Learning Instance-Adaptive Low-Rank Orthogonal Subspaces for Clothes-Changing Person Re-Identification

Dongwoo Kim [*1] **Tae-Kyun Kim** [*1]

Abstract

Clothes-changing person re-identification (CC-ReID) aims to recognize individuals despite drastic appearance changes caused by clothing variation. While existing methods rely on adversarial learning to disentangle clothing features, we propose **Ortho-ReID**, which explicitly models a **low-rank** clothing subspace from VLM text descriptions and extracts clothing-invariant representations via direct geometric constraints. A critical component is our transformer-based **Basis Maker**, which refines a shared, low-dimensional clothing prior into an **instance-adaptive low-rank subspace** through cross-attention with image patches, enabling robust clothing feature extraction even under varying visibility conditions. This instance-adaptive subspace is supervised via alignment with clothing text embeddings, while identity features are extracted via a learnable projection head and geometrically constrained to be strictly orthogonal to it. Extensive experiments demonstrate state-of-the-art performance on PRCC (+5.9% top-1), Celeb-reID-light (+3.5%), and LaST (+5.3%), with competitive results on LTCC.

1. Introduction

Recent representation learning research demonstrates that high-level semantic concepts, such as "clothing" or "color", correspond to structured, low-dimensional linear subspaces within the embedding spaces of large pretrained Vision-Language Models (VLMs) (Bhalla et al., 2024; Marks & Tegmark, 2024; Huang et al., 2024). This low-rank geometric structure offers a principled handle for *disentangled* representation learning. By explicitly modeling a concept as a low-rank subspace, it becomes possible to apply geometric constraints, such as orthogonal projection, to suppress or amplify that concept in downstream representations.

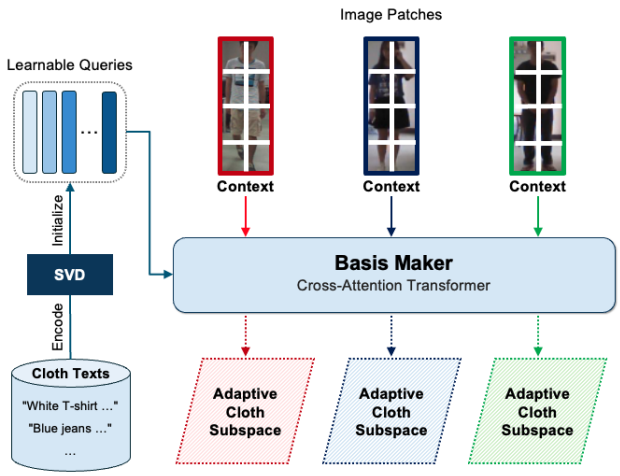


Figure 1. Instance-adaptive low-rank clothing subspace construction. Learnable queries (SVD-initialized from dataset-wide clothing text embeddings) are refined per-image via cross-attention with patch tokens into an instance-adaptive orthonormal basis spanning the low-dimensional clothing subspace.

We instantiate this insight in the challenging task of Clothes-Changing Person Re-Identification (CC-ReID) (Wan et al., 2020). This task requires recognizing individuals across camera views despite drastic clothing variations, a scenario where appearance becomes an unreliable visual cue. Existing methods address this through adversarial disentanglement (Liang & Rawat, 2025; Gu et al., 2022; Xu et al., 2021), causality-based debiasing (Yang et al., 2023b), or architectural modifications (Pathak & Rawat, 2025). However, most existing approaches do not explicitly exploit the low-rank structure of VLM representations. Instead, they rely on learned linear projections or adversarial objectives that often lack geometric interpretability.

We propose **Ortho-ReID**, a framework that treats the clothing concept as a learnable, instance-adaptive low-rank subspace and enforces identity features to be geometrically orthogonal to it. Our approach leverages two key properties of VLM representations. First, clothing semantics can

[1]School of Computing, KAIST, Daejeon, South Korea. Correspondence to: Dongwoo Kim <dwkim8155@kaist.ac.kr>, Tae-Kyun Kim <kimtaekyun@kaist.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

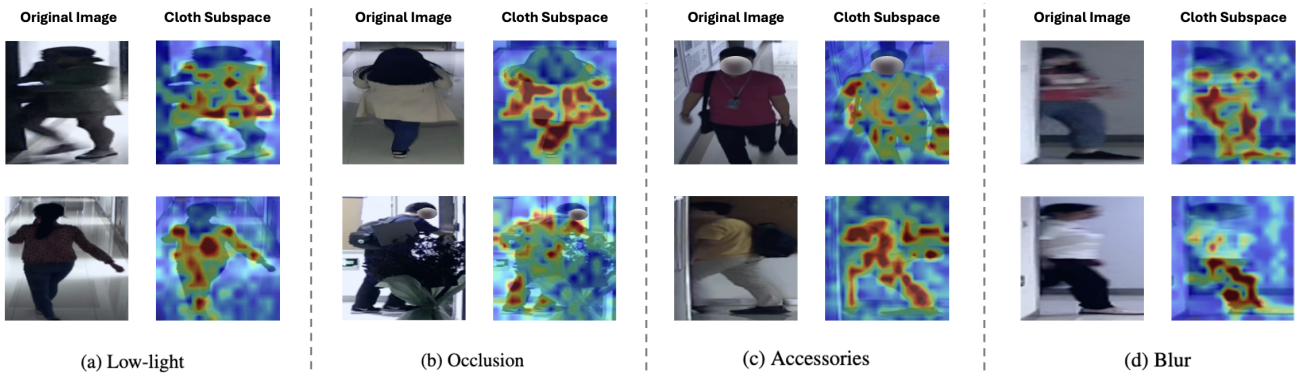


Figure 2. Visualizing the Basis Maker's spatial attention under challenging conditions in the LTCC gallery. Despite severe variations in (a) lighting, (b) occlusions, (c) accessories, and (d) blur, the instance-adaptive mechanism successfully and naturally focuses on the visible clothing regions.

be captured by the principal components of CLIP text embeddings. We obtain these components via Singular Value Decomposition (SVD) of dataset-wide clothing descriptions, which provides a semantically grounded prior over the clothing subspace. Second, this global prior can be adaptively refined per-instance via cross-attention with local image patch tokens. This refinement yields a instance-adaptive low-rank model of the clothing in each query image.

Concretely, our transformer-based **Basis Maker** takes SVD-initialized learnable queries and attends to patch tokens to output an orthonormal basis for the clothing subspace via QR decomposition. The identity representation is then constrained to be orthogonal to this basis during training. This ensures an effective geometric separation of clothing and identity information without relying on adversarial objectives. The most closely related prior work is DIFFER (Liang & Rawat, 2025), which also uses frozen CLIP text encoders for clothing supervision, but relies on adversarial learning rather than explicit low-rank subspace modeling.

2. Related Work

2.1. Clothes-Changing Person Re-Identification

CC-ReID addresses recognizing individuals despite drastic changes in apparel. Existing approaches fall into two streams. The first, clothes-invariant feature extraction, leverages external modalities such as 3D body shape (Chen et al., 2021; Liu et al., 2023), gait (Jin et al., 2022), or pose (Qian et al., 2020a). The second, feature disentanglement, employs adversarial learning (Xu et al., 2021; Gu et al., 2022), causality-inspired debiasing (Yang et al., 2023b), and architectural modifications (Pathak & Rawat, 2025). Recently, text-guided strategies have emerged, leveraging CLIP-based priors (Li et al., 2023b) or explicit textual supervision (Liang & Rawat, 2025; He et al., 2024; Li et al., 2024). Despite these advances, existing methods lack effective mechanisms for learning explicit, instance-adaptive clothing models.

2.2. Linear Concept Representations in VLMs

Recent studies demonstrate that high-level concepts can be approximated as linear directions or low-dimensional subspaces in large pretrained models (Bhalla et al., 2024; Huang et al., 2024; Marks & Tegmark, 2024). With vision-language models, concept-aligned directions in image representations can be derived directly from textual descriptions (Moayeri et al., 2023; Yang et al., 2023a; Bhalla et al., 2024). Recent VLM-based CC-ReID methods (Liang & Rawat, 2025; Li et al., 2024) employ linear projections to model clothing subspaces, but construct them using simple linear matrices. We introduce a cross-attention based Basis Maker for more flexible and expressive instance-adaptive subspace modeling.

2.3. Orthogonality for Re-ID Disentanglement

Projection-based methods (Ravfogel et al., 2020; 2022; Belrose et al., 2023) apply fixed orthogonal projections post-hoc. Enforcing orthogonality via in-training constraints has proven effective for disentangled feature learning. Prior ReID work applies orthogonality to separate shape (Feng et al., 2023) or color (Pathak & Rawat, 2025) from identity. We adopt a similar training-time geometric constraint but apply it to a learned, text-grounded, instance-adaptive clothing subspace.

3. Methodology

Our framework has three main components: (1) a CLIP image encoder extracting global CLS token $\mathbf{f}_{\text{cls}}$ and local patch tokens $\mathbf{F}_{\text{patch}}$, (2) a Basis Maker learning instance-adaptive clothing subspaces via cross-attention, and (3) orthogonal projection for disentanglement. The Basis Maker uses learnable queries initialized from SVD of clothing text

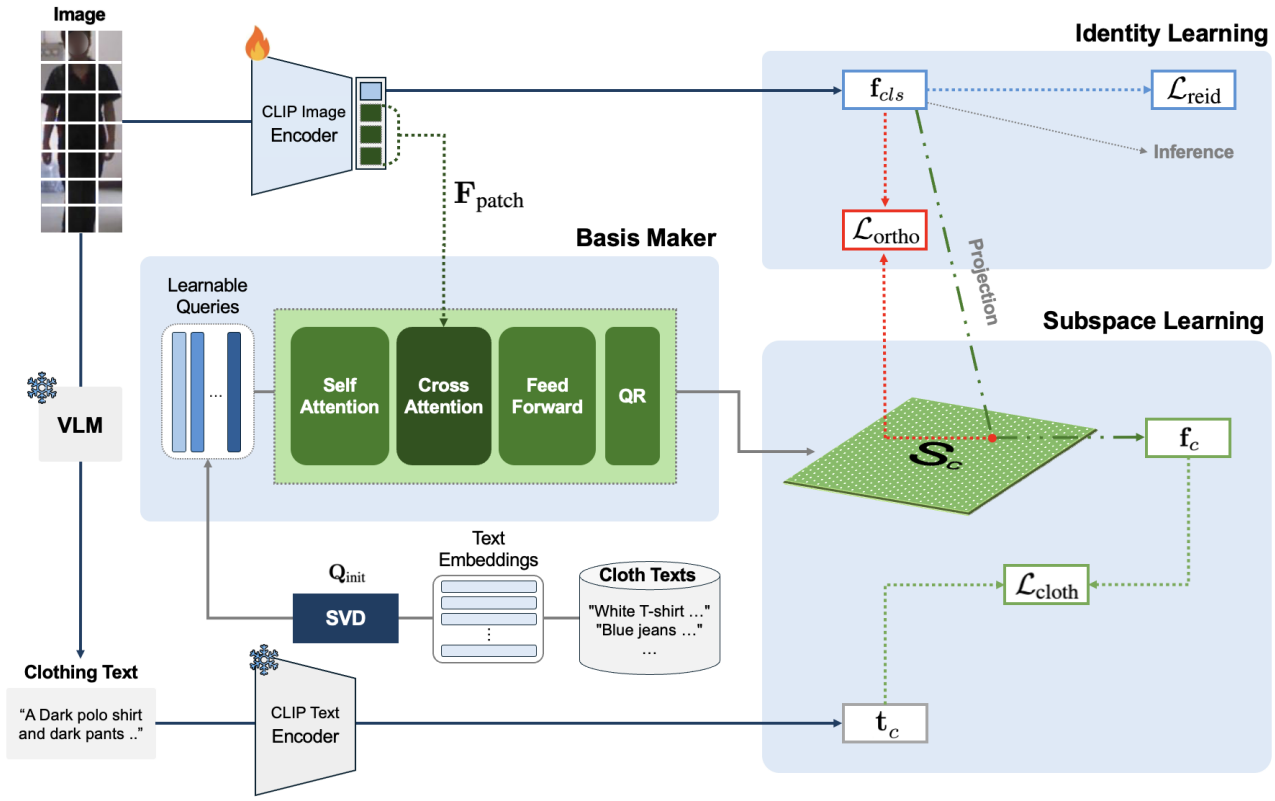


Figure 3. Overview of Ortho-ReID. The CLIP image encoder produces patch features $\mathbf{F}_{\text{patch}}$ and global feature $\mathbf{f}_{\text{cls}}$. Learnable queries are initialized via SVD of dataset-wide clothing text embeddings ($\mathbf{Q}_{\text{init}}$) and refined by the Basis Maker through cross-attention with $\mathbf{F}_{\text{patch}}$. The refined queries undergo QR decomposition to form an orthonormal basis spanning the clothing subspace $\mathbf{S}_c$. Three losses are used: $\mathcal{L}_{\text{cloth}}$ aligns projected features $\mathbf{f}_c$ with clothing text $\mathbf{t}_c$, $\mathcal{L}_{\text{reid}}$ for identity discrimination, and $\mathcal{L}_{\text{ortho}}$ enforces geometric orthogonality to remove clothing from $\mathbf{f}_{\text{cls}}$.

embeddings, refined via self-attention and cross-attention with $\mathbf{F}_{\text{patch}}$ into orthonormal bases $\mathbf{G}$ spanning $\mathcal{S}_c$. The Basis Maker is trained by projecting $\mathbf{f}_{\text{cls}}$ onto $\mathcal{S}_c$ to extract $\mathbf{f}_c = \mathbf{G}\mathbf{G}^\top \mathbf{f}_{\text{cls}}$ and aligning it with clothing text embeddings via $\mathcal{L}_{\text{cloth}}$. Clothing-invariant identity features are obtained via orthogonality loss $\mathcal{L}_{\text{ortho}}$, which trains the image encoder to produce features orthogonal to the clothing subspace.

3.1. CLIP Encoders

We employ EVA-02-CLIP (Fang et al., 2024) as the image encoder, processing input image $I \in \mathbb{R}^{H\times W\times 3}$ to produce $F = [\mathbf{f}_{cls}, \mathbf{f}_1, \ldots, \mathbf{f}_N] \in \mathbb{R}^{(N+1)\times d}$. The global CLS token $\mathbf{f}_{cls}$ serves as the primary feature for the final ReID task, while local patch tokens $\mathbf{F}_{\text{patch}} = \{\mathbf{f}_i\}_{i=1}^N$ provide rich visual details for instance-adaptive subspace learning. The text encoder remains frozen to maintain stable semantic references.

3.2. Basis Maker

3.2.1. Architecture

The Basis Maker employs a standard Transformer decoder (Vaswani et al., 2017) with K learnable queries $\mathbf{Q} \in \mathbb{R}^{K\times d}$ serving as initial basis vectors. Inspired by Q-Former (Li et al., 2023a), these queries dynamically extract task-relevant features through interaction with visual inputs. To specifically adapt this architecture for modeling a strict geometric clothing subspace, we introduce SVD initialization to provide a semantic prior and QR decomposition to ensure geometric independence. The Basis Maker processes queries through alternating self-attention and cross-attention layers. Queries first undergo self-attention for information exchange among bases, then attend to $\mathbf{F}_{\text{patch}}$ through multi-head cross-attention. Using patch tokens retains fine-grained spatial locality, enabling the model to capture clothing characteristics across body regions. After feed-forward networks, the refined queries $\mathbf{Q}' \in \mathbb{R}^{K\times d}$ are orthonormalized via QR decomposition:

$$\mathbf{G} = \text{QR}(\mathbf{Q}'^\top) \quad \text{s.t.} \quad \mathbf{G}^\top\mathbf{G} = \mathbf{I} \tag{1}$$

where $\mathbf{G} \in \mathbb{R}^{d \times K}$ forms an orthonormal basis set for the clothing subspace $\mathcal{S}_c$. This orthonormality constraint encourages each basis vector to capture independent, non-overlapping clothing attributes (e.g., color, texture, style), preventing redundancy.

3.2.2. Text-Guided Query Initialization

To leverage CLIP's semantically aligned text-image space, we initialize the basis learnable queries with clothing text embeddings. We generate natural language descriptions for each dataset using a Vision-Language Model (GPT-4o by default, open-source models yield equivalent performance; see Appendix B) and encode them with the frozen CLIP text encoder to obtain $\{\mathbf{t}_c^{(i)}\}_{i=1}^{M} \in \mathbb{R}^d$. We apply SVD to extract the dominant semantic directions and initialize the learnable queries with the top-K principal components:

$$\mathbf{Q}_{\text{init}} = \text{SVD}_K(\{\mathbf{t}_c^{(i)}\}_{i=1}^{M}) \tag{2}$$

This initialization provides a strong semantic prior ensuring queries reside in CLIP's meaningful space and attend to clothing-relevant features. Importantly, queries remain fully learnable during training, adapting from their text-derived initialization to capture instance-specific visual clothing patterns through self-attention and cross-attention with image patch tokens.

3.2.3. Text-Guided Semantic Alignment

While the Basis Maker learns the clothing subspace through interaction with patch tokens, the ReID task relies on the CLS token as the global image representation. To ensure the learned clothing subspace disentangles clothing features from the CLS token, we project $\mathbf{f}_{\text{cls}}$ onto $\mathcal{S}_c$:

$$\mathbf{f}_c = \mathbf{G}\mathbf{G}^\top \mathbf{f}_{\text{cls}} \tag{3}$$

Critically, we *detach* gradients of $\mathbf{f}_{\text{cls}}$ when computing this projection, ensuring only the Basis Maker parameters are updated—not the image encoder. This gradient isolation means the Basis Maker is trained to capture clothing information as represented by the *current* image encoder, while the image encoder is separately encouraged to remove clothing via $\mathcal{L}_{\text{ortho}}$. We then align projected features with clothing text embeddings through contrastive learning:

$$\mathcal{L}_{\text{cloth}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\cos(\mathbf{f}_c^{(i)}, \mathbf{t}_c^{(i)})/\tau)}{\sum_{j=1}^{B} \exp(\cos(\mathbf{f}_c^{(i)}, \mathbf{t}_c^{(j)})/\tau)} \tag{4}$$

3.3. Orthogonal Projection for Disentanglement

With the clothing subspace $\mathcal{S}_c$ learned by the Basis Maker, we apply orthogonal projection to separate clothing from identity. We obtain clothing-invariant identity features by minimizing the similarity between $\mathbf{f}_{\text{cls}}$ and the clothing component $\mathbf{f}_c$ (with detached basis $\mathbf{G}$):

$$\mathcal{L}_{\text{ortho}} = \left\langle \frac{\mathbf{f}_{\text{cls}}}{||\mathbf{f}_{\text{cls}}||_2}, \frac{\mathbf{f}_c}{||\mathbf{f}_c||_2} \right\rangle^2 \tag{5}$$

By minimizing the normalized inner product, this loss encourages the image encoder to learn clothing-invariant representations during training. At inference time, the Basis Maker is not needed and the encoder efficiently produces clothing-invariant features without additional computational overhead.

3.4. Training Objective

The total loss jointly optimizes the Basis Maker and image encoder:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cloth}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{reid}}\mathcal{L}_{\text{reid}} \tag{6}$$

The ReID loss combines standard objectives with our proposed intra-modal contrastive loss:

$$\mathcal{L}_{\text{reid}} = \lambda_{id}\mathcal{L}_{id} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{intra}\mathcal{L}_{intra} \tag{7}$$

where $\mathcal{L}_{id}$ is cross-entropy for identity prediction, $\mathcal{L}_{tri}$ is batch-hard triplet loss (Hermans et al., 2017), and $\mathcal{L}_{intra}$ treats all clothing variations of the same identity as uniformly positive:

$$\mathcal{L}_{intra} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j \neq i} q_{ij} \log \frac{\exp(\mathbf{f}_{id}^{(i)} \cdot \mathbf{f}_{id}^{(j)}/\tau)}{\sum_{k \neq i} \exp(\mathbf{f}_{id}^{(i)} \cdot \mathbf{f}_{id}^{(k)}/\tau)} \tag{8}$$

where $q_{ij} = \frac{1}{K_i}$ if $\text{pid}_i = \text{pid}_j$, else 0. This dense positive supervision is particularly valuable for CC-ReID because it simultaneously tightens intra-identity clusters despite clothing variations. All loss weights are set to $\lambda = 1$ across all experiments.

4. Experiments

4.1. Datasets and Metrics

We evaluate on four CC-ReID benchmarks: PRCC (Yang et al., 2021) (33,698 images, 221 identities, 2 outfits each), LTCC (Qian et al., 2020b) (17,138 images, 152 identities, avg. 5 clothes/person), Celeb-reID-light (Huang et al., 2019) (10,842 images, 590 identities), and LaST (Shu et al., 2022) (228,000 images, 10,862 identities). We report Rank-1 accuracy and mAP. PRCC reports Same-Clothes (SC) and Clothes-Changing (CC) results; LTCC reports CC and General protocols.

4.2. Implementation Details

We use EVA02-CLIP-L (Fang et al., 2024) (patch size 14) as the visual encoder and GPT-4o (Hurst et al., 2024) to

generate clothing descriptions. Input images are resized to 224×224. The Basis Maker comprises 6 Transformer decoder layers with 16 attention heads, and $K = 16$ learnable queries initialized via SVD. The visual encoder uses SGD ($lr = 2 \times 10^{-6}$, weight decay 3×10^{-4}); the Basis Maker uses Adam ($lr = 2 \times 10^{-5}$, weight decay 5×10^{-4}). Both use cosine LR scheduling over 80 epochs with batch size 64 (PRCC) and 32 (others). All experiments run on two NVIDIA A100 GPUs without data augmentation, re-ranking, or post-processing.

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

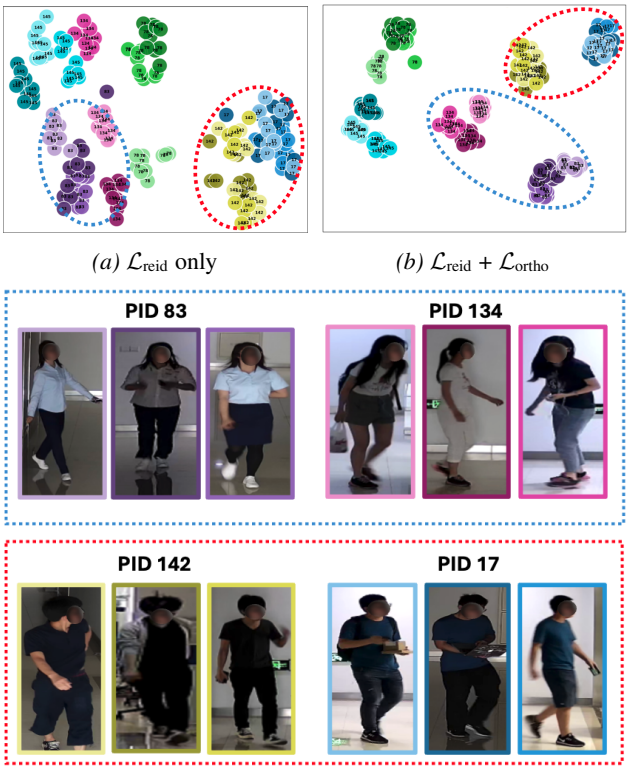
Results on Surveillance Datasets. Table 1 shows results on PRCC and LTCC. On PRCC, Ortho-ReID achieves 74.4% Rank-1 and 70.2% mAP in the CC scenario, surpassing all methods including DIFFER by +5.9% Rank-1 and +5.5% mAP. On LTCC, Ortho-ReID achieves 56.1% Rank-1 (CC), achieving highly competitive performance with DIFFER (58.2%) while outperforming all other existing methods.

Results on Non-Surveillance Datasets. Table 2 shows results on Celeb-reID-light and LaST. On Celeb-reID-light, Ortho-ReID achieves 79.1% Rank-1 and 59.0% mAP, outperforming DIFFER by +3.5% and +4.7%. On LaST (228K identities), we obtain 84.3% Rank-1 and 53.5% mAP, demonstrating +5.3% and +12.6% improvements over MADE. These consistent gains confirm the robustness of our approach across diverse real-world scenarios.

4.4. Ablation Studies

4.4.1. Effectiveness of Orthogonal Loss

Table 3 demonstrates the critical role of the geometric constraint enforced by $\mathcal{L}_{\text{ortho}}$. Adding this orthogonalization step to the standard ReID baseline yields substantial and consistent Rank-1 improvements: +3.8% on PRCC, +3.9% on LTCC, and +3.5% on Celeb-reID-light. This uniform gain across multiple benchmarks confirms that explicitly projecting the image embeddings away from the learned clothing subspace effectively mitigates the severe appearance bias inherent in standard baselines. Furthermore, Figure 4 provides a qualitative visualization of this disentanglement effect using t-SNE. In the baseline model without $\mathcal{L}_{\text{ortho}}$, the feature space is heavily entangled; samples are primarily clustered by their visual appearance and clothing color, failing to group different outfits of the same person. In contrast, applying $\mathcal{L}_{\text{ortho}}$ fundamentally restructures the embedding space. Same-identity samples now form tight, distinct clusters regardless of their drastic clothing variations. This clearly illustrates that the geometric constraint successfully purifies the identity representation by filtering out the clothing-specific components.



(a) $\mathcal{L}_{\text{reid}}$ only *(b)* $\mathcal{L}_{\text{reid}}$ + $\mathcal{L}_{\text{ortho}}$

Figure 4. t-SNE of learned identity features on LTCC. Colors indicate clothing; numbers denote identity. With $\mathcal{L}_{\text{ortho}}$ (right), same-identity samples cluster tightly across diverse clothing, while similar-clothing different-identity pairs are separated.

4.4.2. Attention Mechanism and Initialization

Table 4 studies the variants of attention mechanisms and query initialization in the Basis Maker. Patch-based cross-attention substantially outperforms both CLS-token cross-attention and self-attention alone. This highlights that fine-grained spatial interactions are essential for capturing locally varying clothing patterns. Because clothing details are distributed across different body parts and are often partially occluded, patch-level attention allows the Basis Maker to selectively focus on visible garments, yielding a +2.1% Rank-1 gain on LTCC over the CLS variant. Furthermore, SVD initialization consistently surpasses random initialization across all configurations. This confirms the benefit of starting from a semantically meaningful prior, providing a stable optimization trajectory and ensuring the queries are properly grounded in the clothing text space from the beginning of training.

4.4.3. Query Configuration

Tables 5 and 6 analyze the configuration of the basis queries. As shown in Table 6, making the queries learnable provides a clear advantage over keeping them frozen (+2.8% Rank-1 on LTCC). This demonstrates that while the SVD provides

Architecture	Method	Venue/ Year	Additional Modality	PRCC				LTCC			
				CC		SC		CC		General	
				R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
CNN	CAL (Gu et al., 2022)	CVPR 22	–	55.2	55.8	**100**	99.8	40.1	18.0	74.2	40.8
	AIM (Yang et al., 2023b)	CVPR 23	–	57.9	58.3	**100**	**99.9**	40.6	19.1	76.3	41.1
	CVSL (Nguyen et al., 2024a)	WACV 24	Pose	57.5	56.9	97.5	99.1	44.5	21.3	76.4	41.9
	CLIP3DReID (Liu et al., 2023)	ICCV 23	Text+3D	60.6	59.3	–	–	42.1	21.7	–	–
	CGPG (Nguyen et al., 2024b)	CVPR 24	Silhouette	61.8	58.3	**100**	99.6	46.2	22.9	77.2	42.9
ViT	TransReID (He et al., 2021)	ICCV 21	–	46.6	44.8	**100**	99.0	34.4	17.1	70.4	37.0
	CSCI (Pathak & Rawat, 2025)	ICCV 25	–	66.2	61.3	**100**	**99.9**	47.8	24.4	82.6	48.0
	Instruct-ReID (He et al., 2024)	CVPR 24	Text	54.2	52.3	–	–	–	–	–	–
	MADE (Peng et al., 2024)	TMM 24	Text	64.3	59.1	**100**	98.6	47.4	24.4	84.2	48.2
	DIFFER† (Liang & Rawat, 2025)	CVPR 25	Text	68.5	64.7	99.9	99.5	**58.2**	**31.6**	85.0	**52.8**
ViT	**Ours**	–	Text	**74.4**	**70.2**	**100**	99.0	56.1	30.2	**85.4**	52.0

Table 1. Comparison with state-of-the-art on PRCC and LTCC. **Bold**: best; underlined: second-best. †DIFFER's LTCC result coincides with an unusually high reported baseline (54.6%), substantially above other methods with identical settings (MADE: 45.9%, CSCI: 44.9%), suggesting possible implementation differences.

Method	Venue	Celeb-L		LaST	
		R-1	mAP	R-1	mAP
CAL (Gu et al., 2022)	CVPR 22	–	–	73.7	28.8
MADE (Peng et al., 2024)	TMM 24	72.0	52.3	79.0	40.9
DIFFER (Liang & Rawat, 2025)	CVPR 25	75.6	54.3	–	–
Ours	–	**79.1**	**59.0**	**84.3**	**53.5**

Table 2. Comparison on Celeb-reID-light (Celeb-L) and LaST.

$\mathcal{L}_{\text{reid}}$	$\mathcal{L}_{\text{ortho}}$	PRCC		LTCC		Celeb-L	
		R-1	mAP	R-1	mAP	R-1	mAP
✓		70.6	66.5	52.2	28.7	75.6	56.8
✓	✓	**74.4**	**70.2**	**56.1**	**30.2**	**79.1**	**59.0**

Table 3. Effect of orthogonal loss $\mathcal{L}_{\text{ortho}}$. Adding it consistently improves performance across all three datasets.

a strong global prior, allowing the queries to update during fine-tuning is crucial for adapting to the specific clothing distribution and instance-level variations of the target dataset. Table 5 investigates the optimal number of queries, K, which dictates the rank of the subspace. Performance peaks at $K = 16$. A smaller K restricts the representational capacity, failing to encompass the full diversity of clothing attributes. Conversely, a larger K introduces redundancy and leads to overfitting, potentially causing the subspace to inadvertently capture identity-related or background noise.

Attention Type	Init	PRCC		LTCC	
		R-1	mAP	R-1	mAP
Self-Attn only	Random	71.1	69.0	52.3	29.1
	SVD	72.9	69.5	54.0	29.4
Self + Cross (CLS)	Random	70.4	67.1	52.5	28.6
	SVD	72.5	69.2	53.3	29.8
Self + Cross (Patch)	Random	72.3	69.1	52.5	29.7
	SVD	**74.4**	**70.2**	**56.1**	**30.2**

Table 4. Ablation on Basis Maker attention type and query initialization. SVD initialization (highlighted) consistently outperforms random. Patch-based cross-attention achieves best performance, with the largest gain on LTCC (+2.1%) where clothing diversity is highest.

4.4.4. Effect of QR Decomposition

Table 7 validates the geometric constraint imposed by QR decomposition. Removing this operation results in consistent performance drops across both datasets, including a notable 1.86% decrease in LTCC mAP. Without QR decomposition, the dynamically generated basis vectors are prone to becoming highly correlated or collapsing into a narrower space during training. This confirms that explicitly enforcing orthonormality reduces redundancy among the basis vectors, ensuring a strictly structured and independent clothing subspace. Consequently, this geometric independence is vital for the subsequent orthogonalization step to effectively disentangle clothing features from the identity representation.

K	PRCC		LTCC	
	R-1	mAP	R-1	mAP
1	72.6	68.8	52.8	29.2
4	72.7	69.4	54.1	29.8
8	72.9	69.5	54.8	30.1
16	**74.4**	**70.2**	**56.1**	**30.2**
32	73.6	69.4	55.4	30.3
64	72.9	69.4	53.0	28.6

Table 5. Ablation on the number of learnable queries K. Performance peaks at $K = 16$; small K limits coverage while large K leads to overfitting.

Query Type	PRCC		LTCC	
	R-1	mAP	R-1	mAP
Frozen	73.2	69.5	53.3	29.3
Learnable	**74.4**	**70.2**	**56.1**	**30.2**

Table 6. Effect of query learnability (full Basis Maker config, SVD init). Learnable queries adapt during fine-tuning (+2.8% on LTCC).

Method	PRCC		LTCC	
	R-1	mAP	R-1	mAP
w/o QR	71.18	68.79	51.79	28.34
w/ QR	**74.4**	**70.2**	**56.1**	**30.2**

Table 7. Effect of QR decomposition. Removing it leads to consistent drops, confirming orthonormality reduces redundancy among basis vectors.

5. Conclusion

We present Ortho-ReID, a framework for clothes-changing person re-identification that explicitly models clothing as a text-grounded, instance-adaptive linear subspace and removes it via geometric orthogonal projection. Our Basis Maker, initialized from SVD of clothing text embeddings and refined per-image through cross-attention with patch tokens, enables adaptive and accurate clothing feature extraction even under varied visibility conditions. The orthogonality loss then trains the image encoder to produce representations that are geometrically decoupled from the clothing subspace, leading to consistent state-of-the-art gains across diverse CC-ReID benchmarks.

A current limitation is that our method focuses exclusively on clothing invariance: errors arise when different individuals share highly similar non-clothing attributes such as body shape, hairstyle, or posture. Future work should extend the subspace disentanglement framework to jointly model and remove multiple distractor attributes beyond clothing.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically reliable long-term person re-identification under clothing variation. The primary societal application is surveillance and public safety. As with all person re-identification research, care must be taken to ensure such systems are not deployed in ways that violate privacy rights or enable mass surveillance without appropriate oversight. We do not anticipate unique ethical concerns beyond those standard to the computer vision and person re-identification research community.

References

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F. P., and Lakkaraju, H. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *NeurIPS*, 2024.

Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., and Zheng, W.-S. Learning 3D shape feature for texture-insensitive person re-identification. In *CVPR*, 2021.

Chen, W., Wang, Z., Cao, Y., et al. InternVL2.5: An open-source multimodal LLM at GPT-4o level for autonomous agents. *arXiv:2412.05271*, 2024.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., and Cao, Y. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 150:105171, 2024.

Feng, J., Wu, A., and Zheng, W.-S. Shape-erased feature learning for visible-infrared person re-identification. In *CVPR*, 2023.

Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., and Chen, X. Clothes-changing person re-identification with RGB modality only. In *CVPR*, 2022.

He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. TransReID: Transformer-based object re-identification. In *ICCV*, 2021.

He, W., Deng, Y., Tang, S., et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *CVPR*, 2024.

Hermans, A., Beyer, L., and Leibe, B. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017.

Huang, Y., Wu, Q., Xu, J., and Zhong, Y. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, 2019.

Huang, Y., Shakeri, F., Dolz, J., Boudiaf, M., Bahig, H., and Ben Ayed, I. LP++: A surprisingly strong linear probe for few-shot CLIP. In *CVPR*, 2024.

Hurst, A., Lerer, A., et al. GPT-4o system card. *arXiv:2410.21276*, 2024.

Jin, X., He, T., Zheng, K., et al. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Li, S., Sun, L., and Li, Q. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, 2023b.

Li, S., Leng, J., Li, G., Gan, J., Gao, X., et al. CLIP-driven cloth-agnostic feature learning for cloth-changing person re-identification. *arXiv:2406.09198*, 2024.

Liang, X. and Rawat, Y. S. DIFFER: Disentangling identity features via semantic cues for clothes-changing person re-id. In *CVPR*, 2025.

Liu, F., Kim, M., Gu, Z., Jain, A., and Liu, X. Learning clothing and pose invariant 3D shape representation for long-term person re-identification. In *ICCV*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *COLM*, 2024.

Moayeri, M., Rezaei, K., Sanjabi, M., and Feizi, S. Text2Concept: Concept activation vectors directly from text. In *CVPR*, 2023.

Nguyen, V. D., Khaldi, K., Nguyen, D., Mantini, P., and Shah, S. Contrastive viewpoint-aware shape learning for long-term person re-identification. In *WACV*, 2024a.

Nguyen, V. D., Mantini, P., and Shah, S. K. Contrastive clothing and pose generation for cloth-changing person re-identification. In *CVPR*, 2024b.

Pathak, P. and Rawat, Y. S. Colors see colors ignore: Clothes changing reid with color disentanglement. In *ICCV*, 2025.

Peng, C., Wang, B., Liu, D., Wang, N., Hu, R., and Gao, X. Masked attribute description embedding for cloth-changing person re-identification. *IEEE Trans. Multimed.*, 2024.

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.-G., and Xue, X. Long-term cloth-changing person re-identification. In *ACCV*, 2020a.

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.-G., and Xue, X. Long-term cloth-changing person re-identification. In *ACCV*, 2020b.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. D. Linear adversarial concept erasure. In *ICML*, 2022.

Shu, X., Wang, X., Zang, X., Zhang, S., Chen, Y., Li, G., and Tian, Q. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Trans. Circuits Syst. Video Technol.*, 32(7):4390–4403, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wan, F., Wu, Y., Qian, X., Chen, Y., and Fu, Y. When person re-identification meets changing clothes. In *CVPRW*, 2020.

Wang, P., Bai, S., Tan, S., et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

Xu, W., Liu, H., Shi, W., Miao, Z., Lu, Z., and Chen, F. Adversarial feature disentanglement for long-term person re-identification. In *IJCAI*, 2021.

Yang, Q., Wu, A., and Zheng, W.-S. Person re-identification by contour sketch under moderate clothing change. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6):2029–2046, 2021.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023a.

Yang, Z., Lin, M., Zhong, X., Wu, Y., and Wang, Z. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, 2023b.

A. Clothing Text Description Generation

We use GPT-4o (Hurst et al., 2024) to generate structured clothing descriptions for each person image. The prompt below, developed with GPT-4o's assistance, instructs the model to extract clothing attributes in a consistent format. We apply this prompt to generate clothing text descriptions for all images in the PRCC (Yang et al., 2021), LTCC (Qian et al., 2020b), Celeb-reID-light (Huang et al., 2019), and LaST (Shu et al., 2022) datasets.

Prompt:

```
You are given an image of an individual.
Your task is to provide a structured
clothing and accessories description.


Instructions:


Clothing & Accessories:
- Upper Garment: type, color, style
- Lower Garment: type, color, style
- Footwear: type and color
- Accessories: visible items or
  "None observed"


Clothing & Accessories sentence:
One concise sentence describing what
the person is wearing.


Output Format:

Clothing & Accessories
- Upper Garment: [description]
- Lower Garment: [description]
- Footwear: [description]
- Accessories: [description]


Clothing & Accessories sentence:
[Summary sentence]
```

Example Output:

Clothing & Accessories

- Upper Garment: Dark long-sleeve shirt over a light T-shirt
- Lower Garment: Dark pants
- Footwear: Dark shoes
- Accessories: None observed

Clothing & Accessories sentence: Wearing a dark long-sleeve shirt over a light T-shirt, dark pants, and dark shoes.

B. VLM Comparison for Clothing Description

We evaluate whether the method is reproducible with open-source VLMs. Table 8 shows that InternVL3.5 (8B) (Chen et al., 2024) and Qwen3-VL (8B) (Wang et al., 2024) achieve performance within 0.2% of GPT-4o, confirming that our method does not require proprietary models.

VLM		PRCC (CC)		LTCC (CC)	
Model	**Params**	**R-1**	**mAP**	**R-1**	**mAP**
InternVL3.5 (Chen et al., 2024)	8B	74.1	70.1	55.9	**30.3**
Qwen3-VL (Wang et al., 2024)	8B	**74.6**	70.0	56.0	30.0
GPT-4o (Hurst et al., 2024)	–	74.4	**70.2**	**56.1**	30.2

Table 8. Performance across VLMs for clothing description generation. Open-source 8B models match GPT-4o performance (<0.2% gap).

C. SVD-Based Query Initialization

Algorithm 1 describes the SVD initialization procedure. We use `torch.linalg.svd` for efficiency. Embeddings are L2-normalized before concatenation, and the matrix is centered before SVD to capture variance not mean direction. SVD is preferred over ICA or K-Means as it provides the optimal low-rank approximation aligned with our subspace learning objective.

Algorithm 1 SVD-based Query Initialization

Require: Dataset $\mathcal{D}$, CLIP text encoder f_{text}, query count K

Ensure: $\mathbf{Q}_{\text{init}} \in \mathbb{R}^{K \times d}$

1: Generate clothing descriptions $\{s_i\}_{i=1}^{M}$ via VLM
2: Encode & L2-normalize: $\mathbf{t}_c^{(i)} = f_{\text{text}}(s_i)/\|f_{\text{text}}(s_i)\|_2$
3: Stack: $\mathbf{T} = [\mathbf{t}_c^{(1)}, \ldots, \mathbf{t}_c^{(M)}]^\top \in \mathbb{R}^{M \times d}$
4: Center: $\mathbf{T}_c = \mathbf{T} - \text{mean}(\mathbf{T}, 0)$
5: SVD: $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top = \text{SVD}(\mathbf{T}_c)$
6: **return** $\mathbf{Q}_{\text{init}} = \mathbf{V}[:, :K]^\top$