AXON: Action Characterization Through Cross-Modal Knowledge Distillation for Neurodiverse Individuals

Siddhant Bikram Shah[1], **Kristina T. Johnson**[1, 2]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
[2]Department of Communication Sciences and Disorders, Northeastern University, Boston, MA, USA
{shah.siddhantb, kri.johnson}@northeastern.edu

Abstract

Understanding the communicative behaviors of non- and minimally-speaking individuals with autism spectrum disorder (ASD) and complex neurodevelopmental disorders (NDDs) remains a critical challenge for both clinical support and machine learning (ML) research. However, developing automated systems for this task is hindered by data scarcity, privacy concerns, heterogeneous and idiosyncratic behaviors, and the significant domain shift from neurotypical to neurodiverse populations. To address these challenges, we first present a large-scale, privacy-preserving action recognition dataset with 2,721 3D skeleton samples capturing in-home interactions of individuals with ASD and complex NDDs. Second, we propose AXON, a cross-modal knowledge distillation method that transfers the rich semantic understanding of CLIP to Hyperformer, a graph-based skeleton classification model, outperforming other cross-modal knowledge distillation baselines in action recognition. We further propose a gradient-based interpretability method to characterize how individuals with ASD and complex NDDs perform communicative actions. Our analysis uncovers both individual- and population-level communicative profiles, tendencies, and biases. Our foundational study helps spur the development of more adaptive and personalized augmentative technologies, aiming to foster greater communicative autonomy and understanding for this underserved population.

Code — https://github.com/SiddhantBikram/AXON

1 Introduction

Autism Spectrum Disorder (ASD) affects approximately 1 in 36 children in the United States, with up to 30% of individuals remaining non- or minimally-speaking throughout their lives (Rose et al. 2016). For these individuals, everyday communication depends on a rich tapestry of nonverbal behaviors—including gestures, body movements, facial expressions, eye gaze, and use of augmentative and alternative communication (AAC) devices—to convey needs, emotions, and social intent (Johnson et al. 2023). Yet these behaviors often diverge from neurotypical patterns, creating barriers to understanding and support by both human caregivers and existing automated systems (Radulski 2022). Understanding these diverse communicative patterns is critical for developing technologies that can support both clinical assessment and daily communication needs.

While machine learning (ML) has shown significant promise in large-scale human behavior analysis (Plonsky et al. 2025; Chen et al. 2024), significant gaps remain in applying these technologies to understand the unique communication patterns of individuals with profound ASD and complex neurodevelopmental disorders (NDDs). The inherent heterogeneity of communicative behaviors across this population necessitates large, well-labeled datasets (Warren Jones and Klin 2009); however, such datasets are exceedingly rare due to stringent privacy concerns surrounding sensitive clinical populations (Shokri and Shmatikov 2015; Shah and Johnson 2025b). Further, the well-documented differences between behaviors exhibited in clinical versus naturalistic home settings undermine the real-world relevance of data captured in controlled laboratories (Leaf et al. 2018).

Skeleton-based action recognition using Graph Neural Networks (GNNs) has emerged as a natural paradigm for understanding human behavior (Liu et al. 2025). This approach models the human body as a topological graph, where joints and bones are modeled as nodes and edges, respectively, allowing models to learn the complex spatiotemporal dynamics of bodily movement. However, the vast majority of research in this area has focused on recognizing overt, clearly defined actions performed by neurotypical individuals in controlled settings (Ren et al. 2024). Further, graph transfer learning is often unfeasible due to the inherent task and domain heterogeneity that GNNs suffer from (Ju et al. 2025), compounded by the domain shift when adapting models trained on neurotypical individuals to individuals with NDDs.

Knowledge distillation presents a promising solution to this problem, enabling the transfer of semantic understanding across deep representations from different modalities (Gou et al. 2021). This allows a specialized student model to inherit the rich knowledge of a larger teacher model without the risk of information loss. While this approach has proven effective for vision-text tasks (Yang et al. 2024), its application to graph-text distillation remains largely unexplored.

Interpreting the communicative behaviors of individuals with NDDs is essential as it underpins the development of effective, individualized interventions and assistive tech-

nologies tailored to their diverse communicative profiles (Shah and Johnson 2025a). However, existing approaches of action characterization rely on simple, shallow architectures that may fail to capture the complex spatiotemporal dynamics of understated communicative actions (Dutt, Goodwin, and Omlin 2024). This limitation is particularly problematic for individuals with NDDs who may express themselves through idiosyncratic and often subtle movements requiring sensitive modeling to detect and interpret.

To address these interconnected challenges of data scarcity, domain shift, population heterogeneity, and limited interpretability, we make three primary contributions:

1. We create a large-scale privacy-preserving action recognition dataset with 3D skeleton data capturing naturalistic communication in individuals with ASD and complex NDDs.
2. We propose a novel knowledge distillation framework, **AXON**: **A**ction Characterization through Cross (**X**)-M**O**dal Distillation for **N**eurodiverse Individuals, to distill CLIP's semantic knowledge into Hyperformer, a graph-based skeleton model, that outperforms other cross-modal knowledge distillation methods.
3. We analyze our data from a gradient-based interpretability perspective to reveal both population-level patterns and individual communication biases, enabling better personalized support in this underserved population.

Together, these contributions help us understand communicative behaviors in this population and how they manifest in each individual. Ultimately, we aim to support the development of AI systems that promote greater understanding and communicative autonomy for non- and minimally-speaking individuals with ASD and complex NDDs.

2 Related Work

Video Analysis in Autism. A significant portion of research at the intersection of motion analysis and autism has been dedicated to the crucial tasks of early detection and diagnostic classification (Yang et al. 2025; Jiang et al. 2024; Simeoli et al. 2024). However, few ML approaches treat ASD as a lifelong developmental condition that requires ongoing understanding and support beyond phenotype classification. Further, most methods are primarily trained and tested on lab-based data, which limits their generalization to the real-world behaviors of these individuals in naturalistic environments (Kommineni et al. 2025; Deng et al. 2024). Multimodal datasets such as MMASD (Li et al. 2023) and MMASD+ (Ravva et al. 2024) have advanced the field, but they focus on recognizing overt and objective actions like performing yoga poses and playing musical instruments. In contrast, our dataset focuses on capturing subjective characterizations of subtle daily-life communicative actions from this population. To this end, we introduce a large-scale privacy-preserving action recognition dataset comprising 3D skeleton data of naturalistic, caregiver-annotated behaviors from non- and minimally-speaking children with ASD and complex NDDs. By capturing subtle, heterogeneous everyday communicative acts in home environments, our dataset shifts the focus of ML methods to understanding and support beyond diagnosis.

Transfer Learning in Graphs. Unlike images and text, transfer learning in graph-based models remains a challenging problem due to the variable topology of graphs across domains, especially in low-resource, real-world downstream tasks (Ju et al. 2025; Wu, He, and Ainsworth 2023). To overcome data scarcity, foundation models have been increasingly employed as teacher models to distill knowledge into smaller student models (Tian et al. 2025; Joshi et al. 2022). Despite a few early works in the field, knowledge distillation in cross-modal text-graph scenarios remains underdeveloped (Luo et al. 2025). Building upon research in cross-modal text-image and text-video methods (Miles and Mikolajczyk 2024; Huang et al. 2024; Sinha et al. 2025), we develop AXON, a novel cross-modal knowledge distillation strategy to transfer the pre-trained knowledge of a teacher CLIP model into Hyperformer, a GNN-based human skeleton classification model. We align Hyperformer's graph embedding space with CLIP's rich natural language embedding space to bridge the severe modality gap while enabling efficient knowledge transfer.

3 Dataset

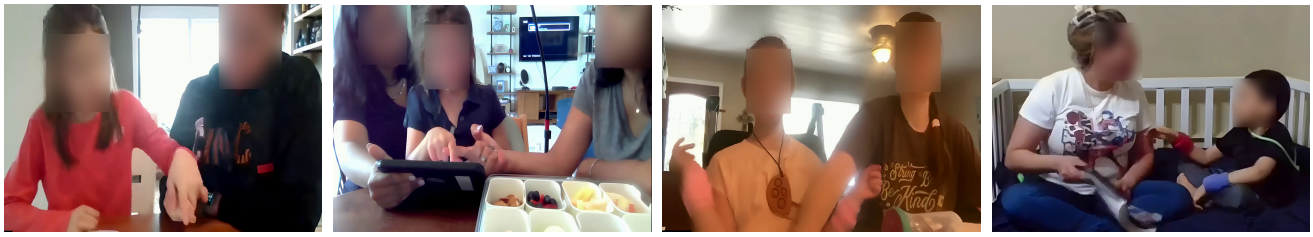
3.1 Data Collection

We use data from the ROSCO (**R**apid **O**nline **S**ample of **CO**mmunication) study, which was designed to capture expressive communicative behaviors in non- and minimally-speaking individuals with ASD and complex NDDs in their home environments. The study produced a video dataset, which was labeled in a unique human-centered method involving both the caregiver and the researcher.

Participants were recruited through a study with the NIH Rare Diseases Clinical Research Network's (RDCRN) Developmental Synaptopathies Consortium (DSC) (Grant 3U54NS092090-10S1), in collaboration with Boston Children's Hospital (BCH). These individuals were both male and female genders, aged 4–12 years old, diagnosed with Phelan-McDermid Syndrome (PMS), PTEN Hamartoma Tumor Syndrome (PHTS), or Tuberous Sclerosis Complex (TSC). The samples were recorded in participants' homes, following the ROSCO study paradigm, capturing their everyday interactions to provide a holistic view of communication, including both verbal and nonverbal elements. This data is especially valuable for analyzing how complex communication unfolds among non- and minimally-verbal individuals with ASD and complex NDDs in daily life scenarios.

The obtained dataset comprises 2,721 video samples (mean duration = 3.11 seconds) obtained from 34 ROSCO sessions with 27 individuals. All recordings were conducted via Zoom[1] video conferencing. While not fully unobtrusive, Zoom-based recordings in a participant's own home with their primary caregiver represent a significant advancement in ecological validity compared to data collected in unfamiliar clinical or laboratory settings. ROSCO captures naturalistic interactions in a familiar environment, which

[1]https://www.zoom.com



(a) **Gestures** for social communication. (b) Requesting through **AAC** device usage. (c) Emotional expression through **body** movement. (d) Using **vocalizations** for commenting.

Figure 1: Samples from our dataset showcasing **action labels** and diversity in participants, caregivers, and environments.

is critical for observing ecologically valid communicative behaviors. Additionally, remote recording reduces the burden on both caregivers and researchers, facilitating broader participation. Figure 1 presents sample image frames from our video dataset, illustrating the heterogeneity in participants, caregivers, recording environments, and participant-caregiver interactions.

All ROSCO session data were collected under the oversight and approval of an Institutional Review Board (IRB). Informed consent was obtained from legal guardians or caregivers, and assent was sought from participants wherever possible, following the guidelines appropriate for this neurodevelopmentally diverse population.

3.2 Annotation

Within the ROSCO sessions, each observed action is labeled into one of the six action classes:

- **AAC** (201 samples): Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.
- **Body** (375 samples): Use of primarily the body or head. This includes more holistic movements like postural shifts (e.g., leaning in, turning away), or whole-body movements (e.g., rocking, walking away).
- **Face** (84 samples): Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.
- **Gesture** (983 samples): Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.
- **Looking** (166 samples): Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.
- **Vocalization** (912 samples): Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

Caregivers annotated communicative behaviors by re-watching the recorded ROSCO session with a researcher (using the Zoom screen-share feature) and following a structured series of questions for each 10-second snippet of the entire session. In each snippet, they were asked to identify whether their child communicated, and if so, what the communicative behavior's action and function were.

Data were annotated based on the primary, predominant action that conveyed a communicative function—like requesting, rejecting, and commenting—as identified by the caregiver. For example, caregivers might identify an extended open palm facing upward as a gesture that communicated a request. The labels and their definitions were iteratively refined based on real-world feedback; for instance, head and body movements were combined into the body label, as distinguishing between them was often impractical. These action labels are not exhaustive, as other idiosyncratic behaviors may also occur outside this predefined set; rather, they serve as a starting point, facilitating future research in this domain.

For idiosyncratic behaviors unique to non- and minimally-speaking individuals, primary caregivers are the gold standard—they understand personalized communication patterns that may be unfamiliar even to professional clinicians. As the labels were provided directly by the caregivers, they represent the most accurate and reliable interpretation of the individuals' behaviors.

3.3 Preprocessing

Video Enhancement. We performed spatial trimming on the raw Zoom recordings to remove non-essential visual artifacts such as user interface elements. To enhance visual quality, we applied RealESRGAN-x4plus (Wang et al. 2021) to upscale each video's resolution by a factor of four. To standardize the temporal resolution and capture finer-grained motion, we used optical flow-based frame interpolation (Baker et al. 2011) to increase the frame rate of all videos to a uniform 60 frames per second.

2D Pose Estimation. Using skeleton data removes identifiable facial and environmental information, which is essential for privacy-preserving proliferation of clinical data. For 2D human pose estimation, we used AlphaPose (Fang et al. 2022), a multi-person pose estimation framework.

Participant Tracking. In this dataset, we focus solely on the participant with ASD and complex NDDs, discarding all caregiver data. While AlphaPose reliably detects humans within individual frames, it lacks the semantic knowledge to distinguish between specific identities, such as children and adults, instead simply assigning them different indices. Additionally, it often reassigns or swaps these indices in different video samples from the same ROSCO session. To resolve this, we employed Video-LLaMA-3 (Zhang et al.

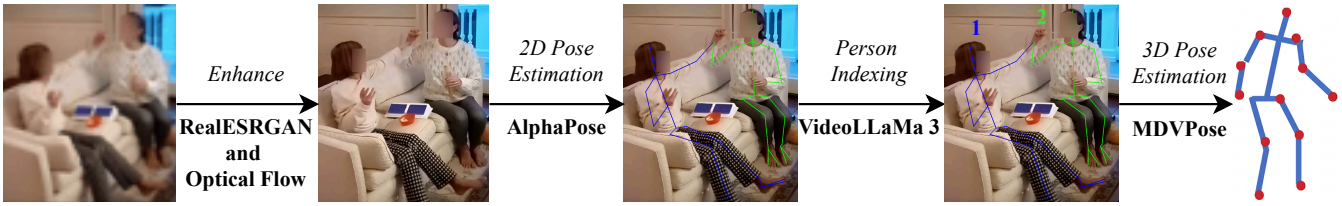


Figure 2: An overview of our data preprocessing *operations* and **tools**.

2025) to semantically identify and track individuals across frames and videos, allowing us to consistently track the required participant for 3D pose estimation.

3D Pose Estimation. To recover 3D human poses from 2D keypoints, we used MDVPose (Gu et al. 2024). Compared to 2D poses, 3D poses offer several advantages: they are largely invariant to viewpoint changes, more robust to occlusions, and provide richer spatial context. To standardize the skeletal structure for compatibility with downstream models, we mapped the 17 joints from the native MDVPose output to the 25-joint NTU RGB+D format (Liu et al. 2019), enabling compatibility and comparison with NTU-based action recognition pipelines.

Human Body as a Graph. The human body naturally forms a topological graph (Feng and Meunier 2022). We formally represent the skeleton at each frame as a graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. The vertices, $V \in \mathbb{R}^{25}$, correspond to the human joints defined by the standardized NTU RGB+D format. The edges in $E \in \mathbb{R}^{24}$ are defined by the natural anatomical connections between these joints, representing the bones of the human skeleton. For each joint vertex $v \in V$, its 3D spatial coordinates (x, y, z) serve as the node feature vector. This process converts an action sequence into a series of graphs, one for each frame, which serves as the final input to our action recognition models. Figure 2 presents an overview of our preprocessing pipeline.

4 Methodology

4.1 Problem Formulation

Let S represent a temporal sequence of 3D human skeleton graphs $G = \{G_1, G_2..., G_T\}$ depicting a single communicative behavior, characterized by an action class label y. Each graph in G contains the vertex set $V \in \mathbb{R}^{25}$ that corresponds to 25 bodily joints, and the edge set $E \in \mathbb{R}^{24}$ represents the natural anatomical connections between them. The state of each joint $v \in V$ at a given time t is described by its 3D spatial coordinates: $v_t \in \mathbb{R}^3$. The task is to learn a representation model $R = f(S)$ that maps the input sequence S to its class label y, i.e., $f : S \rightarrow y$. The model must learn representations that jointly generalize across both cross-subject and cross-view variability while capturing subtle and idiosyncratic motion patterns characteristic of a neurodiverse population.

4.2 Backbone

We employ Hyperformer (Zhou et al. 2022) as our feature encoder for its ability to learn higher-order kinematic dependencies by partitioning joints into multiple hyperedges. The model also shows strong synergy with CLIP for multimodal representation learning (Sinha et al. 2025). Hyperformer learns a student graph embedding x_s^g from the spatiotemporal skeleton sequence S as:

$$x_s^g = \text{Hyperformer}(S) \tag{1}$$

4.3 Feature Distillation

Our framework leverages CLIP as a teacher model to transfer its rich semantic knowledge to our graph-based student, Hyperformer, through cross-modal distillation. For each class c in our dataset, we generate target teacher embeddings x_t^c by encoding its text label t_c $\in$\{*Augmented Communication, Body, Face, Gestures, Looking, Vocalization*\} using CLIP's text encoder ϕ_{text}. We formally define it as:

$$x_t^c = \phi_{\text{text}}(t_c) \tag{2}$$

While prior work on knowledge distillation often employs complex projection techniques (Huang et al. 2024; Yang et al. 2024), we found these methods unsuitable for bridging the cross-modality gap in text-graph distillation. Inspired by findings from Miles and Mikolajczyk (2024), we simplify the distillation architecture into its fundamental components.

Linear Projector. We use a single linear projector to map the student's learned graph embedding to the feature space of the teacher's text embeddings. Unlike complex projectors, this simple projector avoids the risk of information loss caused by decorrelation with the student model. The projected student embedding x_s^p is obtained by passing x_s^g through the linear projection layer ℓ_p:

$$x_s^p = \ell_p(x_s^g) \tag{3}$$

Batch Normalization. We apply batch normalization for both x_t^c and x_s^p before distance calculation between the student and teacher embeddings to stabilize relational information transfer and prevent collapse of the projector's learned weights. For a minibatch of B embeddings, the normalized embedding $\hat{x} \in \{\hat{x}_s, \hat{x}_t\}$, is computed as:

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{4}$$

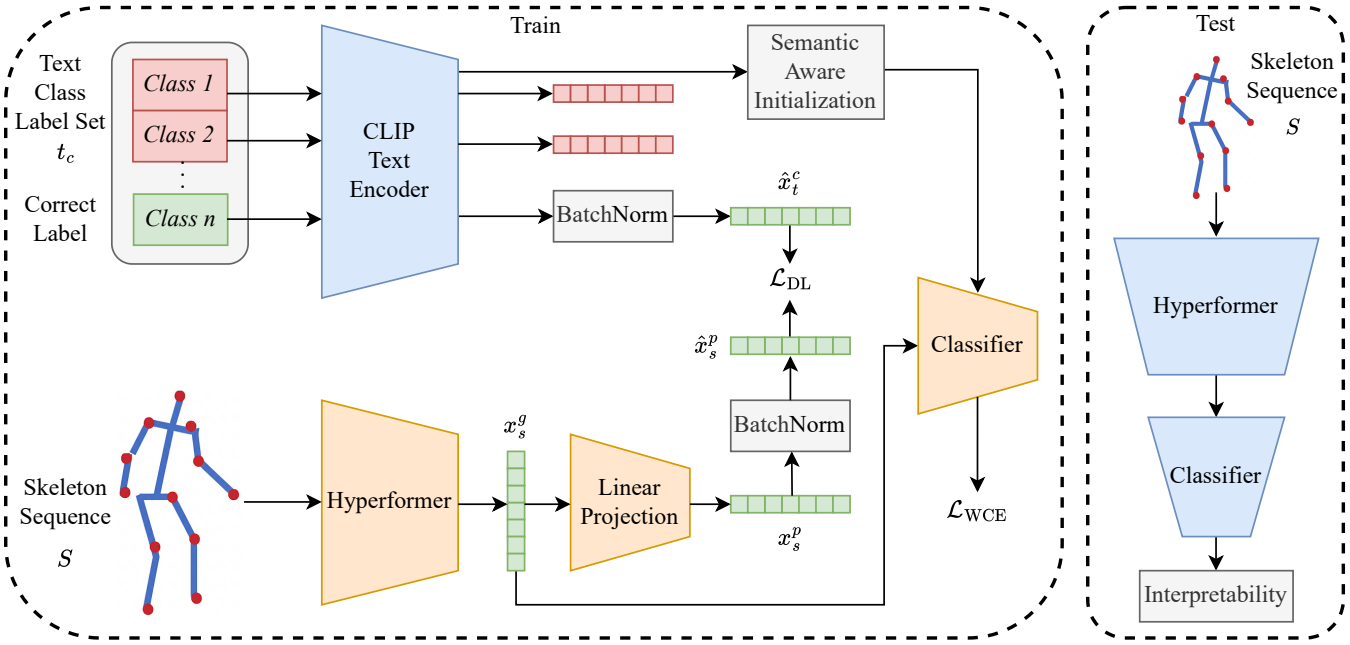


Figure 3: An overview of our proposed model AXON for action recognition through cross-modal text-graph distillation from CLIP's text encoder to Hyperformer.

where μ_B and σ_B^2 are the mean and variance of the mini-batch, respectively, and ϵ is a small constant for numerical stability.

Log-Sum Exponential Loss. The distance between the normalized student embedding $\hat{x}_s^p$ and the corresponding normalized teacher embedding $\hat{x}_t^c$ is calculated using the Log-Sum Exponential loss (Miles and Mikolajczyk 2024). This acts as a soft distance metric to effectively bridge the significant cross-modal representation gap. The distillation loss $\mathcal{L}_{\text{DL}}$ for a single pair of embeddings is defined as:

$$\mathcal{L}_{\text{DL}}(\hat{x}_s^p, \hat{x}_t) = \log\left(\sum_{d=1}^{D} \exp(\hat{x}_s^p - \hat{x}_t^c)\right) \tag{5}$$

4.4 Optimization

Semantic-Aware Classifier Initialization. To further exploit the rich semantic discrimination capabilities of CLIP, we harness the text embeddings of classes using semantic-aware initialization (Shi et al. 2024) to initialize weights of a linear classifier ℓ_{cls}. Specifically, CLIP's text embeddings x_t^c, generated for each class c, are used to initialize the corresponding classifier weight vector w_c. This semantic-aware initialization allows the classifier to start from a space that is already aligned with meaningful inter-class distinctions in natural language, enabling improved generalization to our task.

Training Objective. Our training objective combines classification and distillation losses. For classification, we use weighted cross-entropy to address class imbalance, where each class c receives a weight w_c inversely proportional to its frequency in the training set. We use the original graph embedding x_s^g for classification. The weighted cross-entropy loss for a minibatch of size B is defined as:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{B}\sum_{i=1}^{B} w_{c_i} \cdot \log(\ell_{cls}(x_s^g)) \tag{6}$$

The total loss comprises the losses obtained from equations 5 and 6 with a hyperparameter λ_{DL} that scales the contribution of $\mathcal{L}_{\text{DL}}$ with respect to $\mathcal{L}_{\text{WCE}}$.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{WCE}} + \lambda_{\text{DL}} \cdot \mathcal{L}_{\text{DL}} \tag{7}$$

Figure 3 presents the overall flow of our framework.

5 Experimental Setup

Implementation Details. We conducted all our experiments on Python 3.10.18 with Pytorch 2.7.1 combined with NVIDIA V100, A100, and H200 GPUs. We anchored the seed to 42 for all single-seed experiments and to 42, 100, and 510 for all multi-seed experiments. We used an 80/20 train/test split, set the batch size to 8, and set the learning rate to 10^{-5} for all experiments. We trained each model for 100 epochs and report the best scores for each run. We set the λ_{DL} for AXON to 0.05 and use the default hyperparameters for all other methods. Model performance is reported across Accuracy, Macro F1-Score, and Unweighted Average Recall (UAR).

Baselines. We compare AXON against zero-shot VideoLLaMA 3-7B (Zhang et al. 2025), Hyperformer (Zhou et al. 2022), SkeletonCLIP (Sinha et al. 2025), Optimal Transport (Wu et al. 2022), Inverted Distillation (Auty et al. 2024), Relational Distillation (Park et al. 2019), and Residual Distillation (Huang et al. 2024). For fair comparison, we use Hyperformer with the pre-trained weights for the NTURGB+D

120 X-Subject dataset (Liu et al. 2019) as the backbone for all graph-based methods. We use weighted cross-entropy loss as the default optimization criterion.

6 Results and Analysis

6.1 Performance Comparison

Table 1 presents a comparison of model performance for our dataset. VideoLLaMA 3 performs poorly using raw video data, indicating the difficulty of interpreting the subtle, idiosyncratic actions present in our dataset. Further, the use of RGB video data raises significant privacy concerns, making it a less viable modality for real-world translation. Hyperformer establishes a reference benchmark for action recognition using only skeletal data. However, the substantial domain gap between neurotypical subjects and individuals with complex NDDs, along with differences in actions between the NTU RGB+D dataset and ours, limits the transferability of pre-trained knowledge. SkeletonCLIP achieves competitive performance by using a flexible contrastive loss for skeleton and text data. Optimal transport aligns graph and text representations with a transport-based objective, but lacks the regularization strength of distillation-based methods. In contrast, constrained distillation acts as a powerful regularizer, forcing the model to learn the canonical node-specific features, with residual distillation achieving the highest performance among the baseline methods. AXON outperforms all baselines, indicating its effectiveness in capturing both localized body dynamics within the graph structure and their strong alignment with the linguistic priors.

Model	Peformance Metric		
	Accuracy	**F1**	**UAR**
VideoLLaMA 3-7B	24.07	20.64	25.91
Hyperformer	46.85±1.36	32.28±1.65	31.78±1.71
SkeletonCLIP	46.32±1.07	35.36±1.10	35.15±1.40
Optimal Transport	47.07±0.97	37.02±1.54	36.72±1.53
Inverted Distillation	45.44±1.05	34.21±1.68	33.96±1.72
Relational Distillation	46.67±0.83	36.67±1.51	36.49±1.35
Residual Distillation	48.85±1.50	37.96±2.68	36.38±2.37
AXON (Ours)	**49.17**±1.13	**39.01**±3.42	**38.23**±3.43

Table 1: Comparison of model performance for action classification. The results are in the form of Mean ± Standard Deviation. The best results are highlighted in **bold**.

6.2 Individual-Level Analysis

We trained and evaluated AXON on each participant in our dataset individually to assess its generalization to diverse, idiosyncratic expression styles. This approach standardizes subject-specific traits such as body proportions and bone lengths, and environment-specific variables such as view angle and distance from the recording camera. We used a train/test split of 50/50 to leave enough samples in the test set for each class for heatmap-based interpretability.

Table 2 presents the results for our experiments, revealing a striking diversity in model performance across individuals. The significant variance across all metrics indicates that the model excels at classifying the actions of some individuals while struggling with others. This highlights the challenge of generalizing across the wide spectrum of idiosyncratic ways of expressing communicative acts found within the ASD/NDD umbrella, demonstrating that a one-size-fits-all approach is insufficient for this heterogeneous population, especially with limited data.

Statistic	Accuracy	F1	UAR
Mean	50.38	26.12	31.51
Max	85.71	47.97	50.00
Min	25.92	14.72	16.66
Standard Deviation	14.50	8.63	9.13

Table 2: Summary of AXON's action recognition performance for individual-level analysis.

6.3 Ablation Study

We conduct a systematic ablation study on our framework to identify the contribution of each component to its performance, and the results are presented in Table 3. Our results show a stepwise increase in performance with the sequential addition of Log-Sum Exponential Loss and Semantic-Aware Initialization, with a minor drop with the addition of the linear projector without distillation.

Component				Performance Metric		
HF	**LP**	**LSE**	**SAI**	**Accuracy**	**F1**	**UAR**
✓				47.89	33.49	32.80
✓	✓			45.14	35.73	35.95
✓	✓	✓		48.07	38.63	37.96
✓	✓	✓	✓	**49.36**	**40.80**	**41.02**

Table 3: Results of ablation studies of our framework. The abbreviations HF, LP, LSE, and SAI refer to Hyperformer, Linear Projector, Log-Sum Exponential Loss, and Semantic-Aware Initialization, respectively. The final row corresponds to the complete framework. The best results are highlighted in **bold**.

6.4 Hyperparameter Search for λ_{DL}

We conduct a hyperparameter search to find the best value of λ_{DL} for our model, and our results are shown in Table 4. We start with $\lambda_{\text{DL}} = 0.1$ and search in either direction of this value with a scaling factor of 2. We find that $\lambda_{\text{DL}} = 0.05$ performs the best amongst the tested values.

λ_{DL}	Acc.	F1	UAR
0.025	48.99	38.45	37.30
0.05	49.36	**40.80**	**41.02**
0.1	**49.54**	39.72	38.59
0.2	48.81	39.64	38.47

Table 4: Hyperparameter search for the best value of λ_{DL}. The best results are highlighted in **bold**.

6.5 Per-Class Performance

Given the substantial class imbalance in our dataset, we provide the per-class performance metrics obtained by AXON in Table 5. We observe that explicit actions like AAC, Body, and Gestures were easier to classify than subtle actions like Face and Look. Vocalizations were relatively easier to classify, which may be due to their high sample count.

Class	# Samples	Precision	Recall	F1
AAC	201	52.22±4.50	42.50±5.00	46.74±3.99
BODY	375	47.62±6.64	42.67±7.40	45.00±7.12
FACE	84	20.83±19.09	11.77±10.19	14.90 ±13.09
GEST	983	55.58±1.39	57.19±5.43	56.25±2.32
LOOK	166	17.29±5.07	22.22±8.75	19.41±6.62
VOC	912	50.98±3.18	53.01±5.67	51.76±1.80

Table 5: Per-class performance obtained by AXON for all 6 classes in our dataset.

7 Gradient-based Joint Interpretability

Our interpretability pipeline uses model gradients to determine the body joints that are most influential in predicting a given action. For a given input skeleton sequence S and loss function $\mathcal{L}$, we first compute the gradient of the loss with respect to the input coordinates $G = \frac{\partial \mathcal{L}}{\partial S}$. For each correctly classified test sample i belonging to class c, we aggregate the absolute gradient values for joint vertex v across all frames T and coordinate dimensions D.

$$g_{i,v} = \sum_{t=1}^{T} \sum_{d=1}^{D} |G_{i,t,v,d}| \tag{8}$$

We average these aggregated scores across all M_c samples within the class to get a single importance value per joint. We normalize these average scores across all joints to a $[0, 1]$ scale to compute the final node strength. The resulting scores are visualized as skeletal heatmaps, where warmer colors signify higher node strength. As individuals in our dataset were usually seated throughout the ROSCO protocol, our analysis and visualizations focus on upper-body joints to ensure robustness against lower-body occlusion. The visualizations are not mirrored, meaning individuals are depicted as if viewed from behind; for instance, an individual's left hand will appear on the left side of the heatmap.

7.1 Population-Level Analysis

Figure 4 visualizes the average node strength of each upper-body joint for classifying the six communicative action classes across the entire population. These population-level heatmaps reveal distinct activation patterns that align with intuitive expectations of how different communicative actions manifest physically.

AAC. We observe strong bilateral hand activation with a slight left-hand dominance, reflecting the movements of these joints around a communication device. The relatively low activation in peripheral joints suggests that gross limb motion is not as important as wrist and hand movements for AAC usage.

Body. Strong activations are distributed across the shoulder and elbow joints, indicating broader upper-limb involvement. The heatmap shows bilateral symmetry, indicating that both sides of the body are involved in bodily communication. The basal spine also exhibits moderate importance, likely reflecting changes in posture.

Face. These actions show moderate strength in the head and neck area, but are still characterized by movement throughout the upper body, indicating that facial communication may involve fixed gaze and not by overt movement of the head. Residual importance at the shoulder joints could reflect changes in gaze involving partial shoulder movement.

Gestures. In line with our expectations, this class shows strong activation across the whole body, especially at the wrists, elbows, and shoulders. It also reveals a slight population-level right-sided dominance.

Look. The head and middle spine joint have the most node strength, confirming that head and posture directions are the primary informative features for this class.

Vocalization. This class shows high activation in the basal spine and neck regions, which may reflect postural adjustments linked to speech production. Peripheral joints remain less salient, though activity at the wrists may capture emphatic or co-speech gesture movements.

These population-level patterns establish a baseline understanding of how different action classes typically manifest in the skeletal data for this population. They further indicate that the model has learned anatomically and semantically aligned patterns of joint relevance that mirror the typical kinematics of communicative actions.

7.2 Individual-Level Analysis

Figure 5 visualizes the node strength across different actions for one participant from our dataset. By comparing the heatmaps of individuals to the population average, we can move beyond generalized patterns to identify person-specific, idiosyncratic communication styles. This reveals how the same communicative intent can manifest through vastly different physical actions. Note that not all participants have samples across all 6 classes. The heatmaps demonstrate a distinct profile combining both canonical and idiosyncratic movements. Further, they reveal activations that are more in line with our expectations than the population heatmap, which creates a compressed average of the entire heterogeneous dataset.

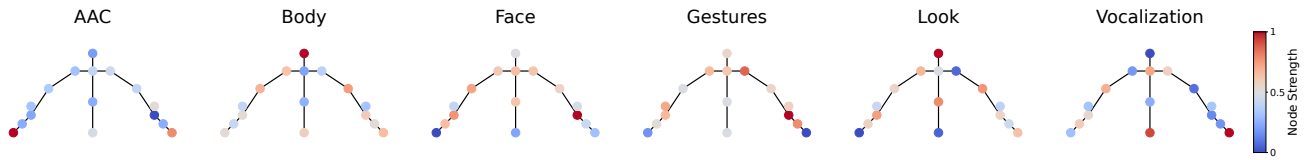


Figure 4: Population-level node strength heatmap across 6 actions for all individuals in our dataset.

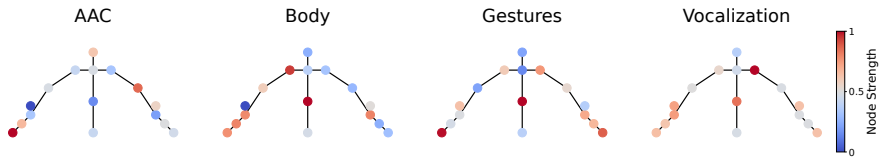


Figure 5: Individual-level node strength heatmap across 4 actions for one person in our dataset.

AAC. During AAC usage, the left wrist shows higher node strength, which may reflect biases in dominant-side interactions with an assistive device for this individual.

Body. These communicative actions engage almost all nodes in the torso, with stronger activity on the left side that may indicate individual motor preferences or situational asymmetry.

Gestures. This class shows widespread and strong activation across the shoulders, wrists, and hands, especially across the right side.

Vocalization. Widespread activation across the middle spine and arms indicates co-speech gestures and movements.

We empirically find that nodes converge closer to the expected activations in classes with a higher number of samples, indicating that a greater sample count and diversity enable the model to learn more robust node-label associations, reducing the influence of spurious correlations.

8 Limitations and Future Work

Our work represents an important but preliminary step in understanding the naturalistic behaviors of non- and minimally-speaking individuals with ASD and complex NDDs. As our dataset captures a brief period of interactions with caregivers, this data collection window may have introduced sampling bias and constrained the diversity of actions. Consequently, this dataset may not strictly expose ML methods to the sufficiently broad range of examples they might need to generalize to heterogeneous, real-world scenarios. In the future, we plan to collect a longitudinal dataset to enable analysis of the development of individuals across time, which would also increase the number of dataset samples from each individual.

The modest performance of the models highlights the complexity of generalizing to subtle cues in cross-subject and cross-view settings, signaling the need for action recognition methods that generalize beyond overt, well-defined actions to subtle affective and communicative behaviors. As multimodal large language models (MLLMs) mimic human perception to an extent (Fu et al. 2024), their suboptimal performance reflects how people in society may interpret the behavior of individuals with complex NDDs. While VideoLLaMA 3-7B performed the worst among all the tested methods, MLLMs provide the most scalability and promise for potential application to real-world support and understanding for this population. Future work should focus on using MLLMs to support real-time interpretation for this population, offering meaningful support to families and individuals navigating complex communication needs.

9 Conclusion

This paper makes a dual contribution to AI and computational healthcare, advancing our understanding of both human communication and machine perception. We investigated the nuanced and idiosyncratic communicative behaviors of individuals with ASD and complex NDDs by challenging ML models to interpret human data in subtle, heterogeneous real-world edge cases. We introduced three core contributions: (1) a large-scale, privacy-preserving action recognition dataset with 3D skeleton data of naturalistic behaviors, (2) AXON, a novel cross-modal text-graph knowledge distillation method that leverages CLIP's semantic knowledge for action recognition in skeleton data, and (3) a gradient-based interpretability analysis that reveals both population-level patterns and individual-specific behaviors.

We demonstrate that by aligning graph-based representations with rich linguistic priors, our model becomes better at classifying complex communicative acts. Our interpretability framework provides a new lens through which to view communication, moving beyond simple classification to characterizing how an individual communicates. This method helps better identify idiosyncratic movement patterns for these individuals, which is a critical step toward empowering truly personalized assistive technologies.

The pipelines developed in this work provide a template for tackling other low-resource, real-world problems beyond our immediate application, particularly those involving underserved communities where data is scarce and highly heterogeneous. Ultimately, our research aims to foster greater communicative autonomy for individuals with complex needs by building a foundation for AI systems that can understand, adapt to, and support the rich diversity of human expression.

Acknowledgements

Special thanks to the participants and families, and to the many researchers and staff who helped bring this project to life, including Dr. Lauren Thompson, Dr. Audrey Thurm, Dr. Carol Wilkinson, Dr. Mustafa Sahin, Bianca Booth, Emine Arcasoy, Isabelle Iannotti, Emma McGonigle, Tsambika Rizas, Miranda Kannisto, Thanh Van Le, and the NIH Rare Disease Clinical Research Network (RDCRN) Developmental Synaptopathies Consortium (DSC). This research was funded, in part, by NIH NINDS TALK Supplement (3U54NS092090-10S1), with support from the Rosamund Stone Zander Translational Neuroscience Center and the Laboratories of Cognitive Neuroscience at Boston Children's Hospital, and the Phelan-McDermid Syndrome Foundation (PSMF) Innovation Award.

Ethics Statement

Potential Risks. We acknowledge the significant ethical considerations involved in developing technologies to interpret the behaviors of vulnerable populations like non- and minimally-speaking individuals with ASD and complex NDDs. The primary risk of this technology is misinterpretation. An automated system that classifies subtle, idiosyncratic communicative acts carries an inherent risk of error. Such errors could lead to incorrect assumptions about an individual's needs or intent, which could negatively impact care, support, and the individual's autonomy. Furthermore, any deployment of in-home behavioral analysis systems must be protected from misuse, such as non-consensual surveillance. We emphasize that the deployment of the methods described in this paper must be governed by strict data privacy, informed consent, and protocols that place the individual's dignity, autonomy, and well-being at the forefront.

Biases. Our experimental results explicitly demonstrate how dataset imbalance significantly impacts model performance. As shown in our per-class analysis (see Table 5), the natural frequency of actions in our dataset led to a large disparity in per-class sample counts in our dataset, which may affect model performance.

This issue is compounded by the inherent heterogeneity of our population. The wide variance in individual-level model performance (see Table 2) highlights this challenge directly: a model that performs well on average may still fail significantly for specific individuals whose idiosyncratic communication styles are not well-represented in the training data. An over-reliance on a model with unaddressed biases could lead to inequitable outcomes, where individuals with more subtle or highly idiosyncratic communication styles are disproportionately misunderstood by the technology.

References

Auty, D.; Miles, R.; Kolbeinsson, B.; and Mikolajczyk, K. 2024. Learning to Project for Cross-Task Knowledge Distillation. *arXiv preprint arXiv:2403.14494*.

Baker, S.; Scharstein, D.; Lewis, J. P.; Roth, S.; Black, M. J.; and Szeliski, R. 2011. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1): 1–31.

Chen, L.-H.; Lu, S.; Zeng, A.; Zhang, H.; Wang, B.; Zhang, R.; and Zhang, L. 2024. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*.

Deng, S.; Kosloski, E. E.; Patel, S.; Barnett, Z. A.; Nan, Y.; Kaplan, A.; Aarukapalli, S.; Doan, W. T.; Wang, M.; Singh, H.; et al. 2024. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia*.

Dutt, M.; Goodwin, M.; and Omlin, C. W. 2024. An interpretable deep learning-based feature reduction in video-based human activity recognition. *IEEE Access*.

Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; and Lu, C. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 7157–7173.

Feng, M.; and Meunier, J. 2022. Skeleton graph-neural-network-based human action recognition: A survey. *Sensors*, 22(6): 2091.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6): 1789–1819.

Gu, R.; Zhu, J.; Si, Y.; Gao, F.; Xu, J.; and Xu, G. 2024. 3D human pose estimation from multiple dynamic views via single-view pretraining with procrustes alignment. In *Proceedings of the 32nd ACM international conference on multimedia*, 10363–10372.

Huang, X.; Zhou, H.; Yao, K.; and Han, K. 2024. FROSTER: Frozen CLIP is A Strong Teacher for Open-Vocabulary Action Recognition. In *ICLR*.

Jiang, Y.; Shen, Q.; Lai, S.; Qi, S.; Zheng, Q.; Yao, L.; Wang, Y.; and Pan, G. 2024. Copiloting diagnosis of autism in real clinical scenarios via LLMs. *arXiv preprint arXiv:2410.05684*.

Johnson, K. T.; Narain, J.; Quatieri, T.; Maes, P.; and Picard, R. W. 2023. ReCANVo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1): 523.

Joshi, C. K.; Liu, F.; Xun, X.; Lin, J.; and Foo, C. S. 2022. On representation knowledge distillation for graph neural networks. *IEEE transactions on neural networks and learning systems*, 35(4): 4656–4667.

Ju, L.; Yang, X.; Li, Q.; and Wang, X. 2025. GraphBridge: Towards Arbitrary Transfer Learning in GNNs. In *The Thirteenth International Conference on Learning Representations*.

Kommineni, A.; Bose, D.; Feng, T.; Kim, S. H.; Tager-Flusberg, H.; Bishop, S.; Lord, C.; Kadiri, S.; and Narayanan, S. 2025. Can Multimodal Foundation Models Help Analyze Child-Inclusive Autism Diagnostic Videos? In *Proc. Interspeech 2025*, 3050–3054.

Leaf, J. B.; Leaf, R.; McEachin, J.; Cihon, J. H.; and Ferguson, J. L. 2018. Advantages and challenges of a home-and clinic-based model of behavioral intervention for individuals diagnosed with autism spectrum disorder. *Journal of Autism and Developmental disorders*, 48(6): 2258–2266.

Li, J.; Chheang, V.; Kullu, P.; Brignac, E.; Guo, Z.; Bhat, A.; Barner, K. E.; and Barmaki, R. L. 2023. Mmasd: A multimodal dataset for autism intervention analysis. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 397–405.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. 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.

Liu, M.; Liu, H.; Hu, Q.; Ren, B.; Yuan, J.; Lin, J.; and Wen, J. 2025. 3D Skeleton-Based Action Recognition: A Review. *arXiv preprint arXiv:2506.00915*.

Luo, B.; Wang, J.; Wang, Z.; Zhu, J.; and Zhao, X. 2025. Graph-Based Cross-Domain Knowledge Distillation for Cross-Dataset Text-to-Image Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 568–576.

Miles, R.; and Mikolajczyk, K. 2024. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4233–4241.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.

Plonsky, O.; Apel, R.; Ert, E.; Tennenholtz, M.; Bourgin, D.; Peterson, J. C.; Reichman, D.; Griffiths, T. L.; Russell, S. J.; Carter, E. C.; et al. 2025. Predicting human decisions with behavioural theories and machine learning. *Nature Human Behaviour*, 1–14.

Radulski, E. M. 2022. Conceptualising autistic masking, camouflaging, and neurotypical privilege: Towards a minority group model of neurodiversity. *Human Development*, 66(2): 113–127.

Ravva, P. U.; Kiafar, B.; Kullu, P.; Li, J.; Bhat, A.; and Barmaki, R. L. 2024. MMASD+: A Novel Dataset for Privacy-Preserving Behavior Analysis of Children with Autism Spectrum Disorder. *arXiv preprint arXiv:2408.15077*.

Ren, B.; Liu, M.; Ding, R.; and Liu, H. 2024. A survey on 3d skeleton-based action recognition using learning method. *Cyborg and Bionic Systems*, 5: 0100.

Rose, V.; Trembath, D.; Keen, D.; and Paynter, J. 2016. The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research*, 60(5): 464–477.

Shah, S. B.; and Johnson, K. T. 2025a. Multi-Feature Audio Fusion for Nonverbal Vocalization Classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Shah, S. B.; and Johnson, K. T. 2025b. N-CORE: N-View Consistency Regularization for Disentangled Representation Learning in Nonverbal Vocalizations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 33362–33379.

Shi, J.-X.; Wei, T.; Zhou, Z.; Shao, J.-J.; Han, X.-Y.; and Li, Y.-F. 2024. Long-Tail Learning with Foundation Model: Heavy Fine-Tuning Hurts. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 45014–45039. PMLR.

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321.

Simeoli, R.; Rega, A.; Cerasuolo, M.; Nappo, R.; and Marocco, D. 2024. Using machine learning for motion analysis to early detect autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, 1–20.

Sinha, A.; Reilly, D.; Bremond, F.; Wang, P.; and Das, S. 2025. SKI Models: Skeleton Induced Vision-Language Embeddings for Understanding Activities of Daily Living. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6931–6939.

Tian, Y.; Pei, S.; Zhang, X.; Zhang, C.; and Chawla, N. V. 2025. Knowledge distillation on graphs: A survey. *ACM Computing Surveys*, 57(8): 1–16.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Warren Jones, B.; and Klin, A. 2009. Heterogeneity and homogeneity across the autism spectrum: The role of development. *J Am Acad Child Adolesc Psychiatry*, 48: 471–3.

Wu, B.; Cheng, R.; Zhang, P.; Gao, T.; Gonzalez, J. E.; and Vajda, P. 2022. Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation. In *International Conference on Learning Representations*.

Wu, J.; He, J.; and Ainsworth, E. 2023. Non-iid transfer learning on graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 10342–10350.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, Z.; Zhang, Y.; Ning, J.; Wang, X.; and Wu, Z. 2025. Early Diagnosis of Autism: A Review of Video-Based Motion Analysis and Deep Learning Techniques. *IEEE Access*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhou, Y.; Cheng, Z.-Q.; Li, C.; Fang, Y.; Geng, Y.; Xie, X.; and Keuper, M. 2022. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*.