空间感知与计算实验室
spAtial Sensing & Computing Lab

CVPR JUNE 3-7, 2026 DENVER COLORADO

Text-guided Feature Disentanglement for Cross-modal Gait Recognition

Zhiyang Lu Ming Cheng*
Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University.
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.

Why gait recognition?

A non-contact biometric for long-range real-world identification

CVPR JUNE 3-7, 2026 DENVER COLORADO

OpenGait

Non-contact

No active cooperation

Long-range

Surveillance-style observation

Hard to disguise

Dynamic walking patterns

Application pull

- Intelligent surveillance and suspect tracking
- Smart buildings and non-intrusive security
- Health monitoring and abnormal gait analysis

Community signal

OpenGait frames gait recognition as a flexible and extensible research platform for benchmarking and practical deployment.

Visual source: OpenGait GitHub / Fan et al., CVPR 2023.

From single modality to cross-modal gait

Real systems are multi-sensor: cameras and LiDAR see complementary cues

2D camera

- Rich appearance and silhouette motion
- Sensitive to illumination and viewpoint/domain shifts

3D LiDAR

- Geometry/depth cues and outdoor robustness
- Different sensor statistics from camera images

Visual sources: Shen et al., CVPR 2023 LidarGait; Han et al., ACM MM 2024 FreeGait.

The bottleneck: modality gap dominates identity cues

The same person looks different across 2D silhouettes and 3D depth maps

Existing alignment strategies

- Synthetic pretraining can introduce **domain bias**
- Prototype / contrastive alignment may **over-compress** classes
- Disentanglement is often a **black box**

Key question

How can we separate sensor-specific modality information while keeping identity-relevant gait cues?

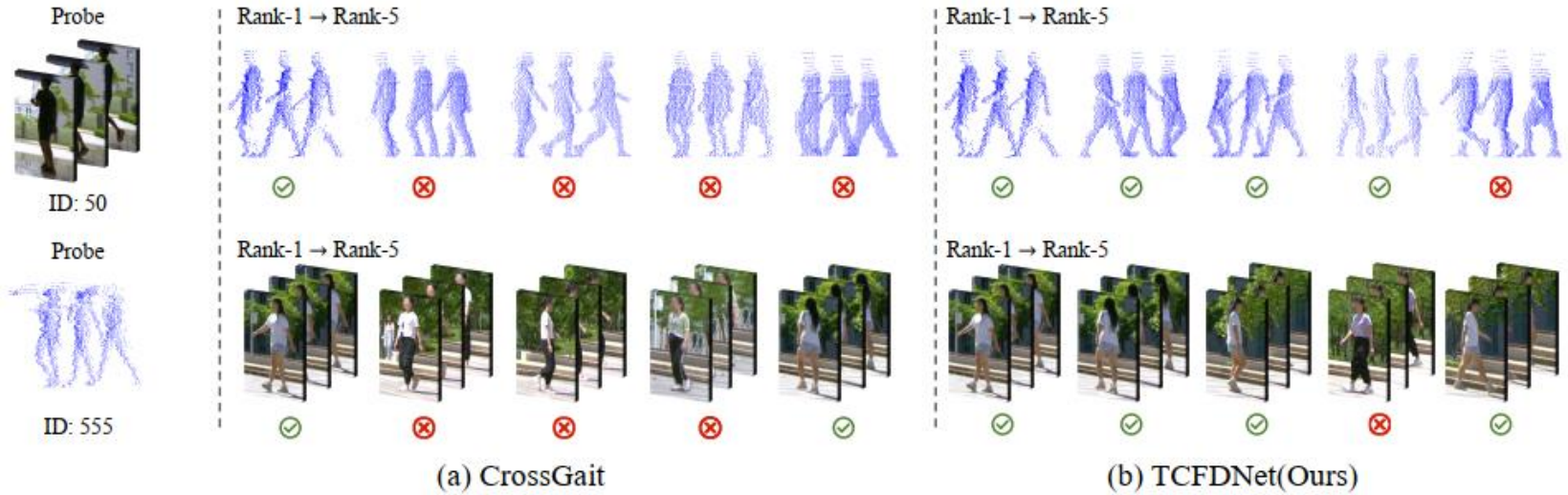

Figure 1. The visualization comparison of cross-modal gait retrieval results includes outcomes from Rank-1 to Rank-5, where a green circle with a checkmark indicates a correct retrieval, and a red circle with a cross denotes an incorrect result.

Source: our CVPR 2026 paper, Sec. 1. External visual: LidarGait project page.

Our idea: use language as a semantic anchor

Text descriptions make modality-specific factors explicit and controllable

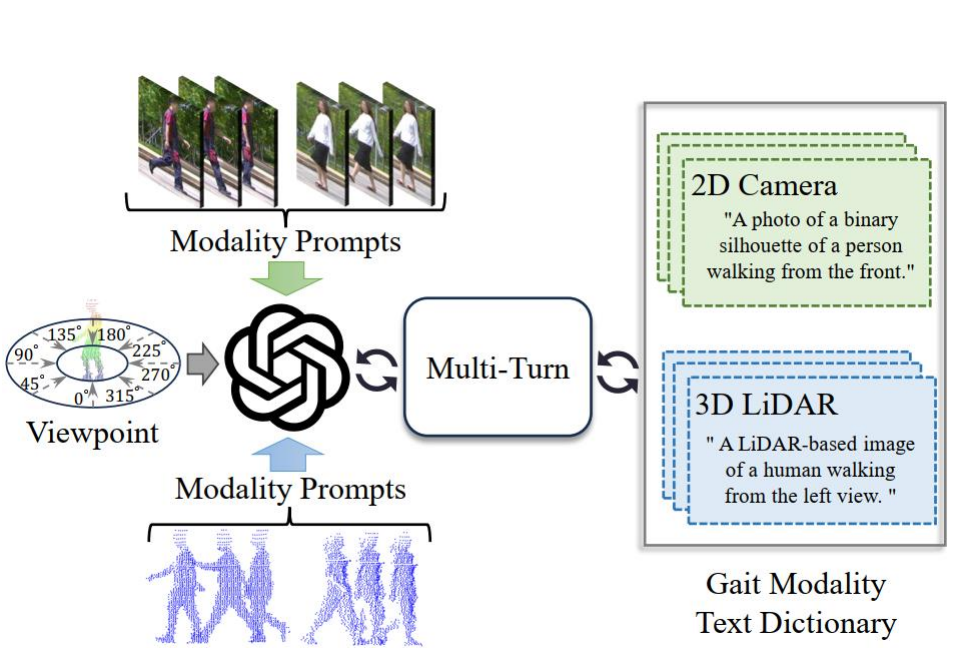

Figure 2. Details of the GMTD construction.

Text prior

modality + viewpoint

Shared gait

identity-consistent

- Generate modality-aware descriptions with LLMs
- Align text and visual gait features in a vision-language space
- Use selected text to reconstruct modality-specific components
- Derive modality-shared features by residual decomposition

TCFDNet at a glance

A complete text-guided feature disentanglement network

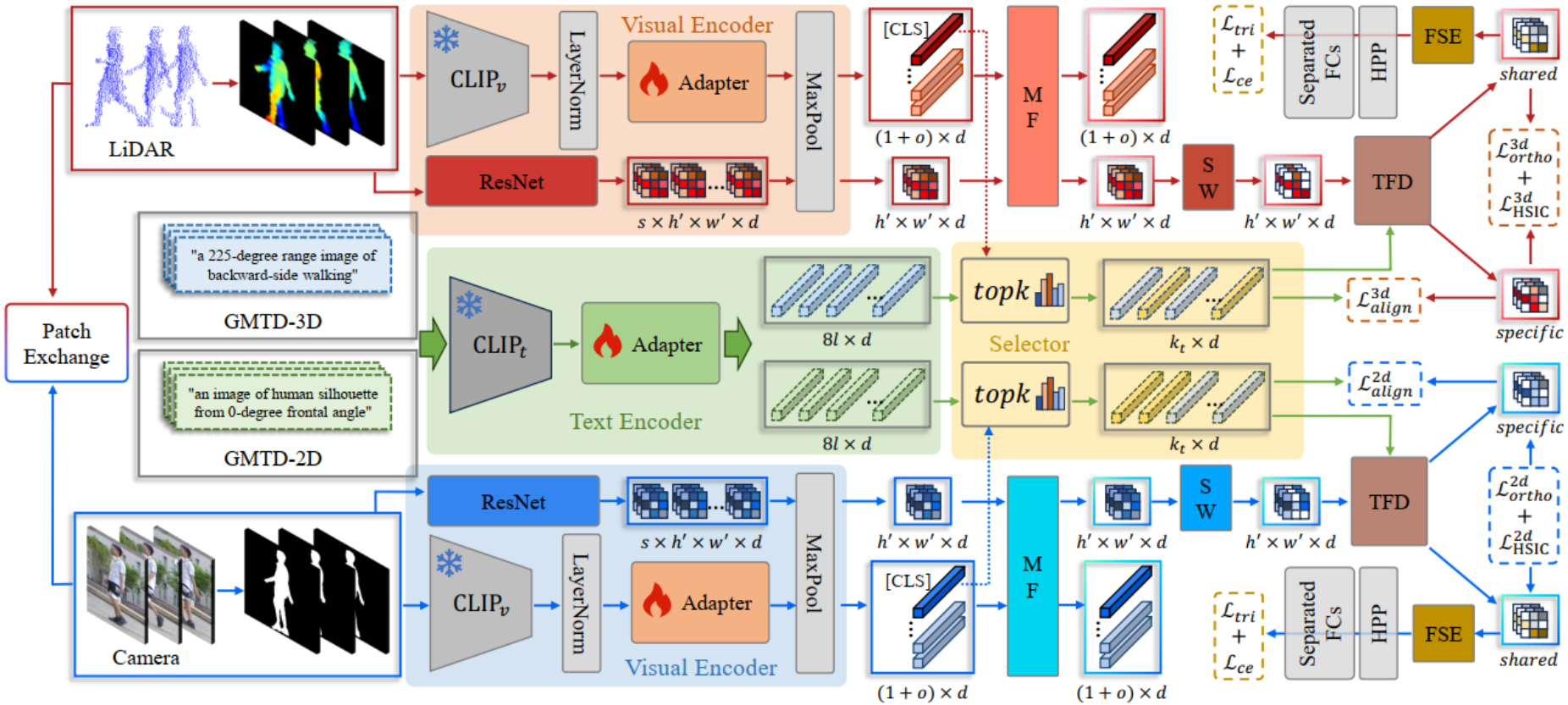

Figure 3. Illustration of the proposed framework

CLIP + ResNet
Multi-grained visual encoder

TFD
Text-guided disentanglement

FSE
Stability refinement

CE + Tri + MA/MO/HSIC
Training constraints

Component 1: Gait Modality Text Dictionary

Explicit semantic priors without extra inference sensors

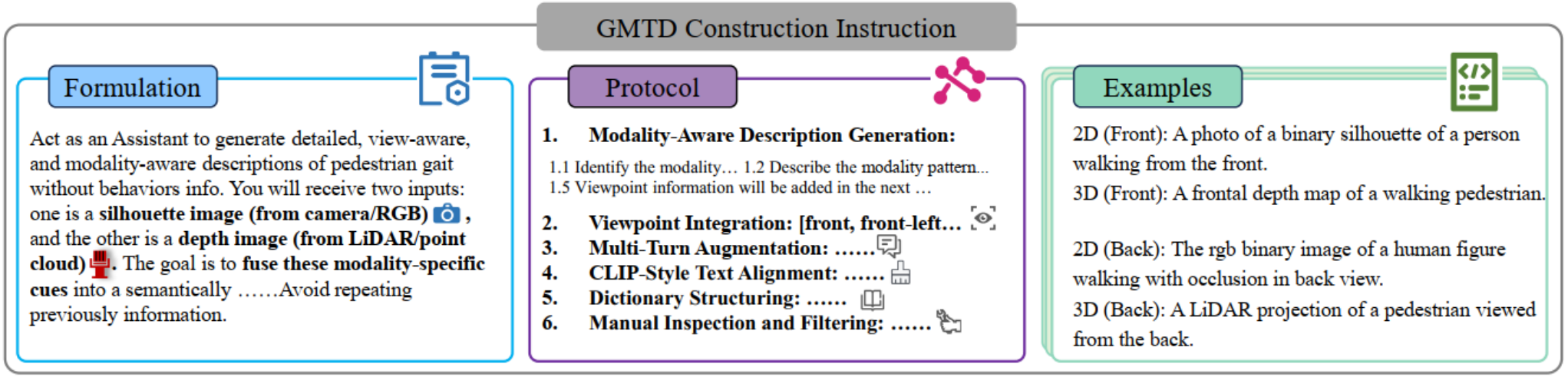

Figure 4. The instruction for GMTD, which consists of three parts: formulation, protocol, and examples. This design encourages LLMs to perform instruction-following, chain-of-thought, and in-context generation.

Top-k retrieval

kt = 16

Fair inference

no GT viewpoint

Modality-aware

not just view cues

Dictionary structure

2 modalities × 8 viewpoints × multi-turn language augmentation

The model automatically retrieves prompts based only on input visual features.

No external viewpoint labels, pose estimators, or auxiliary sensors are required.

Component 2: Text-guided Feature Disentanglement

Reconstruct modality-specific information, then keep the residual shared gait feature

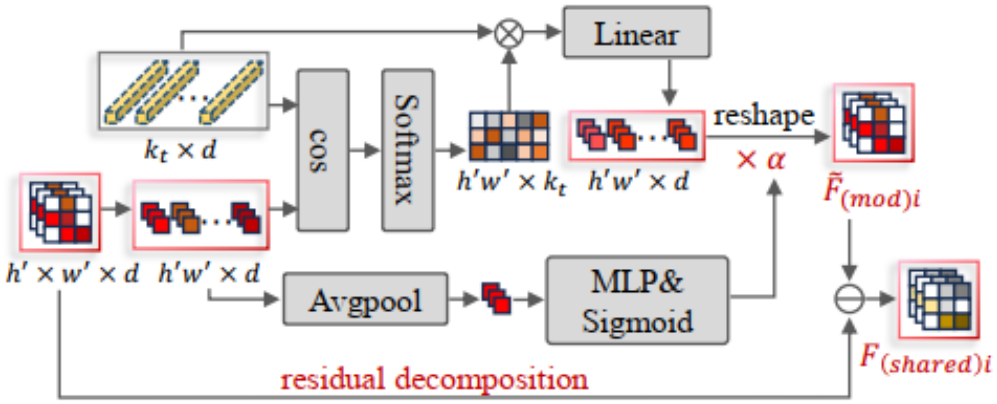

Figure 5. The flowchart of TFD module.

TFD logic

- Select top-k text prototypes by visual–text cosine similarity
- Reconstruct modality-specific component in visual space
- Use gated residual decomposition to obtain shared representation

MA loss align specific feature with text

MO loss orthogonalize specific/shared

HSIC remove nonlinear dependence

Component 3: Robustness after disentanglement

Stabilize the shared feature and expose the model to cross-modal perturbations

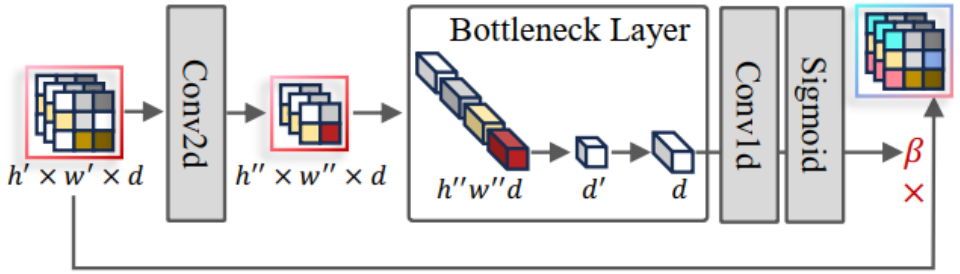

Figure 6. Illustration of the FSE module.

Feature Stability Enhancement

- Local spatial correlation via 2D convolution
- Global channel dependency via bottleneck + 1D convolution

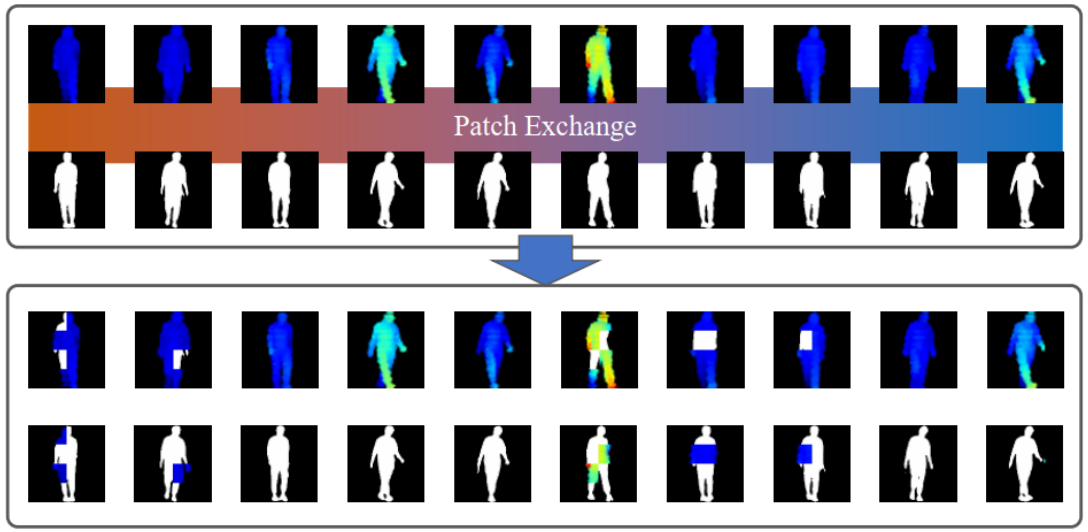

Figure 7. Visualization of the patch exchange data augmentation strategy.

Cross-modal Patch Exchange

- Swap local regions between 2D silhouettes and 3D depth maps
- Best setting: patch size 16 and exchange probability 0.1

State-of-the-art results

Consistent gains on SUSTech1K and FreeGait in both retrieval directions

Table 1. Rank-1 accuracy of cross-modal gait recognition from 2D camera to 3D LiDAR on the SUSTech1K dataset. ♠ indicates that a significant quantity of extra data is required for pre-training in this approach. The best results are indicated in **bold**, second in underline.

Methods		Camera(2D) → LiDAR(3D)								Overall	
		Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	Rank-1	Rank-5
CAJ [42]	ICCV'21	16.4	-	7.5	-	7.4	-	-	2.4	11.3	30.1
SAAI [13]	ICCV'23	22.4	-	14.3	-	14.0	-	-	5.3	23.1	49.5
LidarGait [28]	CVPR'23	18.2	-	3.4	-	3.4	-	-	4.7	9.6	28.1
CL-Gait♠ [15]	ECCV'24	-	-	-	-	-	-	-	-	55.1	77.3
CrossGait [38]	IJCB'24	63.2	-	30.6	-	38.5	-	-	**11.8**	53.6	77.0
IDKL [27]	CVPR'24	60.3	49.8	29.2	48.5	36.9	50.7	64.2	9.4	52.2	75.2
TVI-LFM [18]	NeurIPS'24	61.0	50.3	30.1	50.2	37.5	51.0	66.5	10.0	53.0	76.1
TSKD [33]	PR'25	52.1	43.6	27.9	48.0	32.7	41.1	55.6	6.3	42.6	65.8
SCR [44]	IF'25	61.3	52.9	29.6	53.0	39.1	**53.7**	69.4	10.3	54.9	78.1
TCFDNet	**Ours**	**67.6**	**60.8**	**36.1**	**55.4**	39.1	52.8	**71.2**	11.2	**55.9**	78.1

Table 2. Rank-1 accuracy of cross-modal gait recognition from 3D LiDAR to 2D camera on the SUSTech1K dataset. ♠ denotes methods pre-trained on synthetic data.

Methods		LiDAR(3D) → Camera(2D)								Overall	
		Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	Rank-1	Rank-5
CAJ [42]	ICCV'21	15.3	-	6.4	-	13.0	-	-	2.3	12.3	32.3
SAAI [13]	ICCV'23	26.5	-	21.9	-	23.2	-	-	3.2	26.1	54.1
LidarGait [28]	CVPR'23	23.2	-	14.2	-	24.7	-	-	2.4	18.3	39.6
CL-Gait♠ [15]	ECCV'24	-	-	-	-	-	-	-	-	53.3	75.6
CrossGait [38]	IJCB'24	62.2	-	35.4	-	57.8	-	-	10.3	56.4	79.8
IDKL [27]	CVPR'24	59.6	52.3	31.0	49.5	55.2	56.1	65.3	7.9	54.8	77.1
TVI-LFM [18]	NeurIPS'24	60.4	53.0	32.7	51.6	56.4	55.8	69.2	9.1	55.7	78.5
TSKD [33]	PR'25	50.1	41.3	27.7	42.8	45.9	46.2	52.5	7.8	47.2	68.1
SCR [44]	IF'25	61.6	54.1	35.8	52.0	58.1	55.9	72.6	10.2	57.7	79.5
TCFDNet	**Ours**	**70.9**	**64.8**	**36.7**	**59.1**	**63.3**	**64.3**	**78.7**	**11.1**	**61.7**	**82.5**

55.9 R-1

SUSTech1K Camera → LiDAR

Rank-5: 78.1

61.7 R-1

SUSTech1K LiDAR → Camera

Rank-5: 82.5

Table 3. Accuracy of cross-modal gait recognition on the FreeGait.

Methods		2D → 3D		3D → 2D	
		R-1	R-5	R-1	R-5
HMRNet[17]	MM'24	23.5	55.7	25.1	57.0
CrossGait[38]	IJCB'24	29.6	60.8	32.3	65.9
IDKL [27]	CVPR'24	36.7	67.4	39.5	70.3
TVI-LFM [18]	NeurIPS24	38.9	69.1	41.0	71.8
TSKD [33]	PR'25	25.1	57.9	26.7	60.8
SCR [44]	IF'25	40.1	72.0	43.3	75.9
TCFDNet	**Ours**	**52.1**	**85.3**	**57.9**	**87.2**

52.1 R-1

FreeGait Camera → LiDAR

Rank-5: 85.3

57.9 R-1

FreeGait LiDAR → Camera

Rank-5: 87.2

Why does it work?

Ablations and rebuttal experiments isolate the source of improvement

Table 4. Ablation study on SUSTech1K dataset for cross-modal gait recognition (LiDAR → Camera). At each step, only one functional group is modified while others remain fully integrated.

Text	Visual Backbone				Decoupling		Overall	
GMTD	ViT	ResNet	MF	SW	TFD	FSE	**R-1**	**R-5**
×	*full integration*				*full integration*		56.2	77.3
✓	✓	×	×	×	*full integration*		54.9	74.6
✓	×	✓	×	✓			58.4	78.9
✓	×	✓	×	×			56.7	76.5
✓	*full integration*				✓	×	59.8	80.6
✓					×	×	58.9	79.3
✓	✓	✓	✓	✓	✓	✓	**61.7**	**82.5**

Table 5. **Backbone Generalization Analysis** on SUSTech1K (L→C). We compare the default CLIP backbone with a stronger VLM (SigLIP) and a unimodal self-supervised model (DINOv2) to validate scalability and independence.

Vision Backbone / Text Encoder	Rank-1 (%)
CLIP (ViT-B/16) / CLIP (Default)	**61.7%**
SigLIP (ViT-B/16) / SigLIP	62.1%
DINOv2 (ViT-B/14) / CLIP	60.5%

+5.5% No Text 56.2 → Full 61.7

Text is necessary.

<1.5% drop under offset / shuffle

Semantic anchors tolerate noise.

60.5% DINOv2 backbone

Not only CLIP initialization.

Takeaways

Text-guided disentanglement as a new paradigm for multi-sensor gait recognition

1 Language makes modality-specific factors explicit

2 Residual decomposition isolates shared gait cues

3 FSE + patch exchange improve robustness

4 SOTA accuracy on SUSTech1K and FreeGait

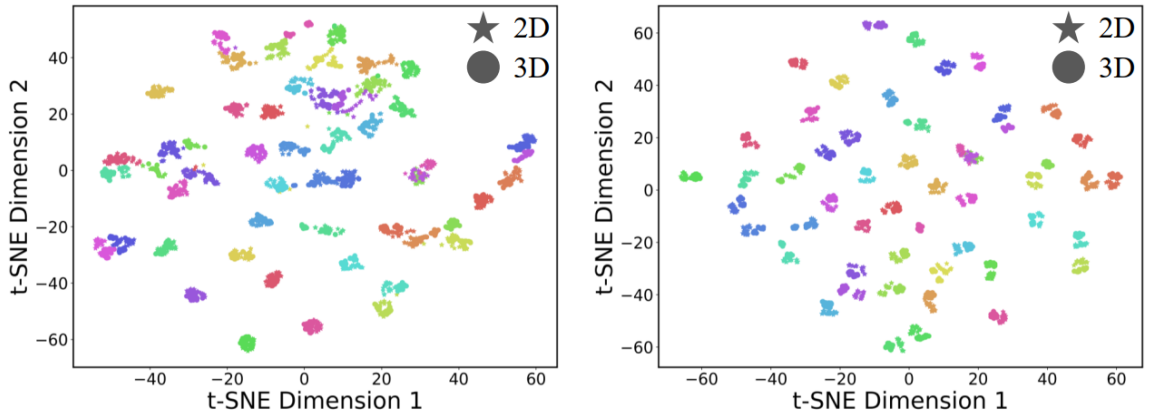

Figure 8. t-SNE visualization of cross-modal 2D and 3D features. Zooming in for details.

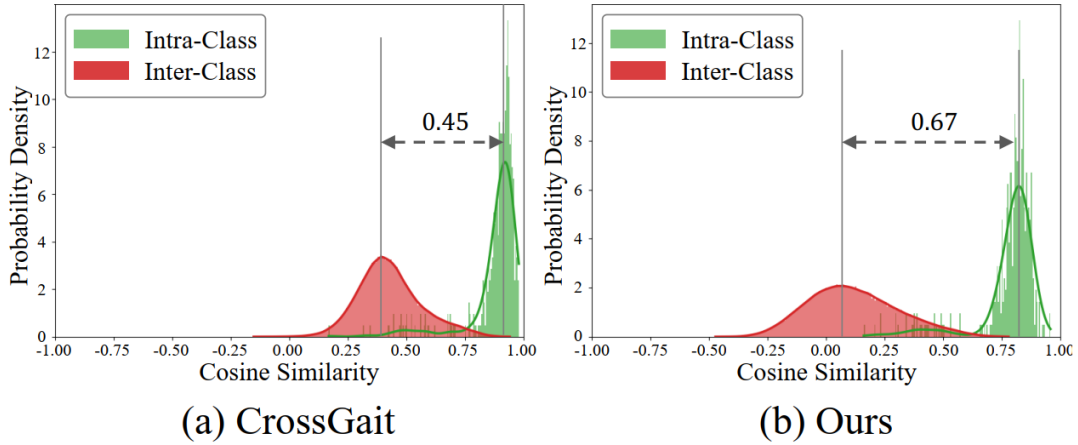

Figure 9. Visualization of cross-modal intra/inter-class cosine similarity distribution.

Thank you!