ONE FOR ALL: GENERALIST FOUNDATION MODEL FOR CROSS-SENSOR SKELETON REPRESENTATION LEARNING

Jeonghyeok Do **Yun Chen** **Munchurl Kim** *
Korea Advanced Institute of Science and Technology (KAIST)
{ehwjdgur0913,cyruby,mkimee}@kaist.ac.kr

Project Page: https://kaist-viclab.github.io/SOfA_site/

ABSTRACT

For learning generalizable motion representations from large-scale unlabeled data, Self-supervised learning (SSL) has become a widely adopted methodology. However, existing approaches are primarily limited by the inherent heterogeneity of skeleton data—characterized by varying joint counts, indexing protocols, and topological structures across different sensors—which typically necessitates training separate, sensor-specific, or even entirely dataset-specific models. To overcome this, we introduce **SOfA** (**S**keleton **O**ne **f**or **A**ll), the *first* generalist foundation model designed to achieve *sensor-unified* skeleton representation learning across diverse sensors. To accommodate the dimensional gap caused by varying joint counts, we introduce a fixed-size set of learnable *Canonical Joint Slots*, acting as a universal vessel that seamlessly accommodates arbitrary skeletal topologies. SOfA fills these slots via an attention mechanism that dynamically aggregates skeletal information from sensor-specific inputs. Furthermore, we resolve joint index misalignment between various sensors by introducing a *Semantic Joint Embedding* derived from a pre-trained text encoder, rather than relying on absolute positional embeddings. To validate our approach, we standardized ten 3D skeleton datasets for unified training. Extensive experiments demonstrate that SOfA can serve as a truly universal encoder, achieving state-of-the-art (SOTA) performance across a wide range of downstream tasks and sensor types, often outperforming dataset-specific models with a single foundation model.

Table 1: **Conceptual comparison of skeleton representation learning paradigms.** Standard SSL methods are constrained by fixed topologies, and recent approach, HSP (Wang et al., 2025a), is limited to intra-scene multi-modal fusion using paired data. In contrast, our SOfA operates as the *first* generalist foundation model, achieving universal generalization across entirely independent datasets and arbitrary sensor configurations.

Property	Standard SSL	HSP (CVPR'25)	SOfA (Ours)
Cross-Sensor Capability	✗	Intra-scene (Paired)	Universal (Unpaired)
Alignment Strategy	✗	Skeletal interpolation	Canonical Joint Slots + Semantic Joint Embedding
Sensor Flexibility	✗	Pre-defined 2D-3D pair	Arbitrary 3D sensor
Dataset Dependency	Homogeneous	Heterogeneous (Paired)	Heterogeneous (Independent)
Representation Scope	Domain-specific	Multi-modal fusion	Generalist foundation

1 INTRODUCTION

Skeleton data provides rich and dense information for human action understanding (Sun et al., 2022; Zhang et al., 2026; Kong & Fu, 2022). Unlike RGB imagery, it reduces reliance on appearance cues, making it less sensitive to complex backgrounds, illumination changes, and partial occlusions. Thus, it has been widely adopted across a variety of vision tasks, including action recognition, action segmentation, and action retrieval. Early advancements in skeleton-based research primarily relied on fully supervised learning (Duan et al., 2022; Chen et al., 2021; Do & Kim, 2024), which yielded remarkable successes. However, this perspective is fundamentally bottlenecked by the prohibitive cost of large-scale, frame-level annotation. Furthermore, supervised models frequently suffer from

*Corresponding author.

severe overfitting to specific domain distributions, resulting in a significant degradation of generalization performance in real-world scenarios.

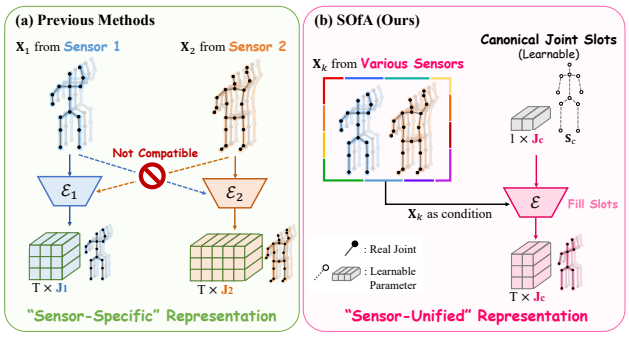

Figure 1: **Motivation of SOfA.** (a) Previous methods require isolated encoders for varying sensor topologies, yielding incompatible *sensor-specific* representation. (b) SOfA (Ours) introduces fixed-size set of learnable Canonical Joint Slots, $\mathbf{S}_c$. By treating arbitrary sensor sequences as conditions, a single shared encoder dynamically fills these slots to extract a standardized, *sensor-unified* representation, regardless of original joint counts.

To overcome the limitations of supervised learning, recent works in skeleton analysis have rapidly shifted toward Self-Supervised Learning (SSL) (Oquab et al., 2023; Chen et al., 2020; He et al., 2022) to build foundation models capable of extracting universal motion representations from abundant unlabeled data. Fig. 1 illustrates a conceptual comparison of skeleton representation learning paradigms. As shown in Fig. 1-(a), current skeleton SSL models fall short of being *true* foundation models. While existing approaches (Mao et al., 2023; Lin et al., 2023; Wang et al., 2025a; Abdelfattah & Alahi, 2024; Sun et al., 2025) are individually trained and optimized for specific dataset structures, they claim foundation-level capabilities solely because a single encoder can be fine-tuned for various downstream tasks, even though the resulting representations remain *strictly sensor- and dataset-specific*. Real-world skeleton sensors (e.g., Kinect (Zhang, 2012), LiDAR, and wearable sensors) exhibit vast diversity, yielding incompatible skeletal configurations and inconsistent anatomical definitions. Because current architectures are tightly coupled to a fixed number of joints and rigid absolute positional embeddings, they suffer from severe fragmentation; they are structurally incapable of processing out-of-distribution sensor data without being entirely redesigned and trained from scratch.

This fragmentation distinctly contrasts with the evolution of foundation models in other domains. In remote sensing (Guo et al., 2024; Wang et al., 2025b), for instance, foundation models successfully integrate heterogeneous modalities and varying resolutions from different satellites into a single, unified geo-spatial representation. The skeleton domain must similarly evolve toward a generalist architecture capable of encompassing diverse sensor configurations. At its core, a human action is defined by its underlying motion rather than the tool used to measure it; a "jump" remains the same movement whether it is recorded by 15 or 25 joints. Table 1 provides a detailed comparison of properties across different skeleton representation learning frameworks. As shown, while a recent attempt, HSP (Wang et al., 2025a), seeks to accommodate diverse skeleton formats, it remains limited to paired intra-scene multi-modal fusion rather than sensor-unified representation learning across independent sensors.

To address these fundamental limitations, we propose **SOfA** (**S**keleton **O**ne **f**or **A**ll), the *first* generalist foundation model that achieves a *sensor-unified representation* across *arbitrary* configurations. To ensure universal adaptability, SOfA introduces a fixed-size set of learnable *Canonical Joint Slots*. Conceptually inspired by the refinement process of diffusion models—where noise is iteratively distilled into a target output guided by conditioning signals—these slots act as a universal template that is dynamically populated by diverse, real-world skeleton sequences. Through an attention mechanism, SOfA distills disparate kinematic signals from various sensors into these standardized slots, ultimately yielding a unified representation that remains consistent regardless of the input's original topology as shown in Fig. 1-(b).

A remaining challenge is the misalignment of joint indices across diverse sensors. Traditional absolute positional embeddings assign rigid numerical indices that are essentially unreasonable for multi-sensor settings, as the same indices often refer to different anatomical joints across datasets, making it difficult for the model to learn a consistent structural representation. To resolve this, we leverage a pre-trained text encoder to introduce *Semantic Joint Embedding*, which utilizes explicit textual names (e.g., "Wrist") from sensor metadata. This semantic-driven positional embedding enables the model to intrinsically capture anatomical correspondences, mitigating the impact of

mismatched joint indexing between sensors. To facilitate cross-sensor research, we unify ten 3D skeleton datasets into a standardized benchmark. Our contributions are summarized as follows:

- We propose **SOfA**, the *sensor-unified foundation model* for skeleton representation learning capable of processing heterogeneous data across varying sensor types, joint counts, and indexing protocols within a single network.
- We introduce a novel paradigm to tackle topological heterogeneity via attention-based *Canonical Joint Slots*, which are dynamically populated using arbitrary sensor inputs as conditioning signals.
- To align joint indices between heterogeneous sensors, we present *Semantic Joint Embedding*, replacing rigid absolute positional embeddings with semantic features derived from textual labels via a pre-trained text encoder.
- Extensive experiments demonstrate that SOfA, functioning as a single universal encoder, achieves SOTA performance across diverse datasets and downstream tasks, yielding results that are highly comparable to or even outperform those of sensor-specific models.

2 Related Works

2.1 SSL for Skeleton Representations

Recent advancements in self-supervised learning (Caron et al., 2021; Oquab et al., 2023; Darcet et al., 2023; Chen et al., 2020; He et al., 2022) for skeleton data have evolved to extract rich motion features through strategies like Contrastive Learning (CL) (Chen et al., 2020) and Masked autoencoder (MAE) (He et al., 2022). CL-based approaches focus on capturing instance-level discrimination, utilizing methods ranging from augmented view agreement (Li et al., 2021b; Guo et al., 2022)) to advanced semantic mining (Lin et al., 2023; Shah et al., 2023). Alternatively, MAE-based methods (Mao et al., 2023; Wu et al., 2023; Abdelfattah & Alahi, 2024; Sun et al., 2025; 2026) encourage models to capture local details by reconstructing raw coordinates (Wu et al., 2023; Mao et al., 2023) or latent features (Abdelfattah & Alahi, 2024; Sun et al., 2025; 2026). While these methodologies have driven significant progress, they share a critical structural bottleneck: they depend on fixed joint topologies and absolute positional embeddings. Consequently, they become structurally incompatible when an entirely new sensor with a different joint count is introduced, making them unable to serve as sensor-unified foundation models.

To tackle this data heterogeneity, a recent approach, HSP (Wang et al., 2025a), attempts to integrate diverse skeleton formats by fusing 2D and 3D sequences. However, as highlighted in Table 1, HSP (Wang et al., 2025a) focuses on intra-scene multi-modal fusion strategy that relies on paired data. To align differing joint coordinates, it uses naive skeletal interpolation, which distorts precise anatomical geometry and restricts its scalability to independent real-world sensors. In contrast, our proposed SOfA seamlessly aligns topological heterogeneity without geometric distortion. With Canonical Joint Slots and Semantic Joint Embedding, SOfA establishes the first sensor-unified foundation model capable of universal generalization across completely independent sensors. To bridge this gap, SOfA establishes a sensor-unified foundation model that captures the intrinsic kinematic essence of human motion across diverse sensor configurations.

2.2 Foundation Models in Other Fields

The innovative success of foundation models in natural language processing (NLP) and 2D vision establishes a powerful example for unifying fragmented datasets through large-scale representation learning. In NLP, Large Language Models (LLMs) (Zhao et al., 2023) achieve exceptional flexibility by mapping highly diverse texts—regardless of languages, lengths, or structures—into a standardized sequence of sub-word tokens. Similarly, vision foundation models, such as DINO (Caron et al., 2021; Oquab et al., 2023) and SAM (Kirillov et al., 2023), achieve powerful zero-shot generalization by decoupling their architectures from fixed input resolutions. By treating images as flexible sequences of patch tokens rather than rigid pixel grids, they can seamlessly process inputs of arbitrary sizes and formats. In the 3D point cloud (Pang et al., 2022a; Yu et al., 2022) and remote sensing (Guo et al., 2024; Wang et al., 2025b) domains, recent frameworks handle highly unconstrained and heterogeneous data by projecting them into a shared latent space.

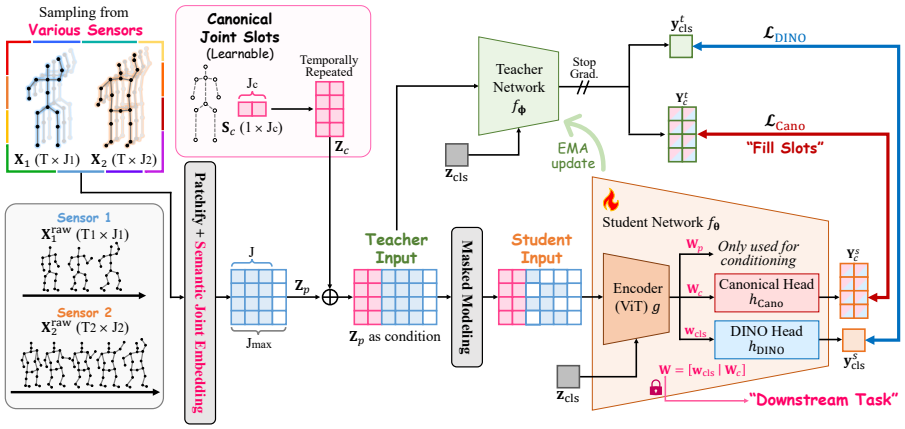

Figure 2: **Overview of SOfA.** Our framework establishes a sensor-unified representation within a teacher–student architecture, where both networks dynamically fill a fixed-size set of Canonical Joint Slots. Processing a masked view, the student encoder minimizes the canonical reconstruction error ($\mathcal{L}_{\text{Cano}}$) to match the slot representations of the unmasked teacher, alongside a global distillation loss ($\mathcal{L}_{\text{DINO}}$) to align high-level semantics. By treating varying sensor tokens w_p as dynamic conditions, the model exclusively extracts the unified canonical representation $\mathbf{w}$ for downstream tasks.

Parallel to these advancements, human motion is fundamentally invariant across sensor modalities; the underlying dynamics of an action remain constant despite variations in the capture devices. To the best of our knowledge, none of the existing frameworks can effectively reconcile the structural and semantic differences arising from multi-sensor skeleton data within a single architecture.

3 Skeleton One for All (SOfA)

3.1 SOfA Framework

Overview. Fig. 2 illustrates the overall architecture of **SOfA**, a sensor-unified framework built upon a teacher–student distillation paradigm (Tarvainen & Valpola, 2017). In this framework, both the student network ($f_{\boldsymbol{\theta}}$) and teacher network ($f_{\boldsymbol{\phi}}$) share an identical encoder g with lightweight (3-layer MLP) prediction heads h_{Cano} and h_{DINO}. While the student network parameters $\boldsymbol{\theta}$ are updated via back-propagation, the teacher network parameters $\boldsymbol{\phi}$ evolves smoothly through an exponential moving average (EMA) of the student: $\boldsymbol{\phi} \leftarrow \tau\boldsymbol{\phi}+(1-\tau)\boldsymbol{\theta}$, governed by momentum τ. Furthermore, SOfA addresses the limitations of sensor-specific models by bridging two structural barriers in multi-sensor data: (i) the *joint index misalignment* (sensor-specific joint indexing); and (ii) the *joint count discrepancy* (varying numbers of joints). Joint index misalignment is resolved through Semantic Joint Embedding (Sec. 3.2), which aligns joints semantically across sensors. Joint count discrepancy is handled by a fixed-size set of Canonical Joint Slots (Sec. 3.3), which treats arbitrary sensor input as a dynamic condition to extract a unified canonical representation. This serves as the final, universally compatible output, regardless of the original sensor topology.

Unified dataset formulation. To ensure geometric consistency across heterogeneous sources, we first align all datasets into a standardized 3D coordinate system. For each sample, the "hip" joint of the initial frame—a universally shared anatomical landmark—is designated as the origin. In addition, the vertical body axis (head-to-foot) is aligned with the y-axis, and all coordinates are normalized to meters to maintain physical consistency. Formally, for an input sequence $\mathbf{X}$, we define the multi-sensor corpus as $\mathcal{D} = \{(\mathbf{X}_i^{\text{raw}}, s_i)\}_{i=1}^{N}$, where $\mathbf{X}_i^{\text{raw}} \in \mathbb{R}^{T_i \times J_i \times 3}$ is i-th input skeleton sequence with temporal length T_i and J_i sensor-specific 3D joints. Each sensor identifier s_i is associated with a metadata dictionary specifying the mapping between semantic joint names and their sensor-specific numerical indices. To enable batch-wise computation, we randomly sample T frames and zero-pad the sequences along the joint dimension to J_{max}. Here, J_{max} denotes the maximum number of joints across the entire corpus. This results in a standardized tensor $\mathbf{X} \in \mathbb{R}^{T \times J_{\text{max}} \times 3}$, providing a unified input format for our architecture.

Architecture of the encoder. Our framework is built upon a Vision Transformer (ViT) (Dosovitskiy et al., 2020) encoder g, which explicitly casts the raw skeleton sequence $\mathbf{X}$ as a *dynamic condition input*. First, $\mathbf{X}$ is partitioned into a sequence of flattened 2D patches $\mathbf{X}_p \in \mathbb{R}^{N \times D_p}$, where $N = (T/P_T) \times (J_{\text{max}}/P_J)$ is the total number of patches with skeletal-temporal resolution (P_T, P_J), and $D_p = P_T \cdot P_J \cdot 3$. To map these diverse sensor inputs into a shared space, each patch undergoes a linear projection $\mathbf{E} \in \mathbb{R}^{D_p \times D}$. Crucially, instead of absolute positional embedding, we inject our Semantic Joint Embedding (`SJE`, detailed in Sec. 3.2) to assign explicit anatomical identities to condition tokens $\mathbf{Z}_p \in \mathbb{R}^{N \times D}$ as:

$$\mathbf{Z}_p = \left[\mathbf{x}_p^1\mathbf{E} \mid \mathbf{x}_p^2\mathbf{E} \mid \cdots \mid \mathbf{x}_p^N\mathbf{E}\right] + \texttt{SJE}. \tag{1}$$

Concurrently, we introduce a set of learnable Canonical Joint Slots (`CJS`, detailed in Sec. 3.3), denoted as $\mathbf{S}_c \in \mathbb{R}^{1 \times J_c \times D}$. To ensure temporal alignment of the condition input $\mathbf{Z}_p$, $\mathbf{S}_c$ is replicated T/P_T times along the time axis, forming the canonical token $\mathbf{Z}_c \in \mathbb{R}^{N_c \times D}$ with $N_c = (T/P_T) \times J_c$. A learnable class token $\mathbf{z}_{\texttt{cls}}$ and canonical tokens $\mathbf{Z}_c$ are then prepended to the patch tokens $\mathbf{Z}_p$ to initialize the input stream as $\mathbf{Z}_0 = [\mathbf{z}_{\texttt{cls}} \mid \mathbf{Z}_c \mid \mathbf{Z}_p]$. To capture sequence-level dependencies, 1D temporal Rotary Positional Embeddings (RoPE) (Su et al., 2024) are applied directly within the attention mechanism. The forward propagation across L layers is formulated as:

$$\begin{aligned}
&\mathbf{Z}_0 = [\mathbf{z}_{\texttt{cls}} \mid \mathbf{Z}_c \mid \mathbf{Z}_p], && \textit{Input initialization} \\
&\mathbf{Z}'_l = \mathsf{MSA}\left(\mathsf{LN}\left(\mathbf{Z}_{l-1}\right)\right) + \mathbf{Z}_{l-1}, && \textit{Attention w/ temporal RoPE} \\
&\mathbf{Z}_l = \mathsf{MLP}(\mathsf{LN}\left(\mathbf{Z}'_l\right)) + \mathbf{Z}'_l && \\
&[\mathbf{w}_{\texttt{cls}} \mid \mathbf{W}_c \mid \mathbf{W}_p] = \mathsf{LN}\left(\mathbf{Z}_L\right), && \textit{Encoder output} \\
&\mathbf{W} = [\mathbf{w}_{\texttt{cls}} \mid \mathbf{W}_c] && \textit{Final skeleton representation}
\end{aligned} \tag{2}$$

where MSA, LN, and MLP denote Multi-Head Self-Attention, Layer Normalization, and the feed-forward network, respectively. After encoding, the sensor-specific tokens $\mathbf{W}_p$ are discarded, retaining only the unified canonical representation $\mathbf{W}$ for downstream tasks.

Prediction heads. The model employs two task-specific prediction heads, the canonical head h_{Cano} and the contrastive head h_{DINO} (depicted in Fig. 2), to process the encoder outputs. h_{Cano} maps the sensor-unified local representations $\mathbf{W}_c$ to target dimensions, yielding $\mathbf{Y}_c = h_{\text{Cano}}(\mathbf{W}_c)$. h_{DINO} projects the global class token $\mathbf{w}_{\texttt{cls}}$ for instance discrimination, producing $\mathbf{y}_{\texttt{cls}} = h_{\text{DINO}}(\mathbf{w}_{\texttt{cls}})$. During training, the teacher network $f_{\boldsymbol{\phi}}$ processes the full, unmasked sequence to generate stable target representations: $[\mathbf{y}^t_{\texttt{cls}} \mid \mathbf{Y}^t_c] = f_{\boldsymbol{\phi}}\left([\mathbf{z}_{\texttt{cls}} \mid \mathbf{Z}_c \mid \mathbf{Z}_p]\right)$. Conversely, the student network $f_{\boldsymbol{\theta}}$ receives a corrupted view. We apply a binary mask $\mathcal{M}_p \in \{0,1\}^N$ to the patch tokens, defining the masked input $\mathbf{Z}_p^{\texttt{mask}}$ as:

$$\mathbf{Z}_p^{\texttt{mask}} = \mathcal{M}_p \odot \mathbf{Z}_p + (1 - \mathcal{M}_p) \odot \mathbf{e}^b_{\texttt{[mask]}}, \tag{3}$$

where $\mathbf{e}_{\texttt{[mask]}} \in \mathbb{R}^D$ is a learnable mask token, and $\mathbf{e}^b_{\texttt{[mask]}} \in \mathbb{R}^{N \times D}$ represents its broadcasted version to the masked positions. $f_{\boldsymbol{\theta}}$ then predicts features based on this corrupted context as: $[\mathbf{y}^s_{\texttt{cls}} \mid \mathbf{Y}^s_c] = f_{\boldsymbol{\theta}}\left(\left[\mathbf{z}_{\texttt{cls}} \mid \mathbf{Z}_c \mid \mathbf{Z}_p^{\texttt{mask}}\right]\right)$.

Optimization. We optimize SOfA through a dual-objective framework that enforces both global context and local structural alignment. For global context, we adopt a distillation loss inspired by DINO (Oquab et al., 2023), aligning the instance-level representations between $f_{\boldsymbol{\phi}}$ and $f_{\boldsymbol{\theta}}$:

$$\mathcal{L}_{\text{DINO}} = -\sum \mathbf{p}^t_{\texttt{cls}} \log \mathbf{p}^s_{\texttt{cls}}, \tag{4}$$

where $\mathbf{p}^t_{\texttt{cls}}$ and $\mathbf{p}^s_{\texttt{cls}}$ are the softmax probability distributions derived from the class outputs $\mathbf{y}^t_{\texttt{cls}}$ and $\mathbf{y}^s_{\texttt{cls}}$. For the local objective, we adopt the Canonical Slot Reconstruction loss, $\mathcal{L}_{\text{Cano}}$. Crucially, since $f_{\boldsymbol{\theta}}$ processes only a corrupted subset of the sensor input, it must implicitly learn to *fill* the fixed-size $\mathbf{Z}_c$ by inferring the missing kinematics from the visible condition patches. We formulate this slot-filling mechanism as an objective between $f_{\boldsymbol{\theta}}$'s predicted canonical representations and $f_{\boldsymbol{\phi}}$'s stable targets:

$$\mathcal{L}_{\text{Cano}} = -\sum \mathbf{p}^t_c \log \mathbf{p}^s_c, \tag{5}$$

where $\mathbf{p}^t_c$ and $\mathbf{p}^s_c$ denote the softmax distributions of the teacher and student canonical outputs, $\mathbf{y}^t_c$ and $\mathbf{y}^s_c$ respectively. The overall objective $\mathcal{L}_{\text{total}}$ integrates the canonical reconstruction and the global semantic alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Cano}} + \lambda \mathcal{L}_{\text{DINO}}, \tag{6}$$

where λ is a hyperparameter balancing the two components. We empirically set $\lambda = 1.0$ for all experiments.

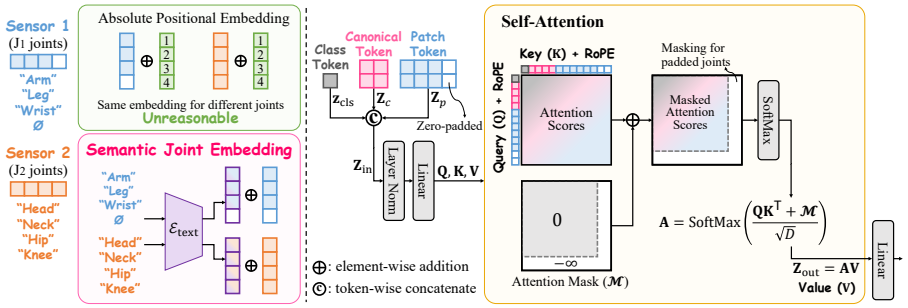

Figure 3: **Semantic Joint Alignment and Canonical Joint Slot aggregation.** (*Left*) to resolve the joint index misalignment, Semantic Joint Embedding (SJE) utilizes a pretrained text encoder to map heterogeneous joint names into a shared latent space, ensuring anatomically consistent representations regardless of the sensor-specific numerical indices. (*Right*) the attention mechanism jointly propagates the concatenated class token, canonical tokens, and condition patch tokens. This allows the model to effectively fill the canonical slots by aggregating information from the visible joints, while an attention mask is applied to exclude the influence of zero-padded tokens.

3.2 Semantic Joint Embedding (SJE)

Fig. 3 illustrates detailed mechanism of SOfA. We introduce SJE to mitigate the *joint index misalignment* caused by sensor-specific topologies as shown in Fig. 3. Standard absolute positional embedding, which rely on arbitrary numerical indices, are fundamentally unreasonable for multi-sensor setups. In such systems, the same index often refers to different anatomical joints across datasets—or even to empty, zero-padded joints—preventing the model from learning a consistent structural representation. To tackle this, we utilize the explicit textual names provided in the metadata of each sensor. For each joint $j \in \{1, \ldots, J_{\max}\}$, we construct a descriptive prompt: $\mathbf{d}_j$ = *"A human skeleton joint of the* $\{n_j\}$*"*, where n_j denotes the j-th joint name (e.g., "Left Wrist", "Head"). These prompts are processed through a pre-trained text encoder $\mathcal{E}_{\text{text}}$ to extract high-dimensional semantic vectors. Formally, the SJE for the j-th joint is defined as:

$$\text{SJE}_j = \begin{cases} \mathcal{E}_{\text{text}}(\mathbf{d}_j) & \text{if } j \le J_i \\ \mathbf{0} & \text{if } j > J_i \text{ (zero-padded)} \end{cases} \tag{7}$$

where J_i is the number of active joints for sensor s_i. We inject SJE into the input $\mathbf{Z}_p$ as described in Eq. 1. This approach ensures that anatomically identical joints share a consistent latent identity across all datasets. Furthermore, using the pre-trained text encoder $\mathcal{E}_{\text{text}}$, the model inherits semantic relationships between joints. For instance, joints sharing the "Left" descriptor are mapped to a similar feature space, helping the network recognize shared functional roles within the human body.

3.3 Canonical Joint Slots (CJS) and Self-Attention Mechanism

We address joint count discrepancy by introducing Canonical Joint Slots (CJS), a fixed-size set of learnable tokens $\mathbf{S}_c \in \mathbb{R}^{1 \times J_c \times D}$ that act as a universal vessel for all skeleton data. Inspired by multi-modal Diffusion Transformer (MM-DiT) (Peebles & Xie, 2023; Esser et al., 2024), which simultaneously processes conditional signals and target features, we propagate the canonical tokens alongside the sensor-specific patch tokens. As shown in Fig. 3, the input stream is constructed by concatenating all tokens (Eq. 2). We implement a *slot-filling* mechanism via masked self-attention. Here, the canonical slots $\mathbf{Z}_c$ are dynamically populated by aggregating skeletal information from the condition tokens $\mathbf{Z}_p$. We employ an attention mask $\mathcal{M}$ to neutralize the influence of zero-padded joints within the $\mathbf{Z}_p$ dimension:

$$\mathbf{A} = \text{SoftMax}\left(\frac{\mathbf{QK}^\top + \mathcal{M}}{\sqrt{D}}\right), \tag{8}$$

where $\mathcal{M} \in \{0, -\infty\}^{N_{\text{total}} \times N_{\text{total}}}$ is the masking matrix for a total sequence length of $N_{\text{total}} = 1 + N_c + N$, representing the $\mathbf{z}_{\text{cls}}$, $\mathbf{Z}_c$, and $\mathbf{Z}_p$, respectively. $\mathcal{M}$ assigns $-\infty$ to the attention scores

Table 2: **Comparison with recent SSL methods on NTU-60, NTU-120, and PKU-MMD (25-joint).** Results are reported as Top-1 accuracy (%) under the *linear evaluation*. Unlike baseline methods that require independent training for each dataset, SOfA employs a single, unified representation across all benchmarks. **Bold** indicates the best result, and underline indicates the second best. X-View* denotes evaluation on leakage-free test samples only (see *Supplementary Material*).

Method	Publication	NTU-60		NTU-120		PKU-MMD
		X-Sub	X-View	X-Sub	X-Set	X-Sub
***Sensor-Specific Representation*:**						
GL-Transformer (Kim et al., 2022)	ECCV'22	76.3	83.8	66.0	68.7	–
CPM (Zhang et al., 2022)	ECCV'22	78.7	84.9	68.7	69.6	48.3
CMD (Mao et al., 2022)	ECCV'22	79.8	86.9	70.3	71.5	43.0
AimCLR (Guo et al., 2022)	AAAI'22	74.3	79.7	63.4	63.4	–
HYSP (Franco et al., 2023)	ICLR'23	78.2	82.6	61.8	64.6	–
HaLP (Shah et al., 2023)	CVPR'23	79.7	86.8	71.1	72.2	43.5
ActCLR (Lin et al., 2023)	CVPR'23	80.9	86.7	69.0	70.5	–
RVTCLR (Zhu et al., 2023)	ICCV'23	74.7	79.1	–	–	–
SkeletonMAE (Wu et al., 2023)	ICMEW'23	74.8	77.7	72.5	73.5	36.1
MAMP (Mao et al., 2023)	ICCV'23	84.9	89.1	78.6	79.1	53.8
PTSL (Zhou et al., 2023)	AAAI'23	77.3	81.8	66.2	67.7	49.3
S-JEPA (Abdelfattah & Alahi, 2024)	ECCV'24	85.3	89.8	79.6	79.9	53.5
IGM (Lin et al., 2024)	ECCV'24	86.2	91.2	80.0	81.4	–
MacDiff (Wu et al., 2024)	ECCV'24	86.4	91.0	79.4	80.2	–
HSP (Wang et al., 2025a)	CVPR'25	80.7	88.0	71.0	73.2	48.9
USDRL (Weng et al., 2025)	AAAI'25	85.2	91.7	76.6	78.1	54.4
GFP (Sun et al., 2025)	ICCV'25	85.9	92.0	79.1	80.3	56.2
AMR (Sun et al., 2026)	CVPR'26	**87.4**	92.3	**81.1**	81.9	60.3
***Sensor-Unified Representation*:**						
SOfA (Ours)		85.7	92.1*	79.2	80.4	65.4
SOfA-L (Ours)		87.0	**92.6***	79.5	**82.0**	**67.1**

corresponding to padded joint indices, effectively excluding invalid joints. This ensures that the final representation $\mathbf{W}$ is strictly computed from valid skeletal information, regardless of the input's original joint count. Beyond resolving joint-count discrepancy, the slot count J_c offers a tunable output resolution; we analyze its effect and downstream flexibility in the *Supplementary Material*.

4 Experimental Results

4.1 Datasets

To pre-train and rigorously evaluate SOfA, we curate a large-scale corpus from various public benchmark datasets, encompassing a wide spectrum of sensor topologies and joint counts. A key contribution of our work is the standardization of these heterogeneous datasets into a unified coordinate system. Comprehensive profiles for each dataset and technical details regarding sensor-specific preprocessing are provided in the *Supplementary Material*.

Foundation model training. For the large-scale pre-training of our foundation model, we utilize eight datasets with high joint counts (20 and 25 joints) to provide rich, fine-grained skeletal dynamics. This training corpus includes: (i) 25-joint datasets: NTU RGB+D 60 (NTU-60) (Shahroudy et al., 2016), NTU RGB+D 120 (NTU-120) (Liu et al., 2019), PKU-MMD II (PKU-MMD) (Liu et al., 2017a), ETRI-Activity3D (ETRI-Act) (Jang et al., 2020), and ETRI-LivingLab (Jang et al., 2020), which offer the most detailed skeletal topologies for complex action recognition; and (ii) 20-joint datasets: MSR-Action3D (Li et al., 2010), NW-UCLA (Wang et al., 2014), and UT-Kinect (Xia et al., 2012a), which serve to inject structural diversity and broaden sensor-specific variations during the pre-training phase.

Unseen arbitrary sensor evaluation. To validate SOfA's robustness against unseen sensor configurations, we reserve two 15-joint datasets for unseen sensor evaluation: Florence (Seidenari et al., 2013a) and SBU-Kinect-Interaction (SBU-Inter) (Yun et al., 2012). By evaluating these disparate sensor topologies with a frozen encoder g, we explicitly demonstrate the model's generalizability and adaptability.

Pre-training corpus and train–test separation. To prevent train–test contamination, we enforce sequence-level separation between pre-training and evaluation. During pre-training, since NTU-

120 contains all NTU-60 sequences, we use only NTU-120 as the NTU source and retain only the intersection of the NTU-120 training splits (X-Sub ∩ X-Set). Evaluation on NTU-60 X-View requires an additional safeguard: X-View partitions sequences by camera ID, whereas NTU-120 X-Set partitions them by setup ID. The official X-View test split therefore cannot be used directly, as it contains some sequences present in the pre-training corpus. We exclude all such sequences before evaluation. Table 2 reports accuracy on the remaining sequence-disjoint subset, denoted X-View*. More details are in the *Supplementary Material*.

4.2 Experiment Details

To ensure a fair comparison with previous sensor-specific methods, we employ an 8-layer (SOfA) and a 12-layer (SOfA-L) ViT (Dosovitskiy et al., 2020) as our backbone encoder g, with complexity comparable to that of previous methods. The encoder g is configured with a hidden dimension of $D = 256$ with 8 attention heads for SOfA, and $D = 384$ with 12 attention heads for SOfA-L. Input skeleton sequences are fixed to a temporal length of $T = 64$ frames. For patch embedding, we use $(P_T, P_J) = (8, 1)$. We initialize $\mathcal{E}_{\text{text}}$ with a pretrained T5 text encoder (Raffel et al., 2020). Setting the maximum sensor size $J_{\max} = 25$, we fix the number of `CJS` to $J_c = 5$. SOfA is pre-trained for 120 epochs within the teacher–student distillation framework (Tarvainen & Valpola, 2017). The f_ϕ's EMA momentum τ is gradually updated from 0.994 to 1. Optimization is performed using AdamW (Loshchilov & Hutter, 2017) with a base learning rate of 2.0×10^{-4}, incorporating a 20-epoch linear warmup followed by a cosine decay (Loshchilov & Hutter, 2016) down to 1.0×10^{-6}. We train the model utilizing a total batch size of 1,536 across eight NVIDIA RTX A6000 GPUs.

4.3 Evaluation on Heterogeneous Benchmarks

Table 3: **Linear evaluation on seen benchmarks with varying joint configurations.** (a) ETRI-Act (25-joint) and (b) NW-UCLA (20-joint) are included in the pre-training domain. Results are Top-1 accuracy (%) under the *linear evaluation* protocol.

(a) ETRI-Act (25-joint)

Methods	ETRI-Act
***Fully-supervised*:**	
IndRNN (Li et al., 2018b)	73.9
Beyond Joint (Wang & Wang, 2018)	79.1
SK-CNN (Li et al., 2017)	83.6
ST-GCN (Yan et al., 2018)	86.8
Ensem-NN (Xu et al., 2018)	83.0
MANs (Li et al., 2021a)	82.4
HCN (Li et al., 2018a)	88.0
FSA-CNN (Jang et al., 2020)	90.6
***Un-supervised* (SSL):**	
SOfA (Ours)	87.9

(b) NW-UCLA (20-joint)

Methods	NW-UCLA
***Sensor-Specific*:**	
LongT-GAN (Zheng et al., 2018)	74.3
P&C (Su et al., 2020)	84.9
MCAE-MP (Xu et al., 2021)	84.9
SeBiReNet (Nie et al., 2020)	80.3
Colorization (Yang et al., 2021)	91.1
GL-Transformer (Kim et al., 2022)	90.4
Masked-Color (Yang et al., 2023)	92.0
***Sensor-Unified*:**	
SOfA (Ours)	92.9

Table 2 and Table 3 summarize the linear evaluation results on both 25-joint and 20-joint benchmarks. As shown in Table 2, SOfA achieves competitive or superior performance on large-scale benchmarks like NTU-60 and NTU-120. To ensure a fair comparison, we report all results using the joint modality alone, without any multi-stream ensemble. Notably, existing skeleton-based SSL methods are inherently sensor-specific and even dataset-specific. They require training independent, isolated models for each unique protocol. In contrast, SOfA is a single, unified model trained across all datasets simultaneously. We provide extensive experimental results in the *Supplementary Material*, including evaluations on semi-supervised learning, action retrieval, robustness across various sensor configurations, and a detailed computational-efficiency analysis showing that SOfA attains a 6.59× reduction in inference GFLOPs (4.30 vs. 28.32) over mainstream MAE-based SSL methods through aggressive temporal compression.

Knowledge transfer in data-scarce scenarios. On PKU-MMD, SOfA-L surpasses the previous best specialist (AMR, 60.3%) by +6.8%-point, reaching 67.1%. While data-rich benchmarks like NTU-120 let specialists saturate by overfitting to massive sample sizes, the smaller PKU-MMD limits isolated models; SOfA bridges this gap by transferring robust skeletal priors from our diverse corpus, internalizing generalized human kinetics rather than memorizing sensor-specific patterns.

Competitive strength against fully-supervised methods. SOfA is also strong against fully-supervised models: as shown in Table 3(a), it reaches 87.9% on ETRI-Act, within a close margin of high-performing supervised models—despite learning these features self-supervised, without any action labels during pre-training.

Adaptability to various joint counts. Table 3(b) shows SOfA's versatility across diverse skeletal topologies. On the 20-joint NW-UCLA dataset, SOfA achieves 92.9% accuracy, outperforming all existing sensor-specific models. Crucially, this result is obtained using the exact same model weights employed for the 25-joint benchmarks in Table 2 and Table 3(a). This success across various joint counts rigorously validates SOfA's capability as a unified foundation model that can seamlessly navigate heterogeneous sensor protocols without the need for any per-dataset reconfiguration.

4.4 Generalization to Unseen Skeletal Protocols

To evaluate the true robustness of SOfA as a foundation model, we test it on arbitrary out-of-distribution (OOD) sensors completely excluded from pre-training. Table 4(a) and Table 4(b) report results on the unseen Florence and SBU-Inter datasets. Without any fine-tuning, SOfA reaches 91.7% on Florence and 97.0% on SBU-Inter, whereas sensor-specific SSL methods such as SkeletonMAE and MAMP degrade sharply (e.g., 66.8% and 73.7% on Florence). Notably, SOfA matches fully-supervised models trained end-to-end with full label access, providing solid evidence that it transcends sensor-specific limitations by internalizing universal skeletal rules.

Table 4: **Linear evaluation on unseen cross-domain benchmarks.** (a) Florence (15-joint) and (b) SBU-Inter (15-joint) are excluded from pre-training and evaluates sensor-unified representation generalization. Results are Top-1 accuracy (%) under the *linear evaluation*.

(a) Florence (15-joint)

Methods	Florence
Seen+*Fully-supervised*:	
Seidenari *et al.* (2013b)	82.0
Devanne *et al.* (2014)	87.0
Vemulapalli *et al.* (2014)	90.9
HarSkel (Luvizon et al., 2017)	94.4
Unseen+*Sensor-Specific*:	
SkeletonMAE (Wu et al., 2023)	66.8
MAMP (Mao et al., 2023)	73.7
Unseen+*Sensor-Unified*:	
SOfA (Ours)	91.7

(b) SBU-Inter (15-joint)

Methods	SBU-Inter
Seen+*Fully-supervised*:	
Co-LSTM (Zhu et al., 2016)	90.4
ST-LSTM (Liu et al., 2016)	93.3
VA-LSTM (Zhang et al., 2017)	97.2
GCA (Liu et al., 2017b)	94.9
LSTM-IRN (Perez et al., 2021)	98.2
IGFormer (Pang et al., 2022b)	98.4
ISTA-Net (Wen et al., 2023)	98.5
Unseen+*Sensor-Specific*:	
SkeletonMAE (Wu et al., 2023)	73.1
MAMP (Mao et al., 2023)	80.2
Unseen+*Sensor-Unified*:	
SOfA (Ours)	97.0

4.5 Ablation Studies

Importance of canonical space via CJS. Without CJS, the model must aggregate all information from arbitrary sensor protocols into a single class token $\mathbf{z}_{\texttt{cls}}$, creating an information bottleneck that cannot capture the complex spatio-temporal dynamics of diverse skeletal structures. As shown in Table 5, removing CJS sharply degrades performance across all benchmarks (average drop of 8.48%), confirming that our canonical slots provide an essential structured, high-resolution workspace for universal motion representation.

Geometric reasoning via SJE. The role of SJE is particularly critical for heterogeneous topologies. Removing SJE causes a catastrophic drop on 20-joint datasets (e.g., NW-UCLA from 92.9% to 72.3%), while the impact on 25-joint datasets is moderate. This stems from the data distribution: roughly 90% of our corpus is 25-joint Kinect-v2 data, so without SJE the network becomes biased toward the 25-joint indexing, and SJE acts as the semantic anchor bridging this index discrepancy.

Table 5: **Ablation of SOfA components.** We isolate the contributions of CJS and SJE across heterogeneous topologies.

Modules		25-Joint Datasets			20-Joint Datasets	
CJS	SJE	NTU-60	NTU-120	PKU-MMD	NW-UCLA	UT-Kinect
✗	✗	75.2	70.7	50.9	68.9	77.0
✓	✗	84.9	78.7	64.7	72.3	80.0
✗	✓	76.4	71.4	52.3	86.7	91.0
✓	✓	85.7	79.2	65.4	92.9	97.0

5 Conclusion

We present **SOfA**, the first sensor-unified foundation model for skeleton representation learning that bridges heterogeneous sensor types, joint counts, and indexing protocols. Shifting from dataset-specific modeling to a unified architecture, SOfA addresses topological heterogeneity through attention-based *Canonical Joint Slots* and *Semantic Joint Embeddings*, reaching state-of-the-art performance across diverse benchmarks with a single unified encoder.

APPENDIX

This *Supplementary Material* provides additional technical details, dataset profiles, and experimental results that complement the main paper. First, Sec. A provides further empirical validation, including network scalability, an extended synthetic topology, semi-supervised learning, action retrieval, and linear evaluation on 20-joint datasets, together with a computational-complexity comparison against state-of-the-art methods. Sec. B presents additional ablation studies on the number of Canonical Joint Slots (`CJS`) and analyzes their interpretability. Sec. C details our pre-training corpus construction and the leakage-free X-View* protocol. Sec. D offers dataset summaries, coordinate statistics, and preprocessing details for the ten heterogeneous datasets. Finally, Sec. E elaborates on implementation specifics.

Table 6: Overview of the *Supplementary Material*.

Section	Contents
Sec. A	Additional results and discussion
Sec. B	More ablation studies on `CJS`
Sec. C	Pre-training corpus and data leakage
Sec. D	Datasets and preprocessing
Sec. E	Implementation details

A ADDITIONAL RESULTS AND DISCUSSION

A.1 NETWORK SCALABILITY

Table 7 presents the effect of scaling the encoder g. To evaluate the scalability of SOfA, we enlarge the backbone from the *base configuration* (**SOfA**: 8 Transformer layers, hidden dimension 256) to a *deeper and wider variant* (**SOfA-L**: 12 Transformer layers, hidden dimension 384). While the base model is adopted for fair comparison with prior methods (Mao et al., 2023; Sun et al., 2025; Wu et al., 2023; Abdelfattah & Alahi, 2024), our pre-training corpus is substantially larger and more diverse than any single benchmark, covering eight heterogeneous datasets and roughly 2.5$\times$ more samples than NTU-120 alone. This motivates investigating whether SOfA can further benefit from increased model capacity. As shown in Table 7, scaling the encoder consistently improves performance across all benchmarks, with SOfA-L yielding clear gains on NTU (e.g., NTU-60 X-Sub 85.7% $\rightarrow$ 87.0%, NTU-120 X-Set 80.4% $\rightarrow$ 82.0%). Due to GPU memory constraints, we reduce the total training batch size from 1,536 to 1,024 for SOfA-L. These results indicate that SOfA scales favorably with capacity and that the unified corpus is large enough to support stronger backbones, suggesting that the modest NTU gains of the base model stem from a capacity bottleneck rather than a limitation of the framework.

A.2 EXTENDED SYNTHETIC TOPOLOGY

Zero-padding to $J_{\text{max}} = 25$ is strictly a batching convenience, not an architectural ceiling. SOfA processes unseen topologies with more than 25 joints without retraining, owing to (i) an attention backbone that treats joints as a variable-length sequence (unlike GCNs with fixed adjacency matrices), and (ii) `SJE`, which dynamically generates valid embeddings for novel joints on the fly from their textual descriptions. We empirically verify this on an extended synthetic topology ($J = 30$), created by interpolating pairs of joints (J_a and J_b) and feeding their textual descriptions (e.g., "midpoint of J_a and J_b") to `SJE`. As reported in Table 7, the J=30 variant attains performance comparable to the standard J=25 setting, demonstrating that the architecture is agnostic to the padded joint budget.

A.3 COMPUTATIONAL COMPLEXITY

Table 8 compares the computational cost of SOfA against mainstream MAE-based SSL methods (Wu et al., 2023; Mao et al., 2023; Abdelfattah & Alahi, 2024; Sun et al., 2025). Although SOfA introduces additional canonical tokens along the joint dimension, its aggressive temporal compression

Table 7: Network scalability on the 25-joint benchmarks (NTU-60, NTU-120, PKU-MMD). Results are Top-1 accuracy (%) under the *linear evaluation* protocol using the joint (J) modality. **SOfA** is the base model, **SOfA** (J=30) is evaluated on an extended synthetic topology, and **SOfA-L** is the deeper and wider variant. **Bold** denotes the best result.

Method	Modality	NTU-60		NTU-120		PKU-MMD
		X-Sub	X-View	X-Sub	X-Set	X-Sub
SOfA	J	85.7	92.1*	79.2	80.4	65.4
SOfA (J=30)	J	85.5	91.9*	79.0	80.1	65.6
SOfA-L	J	**87.0**	**92.6***	**79.5**	**82.0**	**67.1**

Table 8: Computational complexity comparison on PKU-MMD. We report the number of tokens, training and inference GFLOPs, and Top-1 accuracy (X-Sub). SOfA achieves state-of-the-art accuracy while substantially reducing inference cost.

Method	Tokens	GFLOPs		PKU-MMD
	$T \times J$	Train	Inference	X-Sub
***Sensor-Specific*:**				
SkeletonMAE (Wu et al., 2023)	30×25	19.67	28.32	36.1
MAMP (Mao et al., 2023)	30×25	19.67	28.32	53.8
S-JEPA (Abdelfattah & Alahi, 2024)	30×25	47.99	28.32	53.5
GFP (Sun et al., 2025)	30×25	4.18	28.32	56.2
***Sensor-Unified*:**				
SOfA	$8 \times (25{+}5)$	8.60	**4.30**	65.4
SOfA-L	$8 \times (25{+}5)$	14.90	7.45	**67.1**

yields a substantially lower overall cost: SOfA requires only 4.30 GFLOPs at inference, a 6.59× reduction relative to the 28.32 GFLOPs of MAE-based baselines, while simultaneously improving Top-1 accuracy on PKU-MMD by a large margin (65.4% vs. 56.2% for the strongest baseline). Even the larger SOfA-L, at 7.45 GFLOPs, remains far cheaper than the baselines. For downstream deployment, the model can be further streamlined to use only the 8×5 canonical tokens, providing a scalable and lightweight backbone for real-time skeleton analysis.

A.4 Semi-Supervised Learning

Table 9 shows the semi-supervised comparison on NTU-60 (Shahroudy et al., 2016), using only 1% of the training labels. SOfA achieves 71.6% on X-Sub and 74.2% on X-View, outperforming all previous sensor-specific methods on X-View and remaining highly competitive on X-Sub (only 0.2%-point below GFP (Sun et al., 2025)). This shows that a sensor-unified representation not only generalizes across heterogeneous topologies but also surpasses strong sensor-specific pre-training under extremely label-scarce conditions, indicating that our unified canonical representation captures rich, linearly separable motion semantics even with limited supervision.

A.5 Action Retrieval

Table 10 reports skeleton-based action retrieval on NTU-60 (Shahroudy et al., 2016). SOfA achieves 72.1% on X-Sub and 86.8% on X-View. On X-Sub it outperforms the previous best sensor-specific method, GFP (Sun et al., 2025), by 1.2%-point (72.1% vs. 70.9%), establishing a new state of the art; on X-View it trails GFP by only 0.3%-point. These results further demonstrate that a sensor-unified representation can match or exceed sensor-specific models in instance-level retrieval, indicating a well-structured latent space with strong discriminability.

A.6 Linear Evaluation on 20-Joint Datasets

Tables 11 and 12 report linear evaluation on two 20-joint benchmarks, MSR-Action3D (Li et al., 2010) and UT-Kinect (Xia et al., 2012a). On MSR-Action3D, SOfA achieves 82.2%, comparable to fully supervised methods and only 2.3%-point below the best reported result (84.5%). On

Table 9: *Semi-supervised* results on NTU-60 using only 1% labeled data.

Methods	NTU-60	
	X-Sub	X-View
***Sensor-Specific*:**		
CPM (Zhang et al., 2022)	56.7	57.5
CMD (Mao et al., 2022)	50.6	53.0
HaLP (Shah et al., 2023)	46.6	48.7
HiCo (Dong et al., 2023)	54.4	54.8
UmURL (Sun et al., 2023)	58.1	58.3
SkeletonMAE (Wu et al., 2023)	54.4	54.6
MAMP (Mao et al., 2023)	66.0	68.7
S-JEPA (Abdelfattah & Alahi, 2024)	67.5	69.1
USDRL (Weng et al., 2025)	57.3	60.7
GFP (Sun et al., 2025)	**71.8**	<u>72.9</u>
***Sensor-Unified*:**		
SOfA	<u>71.6</u>	**74.2***

Table 10: *Action retrieval* results on NTU-60 (X-Sub, X-View).

Methods	NTU-60	
	X-Sub	X-View
***Sensor-Specific*:**		
LongT-GAN (Zheng et al., 2018)	39.1	48.1
P&C (Su et al., 2020)	50.7	76.3
ISC (Thoker et al., 2021)	62.5	82.6
HaLP (Shah et al., 2023)	65.8	83.6
HiCo (Dong et al., 2023)	68.3	84.8
MAMP (Mao et al., 2023)	62.0	70.0
GFP (Sun et al., 2025)	<u>70.9</u>	**87.1**
***Sensor-Unified*:**		
SOfA	**72.1**	<u>86.8</u>*

Table 11: Fully supervised comparison on MSR-Action3D (20-joint), *linear eval*.

Methods	MSR-Action3D
***Fully-supervised*:**	
HON4D (Oreifej & Liu, 2013)	82.2
Rahmani *et al.* (Rahmani et al., 2014)	82.7
Tran *et al.* (Tran & Ly, 2013)	84.5
***Un-supervised*:**	
SOfA	82.2

Table 12: Fully supervised comparison on UT-Kinect (20-joint), *linear eval*.

Methods	UT-Kinect
***Fully-supervised*:**	
Xia *et al.* (Xia et al., 2012b)	90.9
Devanne *et al.* (Devanne et al., 2014)	91.5
Wang *et al.* (Wang et al., 2016)	96.5
***Un-supervised*:**	
SOfA	97.0

UT-Kinect, SOfA attains 97.0%, surpassing all compared fully supervised methods, including the previous best (96.5%), by 0.5%-point. These results further demonstrate that SOfA learns transferable, highly discriminative representations even on 20-joint protocols.

B More Ablation Studies

B.1 The Number of Canonical Joint Slots

Table 13 shows the impact of varying the number of Canonical Joint Slots (`CJS`), with $J_c \in \{0, 5, 10, 15\}$. Introducing `CJS` consistently improves performance over the variant without canonical slots, confirming the effectiveness of our sensor-unified canonical space. The best overall performance is achieved at $J_c = 5$, which improves NTU-60 (Shahroudy et al., 2016), NTU-120 (Liu et al., 2019), PKU-MMD (Liu et al., 2017a), NW-UCLA (Wang et al., 2014), and UT-Kinect (Xia et al., 2012a) by 9.3, 7.8, 13.1, 6.2, and 6.0%-point over the $J_c = 0$ baseline, respectively. When the number of slots is further increased to 10 or 15, performance gradually declines on most benchmarks. We believe this is because too many canonical slots *reduce the representation bottleneck effect* and *introduce redundancy across slots*, making the canonical space less compact and less effective for cross-sensor semantic alignment. This suggests that a small set of canonical slots is sufficient to capture shared human motion semantics while preserving transferability across heterogeneous sensors.

B.2 Flexibility of the Slot Count

Beyond resolving joint-count discrepancy, the slot count J_c endows SOfA with a tunable output resolution. By scaling J_c, the model generates a standardized representation tailored to specific downstream needs—ranging from compact prompts for motion diffusion, to richer structural inputs for vision–language models (VLMs), to high-level features for action recognition. As quantified in

Table 13: Ablation study on the number of Canonical Joint Slots (CJS). We report linear evaluation accuracy (%) on 25-joint and 20-joint benchmarks, together with inference GFLOPs.

Slots	25-Joint Datasets			20-Joint Datasets		GFLOPs
CJS	NTU-60	NTU-120	PKU-MMD	NW-UCLA	UT-Kinect	Inference
0	76.4	71.4	52.3	86.7	91.0	3.60
5	85.7	79.2	65.4	92.9	97.0	4.30
10	85.5	79.0	64.5	92.7	98.0	5.41
15	85.3	78.9	63.5	91.8	97.0	6.55

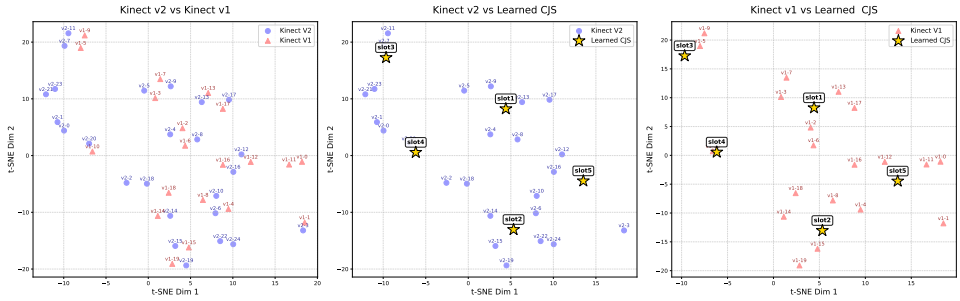

Figure 4: t-SNE visualization of Canonical Joint Slot assignments. The five slots autonomously converge toward the semantic centers of distinct anatomical clusters (torso and four limbs), consistent with the optimal $J_c = 5$ in Table 13.

Table 13, small slot counts already capture the shared human-motion semantics, while overly large counts dilute the bottleneck and introduce redundancy. This flexibility establishes SOfA as a versatile skeletal interface that adapts its output density to the computational and semantic requirements of diverse applications, without any architectural change.

B.3 Interpretability of Canonical Joint Slots

Using CJS to resolve cross-sensor topological gaps is a novel design, and we find the learned slots to be semantically meaningful rather than a generic bottleneck. Fig. 4 visualizes a t-SNE embedding of the slot assignments: the slots autonomously converge toward the semantic centers of distinct anatomical clusters across diverse datasets. If CJS were merely a generic compression bottleneck, its optimal size would be arbitrary; yet Table 13 shows performance peaks precisely at five slots. This natural alignment with the five major regions of the human body (torso and four limbs) quantitatively indicates that CJS extracts an anatomically meaningful space rather than performing arbitrary data compression.

C Pre-training Corpus and Evaluation Protocol for NTU-60/120

Why special handling is needed. Previous skeleton representation methods are sensor-specific and train a separate model for each benchmark. In contrast, SOfA is the first sensor-unified approach that pre-trains a single encoder on a combined skeleton corpus. When constructing this corpus, NTU-60 Shahroudy et al. (2016) and NTU-120 Liu et al. (2019) require special handling because NTU-120 is an extension of NTU-60 and contains all NTU-60 sequences.

Official protocols and splits. NTU-60 contains 56,880 raw action sequences from 60 classes and 40 subjects, whereas NTU-120 extends it to 114,480 sequences, 120 classes, and 106 subjects. Both datasets provide 3D skeleton sequences with 25 body joints captured using Kinect v2 sensors.

NTU-60 defines two official evaluation protocols: X-Sub and X-View. X-Sub divides its 40 subjects into 20 training and 20 test subjects, producing 40,320 training samples and 16,560 test samples. X-View splits samples by camera ID: cameras 2 and 3 are used for training, while camera 1 is used for testing. The resulting sets contain 37,920 training samples and 18,960 test samples. Likewise,

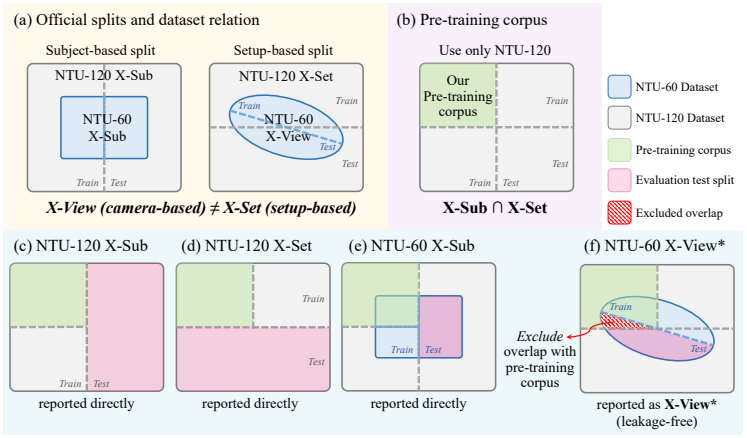

Figure 5: **Construction of the NTU pre-training corpus and evaluation sets.** NTU-120 contains all NTU-60 sequences. NTU-60 and NTU-120 use the same X-Sub assignment for all NTU-60 subjects. However, NTU-60 X-View splits samples by camera ID, while NTU-120 X-Set splits them by setup ID. (b) Pre-training uses only sequences assigned to training by both NTU-120 X-Sub and X-Set. (c–e) The official NTU-120 X-Sub, NTU-120 X-Set, and NTU-60 X-Sub test sets are used directly because they do not overlap with the pre-training corpus. (f) For NTU-60 X-View, overlapping test sequences are removed before evaluation, and the remaining leakage-free subset is denoted X-View*.

Table 14: Summary of the ten heterogeneous 3D skeleton datasets. Datasets are grouped by tracking protocol and skeletal topology; the 15-joint protocols are reserved exclusively for unseen evaluation.

Dataset	Tracker	Joints	Classes	Samples	Evaluation Setting
NTU-60 (Shahroudy et al., 2016)	Kinect v2	25	60	56,880	Pre-train & Downstream
NTU-120 (Liu et al., 2019)	Kinect v2	25	120	114,480	Pre-train & Downstream
PKU-MMD (Liu et al., 2017a)	Kinect v2	25	51	7,096	Pre-train & Downstream
ETRI-Act (Jang et al., 2020)	Kinect v2	25	55	112,620	Pre-train & Downstream
ETRI-LivingLab (Jang et al., 2020)	Kinect v2	25	55	8,605	Pre-train & Downstream
MSR-Action3D (Li et al., 2010)	Kinect v1	20	20	567	Pre-train & Downstream
NW-UCLA (Wang et al., 2014)	Kinect v1	20	10	1,494	Pre-train & Downstream
UT-Kinect (Xia et al., 2012a)	Kinect v1	20	10	199	Pre-train & Downstream
SBU-Inter (Yun et al., 2012)	Custom Tracker 1	15	8	282	Unseen (OOD)
Florence (Seidenari et al., 2013a)	Custom Tracker 2	15	9	215	Unseen (OOD)

NTU-120 defines two official evaluation protocols, X-Sub and X-Set: both datasets use X-Sub, but their second protocols differ. NTU-120 X-Sub divides the 106 subjects into 53 training and 53 test subjects, while X-Set uses the 16 even-numbered setups for training and the 16 odd-numbered setups for testing.

For the original 40 subjects, NTU-60 X-Sub and NTU-120 X-Sub use exactly the same train–test assignment. In other words, a subject used for training in NTU-60 X-Sub is also used for training in NTU-120 X-Sub, and the same holds for testing. The X-View and X-Set splits follow different rules. X-View splits sequences by camera ID, while X-Set splits them by setup ID. Therefore, the same sequence can belong to the X-Set training set but the X-View test set. Figure 5 gives an overview of these dataset relations and official splits.

Pre-training corpus. For NTU-60 and NTU-120, we use only NTU-120 during pre-training because it already contains all NTU-60 sequences. We keep only the sequences assigned to training by both NTU-120 protocols (**X-Sub ∩ X-Set, i.e., 25,053, only 22% of NTU-120**). If either X-Sub or X-Set assigns a sequence to testing, that sequence is not used for pre-training. Figure 5(b) illustrates this construction.

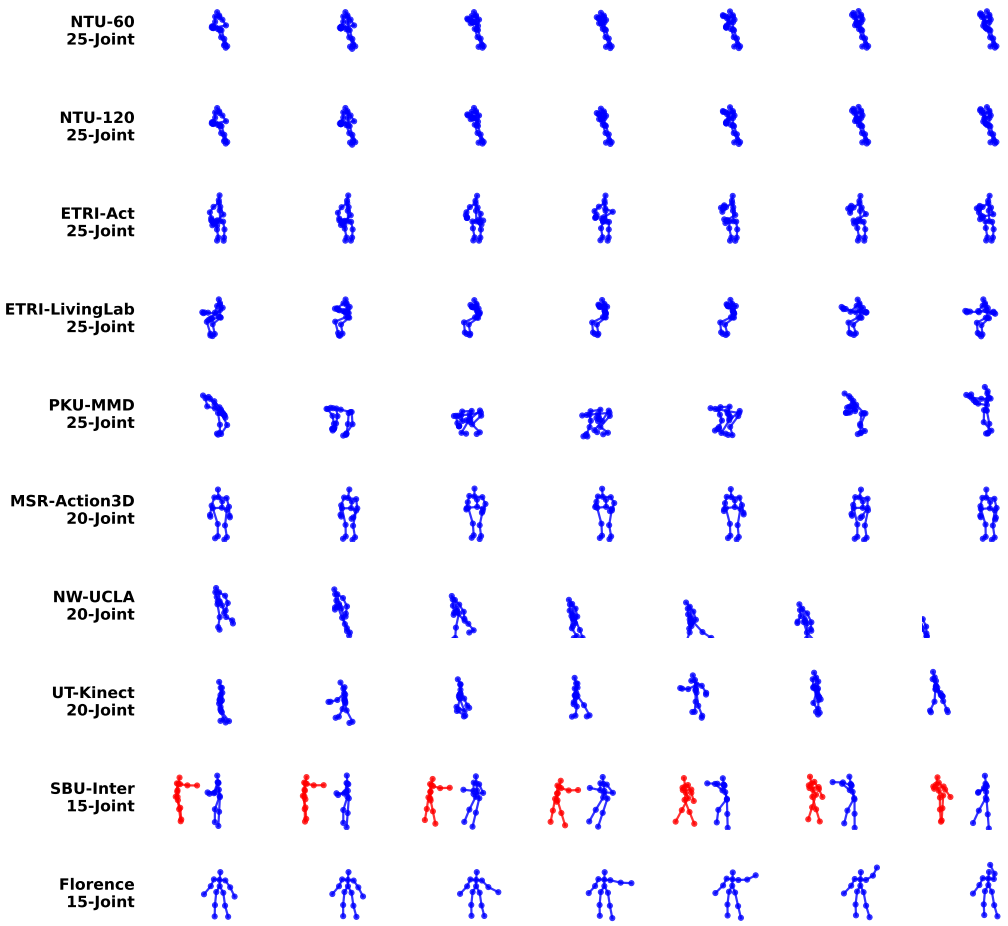

Figure 6: Representative normalized skeleton samples from the ten heterogeneous datasets. Despite variations in joint count, tracking protocol, and acquisition environment, the sequences are consistently aligned to a shared hip-centered coordinate frame.

Evaluation on official test sets. The official NTU-120 X-Sub and X-Set test sets can be used directly because none of their test sequences are included in the pre-training corpus. The official NTU-60 X-Sub test set can also be used directly. For the 40 subjects in NTU-60, NTU-60 and NTU-120 use the same 20 training subjects and the same 20 test subjects under X-Sub. Figure 5(c–e) illustrates these three cases.

NTU-60 X-View*. NTU-60 X-View needs an additional check in our sensor-unified setting. The official X-View split is valid when NTU-60 is used alone; the overlap appears only because our pre-training also uses NTU-120. X-View splits sequences by camera ID, while NTU-120 X-Set splits them by setup ID. Because these rules differ, some sequences in the official X-View test set are also included in our pre-training corpus. Before evaluation, we compare the official X-View test set with the pre-training corpus using sequence IDs and remove every overlapping sequence. We report the remaining leakage-free test set as X-View*. Table 2 in the main paper reports accuracy on this set. This filtering is completed before evaluation and does not depend on model predictions. Figure 5(f) illustrates this process. We remove **5,392** sequences that also appear in the pre-training corpus, leaving **13,568** sequences in X-View*.

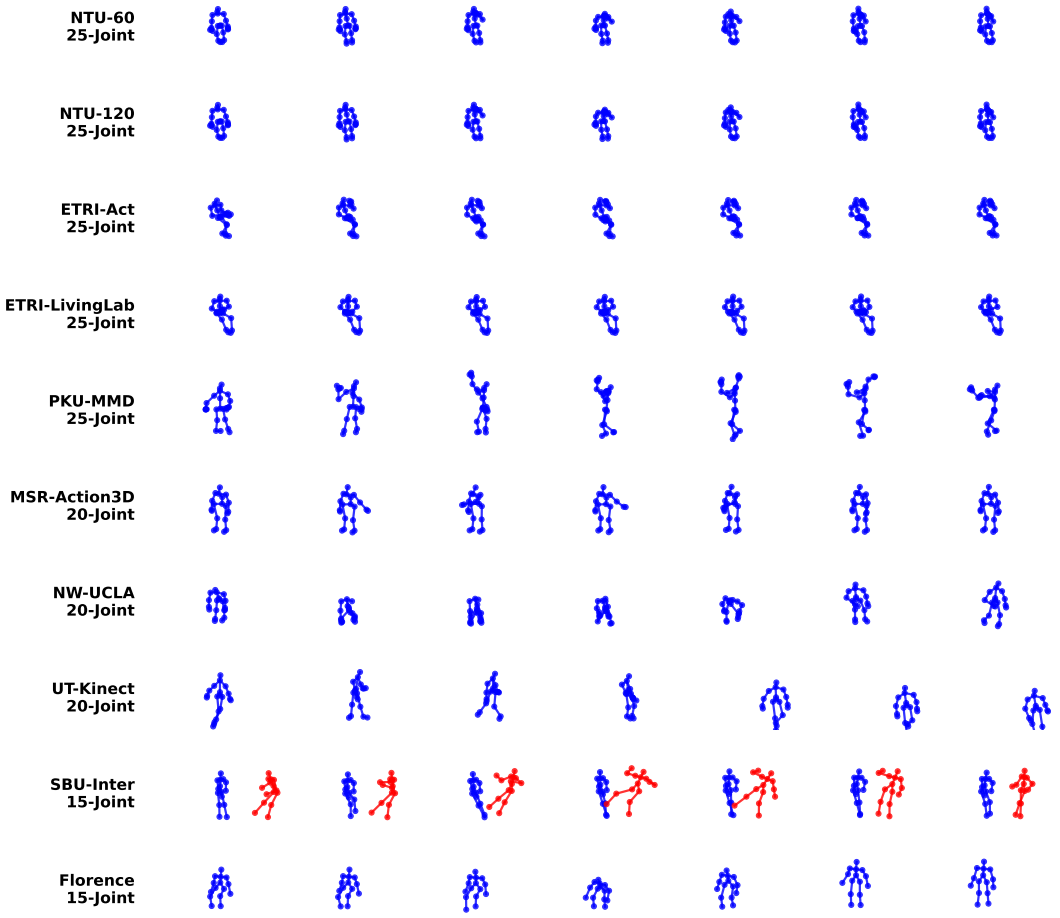

Figure 7: Additional normalized skeleton examples from the ten heterogeneous datasets, illustrating that our preprocessing preserves dataset-specific motion patterns while reducing cross-dataset geometric discrepancies.

D Datasets and Preprocessing

D.1 Datasets

Table 14 summarizes the ten heterogeneous 3D skeleton datasets used in our framework. The practical heterogeneity among datasets arises from differences in tracking protocols, joint indexing conventions, anatomical coverage, and coordinate scales, leading to substantially different skeletal topologies even when some datasets originate from similar hardware families. In our benchmark design, the pre-training corpus consists of eight datasets with seen 20-joint and 25-joint protocols, while two 15-joint datasets are strictly reserved for unseen evaluation, testing whether SOfA generalizes beyond the skeletal structures observed during pre-training. The unseen 15-joint protocols are not merely lower-dimensional versions of the seen datasets but follow different joint layouts and indexing systems, making transfer non-trivial. To further illustrate these structural differences, Fig. 6, Fig. 7, and Fig. 8 visualize representative normalized skeletons from each dataset after preprocessing.

D.2 Preprocessing Details

To reduce geometric inconsistencies across heterogeneous sources, we apply a unified preprocessing pipeline before pre-training and downstream evaluation. First, all coordinates are converted to meters (m) when necessary. Next, each sequence is translated to a shared semantic origin defined

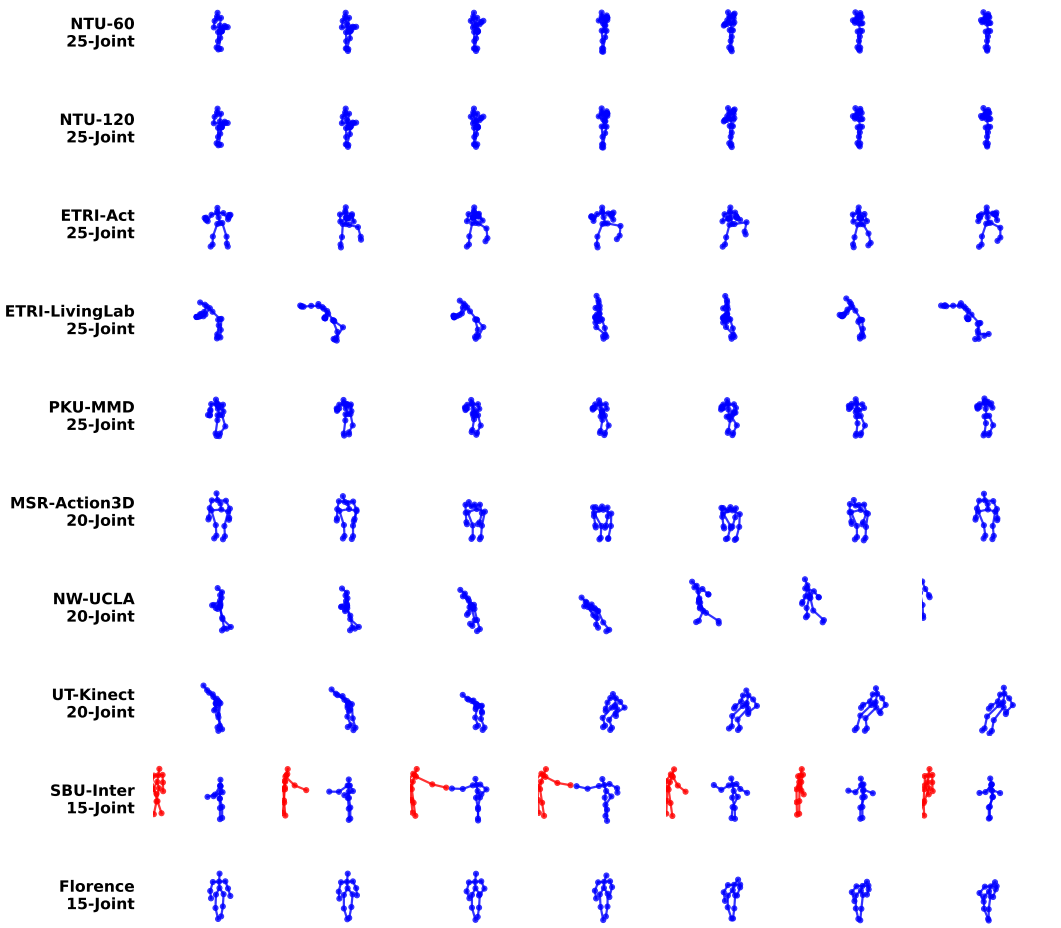

Figure 8: Additional qualitative visualization of normalized skeletons across the ten datasets, including unseen 15-joint protocols. The unified preprocessing remains effective under substantial topological variation and interaction-specific layouts.

from the first frame. For datasets with an explicit "hip" joint, we use that joint as the origin; for 15-joint datasets such as SBU-Inter (Yun et al., 2012) and Florence (Seidenari et al., 2013a), where no explicit "hip" joint is available, we define a virtual hip as the midpoint between the left and right hip joints. All coordinates are then translated by the first-frame origin of the main actor, yielding a consistent hip-centered coordinate frame. We additionally align the vertical body axis to the Y-axis so that skeletons from different tracking protocols share a common upright orientation. Finally, we remove invalid samples, including sequences containing NaN or Inf values, all-zero skeletons, and samples with extreme coordinate outliers beyond a predefined threshold. Table 15 reports per-axis coordinate statistics after preprocessing. Despite the large variation in topology and acquisition settings, the processed datasets exhibit comparable spatial ranges, showing that our standardization successfully neutralizes cross-dataset statistical discrepancies.

E Implementation Details

Encoder architecture. The backbone encoder g is a customized Vision Transformer (ViT) (Dosovitskiy et al., 2020) tailored for skeleton representation learning. It consists of 8 Transformer layers with a hidden dimension of $D = 256$ and 8 attention heads. To alleviate attention artifacts and reduce the risk of feature collapse, we incorporate 4 register tokens ($N_{\text{reg}} = 4$) (Darcet et al., 2023), which absorb outlier features during self-attention. Both the student network $f_{\boldsymbol{\theta}}$ and the teacher network $f_{\boldsymbol{\phi}}$ employ identical 3-layer MLP projection heads, $h_{\texttt{Cano}}$ and $h_{\texttt{DINO}}$, with a hidden dimension of 2,048 and a bottleneck dimension of 256. These heads project the encoded features into

Table 15: Per-axis coordinate statistics of the ten unified datasets after preprocessing (in meters). The heterogeneous datasets are consistently aligned to a shared hip-centered coordinate frame with comparable spatial ranges.

Dataset	Axis	Min	Max	Mean	Median
25-Joint Protocols (Pre-train & Downstream)					
NTU-60 (Shahroudy et al., 2016)	X	-5.3128	5.2244	-0.0060	-0.0022
	Y	-2.7981	2.4020	0.0414	0.0338
	Z	-4.8746	3.8188	-0.0657	-0.0536
NTU-120 (Liu et al., 2019)	X	-5.3128	5.2244	0.0012	-0.0012
	Y	-2.7981	2.4369	0.0422	0.0375
	Z	-4.8746	4.5483	-0.0703	-0.0515
PKU-MMD (Liu et al., 2017a)	X	-2.9480	2.2948	0.0062	0.0021
	Y	-1.7711	2.5059	0.1165	0.1512
	Z	-4.2423	4.6754	-0.0492	-0.0420
ETRI-Act (Jang et al., 2020)	X	-3.0403	3.3122	0.0158	0.0079
	Y	-2.9967	2.4528	0.1052	0.1049
	Z	-3.7749	2.9988	-0.0910	-0.0635
ETRI-LivingLab (Jang et al., 2020)	X	-3.1178	2.2961	0.0071	0.0030
	Y	-1.8460	2.5984	0.1005	0.1201
	Z	-3.0634	2.5608	-0.0704	-0.0564
20-Joint Protocols (Pre-train & Downstream)					
MSR-Action3D (Li et al., 2010)	X	-0.8821	1.0665	-0.0012	-0.0031
	Y	-1.8576	1.1964	-0.1543	-0.0096
	Z	-1.1771	2.7745	-0.0059	0.0132
NW-UCLA (Wang et al., 2014)	X	-3.1459	3.1226	-0.0090	-0.0079
	Y	-1.7658	1.3540	-0.1120	-0.0121
	Z	-2.6875	2.3884	-0.0016	0.0026
UT-Kinect (Xia et al., 2012a)	X	-1.8810	2.1565	-0.0431	-0.0084
	Y	-1.3734	1.2126	-0.1029	-0.0028
	Z	-2.2119	2.2332	-0.0390	-0.0062
15-Joint Protocols (Unseen)					
SBU-Inter (Yun et al., 2012)	X	-2.3407	2.1743	-0.0727	-0.0425
	Y	-1.2912	1.0371	0.0211	0.0989
	Z	-1.0182	0.9315	0.0163	0.0242
Florence (Seidenari et al., 2013a)	X	-0.7025	0.7916	0.0071	0.0116
	Y	-1.6579	1.1580	-0.0384	0.0046
	Z	-0.8494	0.7731	-0.0344	-0.0145

a high-dimensional prototype space with $K = 65{,}536$ prototypes, enabling the model to capture diverse motion patterns. Accordingly, the probability distributions $\mathbf{p}$ in the two distillation losses are K-dimensional.

Training objectives. The framework is optimized with a dual-objective loss combining the Canonical Slot Reconstruction loss $\mathcal{L}_{\text{Cano}}$ and the DINO-based distillation loss $\mathcal{L}_{\text{DINO}}$ (Caron et al., 2021) with an equal weighting factor ($\lambda = 1$). To further regularize the embedding space and encourage features to be uniformly distributed on the unit hypersphere, we additionally employ the KoLeo regularization loss (Oquab et al., 2023) with weight 0.1, and apply the Sinkhorn–Knopp algorithm (Oquab et al., 2023) for prototype-assignment regularization, which prevents mode collapse and promotes balanced usage of the prototype space.

Temporal sampling. To enable batch-wise training across datasets with different sequence lengths and frame rates, we standardize the temporal resolution of all inputs. For each raw sequence, we randomly crop a temporal window spanning 50% to 100% of the original duration and uniformly resample it to a fixed length of $T = 64$ frames, yielding a unified input tensor $\mathbf{X} \in \mathbb{R}^{T \times J_{\max} \times 3}$.

Augmentation and masking. To improve robustness and generalization, we apply rotation, scaling, spatial flipping, and axis dropout, each with probability $p = 0.5$. Our masking serves two purposes. First, for general feature robustness, we apply joint-wise masking as a data augmentation with probability $p = 0.5$. Second, for the Canonical Slot Reconstruction objective ($\mathcal{L}_{\text{Cano}}$), we always apply masking ($p = 1.0$) with a high masking ratio of 90%. By exposing the student network to only a severely corrupted subset of the input, the model is encouraged to internalize human kinematics and infer the missing information required to reconstruct the Canonical Joint Slots from sparse visible cues.

REFERENCES

Mohamed Abdelfattah and Alexandre Alahi. S-jepa: A joint embedding predictive architecture for skeletal action recognition. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13359–13368, 2021.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE transactions on cybernetics*, 45(7):1340–1352, 2014.

Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. In *European Conference on Computer Vision*, pp. 401–420. Springer, 2024.

Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 525–533, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2969–2978, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv preprint arXiv:2303.06242*, 2023.

Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 762–770, 2022.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jinhyeok Jang, Dohyung Kim, Cheonshu Park, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10990–10997. IEEE, 2020.

Boeun Kim, Hyung Jin Chang, Jungho Kim, and Jin Young Choi. Global-local motion transformer for unsupervised skeleton-based action learning. In *European conference on computer vision*, pp. 209–225. Springer, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Ce Li, Chunyu Xie, Baochang Zhang, Jungong Han, Xiantong Zhen, and Jie Chen. Memory attention networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4800–4814, 2021a.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE international conference on multimedia & expo workshops (ICMEW)*, pp. 597–600. IEEE, 2017.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018a.

Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. *arXiv preprint arXiv:2104.14466*, 2021b.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5457–5466, 2018b.

Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 9–14. IEEE, 2010.

Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2363–2372, 2023.

Lilang Lin, Lehong Wu, Jiahang Zhang, and Jiaying Liu. Idempotent unsupervised representation learning for skeleton-based action recognition. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017a.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pp. 816–833. Springer, 2016.

Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017b.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Diogo Carbonera Luvizon, Hedi Tabia, and David Picard. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, 99:13–20, 2017.

Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, pp. 734–752. Springer, 2022.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10181–10191, 2023.

Qiang Nie, Ziwei Liu, and Yunhui Liu. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *European conference on computer vision*, pp. 102–118. Springer, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 716–723, 2013.

Yatian Pang, Wenxiao Wang, Francis E.H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, 2022a.

Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *European Conference on Computer Vision*, pp. 605–622. Springer, 2022b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Mauricio Perez, Jun Liu, and Alex C Kot. Interaction relational network for mutual action recognition. *IEEE Transactions on Multimedia*, 24:366–376, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Real time action recognition using histograms of depth gradients and random decision forests. In *IEEE winter conference on applications of computer vision*, pp. 626–633. IEEE, 2014.

Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 479–485, 2013a.

Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 479–485, 2013b.

Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18846–18856, 2023.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9631–9640, 2020.

Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2973–2984, 2023.

Shengkai Sun, Zefan Zhang, Jianfeng Dong, Zhiyong Cheng, Xiaojun Chang, and Meng Wang. Towards efficient general feature prediction in masked skeleton modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12212–12221, 2025.

Shengkai Sun, Zhiyong Cheng, Zefan Zhang, Jianfeng Dong, Zhihui Li, and Meng Wang. Exploring adaptive masked reconstruction for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13974–13983, 2026.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1655–1663, 2021.

Quang D Tran and Ngoc Q Ly. Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences. In *The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*, pp. 253–258. IEEE, 2013.

Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588–595, 2014.

Chunyu Wang, John Flynn, Yizhou Wang, and Alan Yuille. Recognizing actions in 3d using action-snippets and activated simplices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Hongsong Wang and Liang Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018.

Hongsong Wang, Xiaoyan Ma, Jidong Kuang, and Jie Gui. Heterogeneous skeleton-based action representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19154–19164, 2025a.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2649–2656, 2014.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9888–9899, 2025b.

Yuhang Wen, Zixuan Tang, Yunsheng Pang, Beichen Ding, and Mengyuan Liu. Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7886–7892. IEEE, 2023.

Wanjiang Weng, Hongsong Wang, Junbo Wang, Lei He, and Guo-Sen Xie. Usdrl: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8332–8340, 2025.

Lehong Wu, Lilang Lin, Jiahang Zhang, Yiyang Ma, and Jiaying Liu. Macdiff: Unified skeleton modeling with masked conditional diffusion. In *European Conference on Computer Vision*, pp. 110–128. Springer, 2024.

Wenhan Wu, Yilei Hua, Ce Zheng, Shiqian Wu, Chen Chen, and Aidong Lu. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. In *2023 IEEE international conference on multimedia and expo workshops (ICMEW)*, pp. 224–229. IEEE, 2023.

Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 20–27. IEEE, 2012a.

Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 20–27. IEEE, 2012b.

Yangyang Xu, Jun Cheng, Lei Wang, Haiying Xia, Feng Liu, and Dapeng Tao. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. *IEEE Signal Processing Letters*, 25(7):1044–1048, 2018.

Ziwei Xu, Xudong Shen, Yongkang Wong, and Mohan S Kankanhalli. Unsupervised motion representation learning with capsule autoencoders. *Advances in Neural Information Processing Systems*, 34:3205–3217, 2021.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13423–13433, 2021.

Siyuan Yang, Jun Liu, Shijian Lu, Er Meng Hwa, Yongjian Hu, and Alex C Kot. Self-supervised 3d action representation learning with skeleton cloud colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):509–524, 2023.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19313–19322, 2022.

Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 28–35. IEEE, 2012.

Haoyuan Zhang, Yonghong Hou, Wenjing Zhang, and Wanqing Li. Contrastive positive mining for unsupervised 3d action representation learning. In *European Conference on Computer Vision*, pp. 36–51. Springer, 2022.

Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. Self-supervised skeleton-based action representation learning: A benchmark and beyond: J. zhang et al. *International Journal of Computer Vision*, 134(1):38, 2026.

Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*, pp. 2117–2126, 2017.

Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. Self-supervised action representation learning from partial spatio-temporal skeleton sequences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3825–3833, 2023.

Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13913–13922, 2023.