

Dual the Reasoning, Double the Insight with TambI: A Self-Supervised Framework for Skeleton Action Representation

Mahmoud Ali¹[0009-0002-7658-0446], Snehashis Majhi¹[0000-0002-9101-017X], Di Yang¹[0000-0002-8124-532X], Quan Kong³[0000-0002-4511-4031], Gianpiero Francesca²[0000-0001-6066-9345], and Francois Bremond¹[0000-0003-2988-2142]

¹ Inria Center at Université Côte d’Azur, Sophia-Antipolis, France

² Toyota Motor Europe, Brussels, Belgium

³ Woven by Toyota , Tokyo, Japan

Abstract. Self-supervised learning has shown great promise for skeleton-based action recognition, especially with contrastive methods. However, existing approaches rely on single-stream motion encoders. This may fail to fully capture both spatial and temporal details, which are critical for real-world generalization. To address this, we propose TambI, a novel self-supervised framework for skeleton action representation learning. We introduce a novel Dual reasoning module to learn complementary skeleton motion representations. Subsequently, we further design a dual objective learning for an indirect contrastive strategy: (1) an Instance Consistency Loss that aligns representations across models and preserves motion details, and (2) a refined contrastive loss using multiple positive samples to enhance feature discrimination. Extensive experiments on six benchmark datasets, including laboratory (NTU-RGB+D, PKU-MMD) and real-world (Toyota SmartHome, Penn Action, Posetics) settings, demonstrate state-of-the-art performance and superior generalizability.

Keywords: Skeleton-based action recognition · Self-Supervised learning · Contrastive learning · Representation learning

1 Introduction

Human Action Recognition (HAR) [3, 8, 32, 37, 44] plays a critical role in diverse applications such as healthcare monitoring, daily activity analysis, and sports understanding. Skeleton-based HAR has emerged as a cornerstone in this field due to its ability to capture structured human motion while remaining robust to background noise, lighting variations, and appearance changes [3, 8, 26, 37]. Recent advancements have focused on contrastive-based self-supervised learning (SSL) [11, 22, 28, 30, 35, 38], which offers superior spatiotemporal representation capabilities and real-world transferability compared to fully supervised methods.

Despite these advancements, a key challenge remains: current state-of-the-art skeleton-based SSL models [15, 22, 30, 35, 38] struggle to generalize across variant scenarios. This limitation is evident when deploying these models across

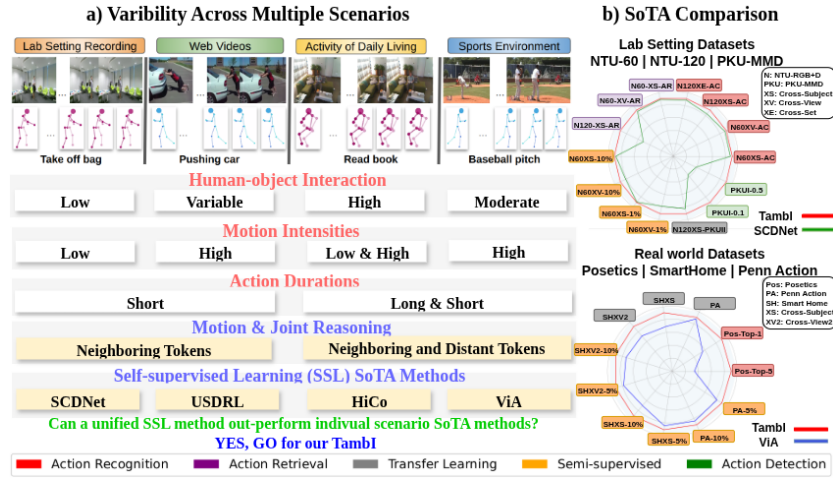


Fig. 1: (a) **Key challenges** across individual skeleton-based HAR scenarios that necessitate tailored motion and joint reasoning strategies to surpass individual SoTA. (b) **TambI emerges as a unified** self-supervised model, demonstrating exceptional generalization to lab settings as well as real-world scenarios and consistently outperforming SoTA benchmarks across multiple evaluation metrics.

datasets that differ significantly in action duration, motion complexity, camera perspectives, and background noise. As shown in Figure 1(a), scenarios such as lab-recorded activities, uncurated web videos, and real-world daily interactions each demands distinct modeling capabilities. The generalization ability of current methods remains far from satisfactory. This raises an important question: *Can we build a unified skeleton-based SSL framework that performs well across all scenarios?*

To answer this question, we analyze recent state-of-the-art methods [22, 30, 35, 38] and identify that the scientific problems underlying this challenge are twofold. Firstly, most existing approaches rely on a single architecture (*e.g.*, Transformers), which offers only one view of the spatiotemporal structure. While effective for modeling fine-grained and short-term motion patterns, such a homogeneous reasoning paradigm lacks the flexibility needed to handle long-term temporal dynamics and distribution shifts, which are common in real-world actions. Secondly, while combining different reasoning architectures (*e.g.*, Transformers and state-space models like Mamba) seems promising, directly merging their outputs in contrastive learning introduces representation misalignment and noisy gradients, as these models process temporal dynamics differently. To tackle these problems, we conduct a comprehensive study on the complementary strengths of Transformer and Mamba-based reasoning. Based on our analysis, we propose TambI, a novel self-supervised framework for learning generalized skeleton-based action representation.

To overcome the bottleneck of single-path reasoning, TambI firstly integrates a novel *dual-path reasoning module*, including a Transformer-based path that captures fine-grained, local motion details, and a Mamba-based path that models

global, long-term temporal dependencies. By jointly leveraging the complementary inductive biases of explicit attention-based reasoning and selective state-space modeling, the two paths are fused into a compact and unified sequence representation. This design enables TambI to precisely discriminate actions across diverse temporal scales within classical contrastive learning framework.

However, due to the divergent reasoning abilities of transformer and mamba, asynchronously training them in contrastive space may induce conflicting and noisy associations. Thereby, to align their representation spaces to ensure stable and effective contrastive learning, TambI secondly proposes a *dual objective indirect contrastive enhancement strategy* to enhance contrastive association via instance consistency contrastive learning. As the first learning objective, we (1) align the representations from these two reasoning paths by introducing the Instance Consistency Loss (ICL). ICL encourages consistency across dual reasoning modules and reduces the conflicting and noisy cues in the contrastive space while enriching the complementary information. This improves the quality of refined features and indirectly boosts both positive and negative samples used in contrastive learning. For the second learning objective, we (2) adopt a Mixing Contrastive Learning (MiCo) strategy that fuses spatial, temporal, and general features from the two encoders into mixed positive pairs. This indirect contrastive setup enables the model to effectively distinguish between similar and dissimilar actions.

We validate TambI on six benchmark datasets, spanning controlled settings (*e.g.*, NTU-RGB+D 60 [25], 120 [18], PKU-MMD [6]) and real-world environments (*e.g.*, Toyota Smarthome [7], Penn Action [43], and Posetics [37]), as shown in Figure 1(b). TambI consistently outperforms state-of-the-art methods [35, 38], demonstrating exceptional generalizability, robustness, few-shot learning capability, and cross-domain transferability across diverse evaluation protocols and challenging scenarios.

In summary, our contributions are: **(i)** Introduced a novel self-supervised method TambI, that showcases superior generalizability and transferability to various scenarios of skeleton-based action recognition. **(ii)** A novel formulation to jointly capture short- and long-term motion dependencies in skeleton sequences via a dual path reasoning module. **(iii)** A new indirect contrastive enhancement using dual objective learning that first aligns the cross-reasoning cues to enrich positive samples and then reinforces discrimination in the contrastive space.

2 Related Work

Skeleton-Based Action Recognition: Skeleton-based action recognition has emerged as a powerful paradigm for modeling structured human motion, offering robustness to appearance, lighting, and background variations. Early methods relied on handcrafted features, later replaced by RNNs [19, 20] and CNN-based pseudo-image encodings [13]. A significant leap came with the introduction of ST-GCN [36] and its successors [3, 8, 10, 37, 44], which utilized graph convolutions tailored to the skeleton’s topology. Despite their success, existing methods face a key limitation: absence of pretraining hinders generalization to real-world scenarios with diverse viewpoints and compositional actions.

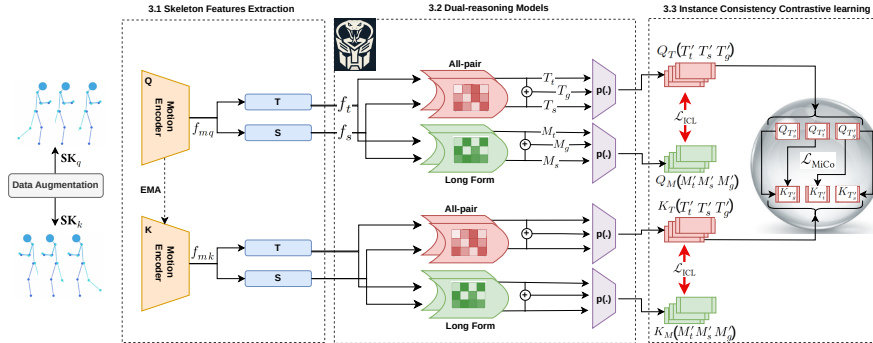


Fig. 2: **Overview of TambI framework.** Given a skeleton sequence SK , we generate augmented views SK_q and SK_k . Motion features f_m are extracted from SK_q and decoupled into temporal (f_t) and spatial (f_s) components. These are refined using Transformer and Mamba to produce T_t, T_s, M_t , and M_s , which are fused into global representations T_g and M_g . All features are projected into a lower-dimensional space and aligned across models using Instance Consistency Loss (ICL). Finally, Mixing Contrastive Loss (MiCo) is applied to Transformer features (Q'_T, Q'_T, Q'_T vs. K'_T, K'_T, K'_T). Note that the Mamba branch is removed at inference time to reduce the computation overhead.

Self-Supervised Skeleton Representation: To bridge this gap, self-supervised learning (SSL) for skeletons typically adopts a two-stage process: pretraining with pretext tasks, followed by supervised fine-tuning. Existing SSL approaches fall into two categories: *generative* [21, 45], which focus on reconstructing input data, and *contrastive* [1, 9, 35], which emphasize pulling positive pairs closer while pushing negatives apart in feature space. Contrastive approaches, often inspired by MoCo v2 [12], leverage momentum encoders and memory banks to enable scalable representation learning. These methods have shown strong performance on curated 3D datasets [33], but their generalization degrades in real-world 2D settings due to noisy joints, occlusions, and view changes [7, 17, 43]. Furthermore, most contrastive methods treat all frames or modalities uniformly, ignoring varying spatiotemporal informativeness.

Spatio-Temporal Modeling: The motion of skeleton joints encodes rich semantic cues for action understanding. Prior work has explored pseudo-labels derived from joint dynamics—such as direction [5], magnitude [14], and motion priors [41]—to guide pretraining. Transformers, with their strong sequence modeling capacity, have been widely adopted; however, their weak inductive bias and data inefficiency pose challenges in low-resource or noisy scenarios. To address this, researchers have explored specialized designs: temporal convolutions [24], graph convolutions [23, 24], and space-time decoupling [27].

In contrast, we propose a unified dual-objective SSL framework combining Mamba-based implicit memory modeling and Transformer-based explicit reasoning. This hybrid setup enables complementary temporal encoding across long and short durations. Our dual-consistency contrastive module enforces intra- and inter-branch alignment, resulting in robust generalization—even on challenging

2D skeleton datasets. The proposed design effectively overcomes rigid inductive biases and modality-agnostic contrastive limitations common in prior work.

3 Proposed Approach

In this section, we introduce **TambI**—a dual-path self-supervised framework, as shown in Fig 2, designed to learn transferable and robust representations for skeleton-based action recognition. TambI is built upon two core ideas: (i) combining complementary temporal reasoning with Transformer and Mamba, and (ii) enforcing consistency and discriminability through a dual-objective contrastive learning strategy. First, we provide theoretical justification for how these mechanisms contribute to improved skeleton-based action representation, and then we proceed to architectural details.

3.1 Theoretical Justification

TambI’s effectiveness is grounded in two principles: **dual-path reasoning** to capture diverse spatial and temporal dynamics, and **dual-objective contrastive learning** to enhance representation alignment and separability.

Complementary Inductive Bias via Dual-Path Reasoning: Let $\mathcal{SK} = \{sk_t\}_{t=1}^T$ denote a skeleton sequence, where $sk_t \in \mathbb{R}^{J \times C}$ represents the coordinates of J joints at time t . The goal is to learn a representation function $f : \mathcal{SK} \rightarrow \mathbb{R}^d$ that captures discriminative spatiotemporal patterns.

Transformers have self-attention to model high-frequency dependencies:

$$\text{Attn}(q, k, v) = \text{softmax}\left(\frac{qk^\top}{\sqrt{d_k}}\right)v, \quad (1)$$

which excels at modeling local, short-term interactions.

Mamba, as a state-space model, resonates long-range dependencies via recurrence-style dynamics:

$$h_t = Ah_{t-1} + Bx_t, \quad y_t = Dh_t, \quad (2)$$

where A , B , and D are learnable parameters. This formulation supports memory-driven reasoning over extended temporal spans.

By constructing representations from both modules, TambI effectively spans the hypothesis space:

$$\mathcal{H}_{\text{TambI}} = \mathcal{H}_{\text{Transformer}} \cup \mathcal{H}_{\text{Mamba}}, \quad (3)$$

allowing it to capture both high-frequency (short-term) and low-frequency (long-term) motion dynamics.

Instance Consistency as Regularization: To encourage coherence between the two reasoning streams, TambI applies an Instance Consistency Loss (ICL). Let $z_T, z_M \in \mathbb{R}^d$ be the Transformer and Mamba embeddings, respectively. The ICL is defined as:

$$\mathcal{L}_{\text{ICL}} = \|Q_T - Q_M\|^2. \quad (4)$$

This acts as a regularizer by reducing divergence between heterogeneous feature spaces. Thereby, it lowers model variance and improves generalization by enforcing agreement under the same input distribution.

3.2 Mixed Contrastive Learning Enhances Separation:

TambI employs a modified InfoNCE loss through Mixing Contrastive Learning (MiCo). For a query Q and a mixed positive key K^+ , the loss is:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k^+ / \tau)}{\sum_{k^-} \exp(q \cdot k^- / \tau)}, \quad (5)$$

where τ is the temperature, k^+ and k^- are the positive and negative keys. In MiCo, all q , k^- , and k^+ are fused representation from Transformer and Mamba features using \mathcal{L}_{ICL} . This strategy enhances intra-class diversities among positives, promoting higher-level invariance and improving representation robustness. It reduces the effect of hard positive mining, resulting in wider margins between classes.

Dual Objective Tightens Generalization Bounds: The total loss in TambI is defined as:

$$\mathcal{L}_{\text{ICCo}} = \mathcal{L}_{\text{ICL}} + \mathcal{L}_{\text{MiCo}}, \quad (6)$$

where $\mathcal{L}_{\text{MiCo}}$ is the sum of InfoNCE losses over spatial, temporal, and global embeddings. This dual-objective optimization can (i) align representations from different inductive biases (Transformer vs. Mamba), (ii) encourage structured discrimination among action classes, and simultaneously (iii) enhance representation consistency and robustness.

Together, all these mechanisms reduce empirical risk and control model complexity, yielding tighter generalization bounds and improved downstream performance.

3.3 Methodological Details

Figure 2 outlines the architecture of **TambI**, which consists of three core components. First, the **Skeleton Feature Extraction** module processes two augmented views of the input sequence to extract spatial and temporal components. These components are independently refined via a **Dual Reasoning** module comprising Transformer and Mamba blocks. Finally, feature alignment and discrimination are optimized through **Instance Consistency Contrastive Learning (ICCo)**, which combines Instance Consistency Loss (ICL) and Mixing Contrastive Loss (MiCo). We describe each component below.

Skeleton Feature Extraction: The input skeleton sequence $\mathcal{SK} \in \mathbb{R}^{t \times J \times C}$, where t is the sequence length, J is the number of joints, and C denotes joint coordinate dimensions (2D or 3D). Two augmented views— \mathcal{SK}_q (query) and \mathcal{SK}_k (key)—are generated using rotation, flipping, and spatiotemporal masking [35] to improve robustness. Both views are encoded using a CTR-GCN backbone [4], producing f_{m_q} and f_{m_k} , respectively. The CTR-GCN consists of spatial and temporal Graph Convolutional Networks (GCNs), modeling intra-frame joint correlations and inter-frame motion dynamics. Inspired by MoCo-v2 [12], the query encoder is updated via backpropagation, while the key encoder is momentum-driven. Both share an identical structure but process different augmentations.

Spatiotemporal Decoupling: Encoder outputs contain tightly coupled spatiotemporal features that are suboptimal for contrastive learning. To disentangle these, we apply MLPs to decouple representations into temporal (f_t) and spatial (f_s) components such that $f_t, f_s = \text{ReLU}(\text{LayerNorm}(\text{Linear}(f_m)))$.

Dual Reasoning Models: Single-stream models struggle to capture both short-term and long-term motion patterns. We introduce a dual-path design using **Transformer** and **Mamba** blocks to enhance feature reasoning.

Transformers attend to local, short-term dependencies using dense self-attention:

$$\mathcal{Q}_T = \text{LN}(x + \text{MHA}(Q, K, V)) + \text{FFN}\left(\text{LN}(x + \text{MHA}(Q, K, V))\right) \quad (7)$$

Mamba models long-term temporal dependencies through state-space operations:

$$\mathcal{Q}_M = \sigma(W_g x) \odot \text{SSM}(\text{Conv1D}(x)) \quad (8)$$

We apply both modules to f_t and f_s , yielding four refined features: T_t, T_s (Transformer) and M_t, M_s (Mamba). Global representations are formed via:

$$T_g = \text{Concat}(T_t, T_s), \quad M_g = \text{Concat}(M_t, M_s)$$

All six features are projected into a contrastive space using a shared projection head $p(\cdot)$:

$$Q_{T'_t}, Q_{T'_s}, Q_{T'_g}, Q_{M'_t}, Q_{M'_s}, Q_{M'_g}$$

The same is applied to the key branch, yielding $K_{T'_t}, \dots, K_{M'_g}$.

Instance Consistency Contrastive Learning: To unify the dual reasoning streams, we introduce a **dual objective** optimization strategy: Instance Consistency Loss (ICL) for feature alignment and MiCo loss for feature discrimination.

Objective 1: *Instance Consistency Loss (ICL)* in equation 4 minimizes redundancy and noise by enforcing alignment between Transformer and Mamba features within each representation branch. This promotes consistent and complementary features across encoders and strengthens subsequent contrastive learning.

Objective 2: *Mixing Contrastive Learning (MiCo)* enhances discrimination in the embedding space by mixing spatial, temporal, and global views from different branches. Following [35], the MiCo loss is defined as:

$$\mathcal{L}_{\text{MiCo}} = \mathcal{L}_{\text{info}}(Q_s, K_g) + \mathcal{L}_{\text{info}}(Q_t, K_g) + \mathcal{L}_{\text{info}}(Q_g, K_s) + \mathcal{L}_{\text{info}}(Q_g, K_t) \quad (9)$$

where Q and K represent query and key features, and $\mathcal{L}_{\text{info}}$ is the InfoNCE loss outlined in equation 5. The InfoNCE loss also takes a positive and negative pair K^+ and K^- from the memory bank.

Final Loss: The total self-supervised loss combines both objectives as in equation 6. This dual strategy ensures that representations are internally consistent and externally discriminative key for robust generalization in downstream tasks.

4 Experiments and Analysis

We conduct extensive experiments to evaluate the effectiveness of TambI under different evaluation protocols. Firstly, we compare TambI with state-of-the-art (SoTA) self-supervised models through various evaluation protocols, including linear evaluation (linear probing), k-NN, semi-supervised learning, cross-dataset transfer learning, action detection and action prediction, on 2D real-world datasets (*i.e.*, **Posetics** [37], **Toyota SmartHome** [7] (SmartHome) and **Penn Action** [43]. and also on 3D datasets (*i.e.*, **NTU-RGB+D 60** [25], **NTU-RGB+D 120** [18], **PKU-MMD** [6]) Subsequently, we provide deeper insights by conducting per-class analysis. Finally, we present a comprehensive ablation study to highlight the impact of each component within the TambI framework. Detailed information about the datasets, along with further studies and experiments, can be found in the **Appendix**. The following experiments demonstrate that TambI achieves a favorable trade-off between real-world and lab-controlled settings, highlighting the strong generalizability and transferability of our model across datasets with both limited and large action diversity.

4.1 Evaluation Protocols

For **Linear Evaluation**, a linear classifier is appended to the frozen encoder and is trained in a supervised manner to predict the corresponding label for the input sequences. For **Semi-supervised Evaluation**, following state-of-the-art practices, we pre-train the encoder using unlabeled data and subsequently fine-tune the entire model using randomly sampled subsets (1%, 5%, and 10%) of the labeled training data. For **Transfer Learning Evaluation**, we investigate the transferability of the model after self-supervised training on a source dataset. For laboratory 3D datasets, we re-train the entire model (Fine-tune) using labeled data following the setting of previous SoTA [15, 35] for fair comparisons. While for real-world 2D datasets, we re-train only the classifier (Linear) when adapting to unseen datasets after large-scale pretraining following the consistent setting of previous SoTA [38] for fair comparisons. We also perform **Action Detection Evaluation** on untrimmed video dataset: PKU-MMD Part I. We freeze the encoder, train a linear classifier, and compute mean Average Precision (mAP) across multiple temporal IoU thresholds.

4.2 Comparisons on Real-world Action Representation: 2D Datasets

In this section, to assess the generalization ability of our model, we evaluate its performance on real-world 2D skeleton datasets that feature higher variability in human poses, occlusions, camera viewpoints, and background clutter conditions that often degrade the performance of models trained in controlled lab environments. These settings serve as a strong benchmark for testing the robustness and transferability of learned representations. The following results reflect the potential of TambI as a versatile and transferable backbone for skeleton-based representation learning.

Linear Evaluation: In Table 1 (*blue*), we report Top-1 and Top-5 accuracy for Posetics dataset. **TambI** achieves a new state-of-the-art performance among

Methods	Linear Evaluation		Transfer Learning Evaluation (After Pre-trained on Posetics)			Semi-supervised Evaluation			
	Posetics		SmartHome		PennAction	SmartHome CS (%)		PennAction	
	Top-1 (%)	Top-5 (%)	CS (%)	CV2 (%)	Top-1 (%)	(5%)	(10%)	(5%)	(10%)
OR-VPE [39] (FG 21)	14.6	31.2	42.7	32.4	78.5	-	-	-	-
3s-CrosSCLR [16](CVPR 21)	18.8	38.1	-	-	-	-	-	-	-
AimCLR [30](AAAI 22)	19.2	39.3	46.6	48.3	-	-	-	-	-
CMD [22](ECCV 22)	20.4	40.5	49.0	52.5	89.4	-	-	-	-
HiCLR [2](AAAI 23)	20.1	39.9	49.1	52.3	88.7	-	-	-	-
HiCo [9](AAAI 23)	21.3	42.1	54.3	54.8	87.6	34.5	46.0	57.2	74.5
PCM ³ [42](ACMMM 23)	20.0	40.3	45.3	46.8	85.6	23.1	30.1	46.3	60.9
Via [38] (IJCV 24)	20.7	40.1	49.5	52.6	90.2	38.6	45.3	65.8	85.2
USDRL [34](AAAI 25)	25.9	48.6	54.3	53.9	89.6	37.1	43.6	58.1	70.3
TambI (Ours)	28.0	51.0	54.8	57.3	91.0	39.6	48.5	69.2	87.4
	(+2.1%)	(+2.4%)	(+0.5%)	(+2.5%)	(+0.8%)	(+1.0%)	(+2.5%)	(+3.4%)	(+2.2%)

Table 1: **Action Recognition** – Combined results for *Linear Evaluation*, *Transfer Learning*, and *Semi-supervised Evaluation* on Real-world datasets (Posetics, SmartHome, and PennAction).

Method	NTU-RGB+D 60		NTU-RGB+D 120	
	CS(%)	CV(%)	CS(%)	CSet(%)
HiCLR [2](AAAI23)	80.4	85.5	70.0	70.4
UmURL [29](ACMMM 23)	82.3	89.8	73.5	74.3
PCM ³ [42](ACMMM 23)	83.9	90.4	76.5	77.5
MAMP [21](ICCV 23)	84.9	89.1	78.6	79.1
HiCo [9](AAAI 23)	81.1	88.6	72.8	74.1
Via [38](IJCV 24)	78.1	85.8	69.2	66.9
SCD-Net [35] (AAAI 24)	86.6	91.7	76.9	80.1
USDRL [34] (AAAI 25)	84.2	90.8	76.0	76.9
Heterogeneous [31](CVPR 25)	80.2	88.0	70.7	73.5
TambI (Our)	87.1	92.4	79.0	80.6

Table 2: **Action Recognition-Linear Evaluation results-** for Lab-setting datasets (NTU-RGB+D 60 and NTU-RGB+D 120)

Method	Transfer to PKU-MMD II	
	NTU-RGB+D 120	CS(%)
HiCo-Transformer [9]	55.4	
CrosSCLR-B [16]	52.8	
CMD [22]	57.0	
UmURL-3 [29]	58.5	
SCD-Net [35]	64.0*	
Heterogeneous [31]	63.1	
TambI (Our)	65.1	

Table 3: **Action Recognition-Transfer Learning Evaluation (Fine-tune) results-** for Lab-setting datasets (from NTU-RGB+D 120 to PKU-MMD II), * means our implementations.

all compared methods and outperforms USDRL [34], by **+2.1%** and **+2.4%**, respectively. The consistent gains over contrastive-based skeleton representation learning methods [2, 38, 42] indicate that TambI learns more discriminative and linearly separable skeletal motion features without supervision.

Transfer Learning Evaluation (Linear): We further assess representation generalization in Table 1 (*green*) by pre-training on Posetics and linearly transferring to downstream benchmarks, including SmartHome and PennAction. **TambI** consistently outperforms prior methods, achieving gains of up to **+2.5%** on SmartHome-CV2 and **+0.8%** on PennAction. These results demonstrate that TambI learns transferable skeleton representations that generalize well across datasets and action distributions.

Semi-supervised Evaluation: Evaluating model generalizability by fine-tuning the pre-trained model with a small number of samples is crucial, particularly on real-world datasets where annotations are challenging. In Table 1 (*yellow*), we report such results using only 5% and 10% of the labeled samples from the SmartHome and Penn Action datasets. TambI generalizes well across diverse datasets and achieves SoTA performance even under limited data settings, demonstrating the robustness of our design. This strong performance is driven by our dual reasoning mechanism, which, together with the two objective losses, enables the model to effectively capture rich spatial-temporal patterns and distinguish fine-grained action classes from skeleton sequences.

Method	CS(%)			CV(%)		
	1%	5%	10%	1%	5%	10%
HiCLR [2]	39.1	63.3	70.7	42.9	68.3	74.8
HiCo [9]	54.4	54.4	73.0	54.8	54.8	78.3
PCM ³ [42]	53.8	-	77.1	53.1	-	82.8
UmURL [29]	58.1	72.5	-	58.3	76.8	-
SCD-Net [35]	<u>69.1</u>	-	<u>82.2</u>	<u>66.8</u>	-	<u>85.8</u>
USDRL (STTR) [34]	55.0	-	76.1	59.1	-	82.0
USDRL (DSTE) [34]	57.3	-	80.2	60.7	-	84.0
Heterogeneous [31]	55.0	<u>76.3</u>	-	55.0	<u>79.1</u>	-
TambI (Ours)	69.2	79.9	83.0	68.7	83.7	87.3

Table 4: **Action Recognition- Semi-supervised Evaluation results-** for Lab-setting datasets (NTU-RGB+D 60)

Method	mAP@tIoU (%)		
	0.1	0.3	0.5
HiCo-Transf. [9]	32.5	31.8	28.6
HiCo-LSTM [9]	39.2	37.2	32.0
HiCo-GRU [9]	50.1	48.6	44.3
LAC [40]	55.2	-	-
SCDNet [35]	<u>65.6</u>	<u>64.7</u>	<u>58.5</u>
TambI (Ours)	72.6	71.7	64.6

Table 5: **Action Detection- Linear Evaluation results-** for Lab-setting datasets (PKU-MMDI)

4.3 Comparisons on Lab-setting Action Representation: 3D Datasets

In this section, we further perform large-scale comparisons with SoTA approaches on variant 3D and lab setting datasets to demonstrate the effectiveness of the proposed TambI. The following results in all settings confirm that our TambI not only captures richer spatio-temporal dependencies but also scales effectively across modalities and dataset protocols.

Linear Evaluation: As skeleton sequence can be represented in different modalities: joint-based (J), focusing on individual joint positions to capture static postures; bone-based (B), encoding the relative positions between connected joints to highlight body part interactions; and motion-based (M), capturing the changes in joint or bone positions over time to emphasize dynamic movement patterns. These modalities offer complementary insights into both static and dynamic aspects of human movement. In Table 2, we report the linear evaluation results compared to SoTA on lab-setting datasets (NTU-RGB+D 60 and NTU-RGB+D 120) using only joint-modality. The multi-modality (J+B+M) results are reported in the **Appendix**. TambI shows a significant improvement compared to SoTA like [35] across all benchmark datasets with different types of modality, particularly in the NTU-RGB+D 120 Cross-Subject protocol, where it improves by +2.1% using only the joint modality.

Transfer Learning Evaluation (Fine-tune): Here, we investigate the transferability of representations learned from the source domain during the pretraining stage to an unseen target dataset. In Table 3, we present our results in comparison to SoTA. We used NTU-RGB+D 120 as the source and PKU-MMD II, a smaller dataset, as the target. TambI outperforms the SoTA results by transferring from NTU-RGB+D 120 to PKU-MMD II, demonstrating the transferability of the learned representations from our model.

Semi-supervised Evaluation: Table 4 shows that TambI achieves SoTA results on NTU-RGB+D 60 under both Cross-Subject (CS) and Cross-View (CV) protocols with 1%, 5%, and 10% labeled data. It consistently outperforms existing methods, such as SCD-Net [35] and Heterogeneous [31], under both the CS and CV protocols. Notably, TambI demonstrates strong performance with minimal supervision in CS with just 1% labeled samples. As the amount of labeled

Activity	Gain from TambI
<i>Long-Duration Activity</i>	
Use tablet	+33.3
Eat at table	+28.0
Read book	+23.9
<i>Short-Duration Activity</i>	
Sit down	+19.1
Walk	+11.6

Table 6: Gain from TambI (Transformer + Mamba) w.r.t. Transformer on SmartHome (Mean per Class).

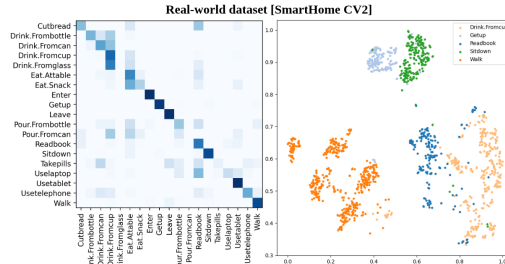


Table 7: Visualization of confusion matrix and TSNE plots showing TambI’s representation learning ability on a real-world dataset.

data increases, TambI scales effectively, reaching 87.3% at 10%, outperforming SCDNet [35] by +1.5%. Other competitors [15] HiCLR [2] show either lower accuracy or inconsistent trends. In contrast, TambI maintains stable gains under the more challenging CV setting, which highlights the robustness and strong few-shot generalization of TambI.

Transfer Learning for Action Detection: As activities in the real world are typically collected in untrimmed and temporally continuous videos, it is essential to evaluate the model in the context of action detection. In Table 5, we report action detection results on PKU-MMD I dataset under a linear evaluation protocol, using mean Average Precision (mAP) at different temporal Intersection over Union (tIoU) thresholds: 0.1, 0.3, and 0.5. For fair comparison with SoTA methods, we use an encoder pretrained on NTU-RGB+D 60 (cross-subject) and attach a linear classifier to predict frame-level action categories, thereby generating the final proposals. We observe that TambI significantly outperforms state-of-the-art self-supervised learning (SSL) methods [9, 35, 40], thanks to its dual reasoning mechanism that effectively captures long-range dependencies and models action transition states.

4.4 Deeper Insights, Analysis and Comparison

To bring additional insights to the significant performance gain of TambI in real-world conditions, we showcase the performance improvements of TambI when applied to various activities compared to a Transformer-only approach in Table 6. It can be observed that TambI demonstrates significant gains, particularly for activities like “Use tablet”, “Eat at table”, and “Read book”. These tasks often involve complex spatial and temporal dynamics, which TambI with dual reasoning models, effectively captures through its diverse reasoning models and improved representation quality. Further, TambI still performs well, with notable improvements for “Sit down” and “Walk”. These short, transitional actions benefit from TambI’s ability to capture fine-grained motion details. Further, Figure 7 shows a discriminative representation space between prominent long and short activities in a real-world dataset. Additional benefits of our method over SoTA w.r.t. long and short duration activities are provided in the **Appendix**.

What makes TambI better than Single reasoning [35]? While **Transformers** are theoretically capable of modeling both short- and long-term skeleton dynamics,

Method	Decoupling	Reasoning	Loss	NTU60	CV	NTU120	CS
Baseline	×	×	NCE	68.2		56.3	
HiCo [9]	✓	×	NCE	88.6		72.8	
SCD-Net [35]	✓	✓ (Single)	NCE	91.7		76.9	
TambI (Ours)	✓	×	NCE	88.9		76.2	
TambI (Ours)	×	✓ (Dual)	ICCo	86.3		73.8	
TambI (Ours)	✓	✓ (Dual)	NCE	90.2		77.1	
TambI (Ours)	✓	✓ (Dual)	ICCo	92.4		79.0	

Table 8: Performance Comparison of Different Methods

Temporal	Spatial	ICL	MiCo	NTU60	CV
Single Reasoning					
Trans	Trans	×	✓		91.7
Mamba	Mamba	×	✓		83.9
Trans	Mamba	×	✓		70.1
Mamba	Trans	×	✓		85.2
Dual Reasoning (Branch-specific)					
Trans-Trans	Mamba-Mamba	✓	✓		72.4
Mamba-Mamba	Trans-Trans	✓	✓		86.9
Dual Reasoning (Ours)					
Trans-Mamba	Trans-Mamba	✓	✓		92.4

Table 9: Ablation Study on Dual Reasoning Design

their dense all-pair token interaction mechanism often introduces redundant relational noise when capturing extended temporal dependencies. This redundancy can obscure critical motion cues and lead to performance degradation, particularly in long horizon skeleton sequences. In contrast, **Mamba** employs a selective state transition strategy that avoids exhaustive token comparisons. Instead, it focuses on propagating salient temporal signals, enabling it to capture long-range dynamics efficiently and precisely. By combining the local precision of Transformers with the global efficiency of Mamba, our **TambI** model achieves a **balanced and complementary temporal representation**. This fusion enhances action recognition across diverse motion scales and improves overall system performance.

In summary, we observe that the criticality between lab settings and complex real-world environments underscores the importance of robust feature learning. TambI consistently improves performance in diverse scenarios, and validates its potential as a universal framework for skeleton-based action recognition.

4.5 Ablation Study

Component-wise Contribution Analysis: Table 8 presents an ablation study isolating the impact of our three core contributions: spatial-temporal decoupling, dual reasoning design, and the ICCo loss function. Starting from a baseline without decoupling or reasoning, we observe significant performance gains when introducing decoupling and single reasoning [9, 35]. Incorporating our dual reasoning design without decoupling yields modest improvements, while combining decoupling with dual reasoning provides a larger boost. Finally, integrating the ICCo loss delivers the highest performance, demonstrating that each of these components decoupling, flexible dual reasoning, and ICCo optimization contributes cumulatively to the overall effectiveness of our model.

Dual Reasoning Design—An Ablation Perspective: In Table 9, we present an ablation study analyzing various configurations of our reasoning modules. Specifically, we compared (i) single reasoning models, where only a Transformer or Mamba is used in isolation, (ii) branch-specific dual designs, where the reasoning models are explicitly assigned to spatial and temporal branches respectively, and (iii) parallel dual reasoning design, where both modules operate concurrently on the same input representation. Our findings indicate that the parallel reasoning design consistently outperforms designs with fixed spatial/temporal assignments. This suggests that enabling the model to discover flexible reasoning pathways,

Reasoning Models	NTU-RGB+D 60	
	CS(%)	CV(%)
Transformer	86.6	91.7
Transformer+ BiGRU	86.4	89.7
Transformer+ LSTM	86.6	89.8
Transformer+ Mamba	87.1	92.4

Table 10: Impact of different reasoning models

rather than imposing a hard separation between spatial and temporal reasoning, leads to superior performance. We attribute these gains to the model’s ability to leverage both short- and long-term dependencies across modalities in a more adaptive manner.

Which Reasoning Model is Complementary to Transformer? The results in Table 10 show that adding BiGRU or LSTM to the Transformer slightly degrades the performance compared to using the Transformer alone. These models limit the Transformer’s global reasoning by forcing the model to process sequences in a constrained, step-by-step manner, leading to suboptimal feature fusion and reduced overall effectiveness. In contrast, Mamba brings complementary reasoning abilities that the Transformer lacks, particularly in modeling long-term dependencies and structured reasoning. The combination of the Transformer’s global attention and Mamba’s efficient state-space modeling leads to richer feature representations, enhancing the overall reasoning effectiveness and achieving better accuracy.

Impact of Instance Consistency Loss (ICL) on Mico loss: In Table 11, we study the effect of different ICL on the contrastive representation. Training a single reasoning model (Transformer (A0) or Mamba (A1)) with MiCo contrastive loss achieves reasonable performance as a baseline, but combining both models in contrastive space without interaction (A2) leads to a degradation. Introducing ICL between models (A3–A6) significantly enhances performance. These results demonstrate that feature alignment across diverse reasoning models plays a pivotal role in enriching positive and negative sample pairings, thereby indirectly boosting contrastive learning effectiveness. Among ICL losses, using MSE with Transformer features (A6) in MiCo yields the best results, achieving state-of-the-art accuracy.

5 Conclusion

In this paper, we introduced TambI, a unified framework for self-supervised skeleton-based action representation learning that combines dual spatial-temporal reasoning with a dual objective contrastive learning strategy. Our approach effectively addresses challenges of generalization and limited supervision in real-world scenarios. Extensive experiments across lab and challenging real-world datasets demonstrate that TambI consistently outperforms state-of-the-art methods under various settings. The robust performance with minimal labeled data and strong generalization across short and long-term actions highlights that TambI represents a generic and effective methodology for real-world deployment in human activity understanding tasks.

Index	MiCo_loss (Contrastive)	ICL (Trans - Mamba)	NTU-RGB+D 60	
			CS(%)	CV(%)
A0:	Transformer	-	86.6	91.7
A1:	Mamba	-	83.0	83.9
A2:	Both	-	82.3	81.0
A3:	Mamba	KL	85.0	88.9
A4:	Mamba	MSE	84.8	90.1
A5:	Transformer	KL	86.3	91.2
A6:	Transformer	MSE	87.1	92.4

Table 11: Impact of ICL on MiCo on NTU-RGB+D 60.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
2. Chen, Y., Zhao, L., Yuan, J., Tian, Y., Xia, Z., Geng, S., Han, L., Metaxas, D.N.: Hierarchically self-supervised transformer for human skeleton representation learning. In: ECCV (2022)
3. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: ICCV (2021)
4. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: CVPR (2021)
5. Cheng, Y.B., Chen, X., Chen, J., Wei, P., Zhang, D., Lin, L.: Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In: ICME (2021)
6. Chunhui, L., Yueyu, H., Yanghao, L., Sijie, S., Jiaying, L.: Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv:1703.07475 (2017)
7. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarhome: Real-world activities of daily living. In: ICCV (2019)
8. Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: ECCV (2024)
9. Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., Wang, X.: Hierarchical contrast for unsupervised skeleton-based action representation learning. In: AAAI (2023)
10. Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: CVPR (2022)
11. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: CVPR (2021)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Kim, B., Chang, H.J., Kim, J., Choi, J.Y.: Global-local motion transformer for unsupervised skeleton-based action learning. In: ECCV (2022)
15. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: CVPR (2021)
16. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: CVPR (2021)
17. Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In: CVPR (2021)
18. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3D human activity understanding. TPAMI (2020)
19. Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G.: Skeleton-based action recognition using spatio-temporal lstm network with trust gates. TPAMI (2017)
20. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: ECCV (2016)
21. Mao, Y., Deng, J., Zhou, W., Fang, Y., Ouyang, W., Li, H.: Masked motion predictors are strong 3d action representation learners. In: ICCV (2023)
22. Mao, Y., Zhou, W., Lu, Z., Deng, J., Li, H.: Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In: ECCV (2022)
23. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. CVIU (2021)

24. Qiu, H., Hou, B., Ren, B., Zhang, X.: Spatio-temporal tuples transformer for skeleton-based action recognition. arXiv preprint arXiv:2201.02849 (2022)
25. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3D human activity analysis. CVPR (2016)
26. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)
27. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: ACCV (2020)
28. Sun, C., Nagrani, A., Tian, Y., Schmid, C.: Composable augmentation encoding for video representation learning. In: ICCV (2021)
29. Sun, S., Liu, D., Dong, J., Qu, X., Gao, J., Yang, X., Wang, X., Wang, M.: Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In: ACM MM (2023)
30. Tianyu, G., Hong, L., Zhan, C., Mengyuan, L., Tao, W., Runwei, D.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: AAAI (2022)
31. Wang, H., Ma, X., Kuang, J., Gui, J.: Heterogeneous skeleton-based action representation learning. In: CVPR (2025)
32. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: CVPR (2023)
33. Wang, Z., Liu, W.: Robustness verification for contrastive learning. In: ICML (2022)
34. Weng, W., Wang, H., Wang, J., He, L., Xie, G.: Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In: AAAI (2025)
35. Wu, C., Wu, X.J., Kittler, J., Xu, T., Atito, S., Awais, M., Feng, Z.: Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In: AAAI (2024)
36. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. AAAI (2018)
37. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., Bremond, F.: Unik: A unified framework for real-world skeleton-based action recognition. In: BMVC (2021)
38. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., Bremond, F.: Via: View-invariant skeleton action representation learning via motion retargeting. IJCV (2024)
39. Yang, D., Wang, Y., Dantcheva, A., Garattoni, L., Francesca, G., Br mond, F.: Self-supervised video pose representation learning for occlusion-robust action recognition. In: FG (2021)
40. Yang, D., Wang, Y., Dantcheva, A., Kong, Q., Garattoni, L., Francesca, G., Bremond, F.: Lac - latent action composition for skeleton-based action segmentation. In: ICCV (2023)
41. Yang, Y., Liu, G., Gao, X.: Motion guided attention learning for self-supervised 3d human action recognition. IEEE TCSVT (2022)
42. Zhang, J., Lin, L., Liu, J.: Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In: ACM MM (2023)
43. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV (2013)
44. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: ICCV (2023)
45. Zhu, Y., Min, M.R., Kadav, A., Graf, H.P.: S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In: CVPR (2020)