Low-Resource Multimodal Prediction of Caregiver and Child Interaction

Dengyi Liu, Ph.D. in Computer Science

FACULTY MENTOR: Honggang Wang, Ph.D.

Katz School of Science and Health

Introduction

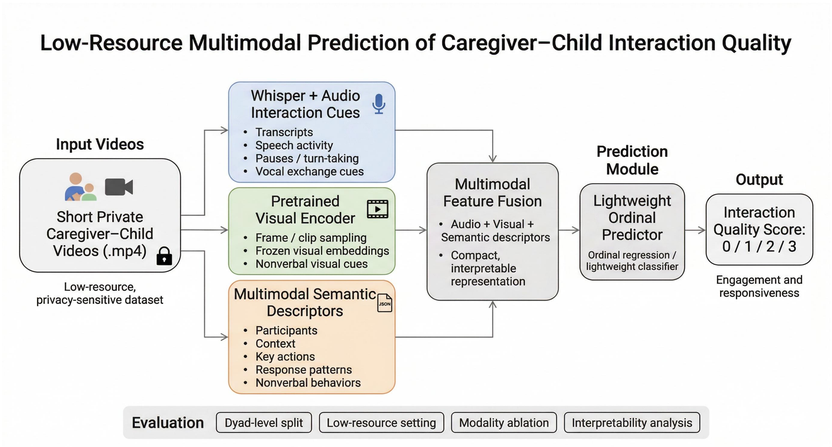
Caregiver-child interaction quality is an important behavioral marker in early childhood and is commonly assessed through expert observation. However, manual coding is time-consuming, costly, and difficult to scale. This challenge is especially significant for short, privately collected videos recorded in naturalistic settings, where dataset size is often very limited and direct supervision of large video models is impractical. Prior work has explored automated parent-child interaction analysis using dialogue, nonverbal behavior, and multimodal cues, but robust prediction under strong data scarcity and privacy constraints remains underexplored.

In this work, we formulated caregiver-child interaction assessment as a 4-level ordinal prediction task (scores 0-3). Instead of training an end-to-end video model, we investigate a low-resource pipeline that leverages pretrained foundation models to convert raw audio-visual signals into compact and semantically meaningful representations. Our goal is to determine whether multimodal pretrained features can support interaction quality prediction and to identity which cues are most informative under severe scarcity.

Method

We propose a modular multimodal framework that converts raw videos into compact features for low-resource ordinal prediction.

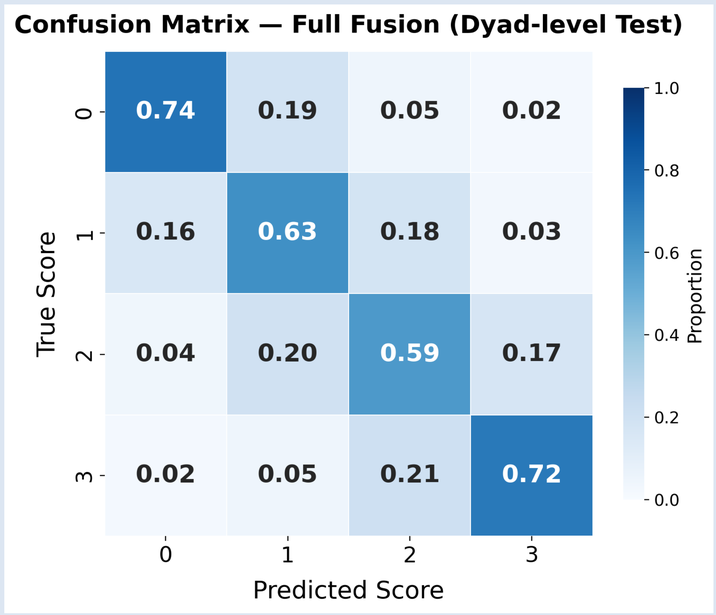
- **Data & task:** 4-level interaction score (0–3); strict participant/dyad-level splits.
- **Audio branch:** Whisper transcripts and speech cues (activity, pauses, turn-taking, vocal exchanges).
- **Visual branch:** Frozen embeddings from a pretrained image/video encoder on sampled frames or short clips.
- **Descriptor branch:** A pretrained multimodal model generates high-level interaction descriptors (participants, context, actions, response patterns, nonverbal behaviors).
- **Prediction:** Fused features feed a lightweight ordinal regressor.



Results

- Strict dyad-level splits test generalization to unseen caregiver–child pairs.
- We report Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Macro-F1, and per-class confusion.
- Multimodal fusion outperforms single-modality baselines, with high-level semantic descriptors providing the strongest individual cue.

Modality	QWK ↑	MAE ↓	Macro-F1 ↑	Accuracy ↑
Audio-only	0.31	0.92	0.34	0.41
Visual-only	0.38	0.85	0.39	0.45
Descriptor-only	0.46	0.78	0.44	0.50
Audio + Visual	0.49	0.74	0.47	0.53
Audio + Descriptor	0.55	0.68	0.52	0.57
Full Fusion	**0.62**	**0.61**	**0.58**	**0.63**

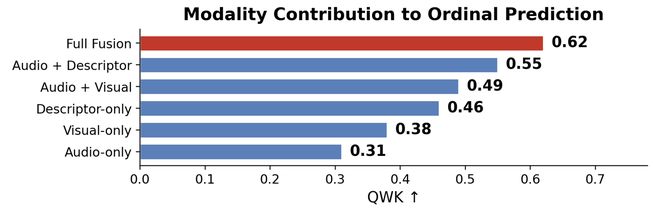


Conclusions

This work presents a low-resource multimodal framework for caregiver-child interaction quality prediction from short private videos. Instead of relying on data-intensive end-to-end video training, our approach leverages pretrained foundation models to extract compact, interpretable audio-visual representations and structured semantic descriptors. This design is well suited for small, privacy-sensitive datasets and supports both prediction and cue-level analysis. Overall, the study suggests that multimodal pretrained representations are a promising direction for scalable and interpretable interaction assessment in early childhood settings.

Key Findings

- Full fusion reaches **QWK 0.62** — a **+0.31** gain over audio-only and **+0.16** over the strongest single modality.
- **Semantic descriptors** (QWK 0.46) are the most informative single cue, suggesting high-level structured signals beat raw audio/visual features in low-resource settings.
- Errors concentrate on **adjacent ordinal classes**; level 0 and level 3 are almost never confused, supporting use as a screening signal.



Limitations & Future Work

- Small dyad count → larger multi-site cohort; explore lightweight adapter / LoRA fine-tuning under privacy constraints.
- Compare against expert inter-rater agreement as a calibration target.

Acknowledgements: Appreciate the help from Dr. Amiya Waldman-Levi, Vanessa Murad, Chana Cunin and the Occupational Therapy department.

References

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters, 140, 325–331.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning, 28492–28518.

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems, 35, 10078–10093.