DisenQ: Disentangling Q-Former for Activity-Biometrics
Supplementary Material

In this supplementary material, we provide details of prompt generation in Section A, along with structured examples of generated descriptions. Then we present additional quantitative results and analysis in Section B and Section C, followed by qualitative analysis in Section D. Then we present the dataset statistics in Section E. Finally, we address some limitations and outline directions for future research in Section F.

A. Details of Prompt Generation

We use the following structured prompt template designed to extract biometrics, non-biometrics, and motion-related details from the key-frame of a given RGB video.

```
Analyze the given image where
action label is <action label>
and extract the following
details: Biometrics: A
<physique/body shape> person
with <posture>, such as
arms/legs positioning. Motion:
Performing the action of <action
label> by <action description>.
Non-biometrics: A <color,
type of clothing> and <other
accessories>.
```

This prompt template is fed into the frozen VLM along with the key-frame, allowing the model to generate structured textual descriptions for each feature category. The output is then parsed into three distinct textual embeddings corresponding to biometrics, motion, and non-biometrics, ensuring explicit separation of identity-related and appearance-based cues. By incorporating structured textual supervision, this approach enhances feature disentanglement, enabling the model to learn identity-relevant representations while mitigating appearance bias. In Figure 7, we present examples of structured textual descriptions generated using a Vision-Language Model (VLM) from a given key-frame and its associated action label.

B. Additional Results

In Table 6 we present performance comparison of our method with existing works and report the rank 5 accuracy. We present the result of our model on the excluding same view evaluation protocol in Table 7. From both of these tables, we observe that our model constantly outperforms all the existing models across all datasets.

C. Additional Analysis

C.1. VLM quality is not a performance scalability bottleneck

We evaluate VLM scalability using two prompt variants: simplified slot-filler prompts (e.g., "`a [body shape] person`"), and fine-grained 4-way disentanglement (biometrics, motion, (upper/lower)-body clothing). As shown in Figure 8, even simple prompts significantly improve performance over no prompts, highlighting the value of semantic structure over linguistic richness. In contrast, granular prompts reduce performance and add complexity. Since all fine-grained features fall within the core axes of biometrics, non-biometrics, and motion, our 3-way setup remains more robust and scalable. Moreover, by restricting VLM use to *training only* and using structured prompts, we reduce noise and ensure that performance does not heavily depend on VLM strength.

C.2. Cross-domain utility of disentangled features.

We evaluated cross-domain generalization from NTU RGB-AB to PKU MMD-AB (Table 8), and found that while biometrics (F_b) transfer well, motion (F_m) showed slightly lower performance due to action variability. Their combination yields the best performance, confirming their complementary strengths (Table 3) even across domains.

C.3. Further analyzing feature disentanglement

We conducted three analyses to validate that the model truly separates features in a visually grounded manner—rather than merely aligning to prompt format.

Mutual Information Analysis: To verify disentanglement, we compute InfoNCE-based mutual information between each feature pair using empirically estimated upper and lower bounds derived from matched (same actor/action) and mismatched (different actor/action) pairs of NTU RGB-AB (Figure 9). The relatively lower InfoNCE for $F_b \leftrightarrow F_m$ falls within a wider range, indicating some mutual information, expected due to identity-linked motion cues such as gait, swing style etc. This highlights F_m as a complementary cue to F_b for identity matching. In contrast, $F_b \leftrightarrow F_{\hat{b}}$ and $F_m \leftrightarrow F_{\hat{b}}$ show consistently higher losses within tighter bounds, confirming minimal shared information and effective disentanglement.

Cross Feature Leakage Test: To further verify disentanglement, we trained classifiers to predict action from F_b (biometrics) and $F_{\hat{b}}$ (non-biometrics) on NTU RGB-AB. F_b achieved 14.4% accuracy, reflecting some posture-related cues embedded in body shape. These cues are expected,

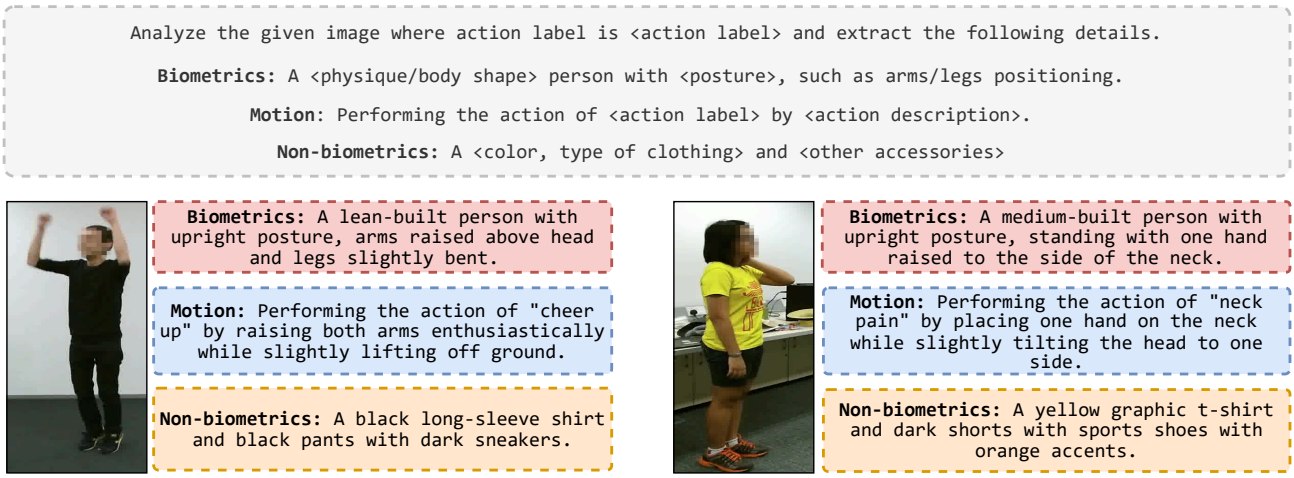

Figure 7. **Examples of generated structured textual descriptions.**

Table 6. **Performance comparison of person reidentification** on NTU RGB-AB, PKU MMD-AB, and Charades-AB datasets. Here we report rank 5 accuracies. † represents results produced in our environment. **Bold** represents best results.

Model	Venue	NTU RGB-AB		PKU MMD-AB		Charades-AB	
		Same	Cross	Same	Cross	Same	Cross
Models with only visual modality							
TSF [28]	AAAI 20	72.9	70.3	78.5	73.5	38.2	32.1
VKD [44]	ECCV 20	68.9	69.2	80.0	74.3	38.9	34.4
BiCnet-TKS [25]	CVPR 21	75.7	70.7	83.0	78.7	41.9	40.6
PSTA [51]	ICCV 21	69.7	67.7	79.1	74.0	45.0	40.5
STMN [15]	ICCV 21	74.8	71.9	79.6	73.3	41.3	35.3
SINet [6]	CVPR 22	71.1	69.1	82.2	78.0	42.3	38.7
CAL [19]	CVPR 22	78.6	76.5	86.0	81.2	48.2	45.3
Video-CAL [19]	CVPR 22	81.3	79.5	83.1	82.5	50.1	48.5
PSTR [7]	CVPR 22	71.2	69.3	85.2	80.0	40.2	37.2
AIM [55]	CVPR 23	73.4	71.8	83.5	80.4	42.1	37.6
SCNet [20]	ACM MM 23	71.9	70.3	81.4	74.9	34.5	30.2
ABNet [3]	CVPR 24	85.3	81.4	91.4	89.3	51.0	52.0
Models with visual +language modality							
CLIP ReID † [32]	AAAI 23	79.2	77.3	85.0	83.2	46.8	44.6
CCLNet † [10]	ACM MM 23	78.2	77.1	86.7	82.5	45.9	41.7
TF-CLIP † [58]	AAAI 24	79.6	77.0	85.9	84.1	43.7	42.1
TVI-LFM † [26]	NeurIPS 24	78.9	77.5	87.1	83.5	49.5	46.3
Instruct-ReID † [22]	CVPR 24	81.1	79.6	87.3	83.5	47.9	43.1
EVA-CLIP [49]		75.4	72.8	77.2	72.1	41.3	33.8
Ours		**88.5**	**86.4**	**94.7**	**90.5**	**56.8**	**54.1**

as they are stable biometrics components, but do not represent dynamic motion. On the contrary, $F_{\hat{b}}$ scored near random (4.8%), confirming no motion leakage into appearance features. These results supports minimal unintended information transfer across branches. **Causal Intervention:** Additionally, to test if disentanglement stems from prompt structure, we swapped prompt semantics across branches without changing losses. Despite this deliberate mismatch (e.g., motion prompts guiding biometrics features), identity and action performance dropped only marginally (1–2%) on NTU RGB-AB, suggesting that feature separation is guided by visual supervision rather than prompt formatting.

Table 7. Performance comparison of our model on NTU RGB-AB, PKU MMD-AB, and Charades-AB datasets for excluding same view evaluation protocols.

	Eval.	Model	Rank 1	mAP
NTU	Same activity	ABNet [3]	77.8	38.8
		DisenQ	**80.7**	**40.9**
	Cross activity	ABNet [3]	76.4	36.1
		DisenQ	**79.3**	**37.6**
PKU	Same activity	ABNet [3]	81.4	51.7
		DisenQ	**84.2**	**55.1**
	Cross activity	ABNet [3]	79.4	46.3
		DisenQ	**82.4**	**50.5**

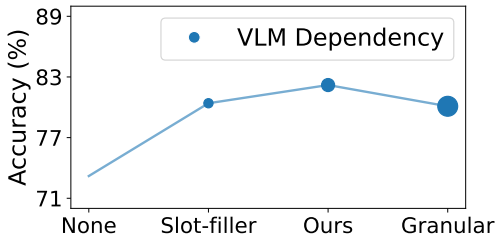

Figure 8. Prompt structure vs. accuracy and VLM reliance (NTU RGB-AB).

Table 8. **Utility of disentangled features** across domain.

Feature	Rank 1
Baseline	61.7
F_b	74.3
F_m	68.1
F_b, F_m	**76.8**

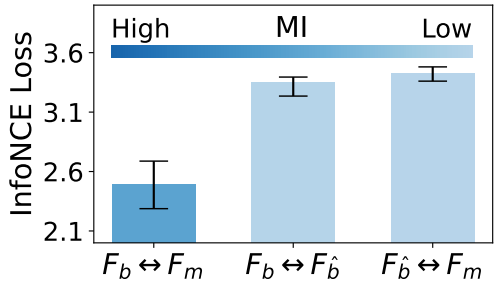

Figure 9. Mutual info (MI) analysis.

C.4. Risk of VLMs' inherent bias propagation

To address potential VLM bias linking appearance with identity, we use the VLM only to generate controlled attributes' descriptions within predefined, structured prompt templates, not free-form text. We further mitigate residual VLM correlations by enforcing orthogonality (Equation 6) and excluding non-biometrics features from identity matching (Equation 7). Together, these steps minimize any implicit bias and maintain clean disentanglement between appearance and identity features. While these design choices aim to mitigate potential sources of bias, we acknowledge that some demographic bias may still persist due to upstream VLM pretraining, which is beyond the scope of this work.

C.5. Non-biometrics branch encodes appearance information to some extent

While the non-biometrics branch lacks explicit supervision, it is guided by appearance-focused prompts and regularized via orthogonality to remain distinct from biometrics. Color histogram analysis of Figure 3 indicates that neighboring pairs in the non-biometrics space tend to have more similar appearance attributes (0.81 vs. 0.54 for random pairs), suggesting that the learned features in this branch reflect clothing-related information to some extent.

D. Qualitative Results

Figure 10 illustrates the top 4 rank retrieval results for a given probe for NTU RGB-AB dataset in both same and cross-activity evaluation setting. This demonstrates the robustness of our model across diverse activities and significant appearance variations. Unlike traditional approaches that struggle with identity retention under clothing changes or motion variations, our method effectively disentangles biometrics, non-biometrics, and motion cues, ensuring accurate identification even when activities differ between the probe and gallery. The strong retrieval performance highlights the effectiveness of our approach in learning identity-consistent representations that generalize across diverse set of real-world activities.

E. Dataset statistics

We evaluate performance under two evaluation protocols: same-activity and cross-activity. In the same-activity setting, all activities are present across both sets, ensuring that each individual is observed performing the same set of actions. In contrast, the cross-activity protocol introduces a more challenging scenario where individuals appear in different activities across the two sets, meaning that activities seen in one set are entirely absent in the other. For datasets with multiple viewpoints, such as NTU RGB-AB and PKU MMD-AB, we further assess two variations: including same view, where all viewpoints are available in both probe and gallery, and excluding same view, where probe viewpoint is excluded from gallery, increasing the difficulty of matching individuals across different perspectives. This allows us to analyze the model's robustness to viewpoint variations. However, for datasets like Charades-AB, which do not contain explicit viewpoints data, only

Figure 10. **Qualitative results**. Here we present the top 4 rank retrieval results for a given probe (left) of our model on same-activity (*top)* and cross-activity (*bottom)* on NTU RGB-AB dataset.

Table 9. **Dataset statistics**

Dataset	Split	#actors	#activities	#samples
NTU RGB-AB	train	85	94	70952
	gallery	21		14192
	probe			3548
PKU MMD-AB	train	53	41	13634
	gallery	13		2727
	probe			681
Charades-AB	train	214	157	45111
	gallery	53		9022
	probe			2256
MEVID	train	104	1	6338 (tracklets)
	gallery	52		316 (tracklets)
	probe	54		1438 (tracklets)

the activity-based protocols are considered. Since, MEVID only contains one activity (e.g. walking), the evaluation of this dataset also falls under the same-activity setting. As MEVID primarily features walking sequences, we assign all tracklets a "walking" label to enable coarse motion supervision, while being consistent with standard re-ID protocols that leverage gait. Since occasional secondary actions are concurrent with walking, it allows us to still use motion supervision without explicit activity labels. A detailed dataset statistics is presented in Table 9.

F. Future Work

While our method demonstrates strong performance in disentangling biometrics, non-biometrics, and motion features for activity-based person identification, there are areas for further exploration. The reliance on structured text supervision ensures effective feature separation, but future work could explore more flexible multimodal alignment techniques to further enhance robustness in unconstrained settings. Additionally, integrating a memory-modeling framework, could enhance identity tracking across much longer activity sequences, ensuring stability even under extreme motion variations or video length.