



Can Multimodal Foundation Models Help Analyze Child-Inclusive Autism Diagnostic Videos?

Aditya Kommineni¹, Digbalay Bose¹, Tiantian Feng¹, So Hyun Kim², Helen Tager-Flusberg³, Somer Bishop⁴, Catherine Lord⁵, Sudarsana Kadiri¹, Shrikanth Narayanan¹

¹Signal Analysis and Interpretation Laboratory, University of Southern California, USA; ²School of Psychology, Korea University, Korea; ³Center for Autism Research Excellence, Boston University, USA; ⁴Department of Psychiatry, University of California San Francisco, USA; ⁵Semel Institute of Neuroscience and Human Behavior, University of California Los Angeles, USA

akommine@usc.edu, dbose@usc.edu, tiantiaf@usc.edu, sophymail@gmail.com, htagerf@bu.edu, Somer.Bishop@ucsf.edu, clord@mednet.ucla.edu, skadiri@usc.edu, shri@usc.edu

Abstract

Multimodal foundation models have paved the way for a paradigm shift in long video understanding. The extent to which these models can help analyze verbal and non-verbal behaviors in the context of human interactions is underexplored, particularly in the challenging settings of clinical diagnosis and treatment. We investigate the use of foundation models across speech, video, and text modalities to analyze child-focused interactions in the context of autism diagnosis. We evaluate model performance in two related tasks, i.e. activity understanding and atypical behavior detection. We further propose a unified methodology for merging information from audio and video streams by leveraging large language models as reasoning agents. Our experiments reveal that, while models perform relatively well for coarse-grained tasks such as activity recognition and over-activity identification, they fail to generalize to fine-grained tasks such as anxiety detection and activity segmentation.

Index Terms: foundation models, autism spectrum disorder, child-inclusive, video language models

1. Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder characterized by challenges with social skills, repetitive behaviors, and nonverbal communication. The condition is highly prevalent globally; for example, according to the Center for Disease Control and Prevention (2023), about 1 in 36 children [1] in the United States are diagnosed with ASD. Currently, diagnosis and behavioral changes related to treatment are evaluated through clinically-validated instruments such as Autism Diagnostic Observation Schedule (ADOS) [2] and Brief Observation of Social Communication Change (BOSCC) [3]. Both involve dyadic interaction sessions between a clinician/caregiver and the child, and incorporate a multitude of complex activities, such as puzzle solving, story creation, toy play, and conversation about emotions, loneliness and social difficulties. These sessions are recorded as long-form videos and contain rich verbal and nonverbal cues. Automatically analyzing such videos at scale using foundation models could better quantify and expand on the behaviors of autistic children.

Multimodal foundation models such as GPT-4o [4] and Gemini [5] have shown impressive performance in visual question answering and spatial reasoning benchmarks. The ability to employ these models in diagnostic interaction settings, such as child-clinician interactions, could provide valuable information in quantifying measures of behavioral patterns related to

the neuro-developmental condition. In medical context, foundation models [5, 6, 7, 8] have shown impressive performance in standard medical examinations and multimodal pathological image analysis providing evidence of their clinical utility. However, the extent to which foundation models could analyze real-world human interactions to yield relevant insights in challenging clinical settings and populations is not well-studied.

In this work, we investigate the potential of multimodal foundation models to analyze human interactions in a child-inclusive diagnostic setting for Autism Spectrum Disorder (ASD). Children on the autism spectrum may exhibit clinically relevant behavior during these sessions, such as inconsistent gaze patterns, repetitive behaviors with toys, and repeated words or phrases (echolalia) [9, 10, 11]. These behaviors are coded either by the clinicians during the assessment or during post-assessment. However, this coding process can still be complemented and supplemented by nuanced interaction details that may be overlooked. Moreover, the whole process requires significant manual effort, which prevents the scaling of the analysis of these diagnostic sessions. Hence, the ability to computationally analyze and interpret these behaviors could help clinicians quantify them more effectively, ultimately enabling the automation of large-scale diagnostic analysis.

Prior work in analyzing child-inclusive videos related to ASD have focused on action recognition [12, 13] and distinguishing autistic from typically developing children [14, 15, 16]. However, a classification label is rarely explainable and does not substantially inform clinicians during diagnosis. Instead, providing measures of specific behaviors such as social affect and restrictive repetitive behaviors would aid clinicians in better conducting the behavior phenotyping. This work represents one of the first studies to explore the utility of multimodal foundation models in a diagnostic setting of child-adult interactions, particularly in the context of ASD. Our main contributions are the following.

- We propose a unified methodology that combines information from video and speech modalities to achieve robust performance across different recording settings.
- We provide a comprehensive analysis of the capabilities of foundation models in analyzing child-adult interactions in autism diagnostic sessions by evaluating two tasks: activity understanding and atypical behavior detection.
- We report noticeable performance gains (~20% relative) while reasoning with Large Language Models (LLMs) and natural language descriptions of video and speech, compared to zero-shot inference of Video-Language Models (VLMs).

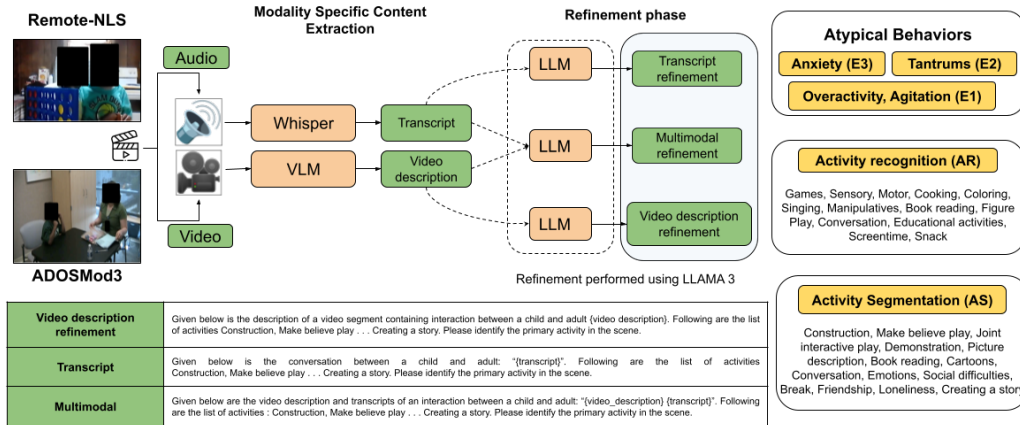


Figure 1: Schematic overview of the proposed multimodal processing pipeline. During the modality-specific content extraction, natural language descriptions of video and speech are obtained. These descriptions are then used for LLM refinement. E1, E2, and E3 are binary classification tasks. The classes for Activity Recognition and Activity Segmentation are as mentioned. Example prompts corresponding to each refinement mode are provided in the bottom table.

Table 1: Details of the datasets used in the analysis. Age is reported in years, Duration of session is reported in minutes.

Dataset	Setting	Adult	Duration	# Sessions	Age	Gender	Tasks
Remote-NLS	Zoom recordings	Parents	~15	89	6.26±1.07	70M, 19F	Activity Recognition
ADOSMod3	Diagnostic	Clinician	~60	83	8.68±2.33	56M, 27F	Activity Segmentation, Atypical behaviors

2. Tasks

To evaluate the performance of foundation models, we consider two exemplary tasks, namely activity understanding and atypical behavior detection. While activity understanding tasks let us evaluate the model’s ability to identify human-object interactions and corresponding nonverbal gestures, atypical behavior tasks offer additional insights in analyzing conversations and interaction-behavioral dynamics.

2.1. Activity Understanding

Diagnostic videos of ASD feature semi-structured interactions with a pre-defined set of activities conducted throughout the session. Each activity explores a target behavior from the child. This makes it important for the models to be able to identify the activity that could further help identify any deviation from expected typical behavior. Activity understanding comprises two subtasks: activity recognition and activity segmentation.

Activity Recognition (AR) is a multilabel task wherein, given a video, the objective is to identify all the activities that occur during the course of the video.

Activity Segmentation (AS) involves identifying specific activities and localizing their timestamps within a video.

2.2. Atypical Behaviors

Atypical behaviors in children refer to atypical verbal or non-verbal actions that are deemed not appropriate for the respective setting or context. During the course of autism diagnostic assessment, clinicians note atypical behaviors that a child might exhibit. These are organized into three constructs, **Overactivity/Agitation (E1)**, **Tantrums (E2)** and **Anxiety (E3)**. For a given clinical interaction, the coding for each construct corresponds to a binary value, i.e. presence/absence of said behavior. The distributions are shown in Table 2.

Overactivity/Aggression (E1) evaluates excessive movement or physical agitation of the child during the ADOS session. This item is coded relative to the participant’s nonverbal mental age. Notable characteristics include standing up from the chair, walking around the room during the session, and fidgeting or moving about in the chair.

Tantrums, Aggression, Disruptive Behavior (E2) refers to any form of anger or disruption by the child during the course of the interaction. This includes behaviors such as occasional mild disruptions in the form of anger, aggression, throwing things, hitting or biting others, and loud screaming or yelling.

Anxiety (E3) is measured through signs displayed by the child during the entire session. Notable signs include worry, upset, or concern from the child, including trembling or jumpiness.

3. Datasets

To evaluate our method, we consider the Remote-NLS (Naturalistic) and the ADOSMod3 (semi-structured) child-adult interaction datasets. Details about the demographics and recording settings of the datasets are provided in Table 1. We comply with the data usage terms mentioned in the IRB and DUAs from the original data owners.

Remote-NLS [17] contains 89 Zoom recordings of 15-minute child-parent interactions focusing on the child’s spontaneous spoken language in a naturalistic context. An example frame is shown in Fig. 1, wherein the child is playing a game of Connect4 with their parents. For this dataset, we explore the Activity Recognition task, where the objective is to identify the subset of 13 activities (mentioned in Fig. 1) at the session level. **ADOSMod3** comprises child-clinician diagnostic interactions for ASD following the ADOS-2 protocol [2, 18]. The dataset contains 83 videos of Module 3 (~1 hour each), designed for verbally fluent children. Each session is administered to elicit spontaneous interaction and observe the child’s verbal and non-

verbal communication and behavior through 14 different activities (Fig. 1). During the session, clinicians evaluate behaviors associated with communication, interactions, and gesturing by the child. For this dataset, we evaluate Activity Segmentation and all three atypical behavior constructs (E1, E2, E3).

4. Methodology

Traditional video processing methods struggle to adapt to long video understanding [19]. Instead, we draw inspiration from recent works in visual reasoning, robotic planning, and navigation domain [20, 21, 22], wherein visual entities such as charts and images are converted to natural language through descriptions. The reasoning agent (LLM) then leverages the enriched prompt with these descriptions to perform single-step or multi-step inference for the desired task. Similarly, rather than performing modality fusion through unified training of vision and speech models, we opt for a training-free alternative wherein task-relevant information from video and speech are extracted from pre-trained models (video-language models and automatic speech recognition (ASR) models, respectively). Following this, the natural language descriptions for each modality are provided to an LLM to perform inference (Fig. 1). In this manner, the reasoning agent leverages the complementary information available in both modalities without explicitly training a combined model. Fig. 1 shows the described pipeline along with example prompts corresponding to the refinement procedure. The textual descriptions of different modalities are integrated into LLM refinement as described in the following.

Video captions are generated by VLMs through prompting for a detailed description of the provided video. The generated description is provided to an LLM to derive the predictions for the specific task. This process allows the VLM to generate a descriptive caption for the video that goes beyond predefined class labels. Following this, the LLM is able to infer the downstream task through the contextually rich description.

Transcriptions are generated from Whisper large-v3 [23], separately for each video. Although the transcriptions would contain errors (WER: Adult \sim 25; Child \sim 55) and would not have the speaker attributions for each utterance, we hypothesize that the information content in transcripts would contain adequate information for downstream tasks. Transcripts are then provided as context to LLMs for the aforementioned activity understanding and atypical behavior detection tasks.

Multimodal setting consists of refinement from LLMs by providing both transcriptions and video captions. The prompt example for this refinement is shown in Fig 1. We assume that LLMs would perform better on the tasks by leveraging the complementary information provided by each stream.

5. Experiments

Due to privacy concerns and the sensitive nature of the data, cloud-based models such as GPT-4 [4] and Gemini [5] cannot be used. Hence, we rely on open-sourced models which can be deployed in secure local servers. For video-language models, we use LLaVA-NeXT-Video 7B DPO, LLaVA-NeXT-Qwen-32B[24] and Video-LLaMA2 7B[25] models. Meta-Llama-3-8B-Instruct[26] is the LLM chosen, owing to its large context length, and Whisper[23] is used for generating the audio transcripts from videos. The 7B/8B models are used at 16-bit precision, and the 32B models are used at 8-bit precision. A single A6000 GPU is used for inference in all experiments.

As a baseline, we show the zero-shot performance of the

Table 2: *Zero-shot and LLM refinement classification results. Activity {Recognition, Segmentation} - AR, AS; overactivity/agitation - E1, tantrums - E2 and anxiety - E3. F1-Macro is reported for AR and AS. PR-AUC is reported for E1, E2 and E3. A refers to absence and P refers to Presence of the atypical behavior. For all metrics, higher is better.*

Model	AR (CAASL)	AS (ADOS)	E1	E2	E3
# Classes	13	14	2 (A - 48) (P - 34)	2 (A - 76) (P - 7)	2 (A - 73) (P - 10)
<i>Zero-shot</i>					
VidL 2 7B ^a	31.6	4.0	84.2	12.4	22.6
L-Next Vid 7B DPO ^b	27.6	3.8	81.5	4.2	19.3
L-Next Vid 32B Qwen ^c	36.8	4.4	91.7	54.2	28.0
<i>LLM Refinement</i>					
↳Video only					
VidL 2 7B ^a	37.7	5.1	78.1	45.5	19.9
L-Next Vid 7B DPO ^b	36.8	4.3	76.8	54.2	23.7
L-Next Vid 32B Qwen ^c	42.8	3.7	61.3	4.2	18.2
↳Transcript only					
whisper-L (16s)	16.3	12.3	83.9	10.7	9.8
whisper-L (64s)	39	22	79.1	10.4	12.0
↳Multimodal					
VidL 2 7B ^a	41.3	12.4	80.1	37.3	16.0
L-Next Vid 7B DPO ^b	38.7	9.6	70.0	9.4	28.8
L-Next Vid 32B Qwen ^c	44.8	9.8	82.2	10.4	12.0

^aVideo-LLaMA2 7B, ^bLLaVA-NeXT-Video 7B DPO, ^cLLaVA-NeXT-Qwen-32B

video-language models on the respective tasks. Then, LLM refinement is performed for three conditions, namely *video only*, *transcript only*, and *multimodal*, as described in the previous section. Zero-shot and video descriptions are extracted for video segments sampled at 1 fps for 16 seconds. A label is generated for each video segment for each task.

Segment-level activity timestamps are unavailable for activity recognition. Hence, evaluation is performed at session level. Session labels are determined based on the segment-level activity predictions. An activity is considered part of the session if it is predicted by the model for at least 90 seconds. This threshold is heuristically set based on the task design. For activity segmentation, metrics are computed at the segment level as segment-level timestamps are available. Macro F1 scores are reported for both Activity Recognition and Segmentation tasks.

For the atypical activities, we have a single label per session. From the video segment predictions, the ratio of number of segments with presence of atypical behavior from predictions is computed. A higher ratio indicates more frequent atypical behavior being observed by the models. This ratio is then used to compute the evaluation metrics. Since E1, E2 and E3 have significant class imbalance, the Precision-Recall Area Under Curve (PR-AUC) metric is reported.

For *transcript-only* refinement, two settings are evaluated, i.e. transcript chunks corresponding to 16-second and 64-second video segments. This is done to ensure a fair comparison for speech modality, as 16 seconds of audio (about 2-3 sentences in total) may not contain enough task-specific cues.

6. Results

6.1. Activity Recognition

Table 2 shows the results for both activity understanding and atypical behavior detection tasks. For activity recognition, we see that both zero-shot and LLM refinement perform significantly better than random chance. Additionally, we see that augmenting the prompt with contextually rich descriptions from

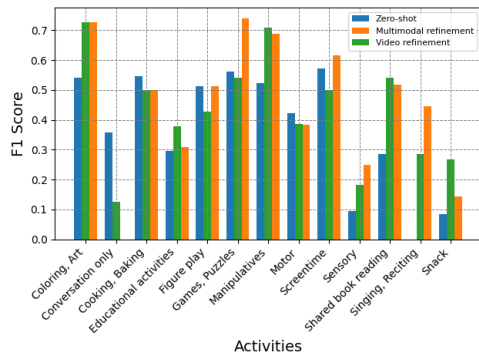


Figure 2: Class-wise Activity Recognition F1 Score for LLaVA-NeXT-Qwen-32B.

videos can improve the LLM refinement performance (~20% relative) compared to zero-shot inference from VLMs across all tested models. Fig. 2 shows the class-wise F1 scores for activity recognition on the best performing VLM, i.e. LLaVA-NeXT-Qwen-32B. Here, we see that the zero-shot inference is unable to generalize well for activities such as *snack*, and *sensory*. This behavior could be explained by VLMs being better at generating descriptions about visual entities but unable to reason the distinctions between these nuanced categories. Hence, LLMs are able to use these descriptive captions of the video to provide better reasoning, thereby better demarcating these classes. Performance gains are observed with multimodal refinement in tasks such as *singing*, *reciting* and *shared book reading*, pointing to LLM being able to leverage complementary information from audio and video to perform more accurate reasoning.

6.2. Activity Segmentation

Activity Segmentation on ADOSMod3 results in chance-level performance for zero-shot and video description-only refinement. This could be explained by two factors. Firstly, the video descriptions generated by the VLMs are unable to capture fine-grained details necessary for the LLM refinement to reason about the downstream tasks. It is worth noting that multiple activities (make-believe play, join interactive play, creating a story) in an ADOS session involve the use of similar toys. The VLM descriptions do not capture these distinctions, thereby causing the LLM refinement to be unable to distinguish these activities. Also, ADOSMod3 dataset contains conversational activities (loneliness, emotions, conversation and reporting), wherein the details are solely captured by speech modality. Hence, for transcript-only or multimodal LLM refinement, we observe a noticeable performance improvement over video-only LLM refinement, indicating a stronger signal for activity segmentation in speech.

6.3. Atypical behaviors

Both zero-shot and LLM refinement are able to capture Overactivity (E1) in the videos, as indicated by the results in Table 2. E1 is often characterized by the child standing up from the chair and moving around the room. Since VLMs are trained on action recognition and movement tasks, they are able to map the construct of overactivity to movement of the child around the room. As far as E2 is concerned, accounting for tantrums and negative behavior requires understanding the voice characteristics (such as loud voice or verbal threats) of the child in addition

to the speech content and actions. This leads to the LLM refinement being unable to generalize well to this task. For E3, we observe that the VLM descriptions incorrectly recognize the child as trembling even though the child exhibits regular hand movements as a part of play. This leads to LLM refinement incorrectly predicting the child appearing to be anxious during these video segments.

6.4. Multimodality vs. Unimodality

The results for activity segmentation and activity recognition show that some modalities might be more informative for a given task. For example, in the case of activity recognition (Remote-NLS), video modality is more informative than audio. However, in the case of activity segmentation, we notice the opposite trend. The proposed approach of combining modality-specific information followed by LLM refinement enables inferring from the more informative modalities, thereby providing improved performance across the different activity tasks. However, while we observe performance gains for multimodal refinement in activity tasks, we do not see a similar trend in atypical behavior tasks. As mentioned earlier, this is primarily because of errors in modality-specific content extraction, i.e., hallucinations of VLMs or ASR models. Currently, multimodal refinement equally weighs all the modality descriptions without providing modality-specific confidence weights. This leads to faulty inferences by multimodal refinement in the presence of errors in modality-specific descriptions.

6.5. Context Scaling and Model Size

We test whether increasing the context length of the input modality results in improved downstream performance. As most video-language models considered in this work are trained for a maximum context of 16 frames, this analysis is applied on audio modality, wherein increasing the context length is equivalent to providing transcripts from longer video segments.

By comparing the rows whisper-L (16s) and whisper-L (64s) in Table 2, we observe that increasing the context length of audio information from 16s to 64s leads to a significant improvement in the performance for activity tasks. This could be the case because transcripts which are 16 seconds long may not provide activity-specific information and might have generic conversations which leads to the model falsely predicting those chunks as conversations. When increasing the context length to 64 seconds, there is a higher chance of capturing information about the primary activity. For the activities task, we see a performance improvement for the 32B llava-next model when compared to the 7B model. However, no consistent trend is seen for atypical behaviors.

7. Conclusion & Future Work

In this study, we explored the utility of foundation models in analyzing clinical videos for ASD diagnosis. The proposed multimodal refinement pipeline provided robustness to unimodal limitations by leveraging complementary information that is present in the additional modalities. We also underscored the limitations of foundation models in supporting reasoning in these complex settings of human interactions. In the future, we plan to expand our work to provide multistep reasoning wherein the LLM reasoning agent would be able to selectively prompt for task-specific information. Additionally, we plan to extend the reasoning capabilities to identification of a larger set of ASD-related behaviors such repetitive behaviors and gestures.

8. Acknowledgements

We gratefully acknowledge support from Simons Foundation (award number: SFI-AR-HUMAN-00004115-03, 655054)

9. References

- [1] M. J. Maenner, “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2020,” *MMWR. Surveillance Summaries*, vol. 72, 2023.
- [2] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, “The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism,” *Journal of autism and developmental disorders*, vol. 30, pp. 205–223, 2000.
- [3] R. Grzadzinski, T. Carr, C. Colombi, K. McGuire, S. Dufek, A. Pickles, and C. Lord, “Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (boscc),” *Journal of autism and developmental disorders*, vol. 46, pp. 2464–2479, 2016.
- [4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [5] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, “Gemini: a family of highly capable multimodal models,” *arXiv preprint arXiv:2312.11805*, 2023.
- [6] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, J. Z. Chaves, *et al.*, “Capabilities of gemini models in medicine,” *ArXiv*, vol. abs/2404.18416, 2024.
- [7] N. Yildirim, H. Richardson, M. T. A. Wetscherek, J. Bajwa, J. Jacob, *et al.*, “Multimodal healthcare ai: Identifying and designing clinically relevant vision-language applications for radiology,” *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.
- [8] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, “Capabilities of gpt-4 on medical challenge problems,” *ArXiv*, vol. abs/2303.13375, 2023.
- [9] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. Jones, R. M. Jones, A. Pickles, M. W. State, *et al.*, “Autism spectrum disorder,” *Nature reviews Disease primers*, vol. 6, no. 1, pp. 1–23, 2020.
- [10] C. Lord, E. Petkova, V. Hus, W. Gan, F. Lu, D. M. Martin, O. Ousley, L. Guy, R. Bernier, J. Gerds, *et al.*, “A multisite study of the clinical diagnosis of different autism spectrum disorders,” *Archives of general psychiatry*, vol. 69, no. 3, pp. 306–313, 2012.
- [11] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, “Autism spectrum disorder,” *The lancet*, vol. 392, no. 10146, pp. 508–520, 2018.
- [12] P. Pandey, A. Prathosh, M. Kohli, and J. Pritchard, “Guided weak supervision for action recognition with scarce data to assess skills of children with autism,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 463–470, 2020.
- [13] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A. C. Murillo, “One-shot action recognition in challenging therapy scenarios,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2777–2785, 2021.
- [14] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, “Discriminative few shot learning of facial dynamics in interview videos for autism trait classification,” *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1110–1124, 2022.
- [15] N. Kojovic, S. Natraj, S. P. Mohanty, T. Maillart, and M. Schaer, “Using 2d video-based pose estimation for automated prediction of autism spectrum disorders in young children,” *Scientific Reports*, vol. 11, no. 1, p. 15069, 2021.
- [16] A. Ali, F. Negin, S. Thümmler, and F. Brémond, “Video-based behavior understanding of children for objective diagnosis of autism,” in *VISIGRAPP*, 2022.
- [17] L. K. Butler, C. La Valle, S. Schwartz, J. B. Palana, C. Liu, N. Peterman, L. Shen, and H. Tager-Flusberg, “Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level,” *Frontiers in Communication*, vol. 7, p. 820564, 2022.
- [18] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, S. Bishop, *et al.*, “Autism diagnostic observation schedule—2nd edition (ados-2),” *Los Angeles, CA: Western Psychological Corporation*, vol. 284, pp. 474–478, 2012.
- [19] C.-Y. Wu and P. Krahenbuhl, “Towards long-form video understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1884–1894, 2021.
- [20] P. Wang, O. Golovneva, A. Aghajanyan, X. Ren, M. Chen, A. Celikyilmaz, and M. Fazel-Zarandi, “Domino: A dual-system for multi-step visual language reasoning,” *arXiv preprint arXiv:2310.02804*, 2023.
- [21] Y. Yang, X. Zhang, J. Xu, and W. Han, “Empowering vision-language models for reasoning ability through large language models,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10056–10060, IEEE, 2024.
- [22] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, “Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning,” *arXiv preprint arXiv:2403.07376*, 2024.
- [23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.
- [24] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, “Llava-next: A strong zero-shot video understanding model,” April 2024.
- [25] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, *et al.*, “Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms,” *arXiv preprint arXiv:2406.07476*, 2024.
- [26] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.