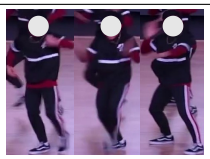
Captions for existing datasets. For existing benchmark datasets (MARS [8], iLIDS-VID [6]), we generate one caption per image using phi-4-mm. Below, we present the prompt sample used for the image captioning task:

> *Write a description about the overall appearance of the person in the image, including the attributes: clothes, shoes, hairstyle, gender, belongings.*
>
> *(Output Examples)*
>
> - *A man is wearing a white short-sleeved T-shirt and black long pants. His shoes are gray and he has short hair.*
> - *A woman with long black hair is dressed in a pink short-sleeved shirt and short navy blue pants. She is wearing pink sandals.*

Additionally, for the MARS dataset, which is relatively larger in scale compared to other datasets, we perform caption-to-caption generation to increase caption diversity. This process creates multiple diverse captions with semantically equivalent content from a single source caption.

Captions for new datasets. For our SportsVReID and DanceVReID datasets, we synthesize captions based on manually assigned annotations created during the dataset creation process. First, the annotation data written in Japanese is translated into English using phi-4-mm. Since the annotation data contain only one description per identity, we perform paraphrasing with phi-4-mm to create variations of the translated sentences. This process expands the data to 10 captions per identity.

Table S.2. Examples of images and captions from the SportsVReID and DanceVReID datasets. Each row corresponds to a different person.

Dataset	Images	Caption
SportsVReID		A female basketball player is wearing a blue uniform with the number 3. She has a ponytail.
		A woman basketball player is seen in a blue uniform with the number 7, and she also wears black shoes.
		A man soccer player, in a white uniform with the number 32, also has black shoes.
DanceVReID		A female is dressed in a white cropped T-shirt with a black inner layer and white track pants. She wears white sneakers and her hair is styled in a half-updo.
		A male is wearing a shirt with red, white, and black stripes on the upper body and red, white, and black striped pants on the lower body. He also has on white socks and white sneakers with black accents, and his hair is short.
		She is a woman dressed in a red, white, and black top and matching red, white, and black pants. She also wears white socks and white sneakers with black accents, and her hair is in a ponytail. She is wearing a short jacket.

A.4. Dataset visualization

Table S.2 shows examples from our SportsVReID and DanceVReID datasets. Each row displays a video sequence of a different person along with the corresponding caption. Both SportsVReID and DanceVReID include individuals wearing nearly identical uniforms or costumes, making them significantly more challenging person ReID datasets than previously available benchmarks (*e.g.*, MARS).

B. Ablation study

B.1. Analysis of the fusion encoder in CMR

We investigate the effectiveness of different approaches for processing the Text Memory and image features in the Caption-guided Memory Refinement (CMR) module. As illustrated in Fig. S.1, we evaluate three configurations: (a) concatenating Text Memory and image features before feeding them into a self-attention layer, (b) applying a self-attention layer to the Text Memory before feeding it into the cross-attention layer with image features, where image features serve as keys and values, and (c) our proposed method, which first applies a cross-attention layer with image fea-

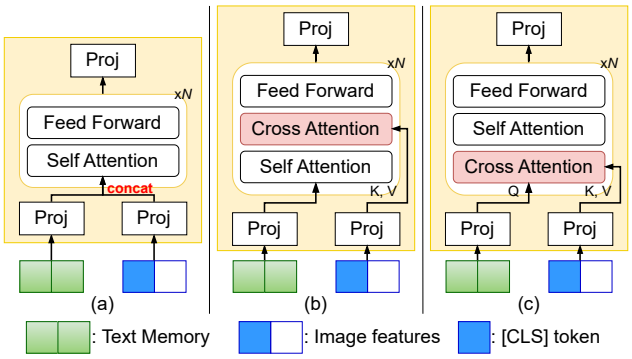


Figure S.1. Illustration of three types of fusion encoders in CMR.

Table S.3. Comparison of different types of fusion encoders in CMR.

Method	MARS		SportsVReID	
	mAP	Rank-1	mAP	Rank-1
(a) Self-attn (concat)	89.6	92.0	76.9	87.1
(b) Self-attn → Cross-attn	89.7	92.4	77.2	88.2
(c) Cross-attn → Self-attn	89.8	92.5	77.7	90.4

Table S.4. Effect of varying the number of Transformer blocks in CMR.

# Blocks	MARS		SportsVReID	
	mAP	Rank-1	mAP	Rank-1
1	89.5	91.9	77.5	90.1
2	89.8	92.5	77.7	90.4
3	89.6	91.9	76.8	87.1
4	89.6	92.4	77.8	88.2

tures as keys and values, followed by a self-attention layer. All configurations use 2 Transformer blocks. As observed in Tab. S.3, our proposed method (c) outperforms the other two configurations on both the MARS and SportsVReID datasets.

We further analyze the impact of varying the number of Transformer blocks in the CMR module. As shown in Tab. S.4, we evaluate configurations with 1, 2, 3, and 4 blocks on MARS and SportsVReID. The results indicate that using 2 blocks achieves the best performance across both datasets, attaining the highest Rank-1 scores. Increasing the number of blocks beyond 2 does not consistently improve performance; an excessive number of blocks (*e.g.*, 3 or 4) may lead to slight performance degradation, likely due to overfitting or increased model complexity.

B.2. Identity-aware text strategies

For existing benchmark datasets, automatically generated captions may suffer from hallucination effects, where MLLMs assign the same or highly similar captions to different individuals. This poses a critical issue for our method, as we use features derived from captions as targets for contrastive learning, which requires these features to be unique to each identity. To encourage identity-unique text features, we investigate two simple strategies summarized in Fig. S.2 on MARS and iLIDS-VID, where captions are fully generated by an MLLM. **(1) ID text.** We append a short identity string to each caption: *"The person's ID is [ID LABEL]."* **(2) ID emb.** We add a learnable identity embedding to the caption feature, analogous to the positional embeddings in Transformers. Table S.5 shows that combining captions with *ID text* yields the best performance on both datasets. This suggests that explicitly injecting identity information

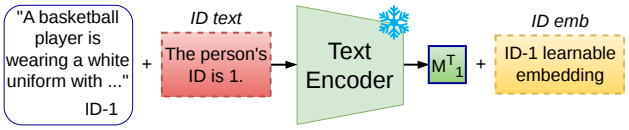


Figure S.2. Illustration of text variants used to construct the Text Memory on existing benchmarks (MARS and iLIDS-VID): caption, *ID text*, and *ID emb*.

Table S.5. Comparison of different types of input texts.

Input text type			MARS		iLIDS-VID	
Caption	*ID text*	*ID emb*	mAP	Rank-1	Rank-1	Rank-5
✓	×	×	89.7	92.3	96.0	99.9
×	✓	×	89.5	92.3	94.7	99.9
×	×	✓	89.6	92.3	95.3	99.3
✓	✓	×	89.8	92.5	96.7	99.9
✓	×	✓	89.8	92.4	96.0	99.9

Table S.6. Effect of different caption sources.

Method	SportsVReID		DanceVReID	
	mAP	Rank-1	mAP	Rank-1
TF-CLIP [7] (w/o caption)	77.3	89.7	51.7	70.8
Ours: MLLM only	77.5	89.3	53.5	74.2
Ours: Manual + MLLM	77.7	90.4	53.8	76.0

is an effective and lightweight way to mitigate caption ambiguity on existing benchmarks.

B.3. Effect of caption sources

In this work, the captions used for training on SportsVReID and DanceVReID are generated based on manually annotated data provided during dataset creation. Specifically, for SportsVReID, accurate annotations including shoe color, sock color, and jersey number are assigned to each player, ensuring high caption quality. However, when applying our method to other datasets, obtaining such precise text annotations can be challenging. In such cases, generating pseudo-captions using MLLMs for image captioning, as we apply to existing video-based person ReID datasets, becomes a practical solution. Therefore, we also conduct training using captions generated solely by MLLMs for both SportsVReID and DanceVReID. In this setting, we also append *ID text* to each pseudo-caption to make the text features identity-discriminative.

As shown in Tab. S.6, we compare the performance of different caption sources. The results indicate that our method, which combines manual annotations with MLLM-generated captions, achieves the best performance on both datasets. Notably, even when using only MLLM-generated captions, the performance remains competitive, demonstrating the feasibility of this approach for datasets where manual annotation is impractical.

Figure S.3. Visualization of top-5 retrieval results. Green and red boxes represent correct and incorrect matches, respectively. Blue-outlined rows show SportsVReID results, and orange-outlined rows show DanceVReID results. Each image shows the first frame of a tracklet.

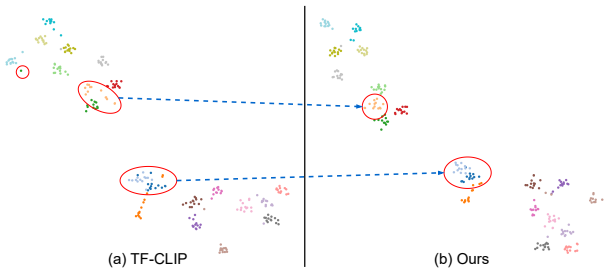


Figure S.4. t-SNE visualization of TF-CLIP and our CG-CLIP on the DanceVReID val set. Different colored dots represent different identities. Best viewed in color.

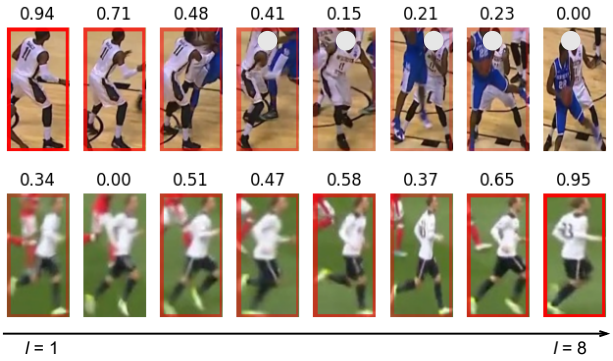


Figure S.5. Visualization of attention weights between the first learnable token and image features from each frame in TFE. Higher values indicate stronger attention to the corresponding frame.

C. Visualization

C.1. Visualization of inference results

To comprehensively analyze the effectiveness of our method compared to the comparative method [7], we visualize the ReID inference results with top-5 rankings. As shown in Fig. S.3, the top two rows with blue outlines present inference results on SportsVReID, while the bottom two rows with orange outlines show examples from DanceVReID. Green and red boxes indicate correct and incorrect matches, respectively. Note that only the first frame of each 8-frame tracklet is displayed for clarity. These visualizations demonstrate our method's superior ability to handle high-difficulty scenarios where many individuals with similar appearances are present.

C.2. t-SNE visualization of feature distributions

To further validate the discriminative capability of our learned features, we perform t-SNE [4] visualization on DanceVReID. We sample 20 identities from three dance groups with high visual similarity and visualize the feature distributions extracted by TF-CLIP [7] and our method in Fig. S.4. As shown in Fig. S.4 (a) and (b), our method produces more compact and well-separated clusters for several identities. The red circles highlight specific examples where our approach achieves tighter intra-class clustering, demonstrating that our method effectively captures identity-specific features even in challenging scenarios with highly similar appearances.

C.3. Visualization of attention in TFE

We visualize the attention weights between the learnable tokens in the Token-based Feature Extraction (TFE) module and the image features from each input frame. Figure S.5 presents these results, where the numerical values above each frame represent the attention weight associated with the first learnable token. The visualization reveals that frames with severe occlusion by other people or blur due to rapid motion receive lower attention weights, while frames where the target person is clearly visible receive higher weights. This indicates that our method effectively selects informative frames for feature aggregation.

Table S.7. Training settings. "Others" include SportsVReID and DanceVReID.

Hyperparameter	Settings MARS	iLIDS-VID	Others
Batch size	32	32	32
# Identities per batch	8	8	8
# Tracklets per identity	4	4	4
# Frames per tracklet	8	8	8
Patch size	16	16	16
Image size ($[H, W]$)	[256, 128]	[256, 128]	[256, 128]
Max. # text tokens	77	77	77
N^Q (learnable tokens)	50	15	15
Momentum factor	0.2	0.2	0.2
L_{v2rm} weight	1.0	1.0	1.0
L_{tri} weight	1.0	1.0	1.0
L_{ce} weight	0.25	0.25	0.25
Epochs	80	60	60
Optimizer	Adam	Adam	Adam
Learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}
Weight decay	1×10^{-4}	1×10^{-4}	2.5×10^{-4}
LR scheduler	StepLR ($\times$0.1 at epochs 30, 50, 70*)		

*Only for MARS dataset.

D. Implementation details

We provide comprehensive training settings for all datasets in Tab. S.7. Our model is implemented using PyTorch and trained on a single NVIDIA A5000 GPU with 24GB memory. All experiments use ViT-B/16 as the image encoder from the pre-trained CLIP model. For data augmentation, we apply random flipping and random erasing [9] during training. The learning rate is warmed up linearly from 5×10^{-7} to 5×10^{-6} over the first 10 epochs.

E. Limitations

While our CG-CLIP framework demonstrates significant improvements in video-based person ReID, we acknowledge several limitations that warrant future investigation.

First, our method's performance is inherently dependent on the quality of generated captions. MLLM-based image captioning is susceptible to hallucinations, particularly when person images have low resolution or poor clarity, which are common in surveillance and sports video scenarios. These inaccuracies in pseudo-captions can propagate through our CMR module and potentially degrade the final person ReID performance. Future work should focus on developing more robust caption generation methods, quality assessment mechanisms, and filtering strategies.

Second, although our method achieves substantial performance improvements over existing approaches in high-difficulty scenarios, we observe that some challenging cases remain, particularly in dance scenes, where individuals exhibit nearly identical visual attributes, making it extremely difficult to verbalize subtle differences through textual descriptions. In such extreme scenarios, language-based descriptions face inherent limitations in expressiveness, as fine-grained visual differences may not be easily captured through natural language. Future approaches could address this limitation by incorporating non-linguistic cues such as facial characteristics, or by developing hybrid frameworks that adaptively balance language-guided and pure visual feature learning based on scenario difficulty.

Finally, while our TFE module improves computational efficiency through its linear complexity with respect to input length, the overall inference speed of our framework is still largely governed by the image encoder. For real-time applications such as online multi-object tracking, future work could explore lightweight architectures or model compression techniques, including pruning and quantization, to accelerate the image encoder while maintaining person ReID accuracy.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 1

[2] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9921–9931, 2023. 1

[3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 4

[5] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20993–21002, 2022. 1

[6] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision (ECCV)*, pages 688–703. Springer, 2014. 1

[7] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6764–6772, 2024. 3, 4

[8] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 868–884. Springer, 2016. 1

[9] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. 5