Toward Automated Behavior Understanding in Autism: A Zero-Shot Vision-Language Model Approach

Supplementary Material

6. Sensitivity Analysis with Class-Balanced Leave-One-Out

6.1. Setup

Our dataset is small and clinically constrained, consisting of 40 curated clips with 10 clips per class across four behavior categories: Normal (NO), Self-Injury (SI), Aggression to Others (AO), and Property Destruction (PD). In this regime, aggregate performance can be sensitive to the particular sample composition. To quantify this effect, we perform a class-balanced leave-one-out (LOO) sensitivity analysis.

Since our method is a zero-shot VLM→LLM pipeline rather than a supervised model trained on this dataset, this analysis should be interpreted as a robustness check under controlled perturbations of the evaluation set rather than as standard cross-validation for model selection.

6.2. Protocol

We use a class-balanced LOO protocol with 10 folds. In each fold, we remove one clip from each class, yielding a held-out set of 4 clips and an evaluation subset of the remaining 36 clips.

For every fold, we rerun the inference pipeline over the anonymized image grids used in the main paper, including VLM prompting and downstream LLM-based zero-shot classification. No predictions are reused across folds. We report the same aggregate metrics as in the main paper: accuracy, macro precision, macro recall, and macro F1-score. In addition, for the DA+IG setting, we report class-wise precision, recall, and F1-score for both Claude Sonnet and GPT-5.

All results are summarized as mean ± standard deviation over the 10 folds.

6.3. Aggregate Sensitivity Across Folds

Table 6 reports aggregate performance across the 10 class-balanced LOO folds for all model and prompt combinations. The relative ordering of prompting strategies is preserved across folds for both models. For Claude Sonnet, performance improves from Generic to DA to DA+IG across all reported aggregate metrics. The same trend holds for GPT-5, with DA+IG remaining the strongest setting overall. Across all prompt settings, GPT-5 consistently outperforms Claude Sonnet in mean accuracy, macro precision, macro recall, and macro F1-score.

Table 6. Class-balanced leave-one-out sensitivity analysis. Results are reported as mean ± standard deviation over 10 folds.

Model	Prompt Setting	Accuracy	Macro Precision	Macro Recall	Macro F1
Claude	Generic	0.350 ± 0.014	0.405 ± 0.026	0.350 ± 0.014	0.298 ± 0.023
Claude	DA	0.375 ± 0.027	0.542 ± 0.063	0.375 ± 0.027	0.336 ± 0.040
Claude	DA+IG	0.525 ± 0.024	0.674 ± 0.043	0.525 ± 0.024	0.500 ± 0.036
GPT	Generic	0.425 ± 0.019	0.528 ± 0.065	0.425 ± 0.019	0.381 ± 0.026
GPT	DA	0.525 ± 0.031	0.662 ± 0.055	0.525 ± 0.031	0.510 ± 0.037
GPT	DA+IG	0.625 ± 0.027	0.741 ± 0.031	0.625 ± 0.027	0.617 ± 0.033

Table 7. Class-wise class-balanced leave-one-out sensitivity analysis for the DA+IG prompting setting. Results are reported as mean ± standard deviation over 10 folds.

Model	Class	Precision	Recall	F1-score
Claude	NO	0.346 ± 0.011	0.900 ± 0.035	0.500 ± 0.012
Claude	SI	0.667 ± 0.136	0.200 ± 0.047	0.306 ± 0.066
Claude	AO	1.000 ± 0.000	0.800 ± 0.047	0.888 ± 0.028
Claude	PD	0.667 ± 0.123	0.200 ± 0.047	0.305 ± 0.060
GPT	NO	0.410 ± 0.017	0.900 ± 0.035	0.563 ± 0.017
GPT	SI	0.750 ± 0.096	0.300 ± 0.054	0.427 ± 0.066
GPT	AO	1.000 ± 0.000	0.900 ± 0.035	0.947 ± 0.019
GPT	PD	0.805 ± 0.073	0.400 ± 0.057	0.531 ± 0.053

The reported standard deviations are nonzero but modest, indicating that performance is sensitive to fold composition, as expected in a dataset of only 40 clips, while the main conclusions of the paper remain stable under class-balanced perturbations of the evaluation set.

6.4. Class-Wise Stability for DA+IG

To localize the source of variation, Table 7 reports class-wise precision, recall, and F1-score for the DA+IG setting across both Claude Sonnet and GPT-5. The class-wise trends observed in the main paper persist under class-balanced LOO evaluation. In particular, AO is the most stable aggression classes across folds for both models, with high recall and relatively small variability. By contrast, SI and PD remain more challenging, exhibiting lower mean recall and larger standard deviations.

This pattern is consistent with the confusion behavior discussed in the main paper, where subtle self-directed or object-directed actions are harder to disambiguate from anonymized image-grid summaries than more overt aggression toward others.

6.5. Discussion

Overall, the class-balanced LOO analysis provides a direct robustness check for the low-data regime considered in this work. Because the dataset contains only 40 clips, even the exclusion of one example per class can affect the aggregate metrics. The results nevertheless show that the main qualitative findings are preserved across folds: domain-aware prompting improves over the generic baseline, DA+IG remains the strongest prompting strategy, and GPT-5 consistently performs better than Claude Sonnet.

At the same time, the fold-level variability highlights the importance of cautious interpretation in low-sample clinical settings, especially for behavior categories such as SI and PD that are intrinsically more ambiguous under anonymized visual summaries. Taken together, Tables 6 and 7 provide a more complete characterization of robustness for the proposed zero-shot framework in this small-data clinical setting.

6.6. Scope of Contribution

We emphasize that the primary contribution of this work is not a new pretrained video backbone or a new supervised learning objective. Rather, the contribution lies in formulating a clinically grounded, label-efficient, and interpretable zero-shot pipeline for behavior analysis in a privacy-sensitive small-data setting. Specifically, our framework combines representative keyframe extraction, behavior-focused VLM description generation, and downstream LLM reasoning, and shows that domain-aware and interaction-guided prompting materially improves downstream classification and explanation quality. We therefore view this work as a feasibility study in clinically meaningful zero-shot behavior understanding, rather than as a replacement for fully supervised video recognition systems trained on large annotated datasets.

6.7. Temporal Information and Its Limitations

The proposed framework does not explicitly model full video dynamics. Instead, temporal information is summarized through representative keyframe selection and joint reasoning over the resulting image grid. This design preserves coarse scene evolution, salient posture changes, and subject–caregiver or subject–object interactions, while substantially reducing redundancy and annotation burden. At the same time, it does not explicitly encode motion trajectories, repetition frequency, onset and offset boundaries, or movement intensity. This limitation is important when distinguishing behavior categories such as self-injury and property destruction, where fine-grained temporal evidence may be necessary to disambiguate subtle self-directed or object-directed actions. The class-wise sensitivity trends in Table 7 are consistent with this limitation, as NO and AO remain comparatively stable while SI and PD exhibit lower recall and greater variability across folds.