CARE-VL: A Domain-Specialized Vision-Language Model for Early ASD Screening

Cheol-Hwan Yoo*, Jang-Hee Yoo, and Jaeyoon Jang

ETRI, Daejeon, Republic of Korea
{ch.yoo, jhy, jangjy}@etri.re.kr

Abstract. We propose an autism spectrum disorder (ASD) screening framework that integrates an expert vision-language model (VLM), CARE-VL, with a large language model (LLM)-based aggregation module to assess children's social interactions and derive subject-level ASD/typical development (TD) classifications. Our framework processes video data collected using social interaction-inducing content, where medical experts annotated predefined query-response (Q-R) intervals based on key social indicators—such as response to name, eye contact, imitation behavior, social smiling, and pointing—by marking correct responses and assigning subject-level ASD/TD classifications. To adapt the general-purpose VLM to the ASD screening domain, we constructed a synthetic instruction-tuning dataset using a label-guided reasoning method on these clinical tags, fine-tuning the model to generate detailed captions and multiple-choice question-answer (MC-QA) pairs, capturing children's critical social behaviors. CARE-VL processes Q-R intervals to produce clip-level MC-QA results and descriptive captions, which are then aggregated by an LLM to derive final ASD/TD classification and clinical reasoning. Our end-to-end framework combines visual understanding and linguistic reasoning, achieving 84.6% accuracy for clip-level response prediction and 75.8% accuracy for subject-level ASD/TD classification. These results demonstrate the potential of our framework as a practical and interpretable tool for early ASD screening and behavioral assessment. The code is publicly available at https://github.com/etri/AI4ASD.

Keywords: ASD · Vision-Language Model · Social Behavior Analysis

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent deficits in social communication and restricted, repetitive behaviors [1, 2]. Early detection and intervention are critical for improving long-term outcomes in children with ASD. However, current clinical diagnostic tools, such as the Autism Diagnostic Observation Schedule (ADOS) [14], are resource-intensive and demand specialized expertise, limiting their accessibility. Recent advancements in large language models (LLMs) have shown promise in clinical

* Corresponding author

applications, including medical diagnosis, computer-aided decision-making, and summarization [19, 25, 20]. Building on this potential, vision language models (VLMs) [24, 11, 21, 9, 13, 8] have also been applied to medical domains. Models such as Med-Flamingo [17] and LLaVA-Med [12] have improved radiology interpretation and pathology image analysis through domain-specific adaptation, underscoring the need for specialized training in clinical tasks. However, as far as we know, their application in ASD screening remains largely unexplored, despite the critical role of social behavior analysis in early diagnosis. While recent research [5] has begun to explore VLMs for analyzing the behaviors of children with autism, distinguishing ASD from typical development (TD) children requires integrating multiple social indicators—such as response to name, pointing, and imitation behavior [7]. Existing study often focuses on isolated features or lacks comprehensive analyses across these critical indicators, limiting their utility in real-world diagnostic scenarios.

To address these challenges, we propose **C**hild **A**utism **R**easoning **E**xpert in **V**ision–**L**anguage (CARE-VL), a domain-specialized VLM designed to analyze children's social interaction behaviors in video data. Our approach begins with a structured system built on Social Interaction-Inducing Content (SIIC), which provides video datasets capturing key social behaviors of ASD and TD children. Medical experts annotate predefined query-response (Q-R) intervals in these videos, labeling each segment for the presence or absence of correct responses and assigning subject-level ASD/TD classifications. Using these expert-provided annotations, we construct a synthetic video instruction-tuning dataset using a label-guided reasoning method. This dataset includes detailed captions and multiple-choice question-answer (MC-QA) pairs tailored to critical behavioral indicators. The fine-tuned CARE-VL model is then capable of generating clip-level MC-QA responses and descriptive captions, which are aggregated by an LLM to provide subject-level ASD/TD classifications and clinical reasoning.

The primary contributions of this work are as follows: (i) Domain-specialized VLM: We introduce CARE-VL, a VLM tailored for ASD screening, incorporating domain-specific knowledge of children's social interaction behaviors. (ii) Synthetic instruction-tuning dataset: We design a structured system to collect SIIC-based videos of ASD/TD children and generate a synthetic instruction-tuning dataset. This dataset is constructed using expert annotations and a label-guided reasoning method, enabling the generation of detailed captions and MC-QA pairs for critical behavioral indicators. (iii) End-to-end workflow for ASD screening: We develop an interpretable pipeline that combines CARE-VL's clip-level predictions with LLM-based aggregation to achieve subject-level ASD/TD classification and provide clinical reasoning.

2 Proposed Method

2.1 SIIC-Based Dataset and Clinical Annotations

As shown in Fig. 1, we collected video recordings of children with ASD and TD peers, each approximately six minutes in length, using SIIC. SIIC is specifically

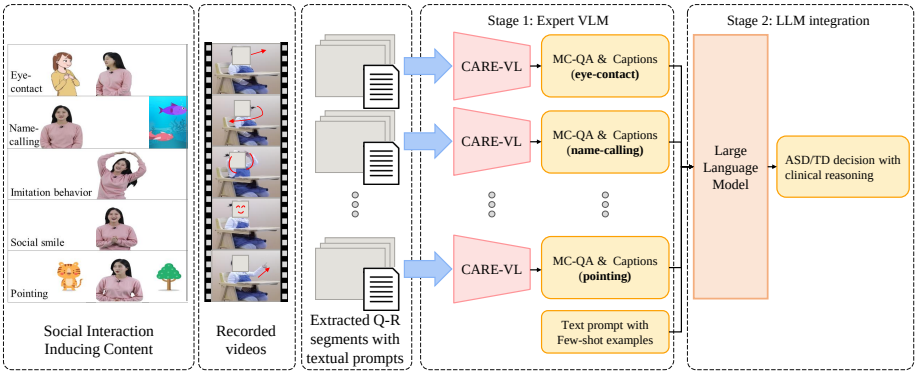


Fig. 1: Overview of the proposed ASD screening pipeline. Videos collected based on SIIC are segmented into predefined QR intervals, which are processed by the expert VLM to generate captions and MC-QA responses. The clip-level outputs are aggregated using an LLM to provide subject-level ASD/TD classification.

designed to elicit core social behaviors—such as response to name, eye contact, imitation behavior, social smiling, and pointing—that are clinically relevant for ASD screening [7], with each behavior corresponding to a predefined Q-R interval. Within these intervals, expert clinicians provided binary labels indicating positive response (PR) or non-positive response (NR), assessing whether the child's behavior met established clinical criteria. In addition to clip-level annotations, each subject was assigned an ASD or TD label based on diagnostic criteria (e.g., DSM-5 [7]). This dual-level annotation allows us to analyze both fine-grained responses at the interval level and global ASD/TD classification at the subject level. Data collection was conducted at two clinical sites, yielding a total of 118 subjects: 85 from Site A, which was used for training, and 33 from Site B, designated as the test benchmark. During the data collection process, contents were displayed across three monitors, positioned in front of the child to ensure engagement and controlled behavioral elicitation. Each subject underwent behavioral assessments across five key social indicators, with video captured from four to five camera views per environment to ensure comprehensive behavioral observation. Additionally, each behavioral indicator was evaluated two to three times per session, resulting in a total of 5,409 video clips for Site A and 1,716 video clips for Site B. This diverse dataset enables a comprehensive evaluation of our proposed ASD screening framework across different environments.

2.2 Synthetic Video Instruction-Tuning Dataset

To adapt a general-purpose VLM to the ASD domain, we adopt a label-guided reasoning approach, similar to [5], to generate a synthetic video instruction-tuning dataset that facilitates fine-tuning for ASD-related tasks. For each social indicator $s_i \in \{\text{response to name}, \text{eye contact}, \text{imitation}, \text{social smiling}, \text{pointing}\}$,

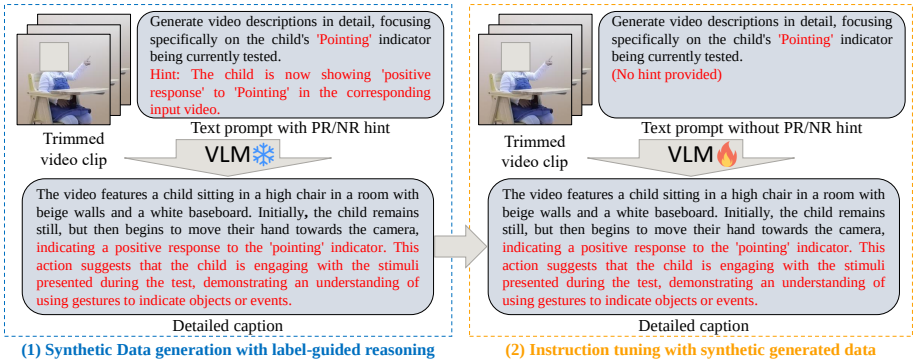


Fig. 2: Synthetic instruction-tuning dataset generation pipeline. The VLM is provided with predefined social indicators and their response labels as hints to generate detailed captions. These generated captions, with instruction prompts without response labels, are used to fine-tune the VLM for ASD screening.

the input video segment corresponding to the Q-R interval, denoted as $V_{\text{Q-R}_i}$, is pre-annotated with a response label $y_i \in \{\text{PR}, \text{NR}\}$. For each $V_{\text{Q-R}_i}$, we provide a baseline VLM (e.g., LLaVA-OneVision [11]) with a textual prompt that includes an instruction I_i, and a hint sentence derived from the pre-annotated label y_i to generate a detailed caption C_i as follows:

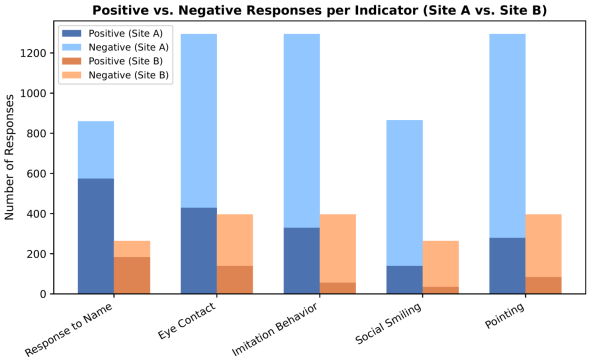
$$C_i = \text{VLM}(V_{\text{Q-R}_i}, I_i, \text{Hint}(y_i)) \tag{1}$$

Here, y_i serves as a guiding label to drive the model to produce a detailed, context-aware description of the child's behavior. An example of this process is illustrated in Fig. 2, where the model generates a detailed caption based on the provided hint. This label-guided reasoning approach shares a conceptual similarity with self-correction [15, 10] and reflection-based learning [18], which have been widely explored in the LLM domain to enhance model reasoning and output reliability. Our approach similarly leverages external guidance to improve model outputs by incorporating structured expert labels before inference.

In addition to caption generation, MC-QA pairs (q_i, a_i) are also created for each $V_{\text{Q-R}_i}$. The question asks whether the child responded appropriately to the given social indicator, with the answer directly derived from the pre-annotated label y_i. By iterating this process for all annotated Q-R intervals in the training set, we construct a synthetic dataset $D_{\text{synthetic}}$:

$$D_{\text{synthetic}} = \{(I_i, C_i, q_i, a_i) \mid i = 1, 2, \ldots, N\} \tag{2}$$

where N represents the total number of Q-R intervals in the data set. The resulting dataset $D_{\text{synthetic}}$ includes both detailed captions and MC-QA pairs tailored to ASD-related social indicators. Thus, the constructed instruction-tuning dataset serves as a specialized resource for fine-tuning a general-purpose VLM,



Data Source	# Video	# Detailed Caption	#MC-QA
Site A	5.4K	5.4K	5.4K
Site B	1.7K	1.7K	1.7K

Fig. 3: Distribution of the synthetic instruction-tuning dataset. Detailed captions and MC-QA pairs are created for each Q-R interval.

transforming it into CARE-VL—an expert model capable of accurately identifying, describing, and evaluating critical social behaviors relevant to ASD screening. Fig. 3 presents the distribution of the generated instruction-tuning dataset.

2.3 LLM Aggregation

As shown in Fig. 1, each test video is segmented into predefined Q-R intervals, allowing the fine-tuned domain-specialized VLM to generate clip-level MC-QA results and detailed captions describing the child's behavior. To achieve subject-level ASD/TD classification with clinical reasoning, we aggregate these clip-level outputs using an LLM, such as LLaMA-3.2 [6]. The LLM takes as input the MC-QA predictions $\{\hat{a}_1, \ldots, \hat{a}_N\}$ and the corresponding captions $\{\hat{C}_1, \ldots, \hat{C}_N\}$, from all Q-R intervals, along with a prompt requesting a final ASD/TD decision. Since the dataset contains a limited number of subject-level samples, we adopt a *few-shot prompting* strategy [3, 16] to guide the LLM. Specifically, a small set of labeled examples (e.g., two ASD and two TD cases from the training dataset) is included in the prompt to guide the LLM in interpreting the MC-QA results and captions without requiring additional fine-tuning. By integrating these inputs, the LLM produces a subject-level classification along with an optional textual summary explaining the reasoning behind its decision. This two-stage pipeline—clip-level analysis followed by LLM aggregation—follows a Divide and Conquer approach [4], where a complex task is decomposed into smaller subproblems that are independently analyzed before final integration. In our framework, the domain-specialized VLM independently processes each Q-R interval, generating localized behavioral descriptions that focus on specific social indicators. The LLM then consolidates these outputs, synthesizing diverse behavioral cues into a holistic subject-level ASD/TD classification, ensuring both interpretability and robustness in clinical decision-making.

3 Experiments

3.1 Experimental Setup

Datasets and Preprocessing. We utilize data collected from two clinical sites, as described in Section 2.1 and Section 2.2. For training, we use the generated synthetic instruction-tuning dataset $D_{\text{synthetic}}$ from Site A, which contains both detailed captions and MC-QA pairs, while datasets from Site B serve as a test benchmark to evaluate generalization ability of the proposed framework.

Implementation Details. For the expert VLM, CARE-VL, we fine-tuned all components of the LLaVA-OneVision-Qwen2-7B-OV [11], including the vision encoder, MLP adapter, and language components. The model was trained for one epoch on 8 NVIDIA A6000 GPUs, with each GPU processing a batch size of 1. The learning rate was set to $1e-5$ using a cosine scheduler. For the LLM aggregation module, we employ an off-the-shelf LLM (e.g., Llama-3.2-3B-Instruct [6]) with a few-shot prompting strategy without additional fine-tuning.

Evaluation Metrics. We evaluate performance in three aspects. First, *clip-level MC-QA accuracy* is the proportion of correctly classified Q-R intervals. Second, we assess *clip-level captions* with LLaVA-Critic [22], which evaluates the descriptive quality of CARE-VL in capturing clinically relevant behaviors. Finally, we report *subject-level ASD/TD classification* with accuracy and F1-score.

3.2 Social Indicator Evaluation at the Clip Level

We assess the ability of different models to identify correct or incorrect responses across all five social indicators (response to name, eye contact, imitation behavior, social smiling, pointing) using clip-level Q-R intervals. As shown in Table 1, we compare our expert model, CARE-VL, with several general-purpose baseline

Table 1: Performance comparison between baseline models and CARE-VL. MC-QA measures correct response identification across social indicators, while caption evaluates descriptive quality.

Model	MC-QA						
	Overall Acc.	Response to Name	Eye Contact	Imitation Behavior	Social Smiling	Pointing	Caption
Chat-UniVi-7B [9]	28.8	**69.7**	35.6	14.6	20.1	21.5	48.8
Video-LLaVA-7B [13]	29.1	**69.7**	35.4	14.1	13.6	21.2	30.5
LLaVA-Video-7B [24]	31.5	62.9	37.1	16.2	17.4	29.8	<u>57.7</u>
LLaVA-NeXT-Video-7B [23]	34.5	32.6	39.1	25.8	38.3	37.6	55.3
LLaVA-OV-0.5B [11]	49.2	60.2	39.6	58.1	34.5	52.5	35.3
LLaVA-OV-7B [11]	<u>61.2</u>	36.4	<u>60.9</u>	<u>68.4</u>	<u>57.6</u>	<u>73.5</u>	53.3
CARE-VL (Ours)	**84.6**	<u>68.9</u>	**72.7**	**94.2**	**92.0**	**92.4**	**69.5**

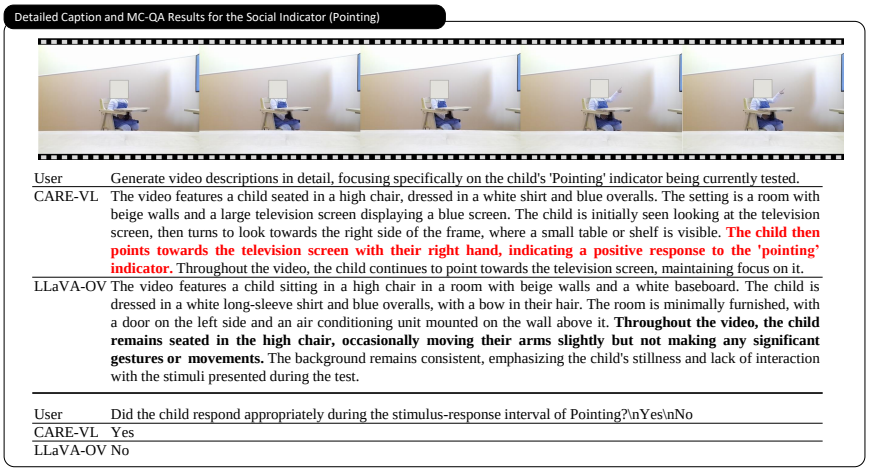


Fig. 4: Comparison of CARE-VL and the general VLM in generating detailed captions and MC-QA responses for the social indicator.

models, including Chat-UniVi [9], Video-LLaVA [13], LLaVA-Video [24], LLaVA-NeXT-Video [23], and LLaVA-OV [11]. CARE-VL achieves the highest average clip-level accuracy of 84.6%, significantly outperforming other baseline models, including LLaVA-NeXT-Video at 34.5% and LLaVA-OV-7B at 61.2%. A breakdown by social indicator further emphasizes the superiority of CARE-VL over baseline models. For instance, in the social smiling category, CARE-VL achieves an accuracy of 92.0%, outperforming LLaVA-OV-7B (57.6%) and LLaVA-NeXT-Video (38.3%). In addition to MC-QA accuracy, we evaluate the descriptive quality of generated captions using an LLM-as-a-judge approach (e.g., LLaVA-Critic [22]), with scores ranging from 0 to 100. CARE-VL achieves an average caption evaluation score of 69.5, outperforming general-purpose models that often provide generic or misleading descriptions. These results confirm the importance of integrating domain-specific knowledge via label-guided reasoning and fine-tuning to enhance the model's ability to accurately capture clinically relevant social behaviors. In Figure 4, we also present a qualitative comparison of the detailed captions and MC-QA responses generated by CARE-VL and the baseline model (LLaVA-OV) for the pointing social indicator. While the baseline model provides a generic and partially incorrect description, CARE-VL generates a clinically accurate caption that captures the child's interaction with the pointing cue. Furthermore, the MC-QA results show that CARE-VL correctly identified the response, whereas the baseline model failed to recognize it, demonstrating its superior understanding of ASD-related social behaviors.

3.3 LLM Aggregation for Subject-Level ASD/TD Classification

To evaluate the effectiveness of the proposed ASD screening framework at the subject level, we aggregate clip-level predictions using an LLM to derive the final

Table 2: Example of LLM aggregation process for ASD/TD classification.

System		You are a helpful assistant for ASD screening. Given Q-R clip observations, you will decide if the child is ASD or TD. (DSM5 CRITERIA)
Few-Shot Examples:		2 ASD, 2 TD cases
Q-R Interval	**Yes/No**	**Generated Caption**
Response to Name	No	"The child is seated in a high chair, wearing a green jacket and blue pants. The room has beige walls and a white door on the left side. The child remains still throughout the video, showing no response to their name being called."
Eye Contact	No	(omitted)
Imitation Behavior	No	(omitted)
Social Smiling	No	(omitted)
Pointing	No	(omitted)
User		"Based on the observations above, determine whether the child is more likely to have ASD or TD. A. TD, B. ASD"
LLM output		**"B. ASD.** The child's behavior and responses to various stimuli in the video suggest that they may have ASD. Here are some reasons why: **Lack of engagement and interaction**: Throughout the video, the child shows minimal engagement and interaction with the stimuli presented, such as their name being called, imitation behavior, and pointing, which is a common characteristic of individuals with ASD. (omitted)"

Table 3: Subject-level ASD/TD classification results. (a) Confusion matrix on test set (Site B), and (b) effect of few-shot prompting on classification accuracy.

(a) Confusion matrix on ASD classification.

Predicted \Actual	ASD	TD
ASD	15	5
TD	3	10

(b) Effect of few-shot prompting on classification accuracy (%).

# Examples	Accuracy	F1-Score
Zero-shot	54.5	70.6
Few-shot	**75.8**	**78.9**

ASD/TD classification. As shown in Table 2, the LLM processes multiple Q-R interval outputs, including MC-QA responses and detailed captions, to generate a subject-level decision with explanatory reasoning. Table 3a presents the confusion matrix of the subject-level classification results on the test set (from Site B). CARE-VL achieves an overall subject-level accuracy of 75.8%, with an F1-score of 78.9% for ASD screening. To analyze the impact of few-shot prompting on classification performance, we conducted experiments comparing zero-shot and few-shot scenarios. As shown in Table 3b, incorporating a small number of few-shot examples significantly improves performance, enhancing the LLM's ability to generalize better to unseen test data. Overall, our subject-level evaluation demonstrates that the proposed ASD screening framework, integrating CARE-VL with LLM-based aggregation, offers promising performance, making it a practical and interpretable solution for real-world ASD screening.

4 Conclusion

In this work, we introduced CARE-VL, a domain-specialized VLM for ASD screening based on social interaction behaviors in video data. By leveraging

SIIC-based video collection and expert-annotated Q-R intervals, we generated a synthetic instruction-tuning dataset using a label-guided reasoning approach to fine-tune a general-purpose VLM for ASD screening. Our experimental results demonstrate that CARE-VL outperforms general-purpose models, providing more clinically relevant descriptions and accurate responses. Furthermore, through few-shot in-context learning, our LLM aggregation approach achieves a subject-level ASD/TD classification accuracy of 75.8%, ensuring both reliable performance and enhanced interpretability of the final predictions. In future work, we plan to refine the aggregation process to further improve subject-level classification, potentially by leveraging stronger LLMs.

Acknowledgments. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2019-II190330, Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders based on Cognition of the Psychological Behavior and Response)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. American Psychiatric Association, D., American Psychiatric Association, D., et al.: Diagnostic and statistical manual of mental disorders: DSM-5, vol. 5. American psychiatric association Washington, DC (2013)
2. Baron-Cohen, S., Ring, H.A., Wheelwright, S., Bullmore, E.T., Brammer, M.J., Simmons, A., Williams, S.C.: Social intelligence in the normal and autistic brain: an fmri study. European journal of neuroscience **11**(6), 1891–1898 (1999)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. MIT press (2022)
5. Deng, S., Kosloski, E.E., Patel, S., Barnett, Z.A., Nan, Y., Kaplan, A., Aarukapalli, S., Doan, W.T., Wang, M., Singh, H., et al.: Hear me, see me, understand me: Audio-visual autism behavior recognition. arXiv preprint arXiv:2406.02554 (2024)
6. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Edition, F., et al.: Diagnostic and statistical manual of mental disorders. Am Psychiatric Assoc **21**(21), 591–643 (2013)
8. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13504–13514 (2024)
9. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)

10. Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J.D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., et al.: Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917 (2024)
11. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
13. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
14. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M.: 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 **30**, 205–223 (2000)
15. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems **36** (2024)
16. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the role of demonstrations: What makes in-context learning work? (2022), https://arxiv.org/abs/2202.12837
17. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
18. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems **36** (2024)
19. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
20. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. Nature Medicine pp. 1–8 (2025)
21. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
22. Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., Li, C.: Llava-critic: Learning to evaluate multimodal models. arXiv preprint arXiv:2410.02712 (2024)
23. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
24. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
25. Zhao, Z., Wang, S., Gu, J., Zhu, Y., Mei, L., Zhuang, Z., Cui, Z., Wang, Q., Shen, D.: Chatcad+: Towards a universal and reliable interactive cad using llms. IEEE Transactions on Medical Imaging (2024)