Neurosymbolic Framework for Concept-Driven Logical Reasoning in Skeleton-Based Human Action Recognition

Talha Ilyas[1,2] , **Deval Mehta**[2,3*] , **Zongyuan Ge**[2,3]
[1]Department of ECSE, Faculty of Engineering, Monash University, Australia
[2]AIM for Health Lab, Faculty of Information Technology, Monash University, Australia
[3]Department of DSAI, Faculty of Information Technology, Monash University, Australia
talha.ilyas@monash.edu; deval.mehta1092@gmail.com

Abstract

Skeleton-based human activity recognition has achieved strong empirical performance, yet most existing models remain black boxes and difficult to interpret. In this work, we introduce a neurosymbolic formulation of skeleton-based HAR that reframes action recognition as concept-driven first-order logical reasoning over motion primitives. Our framework bridges representation learning and symbolic inference by grounding first-order logic predicates in learnable spatial and temporal motion concepts. Specifically, we employ a standard spatio-temporal skeleton encoder to extract latent motion representations, which are then mapped to interpretable concept predicates via a spatio-temporal concept decoder that explicitly separates pose-centric and dynamics-centric abstractions. These concept predicates are composed through differentiable first-order logic layers, enabling the model to learn human-readable logical rules that govern action semantics. To impose semantic structure on the learned concepts, we align skeleton representations with LLM-derived descriptions of atomic motion primitives, establishing a shared conceptual space for perception and reasoning. Extensive experiments on NTU RGB+D 60/120 and NW-UCLA demonstrate that our approach achieves competitive recognition performance while providing explicit, interpretable explanations grounded in logical structure. Our results highlight neurosymbolic reasoning as an effective paradigm for interpretable spatio-temporal action understanding. Code[†].

1 Introduction

Skeleton-based human activity recognition (HAR) provides an efficient, privacy-preserving alternative to video-based methods by focusing on joint coordinates that isolate human structure from environmental noise [Duan *et al.*, 2022; Liu *et al.*, 2023; Asif *et al.*, 2023]. While deep learning has advanced classification accuracy, these models remain black boxes, problematic for high-stakes domains such as healthcare and rehabilitation where transparent decision-making is essential [Ilyas *et al.*, 2025; Mehta *et al.*, 2023].

Neurosymbolic AI has gained traction for combining neural perception with formal logic interpretability [Garcez *et al.*, 2002; Bhuyan *et al.*, 2024]. However, its application to skeleton-based HAR remains largely unexplored due to fundamental integration challenges. This work addresses these challenges to achieve interpretable skeleton-based HAR.

Symbol Grounding. A key challenge is bridging continuous joint trajectories and discrete symbolic predicates. Traditional joint heuristics (e.g., knee angle $< 90^o$ for *bent knee*) [Okamoto and Parmar, 2024] are brittle, as small pose errors or variations in viewpoint and execution can prevent predicate activation for reliable HAR. Inspired by Concept Bottleneck Models (CBMs) [Koh *et al.*, 2020], we instead learn abstract motion concepts and construct an LLM-generated concept bank of spatial primitives (e.g., `arm_raise`, `knee_bend`) that serve as symbolic predicates. We also enforce semantic grounding by aligning GCN-extracted skeleton features with pretrained text encoder based concept embeddings.

Temporal Dynamics. Actions are inherently temporal, varying in speed, duration, and execution order. WEARINGGLASSES and TAKEOFFGLASSES share similar static poses (`hands near the face`, `arm engagement`) but differ in temporal dynamics: the former involves grasping and upward motion, while the latter requires release and downward motion. Purely spatial logic cannot capture such distinctions, and incorporating temporal logical rules into neurosymbolic frameworks is challenging. To address this, we introduce explicit temporal concepts capturing motion direction (e.g., `motion_upward`, `motion_downward`), hand state transitions (e.g., `hand_grasp`, `hand_release`), and execution order (e.g., `seq_hands_first`). Our Spatio-Temporal Concept Decoder (STC-Decoder) then independently decodes spatial and temporal concepts, enabling distingushing semantically similar actions such as (`arm_raise` $\wedge$ `hand_grasp` $\wedge$ `motion_upward`) $\Rightarrow$ WEARGLASSES versus (`arm_lower` $\wedge$ `hand_release` $\wedge$ `motion_downward`) $\Rightarrow$ TAKEOFFGLASSES (Fig 1(b)).

End-to-End Learning. Discrete Boolean operators block gradient propagation, preventing joint optimization of per-

*Corresponding author, now at School of Computing Technologies, RMIT

†Extended version of this paper can be found at: Arxiv

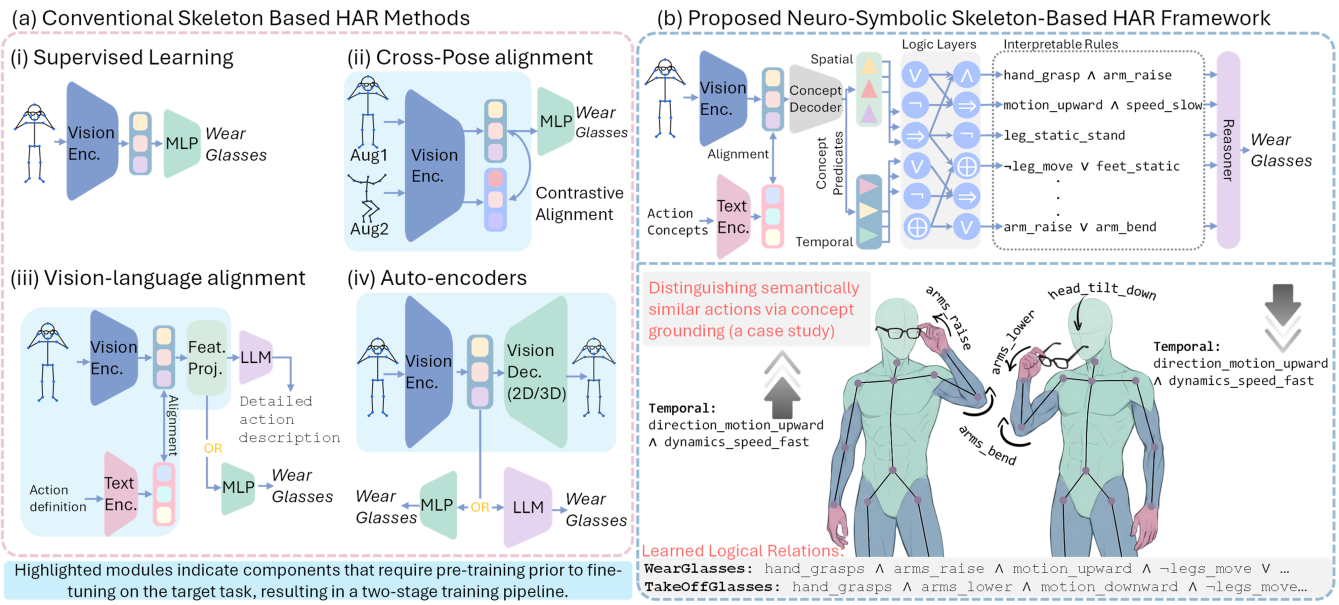

Figure 1: **Our proposed concept-grounded neurosymbolic reasoning for skeleton-based activity recognition:** Grounding spatial–temporal motion concepts in differentiable logic to learn interpretable and compositional action rules.

ception and reasoning components [Schrouff *et al.*, 2021]. While differentiable fuzzy logic [Barbiero *et al.*, 2023] shows promise, scalability to large-scale, fine-grained skeleton-based HAR remains open. We employ differentiable logic layers with soft AND/OR operations [Perfilieva, 2002] and gradient grafting [Wang *et al.*, 2023], enabling discrete inference with end-to-end differentiability. Final predictions aggregate triggered rules via learned weights, defining actions as logical compositions of motion concepts.

Together, these components form an end-to-end neurosymbolic framework (Fig 1(b)) in which a GCN encoder extracts spatio-temporal features with sequence–concept alignment, the STC-Decoder predicts interpretable concepts, and differentiable logic layers compose them via first-order operations ($\wedge$, $\vee$, $\neg$) to derive action predictions for action recognition. Grounding predictions in logical rules over concepts provides both local (concepts) and global interpretability (rules). Experiments on NTU RGB+D 60/120 [Shahroudy *et al.*, 2016; Liu *et al.*, 2019] and NW-UCLA [Wang *et al.*, 2014] demonstrate we achieve competitive performance with transparent explanations. To summarize, our **major contributions** are:

- REASON (**R**ule-based **E**xplainable **A**ction via **S**ymbolic c**ON**cepts), a neurosymbolic framework for skeleton-based HAR that formulates action recognition as logical reasoning over learnable motion concepts and integrates concept-grounded representations with differentiable first-order logic for end-to-end training.
- A Spatio-Temporal Concept Decoder (STC-Decoder) that extracts spatial body-configuration and temporal motion-dynamics concepts, yielding interpretable abstractions robust to execution variability.
- A concept bank of spatial and dynamic concepts for actions available in NTU RGB+D and NW-UCLA.

2 Related Work

Skeleton-based HAR. ST-GCN [Yan *et al.*, 2018] established the GCN paradigm by modeling joints as graph nodes with spatio-temporal convolutions. Subsequent work (Fig 1(a)-(i),(ii)) addresses topology rigidity: 2s-AGCN [Shi *et al.*, 2019] learns adaptive joint relationships, CTR-GCN [Chen *et al.*, 2021] refines topology channel-wise, and InfoGCN [Chi *et al.*, 2022] incorporates information-theoretic objectives. Recent architectures like HD-GCN [Lee *et al.*, 2023] and BlockGCN [Zhou *et al.*, 2024] capture hierarchical interactions. Self-supervised approaches (Fig 1(a)-(iv)) including SkeletonMAE [Wu *et al.*, 2023] and MotionBERT [Zhu *et al.*, 2023] improve generalization via masked reconstruction. Despite strong performance, these models lack interpretable action semantics.

Vision-Language for HAR. Recent works have used LLMs for skeleton-based HAR (Fig 1(a)-(iii),(iv)). SUGAR [Ye *et al.*, 2025] and LLM-AR [Qu *et al.*, 2024] leverage LLM priors for classification or incorporate action prompts [Xiang *et al.*, 2023]. However, these methods yield generic, non-instance-specific explanations and cannot distinguish execution variants (e.g., arms-first vs. legs-first), and are impractical for edge deployment. In contrast, our framework learns efficient, compositional logical rules over motion predicates that capture multiple execution pathways.

Explainable HAR and Neurosymbolic AI. Prior interpretable HAR work has focused on RGB video via post-hoc concept attribution [Kowal *et al.*, 2024] or neurosymbolic pipelines with expert-defined rules [Okamoto and Parmar, 2024], inheriting privacy concerns and sensitivity to environmental noise. While neurosymbolic systems enable traceable reasoning [Hitzler and Sarker, 2022], skeleton data exhibit substantial subject and viewpoint variability, making hand-crafted rules brittle. We address this with learnable concept

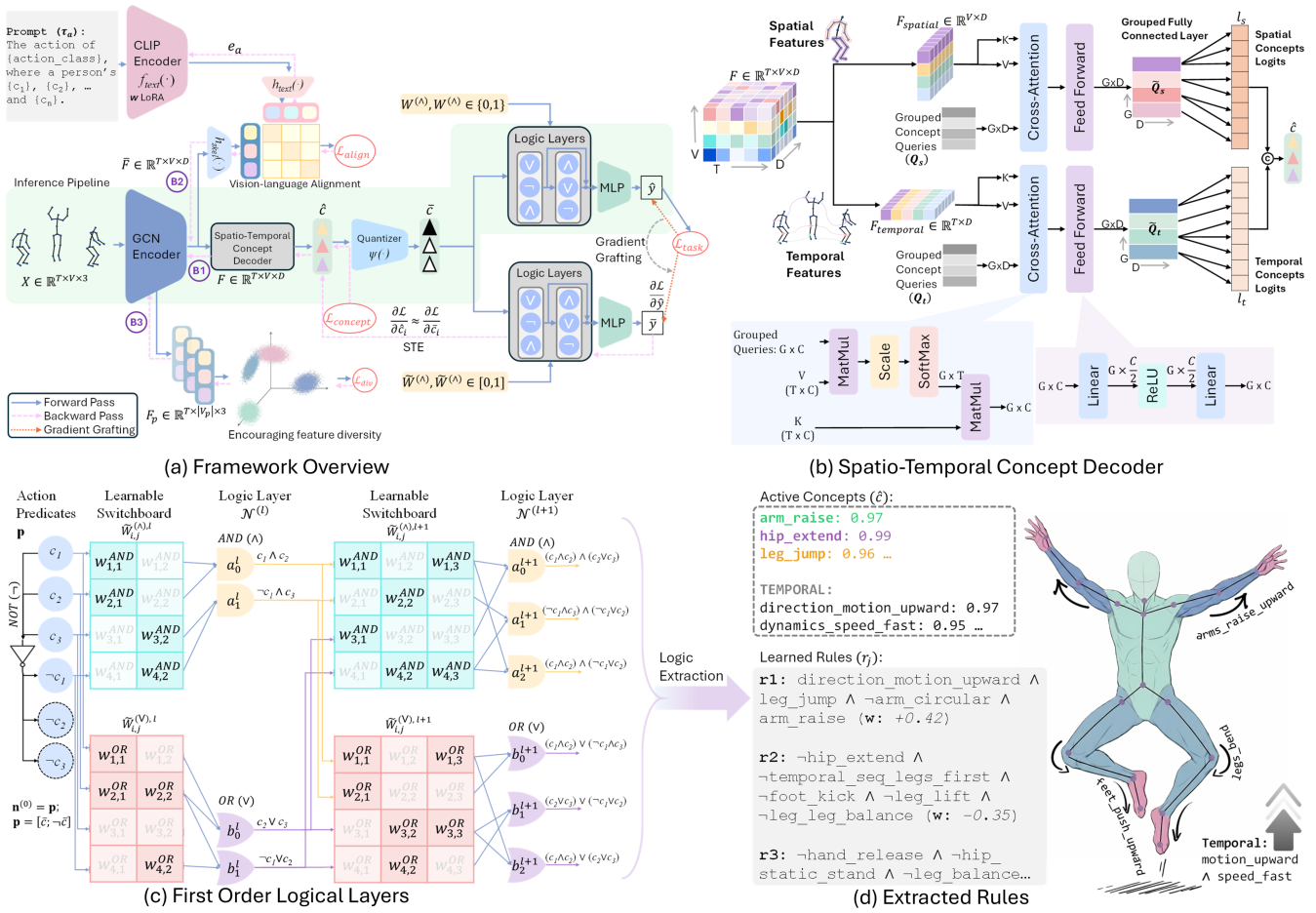

Figure 2: (a) Overall neurosymbolic, concept-grounded framework for skeleton-based human activity recognition. (b) STC-Decoder for decoupling encoder features into spatial and temporal streams to learn spatio-temporal motion concepts. (c) First-order logic layers, for learning logical predicates and compositional rules over predicted concepts. (d) Example visualization for a JUMP action, showing top activated concepts and learned rules.

predicates combined with differentiable logic layers to discover action-defining rules from data.

Concept Bottleneck Models. CBMs constrain representations to interpretable concept layers, enabling intervention and debugging [Koh *et al.*, 2020], with extensions like Logic CBMs [Vemuri *et al.*, 2025] incorporating differentiable logic for concept compositions. Recent work further automates concept bank construction using LLM priors [Mehta *et al.*, 2025; Jiang *et al.*, 2025a], reducing manual effort. Despite success in static image classification tasks [Jiang *et al.*, 2025b], CBMs have not addressed skeleton-based HAR, which requires joint modeling of spatial and temporal dynamics. We bridge this gap by formulating skeleton-based HAR as logical reasoning over learnable motion concepts.

3 Method

Our work centres on the question: **Can concepts serve as an intermediate representation for neurosymbolic skeleton-based HAR?** We address this by first constructing a concept bank of spatial and temporal motion primitives (Sec. 3.1), and then developing a framework to predict concepts from skeleton sequences (Sec. 3.2& 3.3) and composing them via differentiable logic layers (Sec. 3.4& 3.5) for interpretable HAR.

3.1 Concept Bank Generation

Unlike image classification where concepts correspond to static attributes (e.g., a zebra has `stripes`), action recognition requires capturing spatio-temporal dynamics. Furthermore, execution variability within an action (e.g., arm-assisted v/s arm-free JUMP) makes manual concept enumeration infeasible. We address this by constructing a concept bank $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_t \cup \mathcal{C}_{int}$ of spatial, temporal, and interactive motion concepts through a three-stage LLM-based pipeline.

Fine-grained Action Description Generation

Following HAKE [Li *et al.*, 2022], we decompose actions into six body parts: $\mathcal{P} = \{\texttt{head}, \texttt{hand}, \texttt{arm}, \texttt{hip}, \texttt{leg}, \texttt{foot}\}$. We then prompt an LLM (GPT-3) to generate part-wise motion descriptions $\mathcal{D}_a = \{d_a^{(p)}\}_{p\in\mathcal{P}}$ for each action a. For example, JUMP yields: {`Leg`: bends at knees then pushes explosively; `Arm`:

extends overhead; `Foot`: pushes off ground}.

From Action Descriptions to Abstract Concepts

Generated raw descriptions are action-specific. Thus, using them directly could cause vocabulary explosion [Jiang *et al.*, 2025a]. Hence, we convert descriptions into canonical concepts by extracting {`verb-direction-modifier`} patterns from all actions, then embedding them via a sentence encoder ϕ, and clustering them with k-means. Cluster representatives form spatial concepts $\mathcal{C}_{\text{s}}$ (e.g., `arm_raise`, `leg_jump`, `hand_grasp`). Additionally, we define temporal concepts $\mathcal{C}_{\text{t}}$ capturing motion direction, sequence, and dynamics (e.g., `motion_upward`, `hand_first`, `speed_fast`), and an interaction concept for mutual actions.

Concept-Action Association and Refinement

We construct an association matrix $\mathbf{M} \in \{0,1\}^{|\mathcal{A}|\times|\mathcal{C}|}$ where $\mathbf{M}_{a,c} = 1$ indicates concept c is active for action a. Using LLaMA-3.2 7B [Grattafiori *et al.*, 2024], we match each action's descriptions to semantically appropriate concepts, allowing multiple concepts per body part for compositional movements (e.g., JUMP activates both `leg_squat` and `leg_jump`). To ensure discriminability, we verify unique concept signatures and invoke refinement for ambiguous pairs (e.g., WEARGLASSES vs. TAKEOFFGLASSES are distinguished via opposing direction concepts, see Figure 1). Our final concept bank comprises $|\mathcal{C}| = 74$ concepts (52 spatial + 21 temporal + 1 interaction). Implementation details, prompts, pseudocode and the complete vocabulary are provided in Appendix A.

3.2 Skeleton Representation Learning

After constructing the concept bank, we describe the overall framework (Fig 2(a)). Skeleton sequences are first encoded using a GCN and aligned with concept semantics via a concept-grounded text encoder and a contrastive objective.

Skeleton Encoder Let $\mathbf{X} \in \mathbb{R}^{T\times V\times 3}$ denote an input skeleton sequence with T frames, V joints and *(x,y,z)* coordinates. We model the skeleton as a spatio-temporal graph and process it through a GCN backbone (e.g., Hyper-GCN [Zhou *et al.*, 2025]) with multi-scale temporal convolutions. Crucially, we retain full spatio-temporal resolution without pooling, outputting features $\mathbf{F} \in \mathbb{R}^{T\times V\times D}$ to preserve fine-grained motion information for concept prediction.

The encoder feeds into three branches (Fig 2(a)): **(B1)** the STC-Decoder for concept prediction, **(B2)** a text alignment branch where pooled features $\bar{\mathbf{F}} = \texttt{Pool}(\mathbf{F})$ are aligned with concept embeddings, and **(B3)** a part-divergence branch promoting discriminative representations across anatomical regions.

Body-Part Feature Divergence We partition joints into subsets $\{\mathcal{V}_p\}_{p\in\mathcal{P}}$ and extract part-specific features $\mathbf{F}_p \in \mathbb{R}^{T\times|\mathcal{V}_p|\times D}$. To prevent feature collapse across body parts, we introduce a divergence loss that penalizes high cosine similarity between pooled part representations. Let $\bar{\mathbf{f}}_p = \texttt{Pool}(\mathbf{F}_p) \in \mathbb{R}^D$ denote the feature of part (p). The loss minimizes off-diagonal similarities across all part pairs:

$$\mathcal{L}_{\text{div}} = \frac{1}{|\mathcal{P}|(|\mathcal{P}|-1)} \sum_{p\neq q} \text{ReLU}\left(\frac{\bar{\mathbf{f}}_p^\top \bar{\mathbf{f}}_q}{\|\bar{\mathbf{f}}_p\|\,\|\bar{\mathbf{f}}_q\|}\right) \tag{1}$$

This regularization encourages each body part to learn discriminative features aligned with its corresponding concept subset $\mathcal{C}_p \subset \mathcal{C}_{\text{spatial}}$.

Concept-Grounded Text Encoder To ground skeleton features in semantic space, we encode textual descriptions using CLIP [Radford *et al.*, 2021] with LoRA fine-tuning [Hu *et al.*, 2022]. Crucially, rather than generic action labels, we construct concept-grounded prompts incorporating motion primitives from our concept bank. For action a with associated concepts $\{c_i\}$ from matrix $\mathbf{M}$, we construct: *"A person performing [action], where the [part$_1$] [concept$_1$], the [part$_2$] [concept$_2$], ..."*. For example, JUMP yields: "A person performing JUMP, where the `arms` *swing upward*, the `knees` *bend*, and the `body` *moves upward*." This formulation ensures text embeddings $\mathbf{e}_a = f_{\text{text}}(\tau_a)$ encode the same compositional semantics as the concept vocabulary.

We then align skeleton and text representations via contrastive learning. Projected skeleton features $\mathbf{z} = h_{\text{skel}}(\bar{\mathbf{F}})$ and concept embeddings $\mathbf{t} = h_{\text{text}}(\mathbf{e}_a)$ are aligned through:

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{t}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^\top \mathbf{t}_j/\tau)} \tag{2}$$

where τ is a learnable temperature. This objective encourages skeleton features to cluster with their concept-grounded descriptions, reinforcing the compositional structure.

3.3 STC-Decoder

The spatio-temporal concept decoder (STC-Decoder, Fig. 2b) maps continuous skeleton sequences to discrete concept activations, bridging skeleton representations and symbolic predicates. As actions activate multiple concepts, we adopt a multi-label design with separate spatial and temporal pathways.

Spatial-Temporal Feature Decoupling. Given encoder features $\mathbf{F} \in \mathbb{R}^{T\times V\times D}$, we decouple them through dimension-wise aggregation:

$$\mathbf{F}_{\text{s}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}[t,:,:] \in \mathbb{R}^{V\times D},\ \mathbf{F}_{\text{t}} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{F}[:,v,:] \in \mathbb{R}^{T\times D} \tag{3}$$

Spatial features $\mathbf{F}_{\text{s}}$ preserve per-joint information by aggregating across time, while temporal features $\mathbf{F}_{\text{t}}$ retain motion dynamics by aggregating across joints.

Group-Decoding with Concept Queries. Predicting all $|\mathcal{C}|$ concepts jointly poses an optimization challenge: concepts exhibit heterogeneous semantic relationships, some are mutually exclusive (`motion_upward` vs. `motion_downward`), others frequently co-occur (`arm_raise`, `hand_grasp`), and many are conditionally dependent on body-part context. Learning these complex inter-concept relationships in a single framework is difficult.

We address this by partitioning concepts into G groups, where each group predicts $g = \lceil|\mathcal{C}|/G\rceil$ concepts. Rather

than manually assigning semantic roles, we let each group *learn* to specialize through end-to-end training, groups naturally discover coherent concept subsets (e.g., related body-part configurations or temporal patterns) that exhibit homogeneous co-occurrence, making each sub-problem easier to optimize.

We initialize separate learnable query matrices $\mathbf{Q}_\text{s} \in \mathbb{R}^{G_\text{s} \times D}$ and $\mathbf{Q}_\text{t} \in \mathbb{R}^{G_\text{t} \times D}$ for spatial and temporal branches. Each branch performs cross-attention followed by a feed-forward network, where group queries learn to selectively aggregate skeleton information relevant to their respective concept subsets (Appendix C).

Group Fully-Connected Layer. A Group-FC layer [Ridnik *et al.*, 2023] projects the refined group queries to per-concept predictions. Each group $k \in \{0, \ldots, G-1\}$ maintains its own projection $\mathbf{W}_k \in \mathbb{R}^{D \times g}$, responsible for predicting g concepts. For concept i, we determine its group index $k = \lfloor i/g \rfloor$ and position within the group $j = i \mod g$, then compute:

$$\ell_i = \tilde{\mathbf{Q}}[k, :] \cdot \mathbf{W}_k[:, j] \tag{4}$$

This factorization encourages each group query to specialize in its concept subset through gradient-driven learning. Aggregating spatial and temporal logits yields:

$$\hat{\mathbf{c}} = \sigma([\boldsymbol{\ell}_\text{s}; \boldsymbol{\ell}_\text{t}]) \in [0, 1]^{|\mathcal{C}|} \tag{5}$$

where $\hat{c}_i$ denotes the predicted probability of concept c_i, providing soft symbolic grounding for the logic layers.

Concept Prediction Loss. Concept prediction is supervised using the action–concept matrix $\mathbf{M}$. For an action label a, the target concept vector is $\mathbf{c}^* = \mathbf{M}[a, :] \in \{0, 1\}^{|\mathcal{C}|}$, and training minimizes binary cross-entropy $\mathcal{L}_\text{concept} = \text{BCE}(\mathbf{c}^*, \hat{\mathbf{c}})$. This supervision grounds continuous skeleton features into interpretable concept activations $\hat{\mathbf{c}}$, which serve as symbolic predicates for logical composition.

3.4 Differentiable First-Order Logic Reasoning

Symbolic reasoning requires binary predicates, whereas our predicted concepts $\hat{\mathbf{c}} \in [0, 1]^{|\mathcal{C}|}$ provide soft activations over the motion vocabulary. We therefore binarize concept predictions and compose them using first-order logic, while also employing continuous relaxations to enable end-to-end learning and preserve discrete, interpretable rules.

Concept Binarization

Concept leakage [Mahinpei *et al.*, 2021] occurs when downstream layers exploit spurious information in soft activations rather than intended semantics. We therefore employ a concept activation binarizer which applies a threshold $\bar{c}_i = \mathbb{1}[\hat{c}_i > 0.5]$, yielding $\bar{\mathbf{c}} \in \{0, 1\}^{|\mathcal{C}|}$. Furthermore, to enable logical negation of concepts, we augment the predicate space as:

$$\mathbf{p} = [\bar{\mathbf{c}}; \neg\bar{\mathbf{c}}] \in \{0, 1\}^{2|\mathcal{C}|} \tag{6}$$

allowing rules involving both presence and absence of concepts (e.g., `leg_jump` $\wedge$ $\neg$`leg_static`).

Logical Layer Architecture

We compose concept predicates into action-discriminative rules through stacked logical layers (Fig. 2(c)). Each layer $\mathcal{N}^{(l)}$ contains two sub-layers: conjunction (AND) and disjunction (OR), enabling arbitrary propositional formulas in disjunctive normal form [Wang *et al.*, 2023]. Let $\mathbf{n}^{(l-1)}$ denote the input (with $\mathbf{n}^{(0)} = \mathbf{p}$). The conjunction ($\mathcal{A}^{(l)}$) and disjunction layers ($\mathcal{B}^{(l)}$) compute:

$$a_i^{(l)} = \bigwedge_{j: W_{ij}^{(\wedge)}=1} n_j^{(l-1)}, \qquad b_i^{(l)} = \bigvee_{j: W_{ij}^{(\vee)}=1} n_j^{(l-1)} \tag{7}$$

where $\mathbf{W}^{(\wedge)}, \mathbf{W}^{(\vee)} \in \{0, 1\}$ are learnable switchboards (i.e., binary adjacency matrices) specifying which predicates participate in each conjunction/disjunction. By stacking L layers with outputs $\mathbf{n}^{(l)} = [\mathbf{a}^{(l)}; \mathbf{b}^{(l)}]$, the network learns compositional rules of increasing complexity, yielding rule activations $\mathbf{r} \in \{0, 1\}^R$.

Continuous Relaxation and Gradient Grafting

Discrete Boolean operations in logical layer produce zero gradients. For end-to-end training, we adopt differentiable logic relaxations [Barbiero *et al.*, 2023] that approximate AND/OR via product-based formulations. During training, we maintain continuous weights $\tilde{\mathbf{W}}^{(\wedge)}, \tilde{\mathbf{W}}^{(\vee)} \in [0, 1]$ and compute soft operations:

$$\tilde{a}_i^{(l)} = \prod_j \left(1 - \tilde{W}_{ij}^{(\wedge)}(1 - \tilde{n}_j^{(l-1)})\right) \tag{8}$$

$$\tilde{b}_i^{(l)} = 1 - \prod_j \left(1 - \tilde{W}_{ij}^{(\vee)} \tilde{n}_j^{(l-1)}\right) \tag{9}$$

When weights and inputs are binary, these reduce to standard AND/OR (Appendix D). Logical layer training employs gradient grafting [Wang *et al.*, 2023] (Fig 2(a)): the forward pass uses binarized concepts $\bar{\mathbf{c}}$ and discrete weights $\mathbf{W} = \mathbb{1}[\tilde{\mathbf{W}} > 0.5]$ for interpretable rule evaluation, while backpropagation computes gradients through the continuous relaxations (Eq. 8and 9) to update soft weights $\tilde{\mathbf{W}}$. For the concept binarizer, the Straight-Through Estimator [Bengio *et al.*, 2013] passes gradients unchanged: $\partial\mathcal{L}/\partial\hat{c}_i \approx \partial\mathcal{L}/\partial\bar{c}_i$, as shown in Fig 2(a).

3.5 Rule-Based Action Classification

The rule activations $\mathbf{r} \in \{0, 1\}^R$ encode which logical concept combinations are present. A linear classifier maps these to action predictions $\hat{\mathbf{y}} = \mathbf{V}\mathbf{r} + \mathbf{b}$, where $\mathbf{V} \in \mathbb{R}^{|\mathcal{A}| \times R}$, associates rules with actions, and $\mathbf{b} \in \mathbb{R}^{|\mathcal{A}|}$ is a bias term. Post-training, we extract human-readable rules by tracing adjacency matrices back to concept combinations, yielding expressions like r_1 : `leg_jump` $\wedge$ $\neg$`arm_swing` $\wedge$ `motion_upward`. This ensures both *local interpretability* (instance-specific concept activations) and *global interpretability* (logical rules characterizing each action class). The details for post-training rule interpretation are provided in Appendix D.

Overall Training Objective The framework is trained end-to-end by combining all loss terms: $\mathcal{L} = \mathcal{L}_\text{task} + \alpha\mathcal{L}_\text{concept} + \beta\mathcal{L}_\text{align} + \gamma\mathcal{L}_\text{div} + \lambda\|\tilde{\mathbf{W}}\|_1$, where the ℓ_1 regularization on $\tilde{\mathbf{W}}$ enforces sparse, interpretable rules. At inference, concepts are binarized, rules are evaluated discretely, and actions are predicted via triggered rules.

Cat.	Methods	Venue	Mod.	NTU-RGB+D 60		NTU-RGB+D 120		NW-UCLA	Two-stage	Interpret.
				X-Sub	X-View	X-Sub	X-Set			
GCN	ST-GCN	AAAI'18	J+B	81.5	88.3	70.7	73.2	–	N	N
	2s-AGCN	CVPR'19	J+B	88.5	95.1	82.5	84.2	–	N	N
	MS-G3D	CVPR'20	J+B	86.0	94.1	80.2	86.1	–	N	N
	UNIK	BMVC'21	J+B	86.8	94.4	80.8	86.5	–	N	N
	InfoGCN	CVPR'20	J+B	–	–	88.5	89.7	–	N	N
	REASON (Ours)	This work	J+B	89.4	95.6	88.97	89.2	–	N	Y
	DC-GCN+ADG	ECCV'20	J+B+JM+BM	90.8	96.6	86.5	88.1	95.3	N	N
	MS-G3D	CVPR'20	J+B+JM+BM	91.5	96.2	86.9	88.4	–	N	N
	MST-GCN	Access'21	J+B+JM+BM	91.5	96.6	87.5	88.8	–	N	N
	CTR-GCN	ICCV'21	J+B+JM+BM	92.4	96.4	88.9	90.6	96.5	N	N
	EfficientGCN-B4	TPAMI'22	J+B+JM+BM	91.7	95.7	88.3	89.1	–	N	N
	InfoGCN	CVPR'20	J+B+JM+BM	92.7	96.9	89.4	90.7	96.6	N	N
	FR Head	CVPR'23	J+B+JM+BM	92.8	96.8	89.5	90.9	96.8	N	N
	HD-GCN*	ICCV'23	J+B+J'+B'	93.0	97.0	89.8	91.2	96.9	N	N
	DS-GCN	AAAI'24	J+B+JM+BM	93.1	97.5	89.2	90.3	–	N	N
	BlockGCN	CVPR'24	J+B+JM+BM	93.1	97.0	90.3	91.5	96.9	N	N
	Hyper-GNN	TIP'21	J+B+JM+BM	89.5	95.7	–	–	–	N	N
	Selective-HCN	ICMR'22	J+B+JM+BM	90.8	96.6	–	–	–	N	N
	DST-HCN	ICME'23	J+B+JM+BM	92.3	96.8	88.8	90.7	96.6	N	N
	Hyper-GCN	ICCV'25	J+B+JM+BM	93.7	97.8	90.9	92.0	97.6	N	N
	REASON (Ours)	This work	J+B+JM+BM	94.3	97.8	90.1	91.9	97.9	N	Y
Trans.	DSTA-Net	ACCV'20	J+B+JM+BM	89.5	95.7	86.6	89.0	–	N	N
	IIP-Transformer	DATA'23	J+B+JM+BM	89.5	95.7	89.9	90.9	–	N	N
	SkateFormer	ECCV'24	J+B+JM+BM	93.5	97.8	89.8	91.4	**98.3**	N	N
	USDRL	TPAMI'25	J+M+B	87.1	93.2	79.3	80.6	–	Y	N
LLM	GAP	ICCV'23	J+B+JM+BM	92.9	97.0	89.9	91.1	97.2	N	N
	LLM-AR	CVPR'24	J	95.0	**98.4**	88.7	91.5	–	Y	Y
	SUGAR	AAAI'26	J	95.2	97.8	90.1	89.7	–	Y	Y
	REASON+LLM (Ours)	This work	J	**95.6**	98.3	**91.2**	**92.8**	98.1	N	Y

Table 1: Comparison with state-of-the-art methods on NTU RGB+D 60/120, and NW-UCLA. **Cat.**: Categories. **Trans.**: Transformers **Mod.**: input modalities (J: Joint, B: Bone, JM: Joint Motion, BM: Bone Motion). **Two-stage**: requires separate pretraining. **Interpret.**: provides interpretable predictions. Best performance in **bold**, while second-best is underlined.

4 Experiments

We evaluate our neurosymbolic framework on standard skeleton-based HAR benchmarks, focusing on: (i) performance competitiveness, (ii) interpretability of learned rules, and (iii) component-wise contributions.

4.1 Datasets

We evaluate on three benchmarks: **NTU RGB+D 60** (56,880 sequences, 60 classes) under Cross-Subject (X-Sub) and Cross-View (X-View) protocols; **NTU RGB+D 120** (114,480 sequences, 120 classes) under Cross-Subject (X-Sub) and Cross-Setup (X-Set) protocols; and **NW-UCLA** (1,494 sequences, 10 classes) following protocol to train on cameras 1&2, test on camera 3.

4.2 Implementation Details

We follow [Chen *et al.*, 2021] for preprocessing, using random rotation, spatial flip, and temporal cropping with uniform 64-frame sampling. Training uses AdamW for 400 epochs with cosine decay on a single A6000 GPU.

We employ *staggered warmup* to prevent spurious rule formation: the GCN/CLIP/STC-Decoder warm up for 5 epochs while logic layers remain frozen for 15 epochs. This delay ensures logic layers learn from converged concept predictions rather than random activations. Logic layers use $10\times$ higher learning rate (10^{-4} vs. 10^{-5}) to traverse from dense initialization to sparse binary connectivity against ℓ_1 regularization. Full hyperparameters are in Appendix F.

4.3 Comparison with State-of-the-Art

Table 1 compares against GCN-based, Transformer-based, and LLM-augmented methods. Beyond performance, we highlight whether methods require *two-stage* training (separate pretraining) and whether they provide *interpretable* predictions.

Our core neurosymbolic framework (REASON) achieves competitive performance while being the only single-stage, fully interpretable approach. In 4-stream configuration, we attain 94.3% on NTU-60 X-Sub and 92.8% on NTU-120 X-Set, surpassing black-box methods (Hyper-GCN, BlockGCN). Critically, the concept-rule reasoning chain remains fully auditable, distinguishing our approach from prior work.

LLM Re-ranking (Optional). To enable fair comparison with LLM-augmented methods (SUGAR, LLM-AR), we evaluate an optional re-ranking stage. Unlike these methods that use generic action descriptions, we ground a LoRA-finetuned LLaMA-3.2 7B in *instance-specific* concept activations and top-K classifier candidates (Appendix G). With LLM re-ranking, we achieve 95.6% (X-Sub) and 93.2% (X-Set), outperforming two-stage LLM methods, as we reason over interpretable concepts rather than opaque features.

Interpretable Reasoning via Concept-Rule Chains

Beyond accuracy, our framework provides transparent reasoning through learned concept-rule chains. Figure 3 visualizes the complete reasoning pipeline for representative actions. Skeleton poses are rendered with joint sizes scaled by concept activation confidence, revealing which body parts drive predictions. For BRUSHINGTEETH, the model activates intuitive concepts (`hand_lift_to_face`,

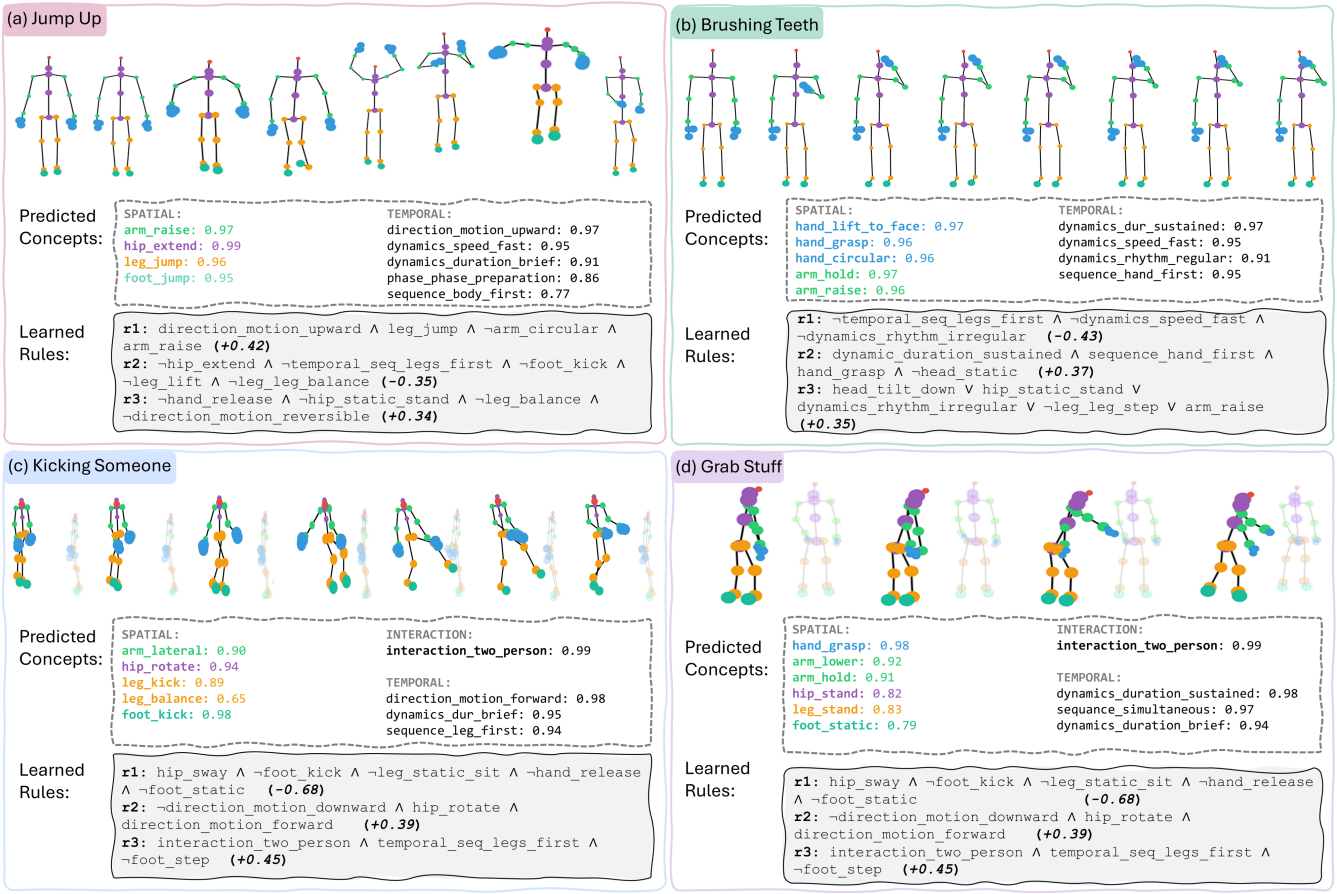

Figure 3: **Visualization of our concept-grounded neurosymbolic reasoning.** Each panel shows temporally sampled skeletons with concept activations, top spatial–temporal concepts (dotted box), and top-3 learned logical rules with weights (gray box).

duration_sustained, rhythm_regular) and applies learned rules such as r_1 : duration_sustained ∧ hand_grasp ∧ ¬head_static (+0.37). This end-to-end transparency, from skeleton input to concept activations to logical rule firings, enables systematic analysis of both correct classifications and failure cases, a capability absent in existing methods.

4.4 Ablation Studies

We conduct ablations on NTU RGB+D 120 (Table 2).

Backbone Generalization (Table 2a). Our framework is encoder-agnostic: across ST-GCN, 2s-AGCN, CTR-GCN, ST-GCN++, and Hyper-GCN, our approach consistently improves performance, demonstrating plug-and-play applicability for existing GCN encoders.

Component Analysis (Table 2b). Each module contributes incrementally: text alignment ($\mathcal{L}_{\text{align}}$) adds +0.9-1.2%, part divergence ($\mathcal{L}_{\text{div}}$) +0.4-1.0%, and STC-Decoder +0.2–1.6%.

STC-Decoder Configuration (Table 2c). Group-decoding with $G_s = 8$, $G_t = 4$ achieves optimal accuracy while cutting cross-attention cost $\sim 6\times$ vs. per-concept queries ($G = 74$).

Concept Vocabulary (Table 2d). Spatial concepts alone achieve 78.39%; adding temporal concepts yields +7.8% (86.15%), demonstrating that motion dynamics are critical for distinguishing actions with similar poses. Interaction concepts provide a further +1.2%.

Logic Layer Capacity (Table 2e). The 128-node configuration is optimal; smaller (64) lacks expressiveness while larger (256/512) overfits. Crucially, disabling negation causes catastrophic failure (49.87%), many actions require ¬concepts (e.g., STANDING needs ¬leg_walk). Skip connections contribute +3.6% by preserving atomic concept access.

Concept Intervention (Table 2f). Replacing mispredicted concepts $\hat{c}_i$ with ground-truth c_i^* improves accuracy monotonically: +1.9% (1 concept), +3.9% (2), +5.7% (3). This confirms that logic layers capture faithful concept–action relationships and errors stem from concept misprediction rather than flawed rules, enabling human-in-the-loop correction. Overall process of concept intervention is in Appendix H.

Feature Space Analysis Figure 4 compares t-SNE embeddings of Hyper-GCN features and STC-Decoder concept activations. Actions overlapping in the Hyper-GCN space (e.g., READING/WRITING, PUTON/TAKEOFFSHOE, THUMBUP/DOWN) are well separated in the concept space by discriminative temporal concepts (e.g., motion_forward vs. motion_backward), with similar separation for mutual actions (e.g., EXCHANGINGTHINGS, SHAKINGHANDS).

Backbone	Accuracy (%)		
	w/o $\mathcal{L}_{\text{align}}$	w/o $\mathcal{L}_{\text{div}}$	w/ both
ST-GCN	80.07	81.15	82.40
2s-AGCN	82.15	83.20	84.85
CTR-GCN	82.90	84.10	86.12
ST-GCN++	83.25	84.65	86.70
Hyper-GCN	84.08	85.27	**87.34**

(a) Backbone ablation

Model Configuration	X-Sub	X-Set
Hyper-GCN	84.28	84.08
+ Text Alignment ($\mathcal{L}_{\text{align}}$)	85.20	85.27
+ Part Divergence ($\mathcal{L}_{\text{div}}$)	86.20	85.70
+ STC-Decoder	86.38	87.34

(b) Component ablation

G_s	G_t	Total G	Acc. (%)	
			X-sub	X-set
Baseline (simple FC layer)			86.20	85.70
4	2	6	85.72	85.12
8	4	12	**86.38**	**87.34**
16	8	24	86.11	87.21
32	16	48	85.81	87.18
51	16	67	86.03	87.09

(c) Group query configuration

Concept Configuration	# Concepts	Accuracy
Spatial only	51	78.39
Spatial + Temporal	66	86.15
+ Interaction concepts	67	87.34

(d) Concept vocabulary composition

Config	NOT	Skip	L1/L2	Final	Acc.
Small	✓	✓	64	390	83.60
Medium	✓	✓	128	646	**87.34**
Large	✓	✓	256	1158	86.18
XLarge	✓	✓	512	2182	81.38
No Skip	✓	✗	128	256	83.77
No NOT	✗	✓	128	579	49.87

(e) Logic layer capacity

Intervention Level	Accuracy (%)	Δ Acc.
No intervention	87.35	–
Correct 1 concept	89.23	+1.88
Correct 2 concepts	91.21	+3.86
Correct 3 concepts	93.02	+5.67

(f) Concept intervention

Table 2: Ablation studies of different components of our proposed framework on NTU-120. (a) Backbone ablation. (b) Component ablation. (c) Group query configuration. (d) Concept vocabulary composition. (e) Logic layer capacity analysis. (f) Concept Intervention Analysis.

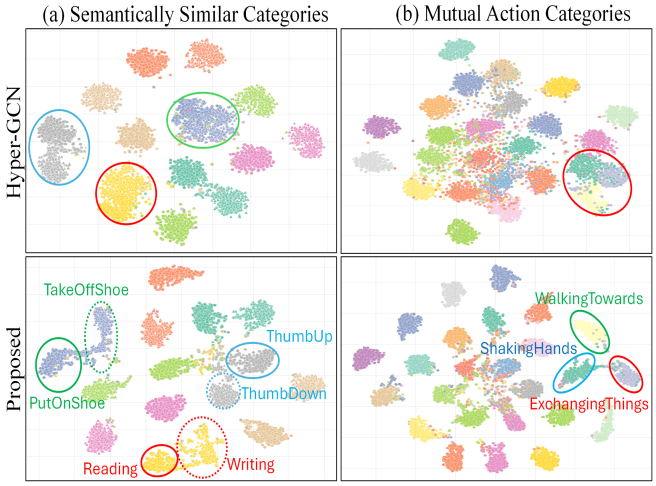

Figure 4: **t-SNE visualization on NTU RGB+D 120**. Hyper-GCN features (top) and our STC-decoder features (bottom). (a) Semantically similar action pairs. (b) Two-person interaction actions.

5 Conclusion

In this work, we presented REASON, a neurosymbolic framework that formulates skeleton-based HAR as logical reasoning over learnable motion concepts. Through concept grounding, spatio-temporal decoding, and differentiable logic, REASON achieves state-of-the-art performance while maintaining fully transparent concept-rule reasoning chains, establishing neurosymbolic learning as a promising paradigm for interpretable skeleton-based HAR.

References

[Asif *et al.*, 2023] Umar Asif, Deval Mehta, Stefan Von Cavallar, Jianbin Tang, and Stefan Harrer. Deepactsnet: A deep ensemble framework combining features from face, hands, and body for action recognition. *Pattern Recognition*, 139:109484, 2023.

[Barbiero *et al.*, 2023] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lió, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, pages 1801–1825. PMLR, 2023.

[Bengio *et al.*, 2013] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[Bhuyan *et al.*, 2024] Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and TP Singh. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21):12809–12844, 2024.

[Chen *et al.*, 2021] 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*, pages 13359–13368, 2021.

[Chi *et al.*, 2022] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20186–20196, 2022.

[Duan *et al.*, 2022] 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*, pages 2969–2978, 2022.

[Garcez *et al.*, 2002] Artur S d'Avila Garcez, Krysia B Broda, and Dov M Gabbay. *Neural-symbolic learning systems*. Springer, 2002.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Hitzler and Sarker, 2022] Pascal Hitzler and Md Kamruzzaman Sarker. Neuro-symbolic artificial intelligence: The state of the art. 2022.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Ilyas *et al.*, 2025] Talha Ilyas, Deval Mehta, Shobi Sivathamboo, Ilma Wijaya, Rob Steele, Hugh Simpson, Lyn Millist, Terence O'Brien, Patrick Kwan, and Zongyuan Ge. Privacy-centric seizure diagnosis via relation-aware fusion of minimally-invasive modalities. In *International Workshop on Applications of Medical AI*, pages 173–183. Springer, 2025.

[Jiang *et al.*, 2025a] Yiwen Jiang, Deval Mehta, Wei Feng, and Zongyuan Ge. Enhancing interpretable image classification through llm agents and conditional concept bottleneck models. *arXiv preprint arXiv:2506.01334*, 2025.

[Jiang *et al.*, 2025b] Yiwen Jiang, Deval Mehta, Siyuan Yan, Yaling Shen, Zimu Wang, and Zongyuan Ge. Wise: Weak-supervision-guided step-by-step explanations for multimodal llms in image classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14685–14696, 2025.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[Kowal *et al.*, 2024] Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G Derpanis, and Pavel Tokmakov. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10946–10956, 2024.

[Lee *et al.*, 2023] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10444–10453, 2023.

[Li *et al.*, 2022] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8494–8506, 2022.

[Liu *et al.*, 2019] 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.

[Liu *et al.*, 2023] Yanan Liu, Hao Zhang, Yanqiu Li, Kangjian He, and Dan Xu. Skeleton-based human action recognition via large-kernel attention graph convolutional network. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2575–2585, 2023.

[Mahinpei *et al.*, 2021] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[Mehta *et al.*, 2023] Deval Mehta, Shobi Sivathamboo, Hugh Simpson, Patrick Kwan, Terence O'Brien, and Zongyuan Ge. Privacy-preserving early detection of epileptic seizures in videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 210–219. Springer, 2023.

[Mehta *et al.*, 2025] Deval Mehta, Yiwen Jiang, Catherine Jan, Mingguang He, Kshitij Jadhav, and Zongyuan Ge. Interpretable few-shot retinal disease diagnosis with concept-guided prompting of vision-language models. In *International Conference on Information Processing in Medical Imaging*, pages 263–277. Springer, 2025.

[Okamoto and Parmar, 2024] Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 3204–3213, 2024.

[Perfilieva, 2002] Irina Perfilieva. Logical approximation. *Soft Computing*, 7(2):73–78, 2002.

[Qu *et al.*, 2024] Haoxuan Qu, Yujun Cai, and Jun Liu. Llms are good action recognizers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18395–18406, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Ridnik *et al.*, 2023] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 32–41, 2023.

[Schrouff *et al.*, 2021] Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*, 2021.

[Shahroudy *et al.*, 2016] 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*, pages 1010–1019, 2016.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[Vemuri *et al.*, 2025] Deepika SN Vemuri, Gautham Bellamkonda, Aditya Pola, and Vineeth N Balasubramanian. Logiccbms: Logic-enhanced concept-based learning. *arXiv preprint arXiv:2512.07383*, 2025.

[Wang *et al.*, 2014] 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*, pages 2649–2656, 2014.

[Wang *et al.*, 2023] Zhuo Wang, Wei Zhang, Ning Liu, and Jianyong Wang. Learning interpretable rules for scalable data representation and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1121–1133, 2023.

[Wu *et al.*, 2023] 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)*, pages 224–229. IEEE, 2023.

[Xiang *et al.*, 2023] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10276–10285, 2023.

[Yan *et al.*, 2018] 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.

[Ye *et al.*, 2025] Qilang Ye, Yu Zhou, Lian He, Jie Zhang, Xuanming Guo, Jiayu Zhang, Mingkui Tan, Weicheng Xie, Yue Sun, Tao Tan, et al. Sugar: Learning skeleton representation with visual-motion knowledge for action recognition. *arXiv preprint arXiv:2511.10091*, 2025.

[Zhou *et al.*, 2024] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024.

[Zhou *et al.*, 2025] Youwei Zhou, Tianyang Xu, Cong Wu, Xiaojun Wu, and Josef Kittler. Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12648–12658, 2025.

[Zhu *et al.*, 2023] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15085–15099, 2023.