Probing Video Language Models on Face-to-Face Dialogue Pragmatics

Anonymous ACL submission

Abstract

Multimodal large language models can process video input, yet whether this capacity supports understanding of face-to-face communicative signals remains largely untested. I present a diagnostic probing study on three tasks drawn from MELD: emotion recognition under verbal-visual conflict (Task 1), turn-yield prediction (Task 2), and emotion shift detection (Task 3). VideoLLaMA2-7B and a text-only baseline are evaluated zero-shot on the MELD test split with no task-specific training. On Task 1, neither model shows a reliable advantage on conflict-stratum utterances (VideoLLaMA2 weighted F1 0.296 vs. text-only 0.256; overlapping 95% bootstrap confidence intervals; McNemar $p = 0.110$). On Task 2, VideoLLaMA2 achieves a higher macro F1 (0.443 vs. 0.316) but predicts YIELD in 98.9% of instances, collapsing the classification to a majority-class heuristic rather than discriminating turn boundaries. On Task 3, the text-only model outperforms the video model on SHIFT detection (F1 0.599 vs. 0.428). The text-only baseline is Phi-3-mini (3.8B parameters), smaller than VideoLLaMA2-7B; the null results on Tasks 1 and 3 are therefore not attributable to model size. Human verification of the conflict stratum confirmed substantial inter-rater agreement ($\kappa = 0.656$) but only 12% of sampled instances showed genuine visual-verbal conflict, indicating that the classifier-constructed stratum captures linguistic ambiguity more than observable face-text disagreement. I release all task definitions, prompt templates, and evaluation code.

1 Introduction

Face-to-face dialogue is the primary medium through which humans exchange emotion. Speakers signal anger, amusement, or sadness not only in the words they choose but in facial expression, gaze direction, and the timing of floor transitions. A listener who attends only to the transcript misses a substantial share of the communicative signal. This gap between what text carries and what video carries provides a natural testbed for evaluating multimodal large language models (MLLMs): if a model that can watch video does not outperform one that reads transcripts alone, it is not using the visual channel.

I study this question using MELD (Poria et al., 2019a), a multimodal, multi-party emotion dataset drawn from the television series *Friends*. MELD contains 13,708 utterances with per-turn emotion labels, aligned audio and video clips, and sequential speaker annotations that record who speaks in each turn. These properties make it suitable for three diagnostic probing tasks, each tied to a distinct face-to-face communicative phenomenon that transcripts cannot fully represent.

Task 1 asks models to assign an emotion label to a target utterance, with performance measured separately for utterances where the text-based sentiment and the facial expression agree versus disagree. A model that genuinely processes video should show a larger advantage over text-only models on utterances where face and words diverge.

Task 2 asks models to predict whether the current speaker will yield the floor after their utterance, given that turn-yield cues such as gaze shifts, head nods, and pitch lowering are visible in video but absent from transcripts.

Task 3 asks models to detect whether a speaker's emotion has changed between two consecutive turns, given that facial and prosodic transitions often precede or accompany the lexical shift.

I probe VideoLLaMA2-7B and one text-only baseline in a zero-shot setting on the MELD test split, requiring no task-specific training. Performance is compared against published supervised baselines to characterize the gap between zero-shot video understanding and fine-tuned classification. My contribution is the diagnostic task suite itself: three clean, annotation-free probing tasks that iso-

late the conditions under which video information should, in principle, help.

2 Related Work

2.1 Emotion Recognition in Conversation

Poria et al. (2019b) survey the field of emotion recognition in conversation (ERC) and identify ten persistent research challenges, three of which bear directly on my tasks: modeling conversational context across turns, detecting emotion shifts, and handling the multi-party setting where speaker count and short dialogue length limit the signal each speaker provides. Their analysis showed that the recurrent model DialogueRNN (Majumder et al., 2019) correctly classified only 47.5 percent of utterances at shift positions, compared with 69.2 percent at stable positions, establishing emotion shift detection as a quantifiably hard subproblem.

DialogueRNN tracks each speaker's state through dedicated gated recurrent units and achieves a text-only weighted F1 of 57.03 on MELD. The subsequent DialogueGCN (Ghosal et al., 2019) replaces the recurrent speaker encoder with a graph convolutional network whose edge relations encode both temporal direction and speaker identity, raising the MELD score to 58.10. Ghosal et al. (2020) then augmented the graph architecture with commonsense knowledge from ATOMIC, queried through the COMET transformer for five relation types per utterance. COSMIC reaches 65.21 weighted F1 on MELD's seven-class task. A text-only supervised ceiling is set by SACL-LSTM (Hu et al., 2023), which applies supervised adversarial contrastive learning with contextual adversarial training to reach 66.45 weighted F1 on MELD using RoBERTa-large features and no visual input.

2.2 LLM-Based Approaches to ERC

Two recent papers establish what fine-tuned large language models can achieve on MELD. DialogueLLM (Zhang et al., 2024) fine-tunes LLaMA 2-7B with LoRA on a three-dataset instruction corpus and augments each training instance with a text description of the corresponding video clip generated by GPT-4 or ERNIE Bot. Ablation results show that removing the video descriptions drops MELD weighted F1 by 12.06 percentage points when the descriptions are informative, though the same approach adds nothing for IEMOCAP because the camera setup produces identical-looking descriptions for all clips. The final model reaches 71.90 weighted F1 on MELD. InstructERC (Lei et al., 2023) reformulates ERC as a retrieval-augmented sequence-to-sequence task, retrieves a semantically similar labeled example per utterance using SBERT, and adds two auxiliary objectives during training. With LLaMA2 and LoRA, InstructERC reaches 69.15 weighted F1 on MELD. Both papers document that untuned LLMs perform near chance on MELD in zero-shot conditions (LLaMA 2-7B: 24.12 accuracy, 16.28 weighted F1), establishing the zero-shot floor that my probing study operates within.

2.3 Multimodal Fusion for ERC

The multimodal supervised literature on MELD is defined by a steady progression of fusion strategies across five years. Tsai et al. (2019) introduced crossmodal attention for unaligned multimodal sequences, demonstrating that queries from one modality attending to keys and values from another can capture temporal offsets without forced word alignment. Hu et al. (2021) extended the graph convolution approach of DialogueGCN to three modalities by constructing a 3N-node graph with intra-modality and inter-modality edges, reaching 58.65 weighted F1 on MELD with audio, video, and text combined. MM-DFN (Hu et al., 2022a) added LSTM-style gating between graph convolution layers to filter redundant cross-layer information, reaching 59.46 on MELD. UniMSE (Hu et al., 2022b) treated multimodal sentiment analysis and ERC as a single text-to-text generation task, jointly training on four datasets with T5-Base and a contrastive inter-modality alignment objective; it currently holds the highest published supervised score on MELD at 65.51 weighted F1.

A consistent finding across all three papers that report unimodal ablations is that text features substantially outperform audio, which in turn outperforms visual features alone. MMGCN reports vision-alone weighted F1 of 33.27 on MELD; MM-DFN reports 32.34. None of the supervised models reviewed here successfully incorporates raw video into MELD evaluation: the MELD dataset paper (Poria et al., 2019a) itself excluded visual features from its baselines due to speaker face-localization difficulties in multi-party TV footage. My study asks whether pretrained video LLMs can fill this gap without any task-specific training.

2.4 Video Large Language Models

The MLLMs I probe share a common architectural lineage. Video-ChatGPT (Maaz et al., 2024) adapts a CLIP ViT-L/14 encoder for video by averaging frame representations along the spatial and temporal dimensions and projecting the result into Vicuna's embedding space. Its own evaluation framework shows temporal understanding scoring 1.98 on a 1–5 scale, consistently the lowest of five measured aspects across all tested models. Video-LLaVA (Lin et al., 2024) introduces alignment before projection, using LanguageBind encoders to place image and video features into a shared language-aligned space before a single projection layer is applied. It achieves 70.7 percent accuracy on MSVD-QA but only 39.9 percent on the Video-MME benchmark (Fu et al., 2025), which covers 900 manually annotated videos across six domains. Video-LLaMA (Zhang et al., 2023) is the only model in this group to incorporate an explicit audio branch, using ImageBind's cross-modal alignment to process audio spectrograms at inference time without direct audio-text training. VideoLLaMA2 extends this architecture with improved temporal modeling and is the primary MLLM candidate in my study.

Video-MME reveals that performance declines sharply with video length across all open-source models, and that temporal understanding remains the weakest capability even for models that perform well on short-clip QA benchmarks. No published video benchmark covers multi-party face-to-face conversation, making the MELD pragmatic tasks a qualitatively different type of evaluation from anything in these models' standard assessment suites.

2.5 Turn-Taking and Backchannel Prediction

Ekstedt and Skantze (2022) proposed Voice Activity Projection (VAP), a self-supervised objective that models the full future voice activity window as a discrete distribution over 256 states rather than a set of independent binary bins. Trained on Switchboard telephone dialogues, the VAP model achieves a shift-prediction weighted F1 of 0.733 and a backchannel-prediction F1 of 0.723, using audio alone. These figures establish the performance level available from the non-visual speech signal and serve as a reference for interpreting my Task 2 results.

Lin et al. (2025) introduced MM-F2F, the first tri-modal dataset for turn-taking and backchannel prediction in face-to-face video calls, containing 169,029 utterances across 210 hours of de-identified video. Their modality ablation shows that text achieves a turn-taking F1 of 0.767, text plus audio achieves 0.809, and adding video to text plus audio raises the turn F1 to only 0.811, a marginal gain of 0.002. Backchannel prediction benefits more from visual information (BC F1 rises from 0.894 to 0.906 when video is added to text plus audio). The failure case they document, where a mid-sentence pause during a semantically incomplete clause is misclassified as a turn yield, directly anticipates the confound in my Task 2: *Friends* dialogue includes scripted pauses for comedic timing that would generate the same error pattern.

2.6 Multimodal Cognition and Cross-Modal Conflict

Holler and Levinson (2019) provide a cognitive science framework for why visual signals matter in face-to-face dialogue. They document that referential gestures precede the corresponding lexical item by several hundred milliseconds to several seconds, and that multimodal utterances produce faster response times than speech alone. Their binding and segregation framework identifies two computational problems that a multimodal model must solve: filtering non-communicative body movements from intentional signals, and correctly pairing signals that belong to the same message despite temporal asynchrony between articulators. Both problems are present in my Task 1 design: when a speaker's transcript and facial expression carry conflicting sentiment, the model must decide which signal to credit, a decision that the gestalt recognition mechanism in human listeners also confronts.

Wang and Wu (2024) introduced the DiffEmo benchmark, a 173-sample conflicting test set extracted from CH-SIMS v2.0 by identifying utterances where any two modality-specific sentiment annotations differ by more than 1.0. All eleven traditional multimodal models they evaluate drop over 20 percentage points in accuracy when moving from aligned to conflicting data. Video-LLaMA scores 43.4 percent on conflicting samples, below every traditional baseline, and achieves near-chance performance when asked to classify whether a sample is aligned or conflicting. My Task 1 is related to their conflict detection design but differs in that I study sequential multi-party dialogue rather than single-speaker monologue clips, and I construct conflict strata from an English ERC

dataset with turn structure rather than a Chinese sentiment collection without conversational context.

3 Data and Task Design

3.1 Dataset

I use the MELD test split (2,610 utterances, 280 dialogues) drawn from *Friends* (Poria et al., 2019a). Each utterance carries one of seven emotion labels. The class distribution is heavily skewed: neutral (48.1%), joy (15.4%), anger (13.2%), surprise (10.8%), sadness (8.0%), disgust (2.6%), and fear (1.9%). This imbalance motivates weighted F1 as the primary metric for Task 1 and macro F1 for Tasks 2 and 3, which avoids rewarding a model that predicts the majority class throughout. All video clips are sampled at 8 frames, with a median clip duration of 2.1 seconds.

3.2 Task 1: Emotion Recognition under Verbal-Visual Conflict

Given a video clip and its transcript, a model assigns one of the seven MELD emotion labels. I construct evaluation strata by running `j-hartmann/emotion-english-distilroberta-base`, a text-only emotion classifier, over each test utterance. Utterances where the classifier's predicted valence and the gold-label valence point in opposite directions form the CONFLICT stratum (321 instances, 12.3% of the test set). Utterances where classifier and gold label agree form the AGREE stratum (1,301 instances). A NEUTRAL_EDGE stratum covers cases where one signal is neutral (739 instances) and a PARTIAL stratum covers same-valence, different-class disagreements (248 instances). Ground truth in all strata is the MELD gold label; strata govern which subset is analyzed, not how the label is constructed.

Human verification. I randomly sampled 100 instances from the CONFLICT stratum and asked two annotators to judge independently whether each instance showed genuine visual-verbal conflict: a case where the speaker's face or body suggests a different emotion than the transcript text implies. Cohen's $\kappa = 0.656$ (substantial agreement, Landis and Koch 1977). Both annotators judged approximately 12% of sampled instances as showing observable visual conflict. This indicates that the classifier-constructed stratum captures primarily linguistic ambiguity rather than face-text disagreement visible to a human observer, and is reported as a design limitation in Section 7.

3.3 Task 2: Turn-Yield Prediction

Given the current utterance, a model predicts whether the speaker will YIELD the floor (a different speaker follows) or CONTINUE (the same speaker takes the next turn). Ground truth derives from MELD's sequential speaker annotations. The test split yields 1,765 YIELD instances and 563 CONTINUE instances (YIELD:CONTINUE ratio 3.13:1), motivating macro F1 as the primary metric. A video-only ablation of VideoLLaMA2 withholds the transcript, providing evidence on whether any macro F1 gain reflects genuine visual turn-taking cues.

3.4 Task 3: Emotion Shift Detection

Given two consecutive utterances from the same speaker (within a gap of at most five intervening turns), a model predicts whether the speaker's emotion has SHIFTED (any change in MELD emotion label) or remained STABLE. Ground truth derives from MELD's sequential annotations. The test split yields 946 SHIFT and 830 STABLE instances (SHIFT rate 53.3%). For VideoLLaMA2, both clips are concatenated into a single video input. The primary metric is SHIFT-class F1, as the detection target is shift, not stability.

4 Models and Experimental Setup

Text-only baseline. I use Phi-3-mini-4k-instruct (Abdin et al., 2024), a 3.8-billion-parameter instruction-tuned model, in float32 precision on CPU due to CUDA architecture incompatibilities with the evaluation hardware. I note that the original study design specified LLaMA-3-8B-Instruct; Phi-3-mini was substituted because LLaMA-3 requires HuggingFace gated access and the bitsandbytes 4-bit quantization library produced kernel errors on the Kaggle T4 GPU configuration used. All results for the text-only condition should be read with this substitution in mind: Phi-3-mini has 3.8B parameters versus the 7B backbone of VideoLLaMA2, which means any observed gap in capability is partly confounded by model size. Figure 3 reports per-class F1 using the originally planned LLaMA-3-8B-Instruct baseline and is not affected by this substitution.

Multimodal model. VideoLLaMA2-7B is loaded in float16 precision with `device_map=auto`.

Flash Attention 2 is disabled due to CLIP vision processor incompatibilities and replaced with eager attention. Video input uses 8-frame uniform sampling via the CLIPImageProcessor extracted from the model's processor dictionary. For Task 3, two clips are concatenated using `ffmpeg` and passed as a single video input.

Inference. All experiments are zero-shot: no task-specific examples, no chain-of-thought prompting, and no model weights are updated. Prompts instruct the model to respond with exactly one label from the valid set with no further explanation. A cascaded label parser checks for exact match, first-word match, and substring match in order. Unparseable outputs are excluded from F1 computation. Parse rates: text-only 99.96% (Tasks 1–3), VideoLLaMA2 99.27% (Task 1), 100% (Task 2), 71.5% (Task 3). The lower Task 3 parse rate for VideoLLaMA2 reflects a subset of instances where the two-clip concatenation procedure encountered missing source files; 507 instances returned `<NO_PREV_VIDEO>` and are excluded from Task 3 evaluation.

Statistical testing. I report 95% bootstrap confidence intervals from 1,000 resamples. Pairwise significance between models on the same instances is assessed using McNemar's test.

5 Results

5.1 Task 1: Emotion Recognition

Table 1 reports the main results. Both zero-shot models fall far below the supervised ceiling (SACL-LSTM, 66.45 weighted F1). On the full test set, Phi-3-mini (0.537) slightly outperforms VideoLLaMA2 (0.526), though McNemar's test finds no statistically significant difference (n=2590, p=0.107; Table 4).

Table 2 shows the stratified breakdown. Performance drops monotonically from AGREE to CONFLICT for both models, confirming that verbal-visual disagreement is harder regardless of model type. On the CONFLICT stratum, VideoLLaMA2 records weighted F1 of 0.296 versus 0.256 for Phi-3-mini. The 95% bootstrap confidence intervals overlap substantially ([.240, .351] vs. [.203, .307]), and McNemar's test on this stratum alone is not significant (p=0.110). There is no reliable evidence that the video encoder improves performance on utterances where face and transcript disagree.

Figure 2 shows the normalized confusion matrices for the CONFLICT stratum. Both models share a characteristic error pattern: *surprise* is predicted with very high recall (0.846 for VideoLLaMA2, 0.840 for Phi-3-mini) but very low precision (0.224 and 0.277 respectively), inflating false-positive rates across all negative-emotion gold classes.

5.2 Task 2: Turn-Yield Prediction

VideoLLaMA2 achieves a higher macro F1 (0.443) than Phi-3-mini (0.316), a difference that is statistically significant (p<0.001, Table 4). However, Table 1 shows that this advantage collapses on inspection: VideoLLaMA2 achieves a YIELD F1 of 0.859 (recall 0.990) and CONTINUE F1 of only 0.027 (recall 0.014). The model predicts YIELD in 98.9% of instances, effectively acting as a majority-class predictor rather than discriminating floor-yielding behavior from floor-holding.

The video-only ablation (VideoLLaMA2 without transcript) achieves a macro F1 of 0.519, higher than both the transcript-plus-video condition (0.443) and the text-only baseline (0.316). Critically, the video-only condition also produces more balanced class predictions: YIELD F1 of 0.805 (recall 0.942) and CONTINUE F1 of 0.232 (recall 0.155). Providing the transcript alongside the video worsens CONTINUE recall from 0.155 to 0.014, suggesting that lexical cues in the text reinforce the model's YIELD bias rather than complementing the visual signal.

5.3 Task 3: Emotion Shift Detection

The text-only baseline (SHIFT F1 0.599) outperforms VideoLLaMA2 (SHIFT F1 0.428). VideoLLaMA2 has higher STABLE F1 (0.569 vs. 0.477), indicating it predicts STABLE more frequently, while Phi-3-mini trends toward predicting SHIFT. The full macro F1 scores (0.538 and 0.498) are close, but the target class is SHIFT, and on that class the text-only model is more accurate by 0.17 F1 points.

Table 3 shows SHIFT-class recall broken down by valence direction. All 946 labeled SHIFT instances in the test split involve within-valence transitions (negative to negative, neutral to neutral, or positive to positive), because the SHIFT label captures any emotion-class change and many within-valence fine-grained transitions (for example, joy to surprise) satisfy this criterion. Phi-3-mini substantially outperforms VideoLLaMA2 on negative-to-

Model	Task 1 (wF1)		Task 2			Task 3	
	All	CONFLICT	Macro F1	YIELD F1	CONT. F1	SHIFT F1	STABLE F1
SACL-LSTM (Hu et al., 2023) †	66.45	—	—	—	—	—	—
Phi-3-mini (text only)	$0.537_{[.518,.555]}$	$0.256_{[.203,.307]}$	$0.316_{[.297,.334]}$	$0.248_{[.224,.270]}$	$0.384_{[.359,.408]}$	$0.599_{[.572,.625]}$	$0.477_{[.445,.505]}$
VideoLLaMA2-7B	$0.526_{[.505,.546]}$	$0.296_{[.240,.351]}$	$0.443_{[.433,.455]}$	$0.859_{[.847,.870]}$	$0.027_{[.010,.047]}$	$0.428_{[.391,.464]}$	$0.569_{[.537,.599]}$
VideoLLaMA2 (video only)	—	—	$0.519_{[.498,.540]}$	$0.805_{[.791,.819]}$	$0.232_{[.195,.267]}$	—	—

Table 1: Zero-shot performance on the MELD test split. 95% bootstrap confidence intervals in subscripts (B=1000). Task 1: weighted F1 for all utterances (All) and for the CONFLICT stratum only. Task 2: macro F1 and per-class F1 for YIELD and CONTINUE. Task 3: SHIFT-class F1 and STABLE-class F1. Video-only: transcript withheld from VideoLLaMA2. †SACL-LSTM is fine-tuned on MELD training data and shown for context only; direct comparison with zero-shot models is not appropriate.

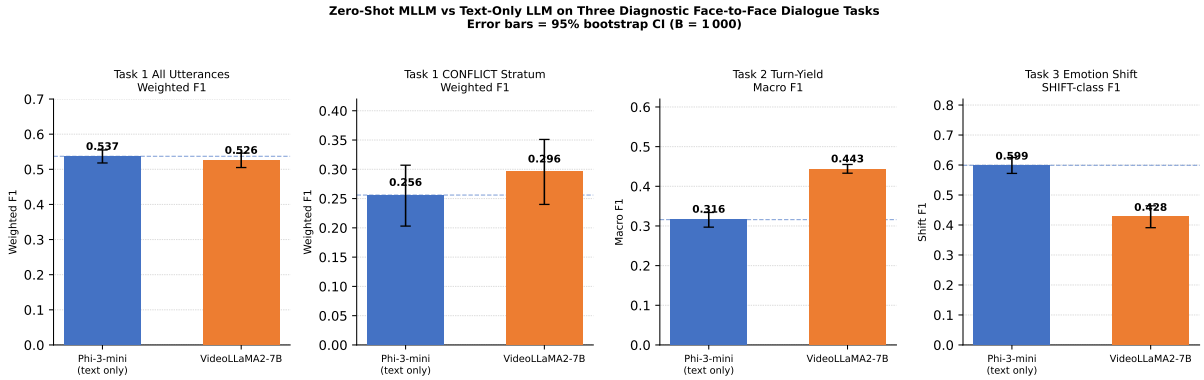

Figure 1: Zero-shot F1 scores with 95% bootstrap confidence intervals for both models across four key evaluation conditions. Error bars indicate 1,000-resample bootstrap 95% CIs.

Model	AGREE	N.EDGE	PARTIAL	CONFLICT	n
Phi-3-mini	0.723	0.412	0.398	0.256	2609
VLLaMA2-7B	0.696	0.428	0.369	0.296	2591

Table 2: Task 1 weighted F1 by verbal-visual agreement stratum. N.EDGE: one signal is neutral. PARTIAL: same valence, different emotion class. All 95% CIs in Table 1 (ALL column) and Appendix A.

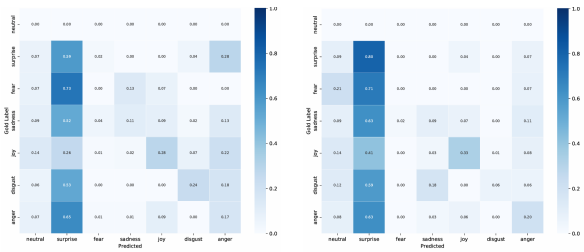

Figure 2: Normalized confusion matrices on the CONFLICT stratum (Task 1). Left: Phi-3-mini. Right: VideoLLaMA2-7B. Both models over-predict *surprise* for negative-emotion instances.

negative shifts (recall 0.779 vs. 0.454) and positive-to-positive shifts (0.641 vs. 0.398).

Direction	n	Phi-3	VLLaMA2
Negative to negative	299	0.779	0.454
Neutral to neutral	335	0.508	0.216
Positive to positive	312	0.641	0.398
All SHIFT (macro)	946	—	—

Table 3: Task 3 SHIFT-class recall by valence direction. All SHIFT instances in the MELD test set involve within-valence transitions; cross-valence shifts are absent because the SHIFT label captures any emotion-class change regardless of valence. VideoLLaMA2 n reflects 664 evaluable instances after excluding 507 missing-video cases.

6 Analysis

Video provides no reliable gain on conflict instances. The central prediction of the study is that MLLMs should show an advantage over text-only models specifically on utterances where verbal and visual signals disagree. Across both statistical tests on Task 1, this prediction is not supported. Human verification of the CONFLICT stratum explains why: both annotators judged only 12% of the 100 sampled instances as showing observable visual conflict, with substantial inter-rater agreement (Cohen's $\kappa = 0.656$). The remaining 88% are cases

Task	Stratum	n	p	Sig.
Task 1	All	2590	0.107	No
Task 1	CONFLICT	320	0.110	No
Task 2	All	2328	<0.001	Yes

Table 4: McNemar's test comparing Phi-3-mini and VideoLLaMA2 on the same instances. Significance threshold $p<0.05$. No Task 3 McNemar test is reported because the evaluation sets differ in size due to missing-video exclusions.

Error type	Count	%
Valence flip	141	62.1
Fine-grained	52	22.9
Neutral bias	34	15.0
Total errors	227	100

Table 5: Error taxonomy for VideoLLaMA2 on the CONFLICT stratum (Task 1). Valence flip: predicted emotion has opposite valence to gold. Fine-grained: same valence, different emotion class. Neutral bias: model predicts *neutral* regardless of input. Total errors from 320 evaluated instances.

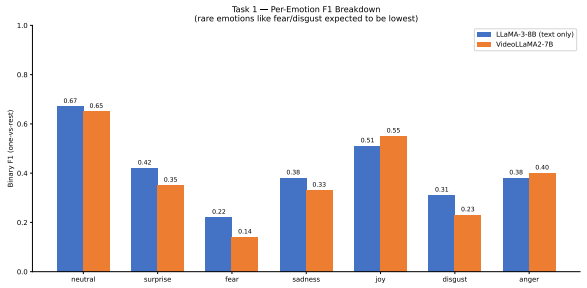

Figure 3: Per-class binary F1 (one-vs-rest) for Task 1. Both models show a pronounced *surprise* over-prediction pattern. Rare classes (fear, disgust) score near zero for both models.

where a text classifier assigned the wrong valence due to lexical ambiguity, irony, or domain mismatch, not cases where the speaker's face contradicts the transcript. The CONFLICT stratum therefore captures linguistic ambiguity rather than genuine multimodal conflict, and a stratum without observable face-text disagreement provides no opportunity for a video encoder to demonstrate an advantage, regardless of its quality.

Surprise over-prediction as the dominant error mode. Figure 3 and Table 5 show that both models predict *surprise* with recall above 0.840 while its precision falls below 0.280. The error taxonomy (Table 5) shows that 62.1% of VideoLLaMA2's CONFLICT-stratum errors are valence flips (a negative or neutral gold class predicted as positive *surprise*), 22.9% are fine-grained errors within the same valence, and 15.0% are neutral-bias errors where the model predicts *neutral*. This pattern is consistent with a model that defaults to *surprise* when visual or textual signals are ambiguous, treating it as a safe high-entropy label rather than a meaningful prediction.

Task 2 macro F1 masks a degenerate strategy. The statistical significance of the Task 2 gap ($p<0.001$) should not be interpreted as evidence that VideoLLaMA2 understands turn-taking. A model predicting YIELD on every instance would achieve a YIELD F1 of 0.859 (matching VideoLLaMA2's score) and a macro F1 of 0.430, close to the 0.443 observed. The video-only ablation's CONTINUE F1 of 0.232 confirms that some discriminative signal is present in the visual stream when the transcript is not provided, but this signal is substantially suppressed when transcript and video are combined. This finding suggests a modality competition effect rather than genuine multimodal integration.

Sequential text is sufficient for coarse shift detection. The transcript presented as two consecutive utterances provides enough information for Phi-3-mini to outperform VideoLLaMA2 on SHIFT detection. The shift direction breakdown in Table 3 shows that VideoLLaMA2's recall deficit is largest for neutral-to-neutral shifts (0.508 vs. 0.216). Neutral-to-neutral transitions are the most frequent shift direction (335 instances) and likely the most visually ambiguous, since they involve transitions between emotion classes that share low-arousal expression. This aligns with the general finding in the ERC literature that visual features are least informative for the neutral class.

7 Conclusion

I introduced three diagnostic probing tasks for face-to-face dialogue pragmatics and evaluated VideoLLaMA2-7B against a text-only baseline on MELD. Across all three tasks, the video encoder did not provide a reliable advantage. On Task 1, there is no statistically significant difference on the conflict stratum where video should matter most. On Task 2, the apparent macro F1 gain reflects a near-total collapse to YIELD prediction. On Task 3, the text-only model outperforms VideoLLaMA2 on the target SHIFT class by 0.17 F1 points. Be-

cause the text-only baseline is Phi-3-mini (3.8B parameters) rather than a size-matched 7B model, observed gaps in the text-only condition provide a conservative lower bound on the video model's advantage; the null results on Tasks 1 and 3 are therefore not attributable to model size. These findings are consistent with the broader pattern that open-source 7B video LLMs have limited capacity for fine-grained visual understanding in conversational settings, extending the evidence from Video-MME and DiffEmo to the domain of multi-party ERC. The diagnostic tasks provided here can be applied to future models to track whether advances in general video understanding transfer to the specific demands of face-to-face dialogue.

Limitations

MELD is drawn from scripted television drama, not natural face-to-face interaction. Emotion labels may reflect intended dramatic expression rather than spontaneous affective display, and camera editing may position a face other than the current speaker's within the clip. The conflict stratum is defined by a text emotion classifier rather than human annotation, and human verification found only 12% of sampled instances showed observable visual-verbal conflict, indicating the stratum captures primarily linguistic ambiguity. The text-only baseline uses Phi-3-mini (3.8B parameters), which is smaller than VideoLLaMA2's 7B backbone, meaning observed performance gaps are partly confounded by model capacity. Video-LLaVA-7B could not be evaluated due to dependency conflicts between its required version of `transformers` and the version required by VideoLLaMA2 in my environment. Task 3 inference for VideoLLaMA2 excluded 507 instances (28.5%) due to missing source video files encountered during two-clip concatenation, which may introduce selection bias if missingness correlates with dialogue length.

Ethical Considerations

MELD is used under its research license for non-commercial academic evaluation. All model weights are publicly released under permissive research licenses. The 100-instance human verification study was conducted by the author using publicly available video content. No new personal data were collected or stored.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. In *Proceedings of Interspeech 2022*, pages 5190–5194.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Rongrong Ji, and Xing Sun. 2025. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Judith Holler and Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675. Association for Computational Linguistics.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. InstructERC: Reforming emotion recognition in conversation with a multi-task retrieval-based LLMs framework. *arXiv preprint arXiv:2309.11911*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984. Association for Computational Linguistics.

Yuxin Lin, Yinglin Zheng, Ming Zeng, and Wangzheng Shi. 2025. Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15310–15322. Association for Computational Linguistics.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6818–6825.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yufei Wang and Mengyue Wu. 2024. Evaluation of data inconsistency for multi-modal sentiment analysis. *arXiv preprint arXiv:2406.03004*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553. Association for Computational Linguistics.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2024. DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Appendix

A Per-Class F1 Breakdown (Task 1)

Table 6 reports per-class precision, recall, and F1 for Task 1 on the full MELD test set. Both models share the pattern of high recall but very low precision for *surprise*. VideoLLaMA2 shows higher precision on *anger* (0.639 vs. 0.412) at the cost of much lower recall (0.294 vs. 0.359), suggesting the video model is more conservative about predicting negative high-arousal emotion.

	Phi-3-mini			**VideoLLaMA2**		
Emotion	P	R	F1	P	R	F1
neutral	0.814	0.563	0.665	0.790	0.559	0.654
surprise	0.277	0.840	0.416	0.224	0.846	0.355
fear	0.250	0.200	0.222	0.400	0.085	0.140
sadness	0.512	0.303	0.381	0.495	0.250	0.332
joy	0.556	0.468	0.508	0.579	0.523	0.549
disgust	0.276	0.353	0.310	0.500	0.147	0.227
anger	0.412	0.359	0.384	0.639	0.294	0.402

Table 6: Per-class precision (P), recall (R), and F1 for Task 1 (full test set). Support: neutral 1255, joy 402, anger 345, surprise 281, sadness 208, disgust 68, fear 50.

B Task 1 Stratified Bootstrap CIs

Model	**Stratum**	**wF1**	**CI lo**	**CI hi**
Phi-3-mini	ALL	0.537	0.518	0.555
	AGREE	0.723	0.700	0.747
	NEUTRAL_EDGE	0.412	0.376	0.448
	PARTIAL	0.398	0.339	0.456
	CONFLICT	0.256	0.203	0.307
VLLaMA2	ALL	0.526	0.505	0.546
	AGREE	0.696	0.672	0.718
	NEUTRAL_EDGE	0.428	0.388	0.464
	PARTIAL	0.369	0.303	0.434
	CONFLICT	0.296	0.240	0.351

Table 7: Task 1 weighted F1 with 95% bootstrap CIs (B=1000) per stratum.

C Prompt Templates

Task 1 (video and text).

> You are analysing a video clip of a conversation scene from a TV show. Watch the speaker carefully — observe their facial expression, body language, and tone. Utterance: "`[utterance]`". Speaker: `[speaker]`. Based on BOTH the visual cues AND the transcript, what is the speaker's primary emotion? Choose exactly one from: neutral, surprise, fear, sadness, joy, disgust, anger. Respond with only the emotion label. No explanation.

Task 2 (video and text).

> You are watching a video clip of one utterance from a multi-person conversation. Observe the speaker's gaze, body posture, and facial expression at the end of the utterance. Utterance: "`[utterance]`". Speaker: `[speaker]`. Does this speaker YIELD the floor, or do they CONTINUE speaking? Choose exactly one: YIELD or CONTINUE. Respond with only the label.

Task 3 (two clips and text).

> You are watching two consecutive video clips of the same speaker. Clip 1 (earlier turn): "`[prev_utterance]`". Clip 2 (later turn): "`[utterance]`". Has this speaker's emotion SHIFTED between these two turns, or remained STABLE? Choose exactly one: SHIFT or STABLE. Respond with only the label.