LATENT CHAIN-OF-THOUGHT NETWORK FOR GAIT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human reasoning advances through intermediate hypotheses and iterative revision, and Chain-of-Thought (CoT) brings this process into large language models by introducing explicit natural-language intermediate steps, achieving impressive performance on complex tasks such as mathematics and code. However, language has intrinsic limitations, including subjectivity, verbosity, and ambiguity, ultimately limiting reasoning capability. Analogously, visual CoT (*e.g.*, thinking with images) relies on intermediate imagery that intensifies the information bottleneck. Accordingly, recent research has pivoted from explicit CoT to latent reasoning. To this end, we present CoTNet that internalizes CoT in latent space beyond explicit language or imagery, emulating a human-like "think-then-speak" mechanism. We take gait recognition as a benchmark where networks aim to identify individuals by capturing motion patterns from videos under cross-covariate conditions. Current gait networks typically perform one-shot temporal aggregation that prematurely collapses the full gait sequence into a single gait representation, thereby blurring intermediate motion patterns. CoTNet introduces a Chain-of-Motion-Thought (CoMT), where Motion Thought with a causal mechanism preserves informative intermediate motion patterns and Motion Chain with a mix mechanism further integrates them to construct robust motion patterns. Experiments on CASIA-B, CCPG, SUSTech1K, CCGR-MINI, Gait3D, and GREW show that CoTNet achieves state-of-the-art performance and enables deeper motion-pattern analysis.

1 INTRODUCTION

"Those who know do not speak, those who speak do not know."

— *Laozi,* Tao Te Ching*, ch. 56*

Human reasoning is not invariably a one-shot process but involves a balance between fast intuition and slow deliberation. The former is efficient in simple or familiar contexts yet prone to heuristic biases under complexity or uncertainty, whereas the latter is more cautious and, through step-by-step reasoning and evidence integration, tends to yield more reliable judgments (Tversky & Kahneman, 1974; Ridderinkhof et al., 2004; De Neys & Raoelison, 2025). Chain-of-Thought (Wei et al., 2022) (CoT) in large language models (LLMs) externalizes a "think then speak" routine as stepwise rationales, allowing long-horizon problems to be decomposed and more readily verified (Chu et al., 2024). Yet human cognition is not confined to discrete symbols and often operates on abstract, continuous, multi-concept features that are difficult to express precisely in language or imagery. Forcing these into fixed words or images often creates an information bottleneck (Chen et al., 2025).

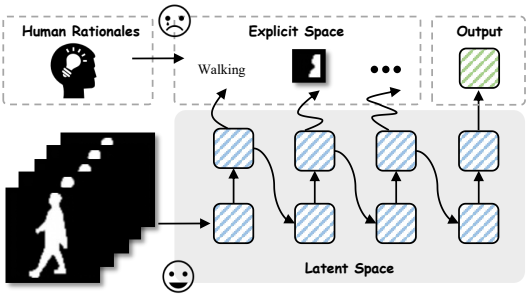


Figure 1: Explicit vs. latent Chain-of-Thought. Explicit CoT uses human rationales in an explicit space (*e.g.*, "walking" in words or images), whereas latent CoT performs multi-step reasoning in hidden states before producing the output.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These limitations have prompted a shift from explicit CoT to reasoning in latent space. Instead of externalizing step-by-step rationales in language or images, latent CoT performs the intermediate reasoning within hidden states (Hao et al., 2024b). Accordingly, we naturally ask ***how to embed latent CoT into a general network.*** We adopt gait recognition as the benchmark, which aims to identify individuals from long-range videos by capturing motion patterns while remaining robust to complex scenarios (*e.g.*, cross-view, cross-carrying, and cross-clothing conditions) (Song et al., 2022). These covariates compress appearance cues and introduce temporal ambiguity, thereby requiring the network to reason over fine-grained dynamic information. As shown in Figure 1, explicit CoT operates in the symbolic space by producing intermediate rationales (*e.g.*, words or images) before output, which may be verbose and introduce an information bottleneck. In contrast, latent CoT performs multi-step reasoning inside the hidden states where the network iteratively refines motion patterns in latent space, preserving informative dynamics. This internal chain yields a compact, noise-resistant representation that better supports the final decision under a complex environment.

We further analyze motion patterns. As illustrated in Figure 2, conventional temporal aggregation in gait networks compresses temporal information before output using max pooling or attention pooling, which often prematurely collapses the sequence into a single descriptor and weakens dynamics. To address this, we propose Chain-of-Thought Network (CoTNet) with a Chain-of-Motion-Thought (CoMT) that follows a "think then speak" paradigm instead of pooling everything at once:

Motion Thought: a causal temporal scan preserves informative intermediate motion patterns.

Motion Chain: a mix mechanism combines intermediate patterns into a robust motion pattern.

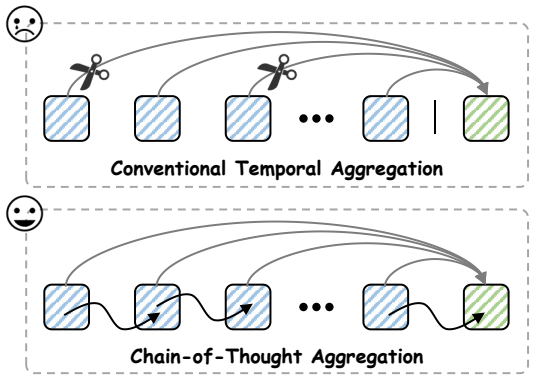


Figure 2: Comparison of temporal aggregation. Conventional temporal aggregation pools all frames and may prematurely collapse motion patterns, whereas Chain-of-Thought aggregation forces the network to preserve intermediate motion patterns for the output.

Two perspectives contextualize latent CoT. **Perspective 1: Recursive model for refinement.** Latent CoT can be viewed as recursive refinement in hidden space. The model forms provisional hypotheses and updates them across a few internal steps before committing to an output. Such looped or iterative latent computation has been shown to improve performance on complex reasoning by consolidating evidence without emitting explicit rationales (Hao et al., 2024b; Zhu et al., 2025; Chen et al., 2025). **Perspective 2: Causal attention for sequential modeling.** Reasoning proceeds under a past-to-present causal prior, where attention curates and stabilizes temporal evidence instead of collapsing the sequence in one pass. This causally constrained view preserves informative mid-range structure and aligns with standard practices in sequence modeling across modalities (Vaswani et al., 2017; Van Den Oord et al., 2016; Bai et al., 2018).

Our main contributions are as follows:

• **Perspective.** We contextualize latent CoT within a generic neural network, extending Chain-of-Thought reasoning beyond explicit language or imagery into continuous hidden representations. This formulation broadens the reasoning capacity of neural networks while alleviating the verbosity, ambiguity, and information bottlenecks of explicit rationales.

• **Architecture.** We propose CoTNet, which introduces a Chain-of-Motion-Thought (CoMT) to preserve and refine informative intermediate motion patterns in latent space. By replacing one-shot temporal aggregation with progressive motion reasoning, CoTNet alleviates premature sequence collapse and constructs more robust motion representations under complex covariates.

• **Experiments.** Extensive evaluations on CASIA-B, CCPG, SUSTech1K, CCGR-MINI, Gait3D, and GREW demonstrate that CoTNet achieves state-of-the-art performance across both controlled and in-the-wild scenarios. Further ablation and visualization analyses validate the effectiveness of latent CoT and provide deeper insights into how intermediate motion patterns are preserved and integrated.

2 Related Work

2.1 Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning augments large language models (LLMs) with intermediate steps instead of directly predicting the final answer. Existing work is often divided into explicit CoT, where the model outputs natural-language rationales, and latent CoT, where the process is kept in hidden states or auxiliary tokens.

Explicit Chain-of-Thought. Chain-of-thought prompting (Wei et al., 2022) guides models to generate stepwise natural-language rationales and substantially improves performance on arithmetic and commonsense benchmarks with only a few CoT exemplars. Zero-shot CoT (Kojima et al., 2022) shows that a brief instruction such as "Let's think step by step" already elicits multi-step reasoning without task-specific demonstrations. Self-consistency decoding (Wang et al., 2022) samples multiple reasoning chains and aggregates their final answers to mitigate spurious trajectories. Tree-of-Thought (Yao et al., 2023) further organizes partial rationales into a search tree so that the model can explore, score and prune alternative thoughts.

Latent Chain-of-Thought. Latent CoT methods move this process from natural language into representation space. Latent-variable CoT training (Phan et al., 2023) treats rationales as unobserved variables and maximizes the marginal likelihood of correct answers, encouraging internal multi-step reasoning from answer-only supervision. Coconut (Chain of Continuous Thought) (Hao et al., 2024a) represents the reasoning state as a continuous hidden vector that is iteratively fed back into the model, enabling multi-step deliberation without emitting intermediate tokens. Compressed CoT (CCoT) (Cheng & Van Durme, 2024) uses a small number of continuous contemplation tokens to summarize long explicit chains, offering a compact alternative to textual CoT. SoftCoT (Xu et al., 2025) introduces soft thought tokens in a continuous space that are reused across decoding steps, yielding CoT-like decomposition with fewer generated tokens. System-1.5 reasoning (Wang et al., 2025b) couples language-space CoT with latent shortcuts, dynamically allocating computation so that models retain CoT-level reasoning ability while emitting far fewer tokens.

2.2 Gait Recognition

Model-based gait recognition. Model-based approaches describe gait in terms of poses, skeletons or parametric meshes and then recognize identities from these structured representations. PoseGait (Liao et al., 2020) encodes 3D joint trajectories with kinematic features so that identity depends mainly on motion patterns. GaitGraph and GaitGraph2 (Teepe et al., 2021; 2022) treat each frame as a skeleton graph and use graph convolutions to capture spatial and temporal dependencies between joints. GaitTR (Zhang et al., 2023) and GaitMixer (Pinyoanuntapong et al., 2023) introduce attention or token mixing to model long-range relations in pose sequences, while GPGait (Fu et al., 2023) learns a unified pose representation for cross-domain transfer. SMPLGait (Zheng et al., 2022a) brings dense SMPL meshes into gait recognition, and SkeletonGait (Fan et al., 2024), HiH (Wang et al., 2023c) and GaitHeat (Fu et al., 2024) convert skeletons to gaussian-like heatmaps that can be processed by standard convolutional backbones.

Appearance-based gait recognition. Appearance-based methods operate directly on silhouette or shape sequences and currently dominate large-scale benchmarks. GaitSet (Chao et al., 2019) aggregates frames as an unordered set, and GaitPart (Fan et al., 2020) decomposes silhouettes into body parts to learn fine-grained motions. GaitGL (Lin et al., 2021) fuses global and local descriptors in a 3D CNN, and GaitBase (Fan et al., 2023) provides a simple yet strong baseline for real-world deployment. Later methods such as DANet (Ma et al., 2023), DyGait (Wang et al., 2023d), HSTL (Wang et al., 2023b), VPNet (Ma et al., 2024), GLGait (Peng et al., 2024a) and GaitMoE (Huang et al., 2024) emphasize local–global modeling, dynamic feature selection and expert mixtures to handle pose changes, occlusion and view variation. Another stream, including GaitGCI (Dou et al., 2023), GaitCSV (Wang et al., 2023a), CLTD (Xiong et al., 2024), GaitC^{3}I (Wang et al., 2025a), QAGait (Wang et al., 2024b), Free Lunch (Wang et al., 2024a) and GaitAttack (Hou et al., 2025), studies confounders and noise to reduce the impact of nuisance and adversarial factors.

Other gait modalities. Beyond silhouettes and skeletons, several methods explore alternative inputs. GaitEdge (Liang et al., 2022) learns gait features from RGB edge maps, and GaitParsing (Wang et al., 2023e), LandmarkGait (Wang et al., 2023f) and ParsingGait (Zheng et al., 2023) leverage

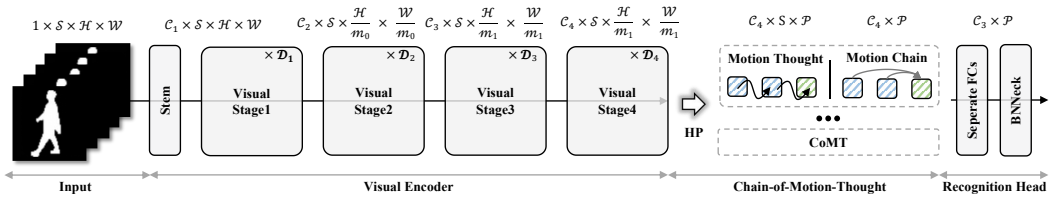


Figure 3: Overview of CoTNet. A gait sequence is first encoded by a visual encoder and horizontally partitioned (HP) into motion patterns, which are then processed by the Chain-of-Motion-Thought (CoMT), where Motion Thought and Motion Chain preserve and mix these intermediate motion patterns before the final representation is fed into the recognition head.

semantic parsing or landmarks to isolate foreground body parts. LidarGait (Shen et al., 2023) uses LiDAR point clouds for illumination-invariant geometry, and MMGaitFormer (Cui & Kang, 2023) and CL-Gait (Guo et al., 2024) integrate multiple modalities or camera views to improve robustness in complex scenes.

3 Methodology

3.1 Overview

As shown in Figure 3, a canonical gait network consists of an input gait sequence $X \in \mathbb{R}^{1\times S\times H\times W}$, where $1, S, H, W$ denote the input channel, sequence length, frame height, and frame width, respectively, and a visual encoder $\mathcal{E}$, a horizontal pooling $\mathcal{P}$, and a recognition head $\mathcal{H}$. $\mathcal{E}$ comprises a stem and several visual stages (*e.g.*, core blocks in ResNet (He et al., 2016)), $\mathcal{P}$ applies global max and average pooling, and $\mathcal{H}$ includes Separate FCs and a BNNeck (*e.g.*, fully connected layers and classifier). The gait sequence X is first processed by $\mathcal{E}$ to extract common frame-wise features, and then $\mathcal{P}$ is applied to each frame to horizontally partition the feature map into body parts Y. Finally, the features of each body part are fed independently into $\mathcal{H}$ to produce the output F. The entire procedure can be formalized as follows:

$$Y = \mathcal{P}(\mathcal{E}(X)) \tag{1}$$

$$F = \mathcal{H}(Y) \tag{2}$$

where $Y \in \mathbb{R}^{C_4\times S\times P}$ denotes the part features, $F \in \mathbb{R}^{C_3\times P}$ represents the final gait features, and C, P denote the channel and the number of body parts. The core module of CoTNet is applied independently to each body part after horizontal pooling to enhance and analyze fine-grained motion patterns. ***Next, we omit the part index for simplicity.***

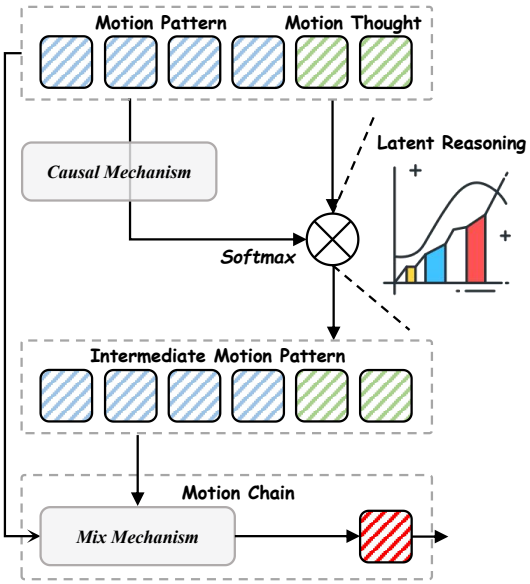


Figure 4: Chain-of-Motion-Thought. The motion thought applies a causal mechanism over motion patterns to produce intermediate motion patterns, and the motion chain then mixes these intermediates into a robust motion pattern.

3.2 Chain-of-Thought Network

Conventional temporal aggregation compresses an entire gait sequence in a single step, which often erases mid-range, phase-aligned dynamics. CoTNet aims to defer global aggregation in favor of a brief latent reasoning that (i) causally preserves informative intermediate motion patterns and (ii) integrates them into robust motion patterns, rather than collapsing the sequence at once. Figure 4 illustrates the core module, Chain-of-Motion-Thought (CoMT), together with two components:

Chain-of-Motion-Thought. The idea is to realize a human-like "think then speak" mechanism entirely in hidden space so that intermediate motion patterns are preserved before global aggregation. As shown in Figure 4, the motion patterns Y and motion thoughts M are concatenated along the temporal axis, and then a causal mechanism produces time-wise weights (softmax over the sequence), yielding intermediate motion patterns. Motion Chain with a mix mechanism then integrates this intermediate signal into a robust motion pattern. The entire procedure can be formalized as follows:

$$\begin{aligned} Y \in \mathbb{R}^{C_4 \times S}, \qquad M \in \mathbb{R}^{C_4 \times l} \\ Y_i = [\, Y \; M \,] \in \mathbb{R}^{C_4 \times (S+l)} \end{aligned} \tag{3}$$

where l is the number of motion thoughts and Y_i is processed by a causal mechanism.

Motion Thought. As shown in Figure 5, we impose a "first-order causal scan" that restricts each time step to its immediate past together with a small set of learnable motion thoughts. This causal constraint induces a fine-grained temporal calculus. The network is forced to explain the current motion pattern by what just occurred rather than consulting distant or future frames. Practically, it (i) aligns phases via step-to-step consistency, (ii) preserves micro-transitions (*e.g.*, heel-strike, toe-off, arm swing) that one-shot pooling often washes out, (iii) suppresses spurious long-range correlations and future leakage, and (iv) provides a stable, transferable inductive bias under cross-view, carrying, and clothing changes. The motion thoughts act as compact memory that summarizes recent evidence. The causal scan then weighs the previous state, the current state, and this memory to highlight informative intermediate motion patterns. Compared with non-causal aggregation, the constraint concentrates capacity on the short-horizon dynamics that actually generate gait, improving the robustness across covariates. The process is formulated as follows:

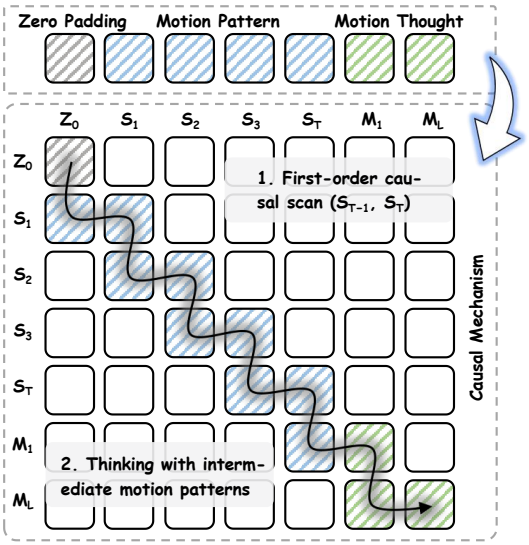


Figure 5: Motion thought with a causal mechanism. A first-order causal scan traverses the zero-padded motion patterns and motion thoughts, generating intermediate motion patterns along the diagonal and then reasoning over them to produce the relations.

$$W_{\mathrm{i}} = \mathrm{Conv1D}_{\mathrm{causal}}(Y_{\mathrm{i}}) \tag{4}$$

$$W_{\mathrm{o}} = \mathrm{softmax}(W_{\mathrm{i}}) \tag{5}$$

$$Y_{\mathrm{o}} = Y_{\mathrm{i}} \otimes W_{\mathrm{o}}^{\top} \tag{6}$$

where $W_{\mathrm{o}} \in \mathbb{R}^{N \times (S+l)}$ denotes the relations among motion patterns, $Y_{\mathrm{o}} \in \mathbb{R}^{C_4 \times N}$ represents the intermediate motion patterns, and N is the number of the intermediates.

Motion Chain. After Motion Thought preserves informative intermediate motion patterns, we compose them into a single descriptor without collapsing temporal evidence. We view the intermediate motion patterns as a small set of complementary motion bases, each capturing a distinct micro dynamic preserved by the causal scan. A lightweight mix mechanism aggregates these bases as a learnable alternative. Therefore, unlike max or average pooling that reduces the whole sequence in one pass, the mechanism consolidates curated evidence, preserves salient intermediates, and suppresses noise and spurious co-activations. The bottleneck structure encourages low-rank composition and smooths phase jitter by fusing nearby bases into stable aggregates, yielding a robust and discriminative motion pattern that retains short horizon transitions and their interactions. Formally:

$$Y_{\mathrm{o}} \in \mathbb{R}^{C_4 \times N} \xrightarrow{\text{Mix Mechanism}} \mathbb{R}^{C_4 \times 1} \tag{7}$$

where the mix mechanism is set to a fully connected layer, integrating the intermediate motion patterns into a robust motion pattern.

Table 1: Datasets and optimization on controlled and in-the-wild benchmarks.

Environment	Dataset	Train Id.	Train Seq.	Test Id.	Test Seq.	Batch Size	Optimizer	LR Milestones (K)
Controlled	CASIA-B (Yu et al., 2006)	74	8,140	50	5,500	(8, 16)	SGD lr = 0.1 dr = 0.1 wd = 0.0005	[20, 40, 50] (60)
	CCPG (Li et al., 2023)	100	8,187	100	8,095	(8, 16)		[20, 40, 50] (60)
	SUSTech1K (Shen et al., 2023)	200	5,988	850	19,228	(8, 8)		[20, 30, 40] (50)
	CCGR-MINI (Zou et al., 2024)	571	27,507	399	20,377	(8, 16)		[30, 55, 65] (80)
In-the-wild	Gait3D (Zheng et al., 2022a)	3,000	18,940	1,000	6,369	(32, 4)		[20, 40, 50] (60)
	GREW (Zhu et al., 2021)	20,000	102,887	6,000	24,000	(32, 4)		[80, 120, 150] (180)

3.3 Training and Inference Details

Training Stage. We adopt standard identity supervision. The network is trained end-to-end with a part-wise triplet loss and cross-entropy loss.

$$\mathcal{L} = \mathcal{L}_{\text{tp}} + \mathcal{L}_{\text{ce}} \tag{8}$$

Inference Stage. As in conventional gait recognition, features after Separate FCs identify the subject.

4 Experiments

4.1 Datasets

Following the common protocol, we evaluate CoTNet on six public gait benchmarks that cover both controlled and In-the-wild benchmarks.

Controlled benchmarks. CASIA-B (Yu et al., 2006) is a classic indoor benchmark where subjects walk along a fixed corridor under three conditions (*i.e.*, normal walking (NM), carrying a bag (BG), and wearing a coat (CL)) and are observed from multiple discrete viewpoints, making it well suited for analyzing view and appearance changes in a clean setup. CCPG (Li et al., 2023) extends this setting with more diverse walking routes, providing more complex clothing and carrying factors. SUSTech1K (Shen et al., 2023) further scales the controlled scenario to around one thousand subjects with richer outfits and carrying status, while still keeping a constrained background. CCGR-MINI (Zou et al., 2024) is a compact subset of CCGR that maintains the same cross-view and cross-scene design, but with reduced identities and sequences, which we mainly use for rapid evaluation. The exact train/test splits and sequence counts for all four datasets are given in Table 1.

Table 2: The performance comparisons on CASIA-B.

Method	Venue	NM	BG	CL	Mean
GaitGraph2 (Teepe et al., 2022)	CVPRW22	80.3	71.4	63.8	71.8
GaitTR (Zhang et al., 2023)	ES23	94.7	89.3	86.7	90.2
GPGait (Fu et al., 2023)	ICCV23	93.6	80.2	69.3	81.0
GaitSet (Chao et al., 2019)	AAAI19	95.0	87.2	70.4	84.2
GaitPart (Fan et al., 2020)	CVPR20	96.2	91.5	78.7	88.8
GLN (Hou et al., 2020)	ECCV20	96.9	94.0	77.5	89.5
GaitGL (Lin et al., 2021)	ICCV21	97.4	94.5	83.6	91.8
QAGait (Wang et al., 2024b)	AAAI24	97.9	94.6	78.2	90.2
GaitBase (Fan et al., 2023)	CVPR23	97.6	94.0	77.4	89.8
DANet (Ma et al., 2023)	CVPR23	98.0	95.9	89.9	94.6
GaitGCI (Dou et al., 2023)	CVPR23	97.9	95.0	86.4	93.1
DyGait (Wang et al., 2023d)	ICCV23	98.4	96.2	87.8	94.1
HSTL (Wang et al., 2023b)	ICCV23	98.1	95.9	88.9	94.3
VPNet (Ma et al., 2024)	CVPR24	98.3	96.3	90.0	94.9
DeepGaitV2 (Fan et al., 2025)	TPAMI25	–	–	–	89.6
CLTD (Xiong et al., 2024)	ECCV24	98.6	96.4	89.3	94.8
Free Lunch (Wang et al., 2024a)	ECCV24	98.1	94.1	77.9	90.0
CoTNet-T	**Ours**	**99.5**	**97.5**	**89.7**	**95.6**

Table 3: The performance comparisons on CCGR-MINI.

Method	Venue	Rank-1	mAP	mINP
GaitSet (Chao et al., 2019)	AAAI19	13.8	15.4	5.8
GaitPart (Fan et al., 2020)	CVPR20	8.0	10.1	3.5
GaitGL (Lin et al., 2021)	ICCV21	17.5	18.1	6.9
GaitBase (Fan et al., 2023)	CVPR23	27.0	24.9	9.7
DeepGaitV2 (Fan et al., 2025)	TPAMI25	39.4	36.0	16.8
CoTNet-M	**Ours**	**40.9**	**37.6**	**23.8**

In-the-wild benchmarks. Gait3D (Zheng et al., 2022a) is collected in a shopping mall with real surveillance cameras, where people are captured under heavy occlusion, crowding and arbitrary walking paths. It contains several thousand identities with a moderate number of sequences per subject. GREW (Zhu et al., 2021) pushes the scale further to tens of thousands of subjects and more than one hundred thousand sequences across hundreds of cameras deployed in a city-scale environment. Both datasets adopt the official train/test

Table 4: The performance comparisons on CCPG with clothing-changing conditions.

Method	Venue	Gait Evaluation Protocol					ReID Evaluation Protocol				
		CL	UP	DN	BG	Mean	CL	UP	DN	BG	Mean
GaitGraph2 (Teepe et al., 2021)	CVPRW22	5.0	5.3	5.8	6.2	5.1	5.0	5.7	7.3	8.8	6.7
Gait-TR (Zhang et al., 2023)	ES23	15.7	18.3	18.5	17.5	17.5	24.3	28.7	31.1	28.1	28.1
MSGG (Peng et al., 2024b)	MTA23	29.0	34.5	37.1	33.3	33.5	43.1	52.9	57.4	49.9	50.8
SkeletonGait (Fan et al., 2024)	AAAI24	40.4	48.5	53.0	61.7	50.9	52.4	65.4	72.8	80.9	67.9
GaitSet (Chao et al., 2019)	AAAI19	60.2	65.2	65.1	68.5	64.8	77.5	85.0	82.9	87.5	83.2
GaitPart (Fan et al., 2020)	CVPR20	64.3	67.8	68.6	71.7	68.1	79.2	85.3	86.5	88.0	84.8
OGBase (Li et al., 2023)	CVPR23	52.1	57.3	60.1	63.3	58.2	70.2	76.9	80.4	83.4	77.7
GaitBase (Fan et al., 2023)	CVPR23	71.6	75.0	76.8	78.6	75.5	88.5	92.7	93.4	93.2	92.0
DeepGaitV2 (Fan et al., 2025)	TPAMI25	78.6	84.8	80.7	89.2	83.3	90.5	96.3	91.4	96.7	93.7
CoTNet-M	**Ours**	**85.2**	**91.8**	**87.6**	**94.4**	**89.8**	**93.0**	**97.8**	**94.6**	**97.6**	**95.8**

Table 5: The performance comparisons on SUSTech1K are reported with Rank-1 and Rank-5 accuracy (%).

Method	Venue	Probe Sequence								Overall	
		Normal	Bag	Clothing	Carrying	Umbrella	Uniform	Occlusion	Night	Rank-1	Rank-5
GaitGraph2 (Teepe et al., 2021)	CVPRW22	22.2	18.2	6.8	18.6	13.4	19.2	27.3	16.4	18.6	40.2
Gait-TR (Zhang et al., 2023)	ES23	33.3	31.5	21.0	30.4	22.7	34.6	44.9	23.5	30.8	56.0
MSGG (Peng et al., 2024b)	MTA23	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	33.8	-
SkeletonGait (Fan et al., 2024)	AAAI24	67.9	63.5	36.5	61.6	58.1	67.2	79.1	50.1	63.0	83.5
GaitSet (Chao et al., 2019)	AAAI19	69.1	68.2	37.4	65.0	63.1	61.0	67.2	23.0	65.0	84.8
GaitPart (Fan et al., 2020)	CVPR20	62.2	62.8	33.1	59.5	57.2	54.8	57.2	21.7	59.2	80.8
GaitGL (Lin et al., 2021)	ICCV21	67.1	66.2	35.9	63.3	61.6	58.1	66.6	17.9	63.1	82.8
GaitBase (Fan et al., 2023)	CVPR23	81.5	77.5	49.6	75.8	75.5	76.7	81.4	25.9	76.1	89.4
DeepGaitV2 (Fan et al., 2025)	TPAMI25	87.4	84.1	53.4	81.3	86.1	84.8	88.5	28.8	82.3	92.5
CoTNet-M	**Ours**	**91.6**	**88.2**	**67.3**	**86.0**	**90.9**	**89.9**	**93.3**	27.7	**87.2**	**94.1**

partition with disjoint identities, and evaluation follows their standard protocol, and refer to Table 1 for the detailed numbers.

4.2 Implementation Details

Input preprocessing. For all datasets, we use binary silhouettes provided by the corresponding benchmarks or by the OpenGait toolbox (Fan et al., 2023) and resize each frame to 64×44 pixels. During training, a fixed-length clip is sampled from each sequence, and the temporal order is preserved. At test time, CoTNet processes the full sequence without frame dropping so as to exploit all available temporal evidence. As in (Fan et al., 2023), each mini-batch is organised as $(\mathcal{I}, \mathcal{J})$, where $\mathcal{I}$ subjects are sampled and $\mathcal{J}$ sequences are drawn for each subject. See Table 1 for details.

Network variants. We instantiate four CoTNet backbones (CoTNet-T/S/M/L) with increasing depth. All variants share a common stem of one 2D Conv_1 layer on inputs of size $S \times C \times H \times W$ (with C=64), followed by four residual stages whose channel widths are $[C, 2C, 4C, 8C]$ and spatial resolutions are gradually reduced to $\frac{H}{4} \times \frac{W}{4}$. CoTNet-T consists of one 2D BasicBlock in Stage 1 and one 3D BottleNeck block in Stages 2–3 (block numbers $[1, 1, 1]$) without Stage 4. CoTNet-S uses one 2D BasicBlock in Stage 1 and three 3D BottleNeck stages with block numbers $[1, 1, 1, 1]$. CoTNet-M keeps a 2D BasicBlock in Stage 1 and adopts P3D BasicBlocks in Stages 2–4, with block numbers $[1, 4, 4, 1]$. CoTNet-L further deepens this P3D design, using block numbers $[2, 4, 4, 2]$ to obtain the largest temporal and spatial receptive field.

Optimization. We train all CoTNet variants from scratch on each dataset using stochastic gradient descent with momentum. Unless otherwise stated, the base learning rate is set to 0.1 with momentum 0.9, dropout rate 0.1 and weight decay 5×10^{-4}. For each benchmark, we adopt a dataset-specific multi-step schedule: the learning rate is kept constant at the base value and then reduced by a factor of 0.1 at three milestones (measured in thousands of iterations), and training stops after a fixed maximum number of iterations. The tuple of milestones and the final iteration count for every dataset are detailed in Table 1.

4.3 CONTROLLED BENCHMARK

CASIA-B. Table 2 shows that recent baselines have largely converged around 94% ∼ 95% mean accuracy. CoTNet-T slightly but consistently pushes this frontier, reaching a mean of 95.6% and improving both normal walking and covariate-heavy conditions, especially under clothing change, which remains the most challenging scenario, demonstrating effective CoT-style refinement across these conditions.

CCPG. As summarized in Table 4, CoTNet-M establishes a new state of the art performance. Under the Gait Evaluation Protocol, it reaches 89.8% mean Rank-1 accuracy, improving upon DeepGaitV2 by 6.5%. The advantage is consistent across all four covariates, and maintains Rank-1 accuracy 85.2% even under full clothing changes. Under the ReID Evaluation Protocol, CoTNet-M further pushes the mean Rank-1 accuracy to 95.8%, higher than DeepGaitV2 by 2.1%, indicating that latent CoT modelling enhances the robustness of motion patterns when clothing and background vary significantly.

SUSTech1K. Table 5 shows that CoTNet-M also achieves the best performance. It reaches 87.2% overall Rank-1 and 94.1% overall Rank-5 accuracy, surpassing the previous strong baseline DeepGaitV2 (82.3% and 92.5%) by 4.9% and 1.6% respectively. The gains are especially notable under clothing changes, where Rank-1 accuracy increases from 53.4% to 67.3%.

CCGR-MINI. As shown in Table 3, CoTNet-M achieves 40.9% Rank-1, 37.6% mAP and 23.8% mINP, surpassing the previous best DeepGaitV2 (39.4% Rank-1, 36.0% mAP, 16.8% mINP) with a particularly large gain on mINP under the long-tail setting.

4.4 IN-THE-WILD BENCHMARK

Gait3D. Table 6 shows that recent methods typically obtain around 60% ∼ 75% Rank-1 accuracy with mAP above 55%. CoTNet-M further lifts this frontier to 77.6% Rank-1 and 68.5% mAP, indicating that CoT-style motion reasoning is effective in complex scenarios.

Table 6: The performance comparisons on Gait3D.

Method	Venue	Rank-1	Rank-5	mAP
GaitSet (Chao et al., 2019)	AAAI19	36.7	58.3	30.0
GaitPart (Fan et al., 2020)	CVPR20	28.2	47.6	21.6
GaitGL (Lin et al., 2021)	ICCV21	29.7	48.5	22.3
SMPLGait (Zheng et al., 2022a)	CVPR22	46.3	64.5	37.2
MTSGait (Zheng et al., 2022b)	MM22	48.7	67.1	37.6
QAGait (Wang et al., 2024b)	AAAI24	67.0	81.5	56.5
GaitBase (Fan et al., 2023)	CVPR23	64.6	–	–
DANet (Ma et al., 2023)	CVPR23	48.0	69.7	–
GaitGCI (Dou et al., 2023)	CVPR23	50.3	68.5	39.5
DyGait (Wang et al., 2023d)	ICCV23	66.3	80.8	56.4
HSTL (Wang et al., 2023b)	ICCV23	61.3	76.3	55.5
VPNet (Ma et al., 2024)	CVPR24	75.4	87.1	–
DeepGaitV2 (Fan et al., 2025)	TPAMI25	74.4	88.0	65.8
CLTD (Xiong et al., 2024)	ECCV24	69.7	85.2	–
GaitMoE (Huang et al., 2024)	ECCV24	73.7	–	66.2
Free Lunch (Wang et al., 2024a)	ECCV24	70.1	–	61.9
CoTNet-M	**Ours**	**77.6**	**88.6**	**68.5**

Table 7: The performance comparisons on GREW.

Method	Venue	Rank-1	Rank-5	Rank-10
GaitSet (Chao et al., 2019)	AAAI19	46.3	63.6	70.3
GaitPart (Fan et al., 2020)	CVPR20	44.0	60.7	67.3
GaitGL (Lin et al., 2021)	ICCV21	47.3	63.6	–
MTSGait (Zheng et al., 2022b)	MM22	55.3	71.3	76.9
QAGait (Wang et al., 2024b)	AAAI24	59.1	74.0	79.2
GaitBase (Fan et al., 2023)	CVPR23	60.1	–	–
GaitGCI (Dou et al., 2023)	CVPR23	68.5	80.8	84.9
DyGait (Wang et al., 2023d)	ICCV23	71.4	83.2	86.8
HSTL (Wang et al., 2023b)	ICCV23	62.7	76.6	81.3
VPNet (Ma et al., 2024)	CVPR24	80.0	89.4	–
DeepGaitV2 (Fan et al., 2025)	TPAMI25	77.7	88.9	91.8
CLTD (Xiong et al., 2024)	ECCV24	78.0	87.8	–
GaitMoE (Huang et al., 2024)	ECCV24	79.6	89.1	–
Free Lunch (Wang et al., 2024a)	ECCV24	65.5	78.7	83.3
CoTNet-L	**Ours**	**82.1**	**90.3**	**92.5**

Table 8: Ablation study on Gait3D and CCPG.

Variant	Gait3D		CCPG				
	Rank-1	mAP	CL	UP	DN	BG	Mean
CoTNet	**77.6**	**68.5**	85.2	**91.8**	**87.6**	94.4	**89.8**
Analysis on core modules							
- w/o CoMT	73.6	65.1	80.1	86.4	83.9	91.4	85.5
- w/o MT	74.5	67.3	85.1	91.3	87.3	94.4	89.5
- w/o MC	69.4	59.1	82.2	88.6	84.1	92.6	86.9
- w/o CM	75.7	68.1	85.2	91.0	87.3	94.3	89.5
Analysis on number of Motion Thought							
l=1	75.5	67.6	85.2	90.8	87.2	93.7	89.2
l=2	75.0	67.0	85.2	91.8	87.6	94.4	89.8
l=3	77.6	68.5	**85.4**	90.9	87.4	**94.5**	89.6
l=4	72.5	63.8	85.0	91.0	87.3	94.3	89.4

GREW. As summarized in Table 7, existing methods largely saturate around 80% Rank-1 and 90% Rank-5. For a stable comparison, CoTNet-L follows the Free Lunch strategy (Wang et al., 2024a), using logits as gait representations at unchanged computational cost.

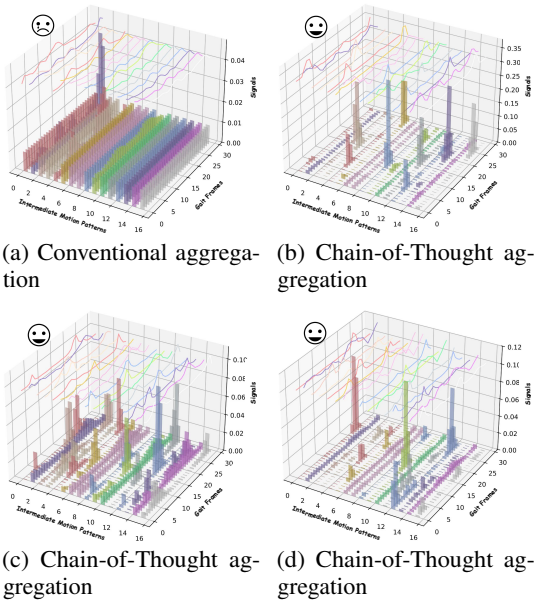
CoTNet-L pushes the performance to 82.1% Rank-1, 90.3% Rank-5 and 92.5% Rank-10, suggesting that the proposed framework scales well to large-scale In-the-wild benchmarks.

4.5 Ablation Study

The analysis on core modules. Table 8 shows that removing the CoMT drops Gait3D Rank-1 from 77.6% to 73.6% and CCPG mean accuracy performance from 89.8% to 85.5%, showing that latent reasoning is crucial for motion patterns under cross-covariate scenarios. Within CoMT, Motion Chain (MC) is the most sensitive part: without it, Gait3D Rank-1 falls to 69.4% and mAP to 59.1%. Motion Thought (MT) and the causal mechanism (CM) also bring steady gains on both benchmarks.

The number of motion thought. As shown in Table 8, the number of motion thought l shows that a small depth already suffices. On Gait3D, Rank-1 improves from 75.5% at l=1 to 77.6% at l=3, while a further increase to l=4 hurts performance. On CCPG, we use l=3 to balance accuracy and computational cost.

The deep analysis of motion patterns. Figure 6 visualizes how temporal aggregation behaves on CCPG. In Fig. 6(a), conventional aggregation with global max pooling produces almost uniform responses over motion patterns with a single dominant spike, indicating that rich dynamics within the sequence are largely collapsed. In contrast, Fig. 6(b)–(d) show that Chain-of-Thought aggregation keeps multiple intermediate motion patterns active over time and assigns larger responses to a few discriminative patterns, suggesting that CoTNet preserves and highlights informative motion evidence instead of over-smoothing it.



(a) Conventional aggregation (b) Chain-of-Thought aggregation (c) Chain-of-Thought aggregation (d) Chain-of-Thought aggregation

Figure 6: Visualizing Chain-of-Thought aggregation.

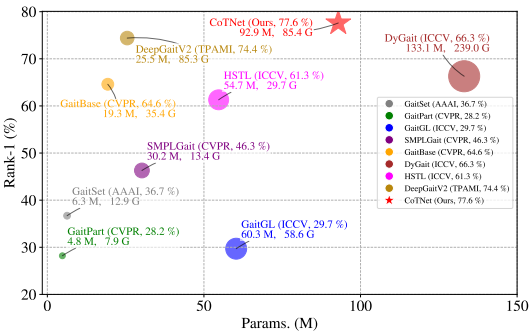


Figure 7: Performance and complexity on Gait3D.

The complexity of CoTNet. Figure 7 plots Rank-1 accuracy against parameter count (marker size encodes FLOPs). Lightweight models such as GaitSet and GaitPart use fewer than 10M parameters but remain below 40% Rank-1, while DyGait improves to 66.3% with 133.1M parameters and 239.0G FLOPs. DeepGaitV2 reaches 74.4% Rank-1 with 25.5M parameters and 85.3G FLOPs, whereas CoTNet attains 77.6% Rank-1 with 92.9M parameters and 85.4G FLOPs, delivering the best accuracy under a similar computational budget and far lower FLOPs than DyGait.

5 Conclusion and Limitations

In this work, we introduced CoTNet, a latent Chain-of-Thought architecture for gait recognition. By inserting a Chain-of-Motion-Thought (CoMT) into a standard silhouette-based pipeline, CoTNet briefly reasons over motion patterns in hidden space before aggregation and yields stronger temporal descriptors. Experiments on six controlled and in-the-wild benchmarks show consistent gains over baselines, suggesting that latent CoT can improve accuracy and motion-pattern interpretability.

Limitations. CoTNet currently relies on binary silhouettes and does not yet exploit richer modalities such as skeletons. In addition, although CoTNet consistently outperforms baselines on six benchmarks, we have not systematically verified whether the latent CoT design also brings stronger cross-dataset or cross-environment generalization, which is left for future study.

AI USE STATEMENT

Generative AI tools were used solely to polish the language and improve the readability of the manuscript. They were not used to develop the method, design or conduct experiments, generate or process experimental data, or interpret the results. The authors take full responsibility for the scientific content and the final text.

ETHICS STATEMENT

This study uses established public gait recognition benchmarks and does not involve new participant recruitment or data collection. Our experiments operate on the provided silhouette sequences and benchmark identity labels, without linking these labels to names or other direct personal identifiers. Dataset use should respect the consent arrangements, access conditions, and usage restrictions established by the original providers. Since gait is a biometric signal, the absence of direct identifiers does not eliminate privacy risks. We advocate privacy safeguards and appropriate authorization for any application of gait recognition, and discourage its use for non-consensual tracking.

REPRODUCIBILITY STATEMENT

We evaluate CoTNet on six public benchmarks: CASIA-B, CCPG, SUSTech1K, CCGR-MINI, Gait3D, and GREW, using their established data splits and evaluation protocols. Section 4.1 and Table 1 specify the datasets, train/test partitions, and optimization schedules. Sections 3.2 and 3.3 describe the proposed modules, training objective, and inference procedure; Section 4.2 provides input preprocessing, network variants, and training hyperparameters. The complete implementation and pretrained models will be released for independent reproduction and further study.

REFERENCES

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8126–8133, 2019.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*, 2025.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1173–1203, 2024.

Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17949–17957, 2023.

Wim De Neys and Matthieu Raoelison. Humans and llms rate deliberation as superior to intuition on complex reasoning tasks. *Communications Psychology*, 3(1):141, 2025.

Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, and Xi Li. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5578–5588, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14225–14233, 2020.

Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9707–9716, 2023.

Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1662–1669, 2024.

Chao Fan, Saihui Hou, Junhao Liang, Chuanfu Shen, Jingzhe Ma, Dongyang Jin, Yongzhen Huang, and Shiqi Yu. Opengait: A comprehensive benchmark study for gait recognition towards better practicality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19595–19604, 2023.

Yang Fu, Saihui Hou, Shibei Meng, Xuecai Hu, Chunshui Cao, Xu Liu, and Yongzhen Huang. Cut out the middleman: Revisiting pose-based gait recognition. In *European Conference on Computer Vision*, pp. 112–128. Springer, 2024.

Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. Camera-lidar cross-modality gait recognition. In *European Conference on Computer Vision*, pp. 439–455. Springer, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024a.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pp. 382–398. Springer, 2020.

Saihui Hou, Zengbin Wang, Man Zhang, Chunshui Cao, Xu Liu, and Yongzhen Huang. Edge-oriented adversarial attack for deep gait recognition. *International Journal of Computer Vision*, 133(4): 1549–1563, 2025.

Panjian Huang, Yunjie Peng, Saihui Hou, Chunshui Cao, Xu Liu, Zhiqiang He, and Yongzhen Huang. Occluded gait recognition with mixture of experts: an action detection perspective. In *European Conference on Computer Vision*, pp. 380–397. Springer, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Yao Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2023.

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 *European conference on computer vision*, pp. 375–390. Springer, 2022.

Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14648–14656, 2021.

Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22076–22085, 2023.

Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 593–603, 2024.

Guozhen Peng, Yunhong Wang, Yuwei Zhao, Shaoxiong Zhang, and Annan Li. Glgait: A global-local temporal receptive field network for gait recognition in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 826–835, 2024a.

Yunjie Peng, Kang Ma, Yang Zhang, and Zhiqiang He. Learning rich features for gait recognition by integrating skeletons and silhouettes. *Multimedia Tools and Applications*, 83(3):7273–7294, 2024b.

Du Phan, Matthew D. Hoffman, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A. Saurous. Training chain-of-thought via latent-variable inference. *arXiv preprint arXiv:2312.02179*, 2023.

Ekkasit Pinyoanuntapong, Ayman Ali, Pu Wang, Minwoo Lee, and Chen Chen. Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

K Richard Ridderinkhof, Markus Ullsperger, Eveline A Crone, and Sander Nieuwenhuis. The role of the medial frontal cortex in cognitive control. *science*, 306(5695):443–447, 2004.

Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1054–1063, 2023.

Chunfeng Song, Yongzhen Huang, Weining Wang, and Liang Wang. Casia-e: a large comprehensive dataset for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):2801–2815, 2022.

Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, pp. 2314–2318. IEEE, 2021.

Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1569–1577, 2022.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12:1, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, and Liang Wang. Causal intervention for sparse-view gait recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 77–85, 2023a.

Jilong Wang, Saihui Hou, Yan Huang, Chunshui Cao, Xu Liu, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Free lunch for gait recognition: A novel relation descriptor. In *European Conference on Computer Vision*, pp. 39–56. Springer, 2024a.

Jilong Wang, Saihui Hou, Xianda Guo, Yan Huang, Yongzhen Huang, Tianzhu Zhang, and Liang Wang. Gaitc 3 i: Robust cross-covariate gait recognition via causal intervention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19582–19592. IEEE, 2023b.

Lei Wang, Yinchi Ma, Peng Luan, Wei Yao, Congcong Li, and Bo Liu. Hih: A multi-modal hierarchy in hierarchy network for unconstrained gait recognition. *arXiv preprint arXiv:2311.11210*, 2023c.

Ming Wang, Xianda Guo, Beibei Lin, Tian Yang, Zheng Zhu, Lincheng Li, Shunli Zhang, and Xin Yu. Dygait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13424–13433, 2023d.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. System-1.5 reasoning: Traversal in language and latent spaces with dynamic shortcuts. *arXiv preprint arXiv:2505.18962*, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain-of-thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gaitparsing: Human semantic parsing for gait recognition. *IEEE Transactions on Multimedia*, 2023e.

Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, and Shibiao Xu. Landmarkgait: intrinsic human parsing for gait recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2305–2314, 2023f.

Zengbin Wang, Saihui Hou, Man Zhang, Xu Liu, Chunshui Cao, Yongzhen Huang, Peipei Li, and Shibiao Xu. Qagait: Revisit gait recognition from a quality perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5785–5793, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Haijun Xiong, Bin Feng, Xinggang Wang, and Wenyu Liu. Causality-inspired discriminative feature learning in triple domains for gait recognition. In *European Conference on Computer Vision*, pp. 251–270. Springer, 2024.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pp. 441–444. IEEE, 2006.

Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023.

Jinkai Zheng, Xinchen Liu, Xiaoyan Gu, Yaoqi Sun, Chuang Gan, Jiyong Zhang, Wu Liu, and Chenggang Yan. Gait recognition in the wild with multi-hop temporal switch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6136–6145, 2022a.

Jinkai Zheng, Xinchen Liu, Xiaoyan Gu, Yaoqi Sun, Chuang Gan, Jiyong Zhang, Wu Liu, and Chenggang Yan. Gait recognition in the wild with multi-hop temporal switch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6136–6145, 2022b.

Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is all you need for accurate gait recognition in the wild. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 116–124, 2023.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14789–14799, 2021.

Shinan Zou, Chao Fan, Jianbo Xiong, Chuanfu Shen, Shiqi Yu, and Jin Tang. Cross-covariate gait recognition: A benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7855–7863, 2024.