DisenQ: Disentangling Q-Former for Activity-Biometrics

Shehreen Azad Yogesh Singh Rawat
Center for Research in Computer Vision
University of Central Florida
https://sacrcv.github.io/DisenQ-website/

Abstract

In this work, we address activity-biometrics, which involves identifying individuals across diverse set of activities. Unlike traditional person identification, this setting introduces additional challenges as identity cues become entangled with motion dynamics and appearance variations, making biometrics feature learning more complex. While additional visual data like pose and/or silhouette help, they often struggle from extraction inaccuracies. To overcome this, we propose a multimodal language-guided framework that replaces reliance on additional visual data with structured textual supervision. At its core, we introduce ***DisenQ*** *(****Disen****tangling* ***Q****-Former), a unified querying transformer that disentangles biometrics, motion, and non-biometrics features by leveraging structured language guidance. This ensures identity cues remain independent of appearance and motion variations, preventing misidentifications. We evaluate our approach on three activity-based video benchmarks, achieving state-of-the-art performance. Additionally, we demonstrate strong generalization to complex real-world scenario with competitive performance on a traditional video-based identification benchmark, showing the effectiveness of our framework.*

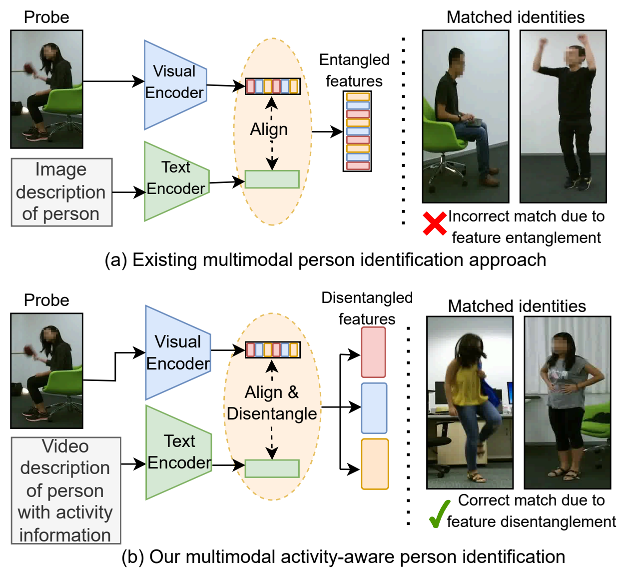


Figure 1. **Comparison of existing and our activity-aware person identification framework:** (a) Existing multimodal methods [32] lacking activity-awareness leads to misidentifications due to entangled biometrics and motion features. (b) Our model disentangles biometrics and motion features using language guidance, enabling activity-awareness and more accurate identification across diverse activities while being appearance invariant.

1. Introduction

Traditional person identification aims to recognize the same individual across different cameras, time and location [57], focusing on facial recognition [1, 38], gait analysis [16, 17, 33, 35], and whole-body biometrics [19, 32, 43, 58]. While effective in gait-based analysis, these approaches struggle when individuals engage in diverse daily activities beyond just standing or walking. Extending beyond conventional methods, activity-biometrics involves identifying individuals from their daily activities [3] by leveraging motion dynamics, making it more suitable for real-world applications like surveillance, healthcare, smart environments where recognition across diverse activities is crucial.

Effective identification in activity biometrics requires disentangling biometric features from appearance-based non-biometric and motion cues to prevent identity bias and fully leverage motion dynamics. Existing method [3] address this using additional visual modalities like silhouettes, but their reliance on accurate extraction limits reliability in real-world settings to aid feature separation. This highlights the need for an approach that ensures effective feature separation without additional visual data, improving generalization across varying conditions and activities.

To reduce reliance on additional visual data, cross-modal learning using text has been explored with CLIP-based contrastive learning methods [22, 32, 58]. While effective in aligning global image-text associations, these methods lack identity-specific feature separation, leading to misidentifications due to appearance variation. Additionally, their failure to maintain temporal consistency makes them less suitable for video-based activity-biometrics where motion dynamics play a critical role (Figure 1). Multimodal Large

Language Models (MLLMs) offer a generative alternative by leveraging structured textual reasoning to achieve better feature separation and robustness in complex scenarios [4, 5, 11, 31, 36, 47, 48]. However, existing MLLM-based person identification approaches [26, 50] rely on additional visual data and are image based, making them ineffective at capturing motion dynamics essential for activity-biometrics.

To enhance activity-based person identification without relying on additional visual data, we introduce a multimodal **Disen**tangling **Q**uerying (**DisenQ**) Transformer based framework that explicitly separates biometrics, non-biometrics, and motion features. Unlike prior works that rely on additional visual modalities, our approach uses structured text descriptions generated from a frozen Vision-Language Model (VLM) to provide semantic supervision for cross-modal feature learning. These language descriptions serve as explicit guidance, enabling robust feature separation by distinguishing identity-related attributes from appearance and motion cues.

At the core of our approach is our proposed **DisenQ**, a Disentangling Querying Transformer, that separates biometrics, non-biometrics and motion features using structured textual guidance while minimizing feature leakage. The disentangled features enhance similarity-based retrieval in a traditional identification pipeline. By leveraging language-driven supervision as an auxiliary modality, DisenQ eliminates the need for additional visual modalities, reduces reliance on biased visual cues, and improves generalization across diverse real-world activities.

We evaluate our framework on three activity-based video benchmarks, demonstrating strong generalization across diverse activities. Additionally, to assess its broader applicability, we evaluate it on a large-scale traditional video-based identification benchmark. Our approach achieves state-of-the-art performance on most datasets and remains competitive on others, highlighting its robustness in real-world scenarios. Our main contributions are as follows:

- We use language guidance for activity-biometrics to explicitly disentangle feature spaces, enabling cross-modal learning without additional visual data.
- We introduce **DisenQ** (Disentangling Q-Former) that effectively separates biometrics, non-biometrics, and motion features for activity-aware person identification.
- Our method achieves state-of-the-art performance on activity-based benchmarks and competitive performance on a traditional video-based person identification benchmark highlighting its generalization ability across diverse real-world scenarios.

2. Related Works

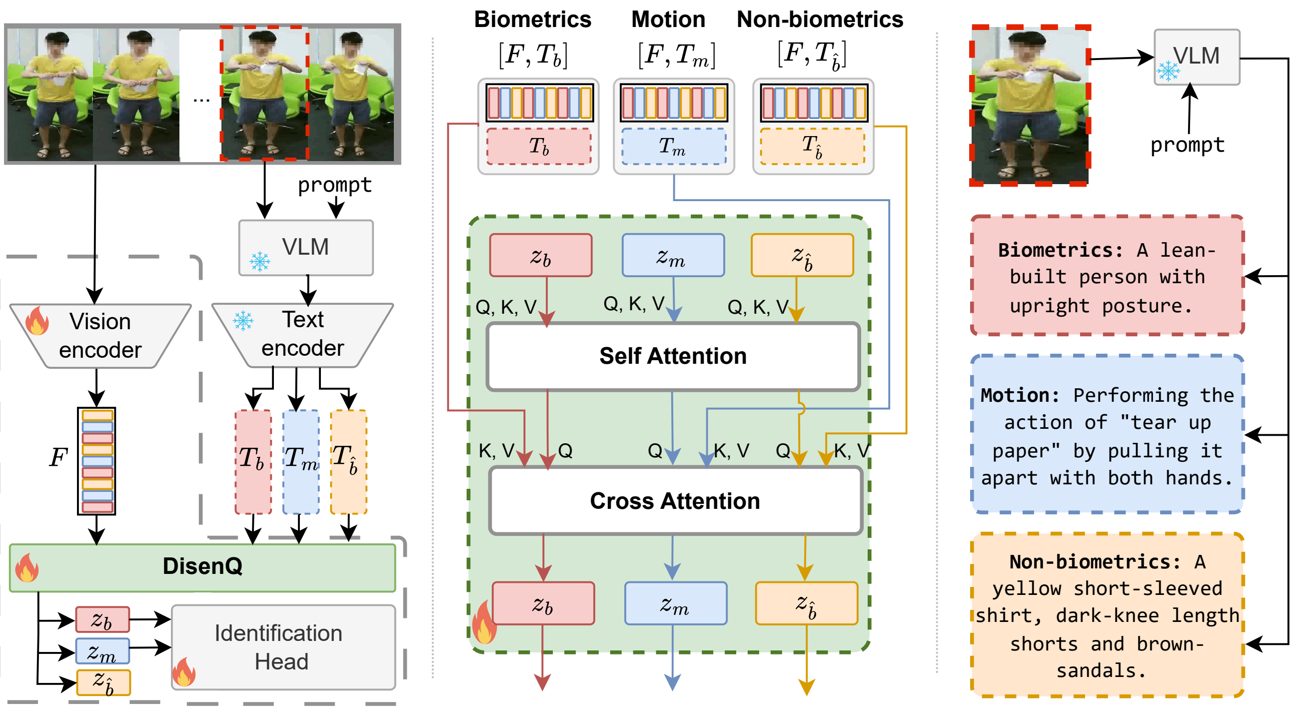
Visual modality based person identification. Traditional person identification primarily rely on image-based approaches [7, 23, 27, 40, 54], focusing on body shape, clothing, and appearance. While some works improve robustness with cloth-invariant representations [19, 20, 55], or additional modalities like silhouettes [29], skeletons [42, 45], or 3D shape [9], they lack spatiotemporal awareness. Video-based approaches [6, 8, 21, 25, 28, 37, 41, 43, 44, 51] integrate temporal cues but remain limited to walking-based identification, making them less effective for activity-based identification. The only prior activity-based identification method [3] relies on silhouette extraction, which is unreliable in challenging conditions. Instead, we leverage structured language supervision to enhance identity learning without requiring additional visual data.

Language modality based person identification. Several CLIP based identification approaches [10, 22, 32, 34, 53, 56, 58] leverage image-text contrastive learning [46] for traditional image-based person identification. However, these methods lack temporal modeling and cannot disentangle motion from biometric features, making them unsuitable for activity-based identification. Similarly, large foundation models for image-based person identification [26, 50] incorporate language but rely on additional visual modalities, as well as lack temporal modeling capabilities. To address this, we leverage language guidance not just for cross-modal learning, but to explicitly disentangle video-based features, enabling a more robust and adaptable framework for activity-aware identification.

Multimodal Large Language Models. Several works [2, 11, 30, 31, 46, 61] have been introduced to enhance image-text alignment by bringing visual features closer to language space. Among these, BLIP-2's Q-Former [31] is a lightweight approach that effectively aligns visual and language modalities. We build our proposed DisenQ based on this architecture to align visual features with language while disentangling biometrics, non-biometrics, and motion through structured textual guidance. To the best of our knowledge, this is the first work to leverage Q-Former for feature disentanglement.

3. Method

We propose a **Disen**tangling **Q**-Former (**DisenQ**) based multimodal framework for activity-biometrics. Our framework consists of a visual encoder to extract visual features from an RGB video and a frozen VLM to generate a detailed prompt which is encoded via a text encoder (Section 3.1). To ensure robust feature disentanglement, DisenQ separates biometrics, non-biometrics, and motion features in the visual domain through language guidance (Section 3.2). Finally, an identification head performs person identification (Section 3.3), ensuring that biometrics features remain clothing-invariant while incorporating motion cues for improved identity matching. Given an RGB video V, with ground truth actor and activity labels (y_{ID} and y_{Action}),



(a) Framework Overview (b) DisenQ Overview (c) Structured Textual Description

Figure 2. **Framework overview:** *(a)* Given an RGB video, our model performs language-guided activity-aware person identification using the proposed **Disen**tangling **Q**-Former (DisenQ). *(b)* DisenQ disentangles biometrics, motion, and non-biometrics visual features using structured language guidance and dedicated learnable queries for each feature type. *(c)* The frozen VLM generates structured textual descriptions from the key-frame (red dotted line) to serve as language supervision for DisenQ. Here, $[F, T_x]$ denotes concatenation of F and T_x. During inference, only the components within the gray dashed box are utilized, removing the need for text generation.

the framework aims to correctly match the identity of the person in a probe video to its corresponding identity in the gallery. The overall approach is illustrated in Figure 2.

3.1. Feature Extraction

Visual feature extraction. Given a sequence of frames from a video V, each frame $v_i \in \mathbb{R}^{H \times W \times 3}$, where H and W represent its height and width, is processed through a visual encoder to extract visual features $f_i \in \mathbb{R}^{N \times D}$. Here, N denotes the number of extracted visual tokens per frame, and D is the hidden dimension of each token. Here, each visual feature f_i has temporal ordering information associated with it through a position embedding layer. Finally, temporal attention pooling is applied on all frame features to get a global video-level feature F.

Prompt generation and textual feature extraction. To generate structured and semantically consistent language description, we use a frozen VLM to generate prompts from the key-frame of the input video during only training, without requiring the VLM during inference. These descriptions are categorized into three distinct components following pre-defined templates: Biometrics prompt (P_b), describing identity-specific traits such as body shape, posture, and notable physical characteristics; Motion prompt (P_m) describing the action label and movement; and Non-biometrics prompt ($P_{\hat{b}}$), describing clothing, and accessories. To maintain consistency, biometrics descriptions are generated only once per unique identity and reused in all subsequent videos of the same actor by storing and iteratively refining it by updating the stored description using a running average. This prevents major description drift and ensures stable identity representation across varied activities and appearances.

The generated prompts are then encoded using a pre-trained frozen text-encoder to obtain textual embeddings ($T_b, T_m, T_{\hat{b}}$) which serve as language-driven supervision for visual feature disentanglement.

3.2. DisenQ

We introduce **DisenQ** (**Disen**tangling **Q**uerying Transformer) to separate biometrics, motion and non-biometrics features in the visual domain by aligning visual representations with structured textual cues. Adapted from the original Q-Former [31], DisenQ introduces three separate sets of learnable queries: z_b (biometrics), z_m (motion) and $z_{\hat{b}}$ (non-biometrics); instead of a single query set, enabling explicit disentanglement. Each query set shares the same self-attention and cross-attention layers while leveraging textual guidance, ensuring effective feature separation. However, they explicitly attend to different information without interaction, preserving distinct feature representations for biometrics, motion and non-biometrics. The learned queries are then utilized for activity-based person identification, improving the model's ability to distinguish individuals based on biometrics while leveraging motion cues and remaining invariant to non-biometrics attributes.

Biometrics feature disentanglement. To extract identity-related features, the biometrics query z_b attends to itself through self-attention to refine itself. Then the refined query performs cross-attention with the visual feature F and biometrics textual supervision features T_b with query, key and value being used as Equation 1.

$$Q_b = W z_b, \;\; K_b = W[F, T_b], \;\; V_b = W[F, T_b]. \tag{1}$$

Here, $[F, T_b]$ denotes concatenation of F and T_b, followed by a linear projection.

Motion feature disentanglement. To extract motion-specific representations, the motion query z_m, similar to biometrics query z_b, first undergoes self-attention, ensuring it refines motion-related patterns independently. Subsequently, the motion query cross-attends to the visual feature F and its corresponding textual feature T_m with query, key and value acting as Equation 2.

$$Q_m = W z_m, \;\; K_m = W[F, T_m], \;\; V_m = W[F, T_m]. \tag{2}$$

Non-biometrics feature disentanglement. To separate non-biometrics features, the non-biometrics query $z_{\hat{b}}$ similar to others, also, first undergoes self-attention, refining itself without influence from other feature categories. Following this, the non-biometrics queries cross-attend to the visual feature F and non-biometrics textual feature $T_{\hat{b}}$ with query, key and value acting as Equation 3.

$$Q_{\hat{b}} = W z_{\hat{b}}, \;\; K_{\hat{b}} = W[F, T_{\hat{b}}], \;\; V_{\hat{b}} = W[F, T_{\hat{b}}]. \tag{3}$$

3.3. Identification Head

The learned query embeddings z_b, z_m and $z_{\hat{b}}$ go through mean pooling to form single vectors, denoted as F_b, $F_{\hat{b}}$, and F_m, among which only F_b and F_m is used for final identification.

Loss Functions. During training, the model is optimized to refine F_b using a combination of standard cross-entropy ($\mathcal{L}_{ID}$), and triplet loss ($\mathcal{L}_{Tri}$) following [3, 19]. These losses are defined as Equation 4 and Equation 5.

$$\mathcal{L}_{ID} = -y \log \hat{y} \simeq \mathcal{L}_{Act}, \tag{4}$$

$$\mathcal{L}_{Tri} = \max\left(\mathcal{D}(F_b^a, F_b^p) - \mathcal{D}(F_b^a, F_b^n) + m, 0\right), \tag{5}$$

Here, y and $\hat{y}$ denote the ground truth and predicted labels. F_b^p and F_b^n represent the positive and negative biometrics features for an anchor biometrics feature F_b^a within the same batch. $\mathcal{D}(\cdot)$ computes the Euclidean distance, and m is the margin in the triplet loss.

Since the motion feature F_m contributes to identity recognition, it is explicitly trained to preserve motion-related information while remaining independent of biometrics attributes. The model is optimized for F_m using the cross-entropy loss ($\mathcal{L}_{Act}$) of Equation 4.

Furthermore, to reinforce the independence of biometrics and non-biometrics features, an orthogonality constraint is imposed between F_b and $F_{\hat{b}}$ as Equation 6.

$$\mathcal{L}_{Orth} = \left\| F_b^T F_{\hat{b}} \right\|. \tag{6}$$

The overall loss function is defined as Equation 7.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}_{Tri} + \lambda_3 \mathcal{L}_{Orth} + \lambda_4 \mathcal{L}_{Act}. \tag{7}$$

Here, $\lambda_{i \in 1,..,4}$, is weighting factor for each loss term.

Identity Similarity Computation. To enhance identity matching, we introduce an adaptive weighting mechanism that integrates motion features into the similarity calculation, unlike traditional methods that rely solely on biometrics. Instead of fixed weights, we use a lightweight MLP to dynamically adjust the contribution of biometrics and motion features based on their relevance. Given a probe identity A and gallery identity B, we compute cosine similarities for both biometrics and motion features, concatenate them, and pass them through the MLP with ReLU activations and a softmax function. This enables the model to leverage motion cues to guide biometrics matching, prioritizing motion when it provides meaningful identity information and relying more on biometrics when motion cues are less discriminative. The final similarity score is computed as Equation 8.

$$Sim(A, B) = \alpha_1 Sim_b(A, B) + \alpha_2 Sim_m(A, B). \tag{8}$$

Here, $\alpha_{i \in 1,2}$ are the weighting factors.

Inference. DisenQ operates without textual supervision during inference, relying solely on the learned query embeddings acquired during training. It utilizes self-attention to retain query-specific information and cross-attention to extract relevant visual features, ensuring effective disentanglement of biometrics, non-biometrics and activity features purely from visual embeddings.

4. Experiments

Datasets: We evaluate our model on NTU RGB-AB, PKU MMD-AB, and Charades-AB, following [3]. NTU RGB-AB consists of 106 actors performing 94 actions across

Table 1. **Performance comparison of activity-based person identification** on NTU RGB-AB, PKU MMD-AB, Charades-AB on same activity (denoted by Same) and cross activity (denoted by Cross) evaluation protocol. R@1 denotes Rank 1 accuracy. † denotes results produced in our environment. **Bold** and underline denotes best and second best results.

Methods	Venue	NTU RGB-AB				PKU MMD-AB				Charades-AB			
		Same		Cross		Same		Cross		Same		Cross	
		R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
Models with only visual modality													
TSF [28]	AAAI 20	71.8	31.8	67.8	26.9	76.4	37.5	71.6	33.2	35.4	21.9	30.2	19.0
VKD [44]	ECCV 20	67.4	35.6	66.3	31.5	78.4	38.5	72.2	34.3	36.3	20.7	31.9	18.8
BiCnet-TKS [25]	CVPR 21	72.7	34.5	69.1	30.2	80.8	38.5	77.1	33.3	40.3	27.3	38.3	23.3
PSTA [51]	ICCV 21	67.4	34.8	65.1	31.4	77.4	50.4	72.4	47.4	42.9	28.3	38.7	24.8
STMN [15]	ICCV 21	73.0	35.1	70.2	30.1	76.6	47.9	71.5	42.2	38.7	24.5	33.9	20.8
SINet [6]	CVPR 22	69.4	30.7	66.2	27.8	79.6	40.8	74.1	26.2	40.3	26.9	37.3	21.9
CAL [19]	CVPR 22	73.8	28.4	70.3	24.0	81.3	49.4	78.3	43.4	43.8	25.8	40.1	21.2
Video-CAL [19]	CVPR 22	75.5	39.9	73.3	31.7	79.6	49.4	77.3	45.7	43.9	28.5	41.5	25.8
PSTR [7]	CVPR 22	69.1	34.1	68.3	32.5	84.3	47.5	78.0	41.2	37.2	24.7	35.1	20.3
AIM [55]	CVPR 23	71.4	35.4	72.8	30.2	82.5	48.9	79.2	44.9	40.1	28.3	35.6	26.7
SCNet [20]	ACM MM 23	69.9	31.5	68.8	26.3	79.5	43.6	73.9	39.7	31.7	21.9	27.4	17.6
ABNet [3]	CVPR 24	<u>78.8</u>	40.3	<u>77.0</u>	<u>37.6</u>	<u>86.8</u>	<u>57.3</u>	81.4	51.8	<u>45.8</u>	<u>31.6</u>	<u>44.8</u>	<u>28.8</u>
Models with visual + language modality													
CLIP ReID [32] †	AAAI 23	77.1	40.2	75.2	33.7	82.3	52.1	81.2	50.8	44.2	31.3	42.1	27.7
CCLNet [10] †	ACM MM 23	75.2	36.1	74.3	33.1	83.2	51.4	80.1	47.5	42.1	29.3	38.8	23.4
TF-CLIP [58] †	AAAI 24	77.3	41.2	74.8	31.3	83.4	52.3	80.8	50.1	40.2	28.1	39.7	26.0
TVI-LFM [26] †	NeurIPS 24	76.2	38.1	75.9	34.1	85.2	53.9	81.5	52.1	45.7	30.1	42.8	28.3
Instruct-ReID [22] †	CVPR 24	78.2	<u>41.5</u>	75.9	33.4	84.3	53.1	<u>81.7</u>	<u>52.3</u>	44.8	28.3	40.1	25.3
EVA-CLIP [49] †		71.2	35.1	69.1	28.3	73.8	46.2	67.4	39.4	38.1	26.1	31.3	21.8
Ours		**82.2**	**43.8**	**80.9**	**41.3**	**89.2**	**59.3**	**84.1**	**56.9**	**49.9**	**34.8**	**48.4**	**32.5**

88.7k samples, while PKU MMD-AB includes 66 actors, 41 actions, and 17k samples. Charades-AB features 267 actors with 157 actions across 9.8k videos, averaging 6.8 activities per video. To assess the generalization capability of our model on more challenging real-world scenarios, we evaluate it on MEVID [12], which includes 158 actors and 8k tracklets, incorporating greater viewpoint, distance, and lighting variations, making it a more complex benchmark for video-based identification.

Evaluation Protocol and Metrics: We follow the same evaluation protocol and dataset splits as [3] for NTU RGB-AB, PKU MMD-AB, and Charades-AB, employing two evaluation protocols: same-activity and cross-activity. Additionally, due to view information explicitly being available for NTU RGB-AB and PKU MMD-AB, we evaluate including and excluding same-view settings too. For MEVID, we use the official protocol and splits. We report rank 1, rank 5 accuracies and mAP as evaluation metrics.

4.1. Implementation Details

We use 8 frames which are randomly selected with a stride of 4 from each original video to create RGB clip. Each frame is resized to 224×224 and horizontal flipping is used for data augmentation, following [3, 18]. We use pre-trained ViT G/14 from EVA-CLIP [49] as the visual encoder and BERT [13] as the frozen text encoder. Additionally, we use LLaVA 1.5 7B [36] as the frozen VLM to generate prompts. We initialize DisenQ with pre-trained weights from InstructBLIP [11]. We train the model for 60 epochs with a batch size of 32 with each batch containing 8 person and 4 clips for each person. AdamW is used as the optimizer with weight decay of $5e-2$ and base learning rate of $1e-4$ with β values as $[0.9, 0.999]$. The triplet loss margin m is set to 0.3, $\lambda_i \in [1, .., 4]$ in Equation 7 is set as 0.01.

4.2. Results

Performance on activity-biometrics benchmarks. Table 1 presents the performance comparison of our framework against other existing methods. Across all datasets, our model outperforms the previous best-performing approach, improving Rank-1 accuracy and mAP across all evaluation protocols on NTU RGB-AB, PKU MMD-AB, and Charades-AB. Notably, we observe an average Rank-1 accuracy improvement of 3.7%, 2.4% and 3.9% respectively on NTU RGB-AB, PKU MMD-AB, and Charades-AB, demonstrating the effectiveness of our approach. We present more results in supplementary.

Generalization to traditional video-based benchmark. Table 2 presents the identification results of our model compared to concurrent methods on MEVID, a large-scale traditional video-based identification dataset primarily focused on walking sequences. Unlike NTU RGB-AB, PKU

Table 2. **Performance comparison of traditional person identification** on MEVID in general evaluation setting. R@1 and R@5 denote rank 1 and rank 5 accuracies. † denotes reproduced results.

Methods	Venue	R@1	R@5	mAP
Models with only visual modality				
Attn-CL [43]	AAAI 20	42.1	56.0	18.6
Attn-CL + rerank [43]	AAAI 20	46.5	59.8	25.9
AP3D [18]	ECCV 20	39.0	56.0	15.9
TCLNet [24]	ECCV 20	48.1	60.1	23.0
BiCnet-TKS [25]	CVPR 21	19.0	35.1	6.3
STMN [14]	ICCV 21	31.0	54.4	11.3
PSTA [51]	ICCV 21	46.9	60.8	21.2
PiT [59]	TII 22	34.2	55.4	13.6
CAL [19]	CVPR 23	52.5	66.5	27.1
ShARc [62]	WACV 24	<u>59.5</u>	**70.3**	29.6
ABNet [3] †	CVPR 24	58.3	68.4	<u>30.1</u>
Models with visual + language modality				
CLIP ReID [32] †	AAAI 23	51.2	64.2	28.3
CCLNet [10] †	ACM MM 23	50.8	60.3	27.1
TVI-LFM [26] †	NeurIPS 24	49.2	61.8	23.7
Instruct-ReID [22] †	CVPR 24	53.8	59.4	28.4
EVA-CLIP [49] †		53.1	59.2	26.9
Ours		**60.7**	**70.3**	**30.4**

Table 3. **Ablation studies for each component.** Here, F_b, $F_{\hat{b}}$ and F_m denote biometrics, non-biometrics and motion features.

Method	NTU RGB-AB		Charades-AB	
	Rank 1	mAP	Rank 1	mAP
Contribution of each component				
Vision encoder	73.2	36.2	40.1	29.2
+ Text encoder	77.7	40.6	46.5	31.8
+ DisenQ	**82.2**	**43.8**	**49.9**	**34.8**
Ablation of different type of feature disentanglement				
No disentanglement	74.2	38.2	42.3	29.9
F_b and $F_{\hat{b}}$	76.6	40.9	44.7	31.9
F_b and F_m	79.2	41.1	48.2	32.9
F_b, $F_{\hat{b}}$ and F_m	**82.2**	**43.8**	**49.9**	**34.8**
Performance of each disentangled feature				
Biometrics	80.4	43.0	48.1	32.0
Non-biometrics	3.8	1.2	1.3	0.1
Motion	76.3	39.4	44.2	27.1
Biometrics + Motion	**82.2**	**43.8**	**49.9**	**34.8**

MMD-AB, and Charades-AB, which contain diverse activities, MEVID lacks activity variability, making activity-based identification less impactful. Despite this, our model remains competitive, achieving a 1.2% improvement in Rank-1 accuracy. This demonstrates that while our framework is designed for activity-biometrics, it generalizes well to traditional video-based identification scenarios by effectively disentangling identity from appearance, ensuring robust performance even in real-world unconstrained settings.

4.3. Ablation Studies

We conduct ablation studies on NTU RGB-AB and Charades-AB datasets on the same activity, including same view evaluation protocol and present the results in Table 3. While NTU RGB-AB provides a controlled setting with diverse clothing and activity variations; Charades-AB contains much more real-world complexity, including varied lighting, occlusions, and higher appearance variations, which better tests model generalization.

Contribution of each component is presented in Table 3 *(top)*. A vision encoder alone struggles due to entangled identity, appearance, and motion features leading to poor performance. Introducing text supervision via cross-attention and projecting features into distinct spaces improves identity retention by mitigating the influence of appearance variability. However, the most substantial gains come from DisenQ, which explicitly separates biometrics, non-biometrics, and motion features. By aligning separate learnable queries with structured textual priors, DisenQ establishes a well-structured feature representation that significantly enhances activity-biometrics performance.

Ablation of different type of feature disentanglement, illustrated in Table 3 (*middle*) presents their individual impact on performance. When biometrics and non-biometrics features are disentangled, the model effectively mitigates clothing bias but struggles with variations in motion, resulting in improved yet suboptimal performance across different actions. Disentangling biometrics and motion features enhances stability by preserving identity-specific movement patterns, which are crucial for reliable identification across activities. The most comprehensive performance is achieved when all three feature types are disentangled, ensuring that identity-related features remain distinct while controlling appearance and motion influences.

Individual performance of each disentangled feature, illustrated in Table 3 (*bottom*), provides further insights into their discriminative power for activity-based identification. Biometrics features alone exhibit the highest performance among each individual feature types, highlighting their intrinsic value in accurately identifying individuals. In contrast, non-biometric features significantly degrade performance, indicating that our disentanglement was effective in removing identity-related information from this feature space. Motion features offer moderate performance, providing additional context but lacking the distinctiveness of biometrics attributes. The synergy between biometric and motion features yields the most effective results, leveraging both identity cues and dynamic movement patterns for robust identification across challenging scenarios.

5. Analysis and Discussion

Effect of disentanglement on feature space. Figure 3 presents the impact of DisenQ on feature disentanglement

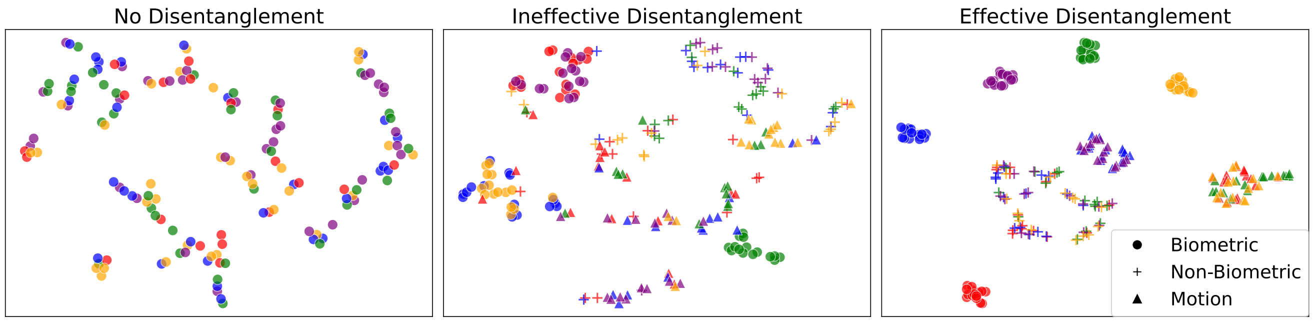


Figure 3. **Impact of DisenQ on feature disentanglement.** The left plot shows the feature space of vision encoder with no feature disentanglement, resulting in poor identity clustering. The middle plot shows an ineffective disentanglement using cross-attention and projection, where biometrics features remain mixed with non-biometrics, causing improper clustering. In contrast, the right plot demonstrates DisenQ-enabled disentanglement, achieving well-separated biometrics clusters, distinctly isolated from non-biometrics and motion features. Colors represent five identities across two activity classes.

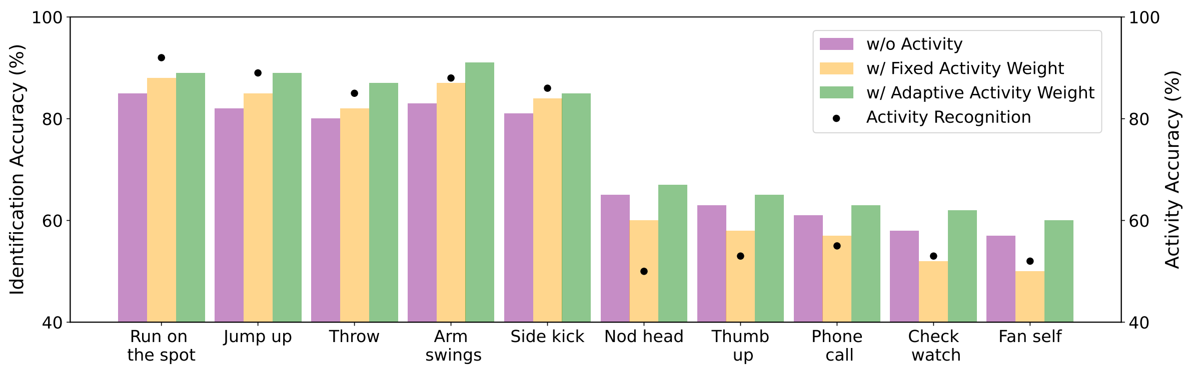


Figure 4. **Performance analysis across activities** (top 5 best and worst) on NTU RGB-AB. Here, bars and dots respectively represent person identification and action recognition accuracy.

by comparing feature spaces before and after its application. Without DisenQ, feature separation relies on simple cross-attention and projection, resulting in overlapping clusters where biometric features mix with non-biometrics from different identities wearing similar clothing. In contrast, DisenQ effectively separates biometrics, non-biometrics, and motion features into distinct spaces with clear boundaries. DisenQ's structured separation ensures that biometrics clusters remain identity-specific and free from confounding appearance-based cues, leading to robust identification.

Impact of design choice for disentanglement. We explore different architectural variations of DisenQ to evaluate the trade-off between complexity and effectiveness. A variant using three independent Q-Formers—each learning biometrics, non-biometrics, or motion features separately—yields only a marginal 0.23% Rank-1 accuracy gain on NTU RGB-AB while tripling the parameter count, suggesting that our original design is already sufficient for disentanglement. To test whether additional parameters could still be beneficial, a deeper DisenQ variant with the same parameter count as the three-Q-Former setup results in a 3.8% drop due to overfitting, indicating that simply increasing model capacity does not guarantee better feature separation. These findings highlight that structured learning is more critical than model size, and our DisenQ architecture strikes an optimal balance between effectiveness, and computational cost for activity-based person identification.

Performance analysis across activities. To examine the impact of different activities on person identification, we analyze performance across activity classes by identifying the five best and worst-performing actions. While activities involving significant body movements (e.g., running, jumping) provide distinctive motion patterns that aid recognition, they can introduce biases if overemphasized. Conversely, subtle activities (e.g., minor hand/head gestures) may lower accuracy due to weaker motion cues. Our findings (Figure 4) show that fixed weighting ($\alpha_1 = \alpha_2 = 0.5$ in Equation 8) of biometric and motion features can negatively affect identification for low-motion activities, whereas adaptive weighting ensures motion features contribute only when beneficial, stabilizing performance. Notably, highly distinctive actions retain high person identification accuracy even without explicit motion cues, confirming that motion serves as a complementary rather than dominant factor. Likewise,

Table 4. **Similarity of generated prompts** across multiple runs on NTU RGB-AB subset. Sim. and Std. Dev. denotes mean cosine similarity and standard deviation.

Ft.	Sim.	St. Dev.
T_b	0.92	0.03
$T_{\hat{b}}$	0.79	0.12
T_m	0.68	0.17

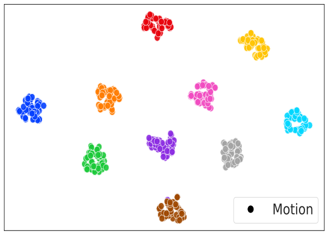


Figure 5. **Feature space of generated motion prompts** across multiple runs (NTU RGB-AB subset).

Table 5. **Performance comparison of different vision encoders and VLMs** for NTU RGB-AB and Charades-AB datasets.

Model	Size	NTU RGB-AB R1	NTU RGB-AB mAP	Charades-AB R1	Charades-AB mAP
Vision encoders					
SigLIP-L [60]	0.3B	80.2	41.6	48.3	33.7
ViT-1B [52]	1B	**83.4**	42.1	49.2	34.7
ViT-G/14 [49]	1.8B	82.2	**43.8**	**49.9**	**34.8**
Visual Language Models (VLMs)					
GPT-4V [39]	-	82.3	43.7	49.8	34.9
InstructBLIP [11]	7B	82.1	43.8	49.7	34.8
LLaVA 1.5 [36]	7B	82.2	43.8	49.9	34.8

challenging activities do not inherently degrade identification performance, as the model prioritizes biometrics features when necessary, ensuring balanced identification.

Utility and quality of the generated prompts. To asses the impact of accurate textual prompts on disentanglement, we replace non-biometrics descriptions with random clothing details, leading to a 9.2% drop in Rank-1 accuracy on NTU RGB-AB, highlighting the necessity of precise appearance descriptions. Additionally, we asses prompt consistency by generating descriptions for the same key-frame over five runs on a subset of NTU RGB-AB (10 identities and 10 action classes) and report the average results in Table 4. Biometrics descriptions remain highly stable, as indicated by high cosine similarity and low standard deviation, ensuring reliable identity representation. Non-biometrics descriptions also exhibit relative consistency, with minor variations. Motion descriptions exhibit the most variability, as different textual descriptions may be generated for the same action label. Figure 5 confirms that semantically similar motion prompts still cluster in the same feature space, ensuring consistency in representation.

Choice of vision encoder and VLM. Our model supports various vision encoder architectures. To identify the best performer, we evaluated three popular vision encoders: SigLIP-L [60], ViT-1B from InternVideo2 [52], and ViT-G/14 from EVA-CLIP [49] and find ViT-G/14 to be the best performing model (Table 5 (*top*)). Additionally, we show robustness of our approach across various VLMs, where we observe that changing the VLM does not contribute to significant changes ((Table 5 (*bottom*))), thus we select LLaVA for its efficiency and open-source property.

Qualitative results. Figure 6 compares the top-2 rank retrieval results of our model with ABNet, the only other exisiting activity-biometrics method. As shown, ABNet often misidentifies individuals when they perform the same activity, indicating an over-reliance on motion cues. In contrast, our model, with adaptive motion weighting and effective disentanglement of biometrics, non-biometrics, and motion features, accurately identifies individuals by prioritizing biometrics features over activities. We present more qualitative example in supplementary.

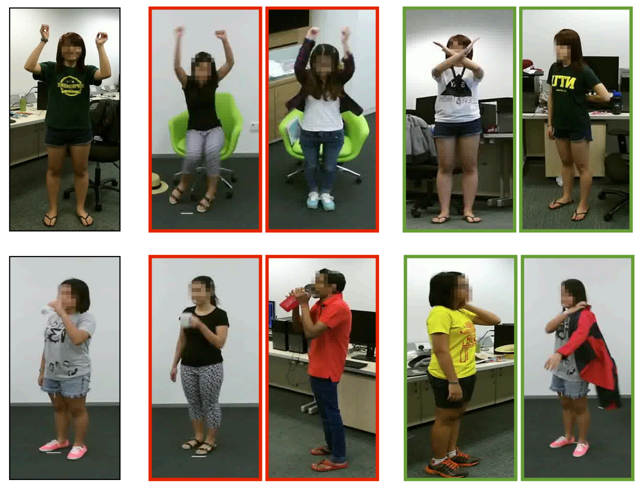


Figure 6. **Top-2 rank retrieval comparison** of our model and ABNet [3]. For a given probe image (*left*), the *middle* two images show incorrect matches retrieved by ABNet due to its over-reliance on activities, while the *right* two images show correct matches retrieved by our model, which effectively disentangles biometrics features from motion cues.

6. Conclusion

In this work, we introduce a multimodal **Disen**tangling **Q**uerying (**DisenQ**) Transformer based framework for activity-biometrics, where individuals are identified across diverse activities. Our framework leverages structured language guidance to disentangle biometrics, non-biometrics, and motion cues without additional visual modalities. By performing feature disentanglement, **DisenQ** ensures that biometrics features remain invariant to non-biometrics information, and motion cues enhance identification without overshadowing stable identity traits. An adaptive weighting mechanism dynamically balances biometrics and motion contributions for reliable identity retrieval. Our approach surpasses existing methods on activity-based benchmarks and generalizes well to traditional video-based identification, demonstrating the importance of effective feature disentanglement in activity-aware person identification.

7. Acknowledgement

This research is based upon work primarily supported by the National Science Foundation under Grant No. 2331319. Any opinions, findings, and conclusions or recommendations contained herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the National Science Foundation.

References

[1] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188, 2020. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736, 2022. 2

[3] Shehreen Azad and Yogesh Singh Rawat. Activity-biometrics: Person identification from daily activities. In *CVPR*, pages 287–296, 2024. 1, 2, 4, 5, 6, 8

[4] Shehreen Azad, Yash Jain, Rishit Garg, Vibhav Vineet, and Yogesh Rawat. Understanding depth and height perception in large visual-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3611–3620, 2025. 2

[5] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8545–8556, 2025. 2

[6] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *CVPR*, pages 7339–7348, 2022. 2, 5

[7] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. Pstr: End-to-end one-step person search with transformers. In *CVPR*, pages 9458–9467, 2022. 2, 5

[8] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning recurrent 3d attention for video-based person re-identification. *IEEE TIP*, 29:6963–6976, 2020. 2

[9] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, pages 8146–8155, 2021. 2

[10] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3667–3675, 2023. 2, 5, 6

[11] 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 preprint arXiv:2305.06500*, 2023. 2, 5, 8

[12] Daniel Davila, Dawei Du, Bryon Lewis, Christopher Funk, Joseph Van Pelt, Roderic Collins, Kellie Corona, Matt Brown, Scott McCloskey, Anthony Hoogs, et al. Mevid: Multi-view extended videos with identities for video person re-identification. In *WACV*, pages 1634–1643, 2023. 5

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[14] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 6

[15] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, pages 12036–12045, 2021. 5

[16] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1

[17] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *CVPR*, pages 9707–9716, 2023. 1

[18] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243. Springer, 2020. 5, 6

[19] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, pages 1060–1069, 2022. 1, 2, 4, 5, 6

[20] Peini Guo, Hong Liu, Jianbing Wu, Guoquan Wang, and Tao Wang. Semantic-aware consistency network for cloth-changing person re-identification. In *ACM MM*, 2023. 2, 5

[21] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1490–1501, 2021. 2

[22] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *CVPR*, pages 17521–17531, 2024. 1, 2, 5, 6

[23] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, pages 10513–10522, 2021. 2

[24] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 388–405. Springer, 2020. 6

[25] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. 2, 5, 6

[26] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. In *NeurIPS*, 2024. 2, 5, 6

[27] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, pages 1–8. IEEE, 2019. 2

[28] Xinyang Jiang, Yifei Gong, Xiaowei Guo, Qize Yang, Feiyue Huang, Wei-Shi Zheng, Feng Zheng, and Xing Sun. Rethinking temporal fusion for video-based person re-identification on semantic and time aspect. In *AAAI*, pages 11133–11140, 2020. 2, 5

[29] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, pages 14278–14287, 2022. 2

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 2, 3

[32] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, pages 1405–1413, 2023. 1, 2, 5, 6

[33] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *ECCV*, pages 375–390. Springer, 2022. 1

[34] Xin Liang and Yogesh S Rawat. Differ: Disentangling identity features via semantic cues for clothes-changing person re-id. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13980–13989, 2025. 2

[35] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 1

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 2, 5, 8

[37] Xuehu Liu, Chenyang Yu, Pingping Zhang, and Huchuan Lu. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[38] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *CVPR*, pages 14225–14234, 2021. 1

[39] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. 8

[40] Priyank Pathak. Fine-grained re-identification. *arXiv preprint arXiv:2011.13475*, 2020. 2

[41] Priyank Pathak and Yogesh S Rawat. Colors see colors ignore: Clothes changing reid with color disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[42] Priyank Pathak and Yogesh S Rawat. Coarse attribute prediction with task agnostic distillation for real world clothes changing reid. *arXiv preprint arXiv:2505.12580*, 2025. 2

[43] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them (student abstract). In *AAAI*, pages 13893–13894, 2020. 1, 2, 6

[44] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *ECCV*, pages 93–110. Springer, 2020. 2, 5

[45] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 2

[46] 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, 2021. 2

[47] Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. Probing conceptual understanding of large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1797–1807, 2024. 2

[48] Madeline Chantry Schiappa, Shehreen Azad, Sachidanand Vs, Yunhao Ge, Ondrej Miksik, Yogesh S Rawat, and Vibhav Vineet. Robustness analysis on foundational segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1786–1796, 2024. 2

[49] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5, 6, 8

[50] Qizao Wang, Bin Li, and Xiangyang Xue. When large vision-language models meet person re-identification. *arXiv preprint arXiv:2411.18111*, 2024. 2

[51] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, pages 12026–12035, 2021. 2, 5, 6

[52] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pages 396–416. Springer, 2024. 8

[53] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023. 2

[54] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 43(6):2029–2046, 2019. 2

[55] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for

cloth-changing person re-identification. In *CVPR*, pages 1472–1481, 2023. 2, 5

[56] Zexian Yang, Dayan Wu, Chenming Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. A pedestrian is worth one prompt: Towards language guidance person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17343–17353, 2024. 2

[57] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 44(6): 2872–2893, 2021. 1

[58] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *AAAI*, pages 6764–6772, 2024. 1, 2, 5

[59] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18 (12):8776–8785, 2022. 6

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 8

[61] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[62] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild. In *WACV*, pages 6290–6300, 2024. 6