ROSCO-Omni: Multimodal LLM-Based Communication Understanding for Non- and Minimally-Speaking Autistic Individuals

Siddhant Bikram Shah[1] **Kristina T. Johnson**[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Communication Sciences and Disorders
Northeastern University, Boston, USA

Abstract

Approximately 30% of autistic individuals remain non- or minimally-speaking throughout their lives, yet communicate richly through gestures, vocalizations, facial expressions, and augmentative devices. Interpreting this communication is an inherently multimodal task: caregivers rely on the simultaneous integration of visual cues, auditory signals, and contextual understanding to infer intent. Despite this natural alignment with multimodal large language models (MLLMs), research in this intersection remains narrowly focused on diagnosis rather than communication understanding. We address this gap by reframing the problem around two complementary dimensions: communicative actions (the physical modality) and communicative functions (the pragmatic intent). We analyze the ROSCO dataset, containing 2,903 caregiver-annotated video samples from 27 non- and minimally-speaking individuals, with multi-label annotations capturing up to three concurrent actions and two functions per sample across 6 action and 6 function classes. We further propose ROSCO-Omni, a teacher-student distillation framework that generates label-guided instruction data from a high-capability teacher MLLM and uses it to finetune a student MLLM for domain-specialized inference. ROSCO-Omni achieves performance comparable to closed-source models, demonstrating that open-source MLLMs can be adapted to understand communication in this underserved population.

1 Introduction

Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental condition affecting approximately 1 in 36 children in the United States (Rose et al., 2016). Within this diverse population, approximately 30% of individuals remain non- or minimally-speaking throughout their lives (Tager-Flusberg and Kasari, 2013). Contrary to misconceptions equating nonverbal or minimally verbal status with communicative absence, these individuals possess rich communicative repertoires, expressing needs, emotions, and social intent through gestures, vocalizations, facial expressions, and Augmentative and Alternative Communication (AAC) devices (Shah and Johnson, 2026; Valle et al., 2021). However, because these idiosyncratic behaviors often diverge from neurotypical patterns, they are frequently misunderstood or overlooked by society at large.

Interpreting this communication is fundamentally a multimodal reasoning task: caregivers simultaneously integrate what they see (a reaching gesture, a facial grimace), what they hear (a vocalization, a laugh), and what they know about the context to infer what their child is communicating. Multimodal Large Language Models (MLLMs), with their capacity for joint reasoning over video, audio, and text, offer a computational pathway to replicate this integrative reasoning at scale (Yin et al., 2024). Yet the vast majority of ML research in ASD remains narrowly focused on diagnosis, classifying individuals as autistic or neurotypical (Ahmed et al., 2025), or detecting deficit-oriented markers such as atypical eye contact or repetitive behaviors (Deng et al., 2024). While valuable for early screening, these approaches answer the question "Is this person autistic?" rather than "What is this person communicating?"–failing to address the lifelong communicative needs of individuals already diagnosed.

Our work pivots entirely toward communication understanding for non- and minimally-speaking autistic individuals. We frame the problem through two complementary dimensions: *communicative actions* (the physical modality, e.g., gesture, vocalization, AAC use) and *communicative functions* (the pragmatic intent, e.g., requesting, rejecting, social engagement). This reframing moves beyond recognizing overt behaviors to interpreting the subtle, subjective nuances of interaction in naturalistic

Findings of the Association for Computational Linguistics: ACL 2026, pages 40453–40469

environments, supporting the autonomy and communicative agency of this underserved population.

To this end, we curate the Rapid Online Sample of Communication (ROSCO) dataset, comprising 2,903 video samples from 27 participants recorded in home environments. ROSCO is the first dataset in this domain with multi-label and multi-class annotations–capturing up to three concurrent actions and two functions per sample–reflecting the reality that communicative acts are rarely isolated (e.g., a child may vocalize while gesturing, simultaneously expressing emotion and making a request). This multi-label formulation is essential for faithfully representing how this population communicates, and it enables models that can scale to the full complexity of real-world interaction. Our caregiver-centric annotation protocol uniquely grounds labels in the child's intent as understood by those who know them best, rather than external clinical interpretation (Johnson et al., 2023).

We further propose ROSCO-Omni, a two-stage distillation framework that uses a high-capability teacher model (Gemini-3-Flash) to generate synthetic instruction data via label-guided reasoning, then distilling this knowledge into an open-source student model (Qwen3-Omni) through finetuning for domain-specialized inference. Our framework's design is motivated by two empirical findings. First, even state-of-the-art MLLMs struggle with this domain zero-shot, as the target behaviors diverge substantially from the neurotypical interactions that dominate their training data. Second, few-shot prompting yields only marginal improvements, as a small set of static samples cannot capture the extreme heterogeneity of communicative behaviors within and across individuals in this population. Together, these findings motivate the need for domain adaptation through finetuning.

ROSCO-Omni outperforms the closed-source Gemini-2.5-Flash baseline on both action and function classification while achieving competitive performance with the teacher model Gemini-3-Flash, demonstrating that targeted domain adaptation enables open-source models to match or exceed closed-source models.

Our contributions are summarized as follows:

- We provide the first comprehensive benchmark of open-source and closed-source MLLMs on nonverbal communication understanding in non- and minimally-speaking autistic individuals, including participant-wise evaluations and curated challenge sets probing zero-shot generalization, cross-session transfer, and instruction-type sensitivity.
- We introduce a novel instruction-tuning dataset derived from the ROSCO study, capturing the multimodal complexity of real-world interaction.
- We propose ROSCO-Omni, a teacher-student distillation pipeline using label-guided reasoning to generate synthetic instruction data, enabling an open-source model, Qwen-3-Omni, to achieve performance comparable to closed-source Gemini Flash models.

2 Related Work

Datasets for Autism Behavior Analysis. The advancement of ML research relies heavily on high-quality datasets, yet existing datasets for ASD have largely been constrained by a diagnostic focus, prioritizing behaviors that distinguish ASD from neurotypical development rather than those that facilitate communication (Khor et al., 2024; Serna-Aguilera et al., 2024). Prominent datasets like MMASD (Li et al., 2023) and AV-ASD (Deng et al., 2024) typically capture overt, objective actions–such as repetitive movements or eye contact avoidance–which, while useful for phenotype characterization, characterize individuals primarily by their deficits rather than their communicative agency. These datasets typically employ single-label classification schemes that fail to capture the multimodal, multi-intent nature of real-world communication.

To address this gap, we curate the ROSCO dataset, the first resource specifically designed for interpreting communicative behavior and intent in non- and minimally-speaking autistic individuals. ROSCO adopts a multi-label annotation schema capturing both *actions* and *functions*, with up to three concurrent action labels and two function labels per sample. This structure reflects the complexity of naturalistic interaction, where a single gesture may simultaneously serve as a *request* and a *social* interaction. Furthermore, our caregiver-centric annotation protocol leverages primary caregivers' intimate knowledge of their child's idiosyncratic behaviors to identify communicative acts, grounding labels in the child's intent rather than external clinical interpretation.

MLLMs for Behavioral Understanding in Videos. Multimodal large language models have demonstrated strong capabilities in video reasoning and instruction following (Huang et al., 2025a). Re-

cent omni-modal extensions jointly encode video, audio, and text for richer contextual understanding (Xu et al., 2025). However, general-purpose MLLMs struggle with specialized domains, particularly when target populations exhibit behaviors diverging from training data (Huang et al., 2025b). Several adjacent methods use MLLM-generated action descriptions to train smaller systems (Xiang et al., 2023; Chen et al., 2026) or combine LLM reasoning with skeleton-based features (Qu et al., 2024). However, these methods operate on clean skeleton inputs, discarding affective content like facial expressions, vocalizations, and caregiver-child dynamics, which are essential in our domain. In autism research specifically, MLLM applications remain focused on diagnostic screening, reducing complex interactions to diagnostic labels while overlooking communicative signals (Yang et al., 2025b).

ROSCO-Omni departs from this paradigm by using MLLMs as interpreters of nonverbal communication rather than diagnostic instruments–reasoning about *why* behaviors are performed, not merely *what* they are.

3 Dataset

3.1 Data Collection

We utilize data collected via the Rapid Online Sample of Communication (ROSCO) paradigm, a data collection framework designed to capture naturalistic communicative behaviors from autistic individuals in the home environment. Unlike traditional laboratory-based assessments, which are resource-intensive and often fail to capture the full range of a child's daily communicative repertoire, ROSCO brings clinical observation directly into the home through remote Zoom video conferencing, offering significant advantages in ecological validity. By recording in a familiar setting with primary caregivers, we minimize the behavioral discrepancies often associated with clinical visits, allowing for the capture of authentic, spontaneous interactions.

ROSCO is intentionally centered on non- and minimally-speaking individuals with profound autism–a population for which no large-scale, naturalistic video resources currently exist. By definition, this subpopulation includes individuals with profound intellectual disabilities, many of whom have zero or a few spoken single words, as well as overlapping genetic neurodevelopmental disorders. Furthermore, the protocol places a strong emphasis on the multimodal nature of communication: while participants are non- or minimally-speaking, the inclusion of audio allows for the analysis of nonverbal vocalizations, which serve as critical predictors of affective and communicative states (Shah and Johnson, 2025a).

The cohort comprised 27 individuals (12 male, 15 female) aged 2–12 years (mean = 6.03 years), diagnosed with Phelan-McDermid Syndrome (PMS), PTEN Hamartoma Tumor Syndrome (PHTS), or Tuberous Sclerosis Complex (TSC)–genetic neurodevelopmental disorders strongly associated with autism, profound intellectual disability, and limited verbal communication. Data collection yielded a total of 2,903 video samples (mean duration = 3.11 seconds) derived from 34 ROSCO sessions.

The ROSCO study protocol balances personalization with standardization, as each activity follows the same structure and duration while allowing caregiver-specific adaptation, controlling many variables in the recorded interaction. To control for language environment, all participants had English as the primary language spoken in their homes. The inclusion criteria allowed for all ethnicities and socioeconomic backgrounds within a larger clinical study in the US, capturing a representative sample of these clinical populations. All data were collected under IRB oversight and approval. Informed consent was obtained from legal guardians or caregivers, and assent was sought from participants wherever possible, adhering to ethical guidelines appropriate for this neurodevelopmentally diverse population. This dataset represents a unique archive of neurodiverse communication and will be released to the research community under an IRB-approved protocol.

3.2 Data Annotation

Unlike neurotypical communication, the highly idiosyncratic nature of non- and minimally-speaking autistic individuals' behaviors means that external observers–including clinicians–may misinterpret communicative acts. Caregivers who have spent years living and communicating daily with these individuals are often the most, and sometimes the only, reliable interpreters of their communication. We therefore employed a caregiver-centric annotation protocol where primary caregivers annotated the video clips by rewatching the ROSCO sessions with a researcher in 10-second snippets. In each snippet, they were asked to identify whether their child communicated, and if so, what the commu-

(a) Giving her mother an item (**gesture**) with her hands, **requesting** her help.

(b) **Looking** at his mother with a happy **facial** expression, engaging in **social** reciprocity.

(c) **Rejecting** an offered item and twisting her **body** while **vocalizing**.

(d) **Self-directed behavior** by brushing her fingers (**gesture**) on the book.

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

nicative behavior's action and function were.

Our annotation schema is both multi-label and multi-class, designed to capture the complexity of real-world interaction. Each video sample was annotated for both Communicative Action (the physical modality used) and Communicative Function (the pragmatic intent). Recognizing that communicative acts are rarely isolated, the schema allows for up to three concurrent action labels (e.g., a child might vocalize while gesturing) and up to two function labels (e.g., a child might display emotion while requesting) for each sample. To ensure the robustness of the training data, samples labeled as *OTHER* or *UNKNOWN* were filtered out, retaining only those with clear, definable communicative intent. Example frames from our dataset are presented in Figure 1.

Each sample is labeled with 1-3 of the six action classes:

1. **Alternate and Augmented Communication (AAC)** (217 samples): Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.
2. **Body** (500 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).
3. **Face** (178 samples): Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.
4. **Gestures** (1176 samples): Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.
5. **Looking** (276 samples): Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.
6. **Vocalization** (973 samples): Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

Each sample is also labeled with 1-2 of the six function classes:

1. **Commenting** (186 samples): Sharing observations, expressing interest, or drawing attention to objects or events.
2. **Emotion** (706 samples): Expressing an internal feeling state like happiness, frustration, excitement, or distress.
3. **Reject** (361 samples): Rejecting items or activities through any modality.
4. **Request** (914 samples): Asking for objects, actions, continuation, or assistance through any modality.
5. **Self-Directed Behavior (SDB)** (283 samples): Behaviors that were not perceived as being intentionally communicative or directed at another person, like self-stimulatory behavior.
6. **Social** (158 samples): Social interactions, such as sharing toys, cooperative play, and expressing greetings.

3.3 Dataset Partitioning and Challenge Sets

Given the significant variability in idiosyncratic behaviors across this population, random splitting risks overfitting to participant-specific features. To rigorously evaluate cross-participant generalization and robustness, we implemented a stratified splitting strategy with curated challenge sets involving six participants probing distinct capabilities:

Zero-Shot Generalization (TALK009). To assess the model's ability to classify behaviors of completely unseen individuals, all data from participant TALK009 was excluded from the training set.

Cross-Session Participant Generalization (TALK015b, TALK026b). To evaluate ro-

bustness across different recording sessions, we held out the entirety of sessions TALK015b and TALK026b from the training set. This tests whether the model can train using data from previous sessions of the same individual (TALK015; TALK026 and TALK026c) to generalize to new, unseen sessions, effectively measuring longitudinal consistency within the same individual.

Instruction-Type Sensitivity (TALK012, TALK025, TALK028). To analyze the impact of semantic instruction granularity, we restricted the training data for these participants to specific description types. TALK012 included only function-oriented descriptions, TALK025 included only action-oriented descriptions, and TALK028 utilized joint action-function descriptions. This ablation allows us to determine if specific semantic prompts drive better communicative understanding.

We stratified an 80/20 train-test split using balanced action and function counts for the remaining participants, yielding training and testing sets of 2001 and 902 videos, respectively. The corresponding training and testing instruction sets consisted of 5485 and 2452 samples, respectively.

4 Methodology

In this section, we describe ROSCO-Omni, our two-stage framework for interpreting the communicative behaviors of non- and minimally-speaking autistic individuals. First, we leverage a high-capability teacher MLLM (Gemini-3-Flash) to generate rich, context-aware synthetic instruction tuning data via a label-guided reasoning approach. Second, we utilize this synthetic dataset to finetune an open-source student MLLM (Qwen3-Omni), enabling domain-specialized inference that is both accessible and deployable without reliance on closed-source APIs. Our overall framework is shown in Figure 2, and MLLM prompts are presented in Appendix A.6.

4.1 Synthetic Instruction Generation via Label-Guided Reasoning

A significant bottleneck in training models for non-verbal behavior analysis is the scarcity of detailed textual descriptions aligned with labels. While ROSCO provides categorical annotations, training an instruction-following model requires richer supervision: natural language descriptions that explain *what* the child is doing and *why*. To generate this supervision at scale, we employed a teacher-student distillation paradigm, utilizing Gemini-3-Flash as the teacher model M_{T} to generate descriptive synthetic captions x_{syn} that serve as instructions for the student model.

We adopted a label-guided reasoning approach to ensure the generated captions were accurate. Following Yoo et al. (2025), we provided the teacher model with the video clip v_{i} alongside a textual hint $H(y_{\text{i}})$ derived from the expert annotations:

$$x_{\text{syn}}^{(\text{i})} \sim P_{M_{\text{T}}}(x \mid v_{\text{i}}, H(y_{\text{i}})) \tag{1}$$

For instance, if a clip was labeled with the action *Gesture*, the following prompt was injected with the hint: *The child's primary communicative action in this video is Gesture: Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.*

This conditioning forces the teacher model to focus its visual perception on the specific features relevant to that label, reducing hallucinations and ensuring the generated text aligns with the outlined action and function definitions (Cai et al., 2025).

We generated three distinct types of instruction data to support our multi-objective analysis:

Action Descriptions. Focused on the physical kinematics of the child, such as "pointing," "rocking," or "using an AAC device," using action labels as hints.

Function Descriptions. Focused on the pragmatic intent behind the behavior, such as "requesting an object" or "social initiation" using function labels as hints.

Joint Descriptions. Integrating both the how (action) and why (function) of the communicative act, providing a holistic view of the interaction using both action and function labels as hints.

To minimize redundancy, each description was generated in a stateless API call, ensuring the model had no access to previously generated descriptions. These synthetic descriptions were then paired with their corresponding video clips to form the synthetic instruction-tuning dataset $D_{\text{syn}} = \{(v_i, x_{\text{syn}}^{(i)}) \mid i = 1, 2, ...N\}$.

4.2 Student Training

For the student model M_{S}, we chose Qwen3-Omni-30B-Thinking, an open-source multimodal model capable of jointly processing video, audio, and text streams. This model was chosen for two reasons:

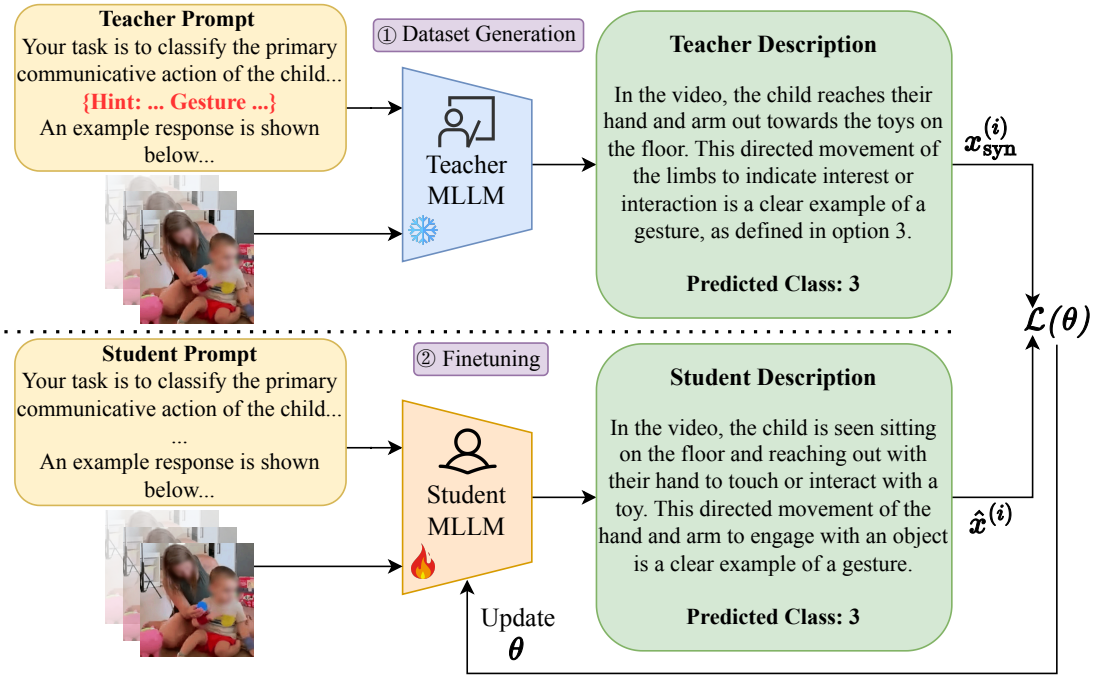

Figure 2: Overview of ROSCO-Omni, our two-stage framework. In Stage 1 (dataset generation), a teacher MLLM receives video input with label-guided hints derived from caregiver annotations and generates synthetic instruction data. In Stage 2 (finetuning), a student MLLM is trained to produce equivalent descriptions without hints, with weights updated via causal language modeling loss $\mathcal{L}(\theta)$.

its omni-modal architecture mirrors the integrative reasoning process that caregivers perform when interpreting communication, and its open-source availability enables domain-specific finetuning and eventual deployment in clinical or caregiving settings without dependence on proprietary APIs.

Finetuning. We employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) to efficiently update the model weights while mitigating catastrophic forgetting. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the update is constrained by a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \quad (2)$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable matrices with rank $r \ll \min(d, k)$. The vision transformer (ViT) encoder, multimodal aligner, and the language model components were updated with this strategy. We targeted all linear modules with a LoRA rank of $r = 8$ and an alpha of $\alpha = 16$, applying a dropout rate of 0.05 for regularization.

Loss function. The optimization objective was to minimize the causal language modeling loss on the synthetic captions x_{syn} given the visual input v_i. The loss function $\mathcal{L}$ for a sequence of length T is defined as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log P_{M_S}(x_t^{(i)} \mid x_{<t}^{(i)}, v_i; \theta) \quad (3)$$

where θ are M_S's LoRA parameters. This objective effectively teaches M_S to map visual inputs to the descriptions and classification labels, effectively transferring the semantic reasoning capabilities of M_T into the domain-specific student model, ROSCO-Omni. This approach shares similarities with reflection-based learning (Kumar et al., 2024) and self-correction in LLMs (Kamoi et al., 2024; Yoo et al., 2025).

4.3 Task Formulation

We formulated the analysis as a multi-label classification problem within a generative framework. The model is prompted to predict the communicative category (Action and/or Function) and provide a reasoning rationale. We define the space of Action labels as Y_A and Function labels as Y_F, each containing 6 classes.

Action Classification. The model predicts $\hat{y}_a \in Y_A = \{\text{AAC, Body, Face, Gesture, Looking, Vocalization}\}$.

Function Classification. The model predicts $\hat{y}_f \in Y_F = \{\text{Commenting, Emotion, Reject, Request, SDB, Social}\}$.

Joint Prediction. The model simultaneously predicts both action and function to maximize the joint probability:

$$(\hat{y}_a, \hat{y}_f) = \underset{y_a \in Y_A, y_f \in Y_F}{\operatorname{argmax}} P(y_a, y_f \mid v_i) \quad (4)$$

5 Results

5.1 Action and Function Classification

We evaluated the performance of ROSCO-Omni against open- and closed-source MLLM baselines on the ROSCO dataset. The comparative results for both action and function classification experiments are presented in Table 1. Zero-shot performance on the full dataset is presented in Appendix A.2, per-class performance is reported in Appendix A.3, and a misclassification analysis is shown in Appendix A.5.

ROSCO-Omni. Our model achieves substantial improvements over the base Qwen3-Omni-30B, notably surpassing the closed-source Gemini-2.5-Flash on both tasks. It also achieves competitive performance with Gemini-3-Flash, the strongest model overall. This demonstrates that for nonverbal communication interpretation in this population, a domain-adapted open-source model, trained with synthetic instructions generated via label-guided reasoning, can compete with larger closed-source systems.

Open-Source MLLMs. Consistent with established scaling laws (Yang et al., 2024), performance generally improved with model size. The performance gain of Qwen3-Omni-30B over Qwen3-VL-30B highlights the importance of audio for this domain–a modality absent in the vision-only Qwen3-VL models.

Closed-Source MLLMs. Gemini-3-Flash achieved the highest overall performance, establishing it as a strong teacher model for ROSCO-Omni. Its substantial improvement over Gemini-2.5-Flash confirms that newer-generation models possess stronger reasoning capabilities for complex video understanding.

Effect of In-Context Learning (ICL). We evaluated few-shot prompting using 6 manually annotated examples per task. Contrary to prior work showing substantial gains from many-shot ICL in MLLMs (Cho et al., 2025), we observed only marginal improvements over zero-shot baselines. This suggests that the extreme heterogeneity of communicative behaviors across individuals limits its few-shot prompting: static examples from one individual's behavioral profile provide little transferable signal for interpreting another's.

Joint Prediction. Predicting both labels simultaneously yielded performance comparable to single-task inference across all baselines, demonstrating that models can handle multi-objective reasoning without substantial degradation. This supports the viability of holistic communication understanding systems that produce comprehensive behavioral interpretations rather than isolated labels. Notably, joint prediction improved function classification in ROSCO-Omni. We hypothesize that multi-task instruction tuning encourages explicit reasoning about observable physical actions as a chain-of-thought scaffold (Wei et al., 2022), which narrows the hypothesis space for function prediction and reduces hallucination, i.e., answering *what is the child doing?* helps the model better infer *why*.

Table 1: Action and function classification on the ROSCO dataset. Best results are in **bold** for both closed- and open-source models.

Model	Action		Function	
	Acc	F1	Acc	F1
Closed-source				
Gemini-2.5-Flash	48.49	33.63	28.61	27.45
↪ ICL	46.99	34.08	30.47	28.51
↪ Joint Prediction	49.60	33.96	29.17	28.26
Gemini-3-Flash	52.12	42.27	33.25	30.08
↪ ICL	**54.24**	**42.78**	**35.50**	**31.88**
↪ Joint Prediction	50.06	39.61	34.85	30.83
Open-source				
Qwen3-VL-8B	39.60	25.03	23.26	20.03
Qwen3-VL-30B	39.71	19.55	26.92	21.57
Qwen3-Omni-30B	48.39	28.83	35.46	**30.88**
↪ Joint Prediction	51.53	28.94	38.57	26.62
ROSCO-Omni	**55.28**	**40.03**	39.43	28.05
↪ Joint Prediction	51.17	35.74	**39.91**	28.36

5.2 Participant-wise performance

To assess generalization across individuals, we computed classification metrics for each participant and report aggregated statistics for action and function classification in Tables 2 and 3, respectively.

ROSCO-Omni achieves the highest accuracy among all evaluated systems for both tasks, even higher than both Gemini Flash models. While

Gemini-3-Flash achieves a higher F1 score, our model performs competitively. These results show that domain adaptation substantially narrows the gap between open-source and closed-source systems, validating our synthetic instruction tuning approach for this challenging domain. The significant standard deviations across all models reflect the substantial behavioral heterogeneity across individuals in this population.

Table 2: Aggregated participant-wise action classification performance. Results report mean ± standard deviation across participants. Best results are in **bold** for both closed- and open-source models.

Model	Acc	F1
Closed-source		
Gemini-2.5-Flash	49.33±17.16	23.64±8.99
Gemini-3-Flash	**51.91**±16.78	**28.77**±12.21
Open-source		
Qwen3-VL-8B	41.44±14.67	17.19±6.25
Qwen3-VL-30B	41.68±15.00	15.46±7.21
Qwen3-Omni-30B	49.53±14.57	21.49±7.87
ROSCO-Omni	**55.55**±14.55	**26.09**±9.57

Table 3: Aggregated participant-wise function classification performance. Results report mean ± standard deviation across participants. Best results are in **bold** for both closed- and open-source models.

Model	Acc	F1
Closed-source		
Gemini-2.5-Flash	26.54±13.61	16.75±11.11
Gemini-3-Flash	**33.12**±14.97	**20.79**±10.86
Open-source		
Qwen3-VL-8B	23.68±10.51	13.77±6.85
Qwen3-VL-30B	27.98±12.82	14.86±6.95
Qwen3-Omni-30B	33.81±15.29	**19.79**±10.54
ROSCO-Omni	**38.52**±18.26	19.76±10.35

5.3 Ablation Studies and Analysis

To isolate the contribution of individual prompt and input components, we conducted ablation studies on the Qwen3-Omni-30B model, and the results are shown in Table 4. We exclude ROSCO-Omni from these experiments as it was trained with prompts referencing autism, formal class definitions, audio data, and joint action-function descriptions that expose action-function correlations.

Table 4: Ablation study using Qwen3-Omni-30B. Best results are in **bold**.

Ablation	Action		Function	
	Acc	F1	Acc	F1
Qwen3-Omni-30B	48.39	28.83	35.46	30.88
↪ No identity	47.50	25.98	35.04	30.28
↪ No definition	43.27	27.79	29.87	24.09
↪ No audio	41.05	20.50	30.11	24.79
↪ Cross-task hint	**51.70**	**30.77**	**45.09**	**32.43**

Domain-Specific Definitions. Removing explicit class definitions caused the most substantial performance degradation across both tasks. This finding reveals that MLLMs cannot reliably map general-purpose interpretations of terms like *Request* or *Social*, learned from neurotypical data, to the specialized meanings used by clinicians and researchers for this population. Domain-specific grounding is essential to align model reasoning with expert interpretations of communicative behaviors.

Identity Disclosure. Omitting references to the child being non- or minimally-speaking and autistic yielded negligible performance differences. This suggests that for recognizing communicative behaviors, explicit diagnostic labels matter less than the visual and auditory features themselves. While identity disclosure may be critical for diagnostic tasks (Yoo et al., 2025), our results indicate it is less important for behavior classification–underscoring that perceptual grounding matters more than diagnostic framing for communication understanding in this population.

Audio Modality. Removing audio input caused notable performance degradation, particularly for action classification. This confirms that vocalizations and auditory cues are integral to characterizing communicative attempts in this population, even when the individual is non- or minimally-speaking (Shah and Johnson, 2025b), validating the choice of an omni-modal architecture for ROSCO-Omni.

Cross-Task Hinting. We investigated the relationship between actions and functions by providing ground-truth labels from one task as hints for the other. This strategy yielded substantial improvements for both tasks, with function classification benefiting most from action hints. The strong bidirectional correlation confirms that physical behav-

Table 5: Participant-wise action classification performance for the challenge sets. Best results are in **bold** for both closed- and open-source models.

Model	TALK009		TALK012		TALK015b		TALK025		TALK026b		TALK028	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Closed-source												
Gemini-2.5-Flash	**46.55**	21.31	60.00	20.12	29.10	21.06	67.35	26.91	52.58	35.25	**57.14**	25.74
Gemini-3-Flash	39.66	**31.47**	**70.00**	**47.64**	**36.03**	**26.15**	**73.68**	**35.57**	**54.74**	**36.11**	**57.14**	**43.86**
Open-source												
Qwen3-VL-8B	46.55	15.71	42.50	13.66	32.59	18.07	38.54	15.30	27.84	14.55	57.14	25.95
Qwen3-VL-30B	32.76	11.46	52.50	12.50	28.89	15.58	41.67	20.31	34.02	21.27	54.29	20.58
Qwen3-Omni-30B	46.55	20.97	57.50	**24.11**	37.78	26.34	52.08	23.78	41.24	23.20	**65.71**	31.11
ROSCO-Omni	**55.17**	**31.43**	**65.00**	24.04	**45.19**	**29.22**	**57.89**	**30.86**	**54.64**	**37.24**	62.86	**31.76**

Table 6: Participant-wise function classification performance for the challenge sets. Best results are in **bold** for both closed- and open-source models.

Model	TALK009		TALK012		TALK015b		TALK025		TALK026b		TALK028	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Closed-source												
Gemini-2.5-Flash	27.45	20.62	39.47	17.06	24.37	**26.24**	25.97	25.58	**32.61**	**15.87**	26.47	8.33
Gemini-3-Flash	**34.00**	**24.34**	**41.67**	**28.11**	**29.91**	25.41	**36.84**	**32.95**	15.73	8.73	**44.12**	**15.55**
Open-source												
Qwen3-VL-8B	21.57	13.23	26.32	8.13	21.49	16.01	27.63	24.74	15.05	13.13	22.86	7.21
Qwen3-VL-30B	33.33	20.34	50.00	17.45	15.70	15.60	22.37	20.86	23.66	12.00	45.71	13.59
Qwen3-Omni-30B	**39.22**	**26.08**	57.89	27.49	26.45	22.52	**28.95**	**26.84**	**33.33**	**16.94**	**68.57**	26.22
ROSCO-Omni	**39.22**	23.90	**76.32**	**34.85**	**36.36**	**25.60**	27.63	23.64	22.58	14.18	65.71	**26.56**

iors are tightly coupled with pragmatic intent in this population, motivating our approach of leveraging one label to disambiguate the other and supporting the joint prediction formulation.

5.4 Challenge Sets

To move beyond aggregate metrics, we evaluated model performance on curated challenge sets designed to probe specific generalization capabilities. The results of the challenge set experiments for action and function classification are shown in Tables 5 and 6, respectively, with detailed analysis in Appendix A.4.

6 Conclusion

We presented ROSCO-Omni, a framework for interpreting nonverbal communication in non- and minimally-speaking autistic individuals–a population whose communicative competence has been historically overlooked. Our contributions include: (1) analysis of the ROSCO dataset, the first multi-label resource capturing communicative actions and pragmatic functions with caregiver-grounded annotations; (2) a systematic MLLM benchmark with participant-wise evaluations and challenge sets; and (3) ROSCO-Omni, a teacher-student distillation framework enabling an efficient open-source model to achieve competitive performance with closed-source alternatives.

By shifting focus from diagnosis to communication, this research supports a more human-centered approach to autism research that foregrounds communicative agency. We hope our work will catalyze further research in this underserved domain, ultimately contributing to tools that help caregivers, educators, clinicians, and the general public better understand this population.

Acknowledgements

We thank Dr. Lauren Thompson, Tsambika Rizas, Miranda Kannisto, Emma McGonigle, Bianca Booth, Emine Arcasoy, Isabelle Iannotti, Dr. Carol Wilkinson, Dr. Audrey Thurm, Dr. Mustafa Safin, and the participants for their contributions to the initial ROSCO dataset. This research was supported in part by NIH NINDS TALK Supplement (3U54NS092090-10S1).

Ethical Considerations

Potential Risks. The primary risk of developing automated communication interpreters is misinterpretation: systems classifying subtle, idiosyncratic communicative acts may produce errors that lead to incorrect assumptions about an individual's needs or intent, potentially compromising care quality and individual autonomy. Additionally, in-home behavioral analysis systems must be safeguarded against misuse, including non-consensual surveillance or data exploitation. Any deployment of these methods must be governed by strict data privacy protocols, informed consent procedures, and frameworks that prioritize the individual's dignity, autonomy, and well-being.

Biases. Our dataset reflects natural communicative distributions, resulting in substantial class imbalance. For example, *Gestures* are significantly more frequent than *Face*, which may bias model predictions toward majority classes. This issue is compounded by population heterogeneity: the high variance in participant-wise performance demonstrates that models performing well on aggregate metrics may fail substantially for individuals whose idiosyncratic communication styles are underrepresented in training data. An over-reliance on systems with unaddressed biases could produce inequitable outcomes, disproportionately misunderstanding individuals with rarer or more subtle communicative patterns.

Reproducibility Statement. We include implementation details and hyperparameter settings for all models in Appendix A.1. Our code is publicly available at https://github.com/SiddhantBikram/ROSCO-Omni.

Limitations

Despite our contributions, we acknowledge several limitations of our research. First, absolute model performance remains modest, indicating substantial room for improvement before clinical deployment.

Second, while ROSCO represents the largest naturalistic video dataset collected for this population, it is limited in scale (2,903 samples from 27 participants) and restricted to three specific syndromes (PMS, PHTS, TSC). Data from individuals with profound autism and complex neurodevelopmental disorders is deeply underrepresented and difficult to collect due to specialized study design requirements, participant recruitment challenges, vulnerability considerations, and stringent IRB requirements. Generalization beyond our cohort to the broader autistic population or other neurodevelopmental disorders remains unvalidated.

Third, nonverbal communication norms vary substantially across cultures: for example, what constitutes a "request" gesture or an appropriate social initiation differs across cultural contexts. Caregiver interpretation of communicative intent is also culturally situated. While our cohort captures demographic diversity within a US-based clinical population, our annotations and trained models may not generalize to populations outside this demographic profile. Similarly, the home recording environment–including available objects, space, noise levels, and the caregiver's interaction style–shapes the types and frequencies of behaviors observed. These uncontrolled contextual factors contribute to the substantial inter-participant variability in our results (Tables 2 and 3) and may affect model generalization to different settings.

Fourth, our caregiver-centric annotation protocol, while providing the most ecologically valid ground truth for this population, captures a single interpretive perspective that may not fully align with interpretations by other caregivers, educators, or clinicians, and the risk of misinterpretation exists for both human and machine interpreters. Developing multi-annotator protocols and cross-rater validation for this population is an important direction for future work.

Finally, ROSCO-Omni focuses exclusively on action and function classification, which represents only a subset of the full communication understanding pipeline. Larger-scale data collection, cross-population validation, and multi-perspective annotation represent promising directions for future work.

References

Masroor Ahmed, Sadam Hussain, Farman Ali, Anna Karen Gárate-Escamilla, Ivan Amaya, Gilberto Ochoa-Ruiz, and José Carlos Ortiz-Bayliss. 2025. Summarizing recent developments on autism spectrum disorder detection and classification through machine learning and deep learning techniques. *Applied Sciences*, 15(14):8056.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. 2025. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 239–249.

Yang Chen, Jingcai Guo, Miaoge Li, Zhijie Rao, and Song Guo. 2026. Star++: Region-aware conditional semantics via interpretable side information for zero-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hyundong Justin Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T Chavez, and Jonathan May. 2025. Can vision language models understand mimed actions? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26744–26759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Shijian Deng, Erin E Kosloski, Siddhi Patel, Zeke A Barnett, Yiyang Nan, Alexander Kaplan, Sisira Aarukapalli, William T Doan, Matthew Wang, Harsh Singh, and 1 others. 2024. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia*.

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

Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. 2025a. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, 133(11):8151–8189.

Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, and 1 others. 2025b. Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning. *arXiv preprint arXiv:2502.16268*.

Kristina T Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W Picard. 2023. Recanvo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Stephen Wen Hwooi Khor, Aznul Qalid Md Sabri, and Alice Othmani. 2024. Autism classification and monitoring from predicted categorical and dimensional emotions of video features. *Signal, Image and Video Processing*, 18(1):191–198.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, and 1 others. 2024. Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In *Proceedings of the eleventh ACM conference on learning@ scale*, pages 86–97.

Jicheng Li, Vuthea Chheang, Pinar Kullu, Eli Brignac, Zhang Guo, Anjana Bhat, Kenneth E Barner, and Roghayeh Leila Barmaki. 2023. Mmasd: A multimodal dataset for autism intervention analysis. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 397–405.

Haoxuan Qu, Yujun Cai, and Jun Liu. 2024. Llms are good action recognizers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18395–18406.

V Rose, David Trembath, Deb Keen, and Jessica Paynter. 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.

Manuel Serna-Aguilera, Xuan Bac Nguyen, Han-Seok Seo, and Khoa Luu. 2024. A novel dataset for video-based autism classification leveraging extra-stimulatory behavior. *arXiv preprint arXiv:2409.04598*.

Siddhant Bikram Shah and Kristina T Johnson. 2025a. Multi-feature audio fusion for nonverbal vocalization classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Siddhant Bikram Shah and Kristina T Johnson. 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*, pages 33362–33379.

Siddhant Bikram Shah and Kristina T Johnson. 2026. Axon: Action characterization through cross-modal knowledge distillation for neurodiverse individuals.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39228–39238.

Helen Tager-Flusberg and Connie Kasari. 2013. Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism research*, 6(6):468–478.

Chelsea La Valle, Karen Chenausky, and Helen Tager-Flusberg. 2021. How do minimally verbal children and adolescents with autism spectrum disorder use communicative gestures to complement their spoken language abilities? *Autism & Developmental Language Impairments*, 6:23969415211035065.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu. 2024. Law of vision representation in mllms. *arXiv preprint arXiv:2408.16357*.

Ziqian Yang, Yuyao Zhang, Jiachuan Ning, Xin Wang, and Zhihui Wu. 2025b. Early diagnosis of autism: A review of video-based motion analysis and deep learning techniques. *IEEE Access*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Cheol-Hwan Yoo, Jang-Hee Yoo, and Jaeyoon Jang. 2025. Care-vl: A domain-specialized vision-language model for early asd screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 57–66. Springer.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

A Appendix

A.1 Experimental Settings

Implementation Details. Model finetuning and inferencing were conducted on a single NVIDIA H200 GPU using the ModelScope Swift (Zhao et al., 2025) framework with Flash Attention 2 (Dao, 2023). We used a batch size of 4 and accumulated gradients over 4 steps. Gemini API was used to generate descriptions for the instruction dataset and for Gemini-based model inference. We processed video inputs at a resolution of $24 \times 28 \times 28 = 18,816$ pixels with a sampling rate of 2 frames per second (FPS), ensuring the model captured sufficient temporal granularity to distinguish subtle movements. ROSCO-Omni was trained for 3 epochs with a learning rate of 2×10^{-5} and a cosine decay scheduler, with each epoch requiring approximately 7 hours.

Baselines. We benchmarked both closed-source and open-source MLLMs. For the closed-source models, we used Gemini-2.5-Flash (Comanici et al., 2025), and Gemini-3-Flash. For the open-source models, we used Qwen3-VL-8B-Thinking (Yang et al., 2025a), Qwen3-VL-30B-A3B-Thinking, and Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025).

A.2 Zero-Shot Evaluation on the Full Dataset

To validate that our test set findings generalize beyond the stratified split, we conducted zero-shot classification experiments on the entire ROSCO dataset, and our results are presented in Table 7. Performance trends on the full dataset closely mirror those observed on the test set, and ablation results on the full dataset support our earlier findings. Joint prediction maintains competitive performance while providing both labels simultaneously, supporting the viability of unified communication understanding systems. The consistency between test set and full dataset results suggests that our stratified splitting strategy produced a representative evaluation set, and that the observed model behaviors reflect genuine capabilities rather than artifacts of data partitioning.

Table 7: Zero-shot action and function classification on the entire ROSCO dataset. Best results are in **bold** for both closed- and open-source models.

Model	Action		Function	
	Acc	F1	Acc	F1
Closed-source				
Gemini-2.5-Flash	50.98	35.60	29.98	27.13
↪ ICL	48.38	35.69	29.54	26.99
Gemini-3-Flash	50.82	40.36	**36.28**	**32.19**
↪ ICL	**53.14**	**41.73**	36.08	31.76
Open-source				
Qwen3-VL-8B	39.79	24.17	24.89	20.46
Qwen3-VL-30B	42.01	21.89	27.14	21.89
Qwen3-Omni-30B	50.65	29.13	34.17	28.92
↪ No identity	51.27	26.73	37.08	31.11
↪ No definition	45.56	28.80	30.98	23.99
↪ No audio	43.59	23.43	31.17	24.61
↪ Cross-task hint	**54.70**	**32.95**	**45.36**	**32.06**
↪ Joint prediction	51.53	28.94	43.03	27.03

A.3 Per-Class Performance Analysis

To understand model behavior across individual categories, we report per-class precision, recall, and F1-score for ROSCO-Omni in Tables 8 and 9.

Action Classification. Performance varies substantially across action classes, reflecting both class imbalance and inherent recognition difficulty. *Gesture*, the most frequent class (1,176 samples), achieves the highest performance, benefiting from abundant training examples and visually distinctive hand/arm movements. *AAC* and *Vocalization* achieve moderate performance, with AAC's high precision suggesting the model reliably identifies device usage when it predicts this class. In contrast, *Looking* and *Face* exhibit poor recall, indicating that the model frequently fails to detect these subtle behaviors–likely because they rely on fine-grained visual cues that are difficult to capture in brief video clips recorded via Zoom.

Function Classification. Function recognition proves more challenging overall, with lower F1-scores across most classes. *Request*, the most frequent class (914 samples), achieves the highest performance, while *Emotion* attains moderate performance. The model struggles substantially with *Social* and *SDB*, both of which are minority classes that require inferring abstract intent rather than observable physical actions. *Commenting* and *Reject* show intermediate performance but with precision-recall trade-offs. The difficulty of function classification, particularly for classes requiring pragmatic reasoning, highlights the inherent challenge of inferring communicative intent from brief video segments and suggests that richer contextual information or multi-turn interaction modeling may be necessary for improvement.

Table 8: Per-class action classification performance for ROSCO-Omni.

Action	Samples	Precision	Recall	F1
AAC	217	65.38	43.59	52.31
Body	500	53.61	25.62	34.67
Face	178	28.57	21.43	24.49
Gesture	1,176	58.35	67.82	62.73
Looking	276	27.59	9.88	14.55
Vocalization	973	56.20	47.39	51.42

Table 9: Per-class function classification performance for ROSCO-Omni.

Function	Samples	Precision	Recall	F1
Commenting	186	17.65	13.04	15.00
Emotion	706	43.53	49.75	46.44
Reject	361	47.76	25.20	32.99
Request	914	46.07	60.27	52.23
SDB	283	16.42	10.09	12.50
Social	158	10.42	8.20	9.17

A.4 Challenge Set Details

The results of the challenge set experiments for action and function classification are shown in Tables 5 and 6.

Zero-Shot Generalization (TALK009). On the held-out participant unseen during training, our model significantly outperformed the Qwen3-Omni baseline and achieved parity with the larger teacher model, demonstrating that our synthetic instruction tuning effectively learns transferable communicative patterns rather than overfitting to participant-specific visual features.

Cross-Session Generalization (TALK015b, TALK026b). Evaluating on held-out sessions from participants with training data in other sessions tests longitudinal consistency. On TALK015b, ROSCO-Omni improved substantially over the baseline for both action and function classification. For TALK026b, while action classification improved, function classification

degraded significantly, suggesting that physical actions transfer more reliably across sessions than communicative intent, which may vary due to contextual or environmental factors.

Instruction-Type Sensitivity (TALK012, TALK025, TALK028). Training with function-oriented descriptions only (TALK012) yielded the strongest improvements over the baseline. This suggests that grounding on communicative intent provides a more robust semantic anchor than physical descriptions alone. In contrast, action-only descriptions (TALK025) produced modest action gains but degraded function understanding, while joint descriptions (TALK028) underperformed single-task baselines on both metrics–suggesting that longer multi-objective prompts may dilute the learning signal when an individual's training data is restricted to a single description type.

A.5 Misclassification analysis

We report the top-5 most frequent misclassifications across Gemini-3-Flash, Qwen3-Omni-30B, and ROSCO-Omni in Table 10.

For action classification, we observed a persistent ambiguity across all evaluated models, where *Vocalization* and *Body* labels were frequently mislabeled as *Gestures*. While ROSCO-Omni reduced the absolute frequency of these errors compared to the base Qwen3-Omni-30B baseline, the structural pattern of confusion remained consistent.

For function classification, the error profiles expose differences in how models perceive neurodiverse intent. The base Qwen3-Omni model frequently misclassified pragmatic intents, such as *Request* or *Emotion*, as *SDB*. This represents a clinically significant failure mode, as it effectively dismisses valid communicative attempts as non-communicative stimming. In contrast, ROSCO-Omni shifted this error distribution, eliminating SDB from its top-5 confusions entirely, and more frequently confusing *Reject* or *SDB* with *Request*. While still erroneous, this qualitative shift indicates that our label-guided instruction tuning successfully biased the model towards recognizing communicative agency, moving away from the base model's tendency to view neurodiverse behavior as passive or non-interactive.

A.6 Prompt Templates

We designed structured prompts to standardize input for ROSCO-Omni and elicit interpretable reasoning through a CoT process. Each prompt establishes the diagnostic context by identifying the subject as a non- or minimally-speaking child with profound autism and complex neurodevelopmental conditions. The model is explicitly instructed to focus on the child and ignore adults, a necessary constraint for naturalistic home videos showcasing child-caregiver interaction. The action, function, and joint classification prompts are shown in Figures 3, 4, and 5, respectively.

Table 10: Top-5 most frequent misclassification cases across Gemini-3-Flash, Qwen3-Omni-30B, and ROSCO-Omni.

Model	Action			Function		
	Actual	**Predicted**	**Frequency**	**Actual**	**Predicted**	**Frequency**
Gemini-3-Flash	Gesture	Looking	67	Request	Social	67
	Vocalization	Gesture	51	Emotion	Social	45
	Vocalization	Looking	44	Reject	Request	39
	Body	Gesture	41	Emotion	Request	31
	Gesture	Vocalization	40	Reject	Social	27
Qwen3-Omni-30B	Vocalization	Gesture	110	Request	SDB	76
	Body	Gesture	89	Emotion	SDB	69
	Gesture	Vocalization	38	Emotion	Request	37
	AAC	Gesture	36	Emotion	Request	34
	Body	Vocalization	35	Emotion	Reject	30
ROSCO-Omni	Vocalization	Gesture	96	Reject	Request	59
	Body	Gesture	84	SDB	Request	49
	Gesture	Vocalization	40	Emotion	Request	49
	AAC	Gesture	36	Request	Emotion	45
	Body	Vocalization	29	SDB	Emotion	36

Action Classification Prompt

Your task is to classify the primary communicative action of the child in this video. The child is non- or minimally-speaking and has profound autism and other complex neurodevelopmental disorders. Choose the most accurate option below.

0: AAC: Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.

1: Body: 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).

2: Face: Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.

3: Gesture: Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.

4: Looking: Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.

5: Vocalization: Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

Reasoning Steps:

1. First, identify the child in the video. You must ignore any adults present.
2. Understand the most significant communicative action performed by the child.
3. Choose the single best-fitting class from the list above.

Carefully think through the answer by briefly detailing the particular movements that you see the child doing. Your output must contain your explanation, and then in a new line, a single integer corresponding to the option you choose. An example response is shown below:

In the video, the child is pointing a single finger toward an object. This is most accurately described by option 3.

3

Figure 3: Action classification prompt

Function Classification Prompt

Your task is to classify the primary communicative function of the child in this video. The child is non- or minimally-speaking and has profound autism and other complex neurodevelopmental disorders. Choose the most accurate option below.

0: Commenting: Sharing observations, expressing interest, or drawing attention to objects or events.

1: Emotion: Expressing an internal feeling state like happiness, frustration, excitement, or distress.

2: Reject: Refusing items, activities, pushing away, head shaking, or vocal protests indicating a desire for an activity to stop.

3: Request: Asking for objects, actions, continuation, or assistance through any modality.

4: Self-Directed Behavior: Behaviors that were not perceived as being intentionally communicative or directed at another person, like self-stimulatory behavior.

5: Social: Initiating social interaction, maintaining social reciprocity, greeting, responding to a name, or joint attention.

Reasoning Steps:

1. First, identify the child in the video. You must ignore any adults present.
2. Understand the most significant communicative function performed by the child.
3. Choose the single best-fitting class from the list above.

Carefully think through the answer by briefly detailing the particular movements that you see the child doing. Your output must contain your explanation, and then in a new line, a single integer corresponding to the option you choose. An example response is shown below:

In the video, the child is pointing a single finger toward an object. This is most accurately described by option 3.

3

Figure 4: Function classification prompt

Joint Action and Function Classification Prompt

Your task is to classify the primary communicative action and function of the child in this video. The child is non- or minimally-speaking and has profound autism and other complex neurodevelopmental disorders. Choose the most accurate option below.

ACTION Class Definitions:

0: Alternate and Augmented Communication: Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.

1: Body: 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).

2: Face: Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.

3: Gesture: Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.

4: Looking: Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.

5: Vocalization: Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

FUNCTION Class Definitions:

0: Commenting: Sharing observations, expressing interest, or drawing attention to objects or events.

1: Emotion: Expressing an internal feeling state like happiness, frustration, excitement, or distress.

2: Reject: Refusing items, activities, pushing away, head shaking, or vocal protests indicating a desire for an activity to stop.

3: Request: Asking for objects, actions, continuation, or assistance through any modality.

4: Self-Directed Behavior: Behaviors that were not perceived as being intentionally communicative or directed at another person, like self-stimulatory behavior.

5: Social: Initiating social interaction, maintaining social reciprocity, greeting, responding to a name, or joint attention.

Reasoning Steps:

1. First, identify the child in the video. You must ignore any adults present.
2. Determine the primary action (how the child is communicating).
3. Determine the primary function (why the child is communicating).
4. Choose the single best-fitting class for each category.

Carefully think through the answer by briefly detailing the particular movements that you see the child doing. After your reasoning, your response MUST end with exactly this format on the last two lines:

```
Action: <single digit 0-5>
Function: <single digit 0-5>
```

Figure 5: Joint classification prompt