Video2Reaction: Mapping Video to Audience Reaction Distribution in the Wild

Trang Nguyen[1]*, Sidong Zhang[1], Shiv Shankar[1], Gauri Jagatap[2], Deepak Chandran[2], Andrea Fanelli[2], and Madalina Fiterau[1]

[1] University of Massachusetts Amherst, 130 Governors Drive, Amherst, MA, USA
tramnguyen@umass.edu, sidongzhang@umass.edu, sshankar@umass.edu, mfiterau@cs.umass.edu

[2] Dolby Laboratories, 1275 Market Street, San Francisco, CA, USA
Gauri.Jagatap@dolby.com, Deepak.Chandran@dolby.com, Andrea.Fanelli@dolby.com

Abstract. Understanding and forecasting audience reactions to video content are crucial for improving content creation, recommendation systems, and media analysis. To enable audience reaction prediction and other content engagement applications, we introduce **Video2Reaction**, a multimodal dataset that maps short movie segments to a distribution of *induced emotions* of viewers in the wild, as expressed through social media. **Video2Reaction** spans more than 10,000 videos and serves as a reliable benchmark as well as a training resource for audience reaction prediction. To enable cost-effective continuous annotations as reactions may change over time, we develop a two-stage multi-agent pipeline using only open-source LLMs, achieving 86% correctness under blind human verification despite the inherently noisy and subjective nature of the task. We establish the first benchmark for video-to-reaction-distribution prediction in the wild and show that pretrained foundation video models fail in zero-shot settings, while finetuning transforms them into state-of-the-art predictors capable of modeling both full reaction distributions and dominant responses from video alone. However, the task remains challenging: even the strongest methods achieve only 77% Top-3 F1 in dominant reaction prediction (LLaVA-Next), highlighting a substantial gap in modeling collective audience reaction. Dataset and code are available at our project page: https://information-fusion-lab-umass.github.io/video2reaction-bench.github.io/.

1 Introduction

Understanding how people react to video content is a crucial yet underexplored aspect of affective computing. Prior work has established a distinction between perceived emotion—the emotion conveyed or expressed by the content itself—and induced emotion—what is experienced by the audience [29]. While existing video sentiment datasets focus on perceived emotions (e.g., emotions of

* Corresponding author

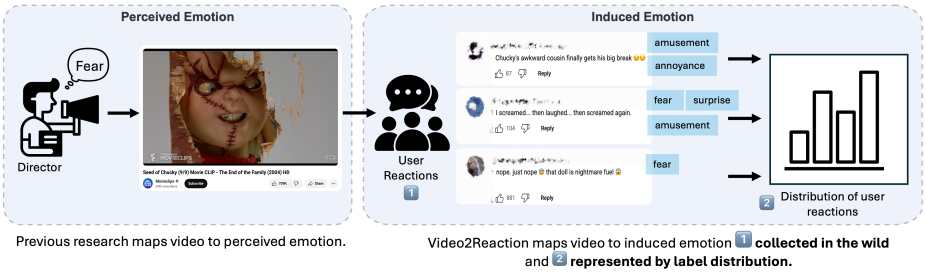


Fig. 1: Video2Reaction is the first benchmark that uses video data to directly learn induced emotion distribution in the wild.

characters in the scene or filmmaker's emotion intent), audience reactions can vary greatly depending on personal, cultural, or temporal context. For example, Figure 1 illustrates how a horror movie scene might be perceived as frightening or suspenseful based on the director's intent, but viewer reactions could diverge significantly-some may experience genuine fear and anxiety, others might laugh at predictable genre tropes, and those with personal trauma related to similar situations might be triggered and distressed. This underscores the need to study induced emotional reactions directly, rather than relying on perceived sentiment, so that the bias between the content creator's intention and possible audience feedback can guide the content creation.

In this paper, we introduce **Video2Reaction**, a multimodal dataset consisting of over 10,000 videos of movie content, paired with audience reaction distribution derived from viewer comments, allowing for a precise mapping between visual content and the emotional reactions it induces. To our knowledge, **Video2Reaction** is the first and largest multimodal dataset capturing distribution of induced emotion from cinematic content in non-controlled environments (in the wild). Unlike perceived emotions, which are typically modeled as unimodal (single-label), induced emotions can vary from person to person, and it is important for content engagement applications to capture this variance rather than simply predict single or multi-label outputs. To address this, we frame audience emotion recognition as a **label distribution learning** (LDL) problem [10], in which each video is mapped to a distribution of emotional reactions across viewers .

Our main contributions are as follows:

- **A large-scale benchmark for induced audience emotion in the wild.** To our knowledge, **Video2Reaction** is the first and largest multimodal dataset capturing *induced emotion* in the wild (10,348 videos spanning approximately 400 hours and 800,000 comments) from cinematic content.
- **A scalable, updatable LLM-based annotation pipeline with extensive quality analysis.** We develop a two-stage multi-agent LLM-based annotation pipeline that enables cost-effective, extensible reaction labeling, paving the way for future dataset updates as new content and evolving au-

dience perspectives emerge. Beyond scalability, we conduct extensive quantitative and human evaluations to characterize annotation reliability, error modes, and robustness. Two complementary human evaluation results show LLM annotations operate within the range of human agreement while enabling large-scale distributional labeling at a fraction of the cost.
– **A new task: distributional video-to-reaction prediction.** We propose a novel and challenging benchmark: predicting audience reaction distributions directly from multimodal video content, and design a comprehensive evaluation framework that captures both distributional prediction and dominant reaction classification. Our extensive benchmark, including classical LDL algorithms (i.e. SA-BFGS [10]), adapted multimodal emotion recognition models (i.e. CubeMLP [26]), and large vision-language models (i.e. Gemini 2.5 [6]), demonstrate that this task is challenging yet learnable.
– **Demonstrating the potential of foundation models for audience reaction forecasting.** We demonstrate that pretrained VLMs fail zero-shot but can be transformed into state-of-the-art audience reaction predictors via lightweight finetuning, with preliminary evidence of cross-dataset transfer.

2 Related Work

2.1 Emotion Video Datasets

Table 1 summarizes key characteristics of existing emotion video datasets and **Video2Reaction**. Most prior datasets like IEMOCAP [3], MELD [25], and CMU-MOSEI [37] focus on *perceived emotions*—the emotions expressed or felt by the characters in a movie scene—rather than the *induced emotions* experienced by the audience. Some datasets like LIRIS-ACCEDE [2], COGNIMUSE [40], and DEAP [19] have attempted to capture induced emotion. However, those datasets only capture emotional response in the 2D valence-arousal space, whereas our dataset captures the full distribution of categorical emotions. Furthermore, prior datasets often rely on controlled environments, such as lab settings or small groups of participants watching content together [23, 29], limiting the impact across larger and more diverse demographics. For instance, Muszynski *et*

Table 1: Comparison of **Video2Reaction** and Existing Emotion Prediction Video Datasets

Dataset	# Videos (Hours)	Emotion Type	Emotion Setting	Emotion Representation
IEMOCAP [3]	7,433 (12 hours)	Perceived	Lab	Single-label
MELD [25]	13,000 (13 hours)	Perceived	Lab	Single-label
CMU-MOSEI [37]	23,453 (66 hours)	Perceived	Lab	Continuous, Single-label
LIRIS-ACCEDE [2, 23]	9, 800 (27 hours)	Perceived, Induced	Lab	Continuous
COGNIMUSE [40]	50 (3.5 hours)	Induced	Lab	Continuous, Single-label
DEAP [19]	120 (2 hours)	Induced	Lab	Continuous
CMSV [35]	8,210 (69 hours)	Induced	Social Media	Single-label
VCE [21]	61,046 (239 hours)	Induced	Lab	Distributional
Video2Reaction (ours)	10,348 (398 hours)	Induced	Social Media	Distributional

al. [23] extended LIRIS-ACCEDE database [2] to use aesthetic features from movie scenes to model induced emotions, but only with 10 participants in a co-viewing setting—unlike the typical media consumption nowadays. CMSV [35] is a recent benchmark that attempts to capture induced emotion in the wild but their task is to predict induced emotion given a pair of video and comments while our benchmark predict induced emotion from video content only.

Additionally, existing emotion video datasets typically aggregate reactions into a single outcome label (e.g., a majority vote or mean score) for each clip [2, 19, 40], failing to reflect the diversity of viewer responses. In contrast, we model the full distribution of emotional reactions, using soft labels derived from real-world audience responses.

Closest to our work is the VCE dataset [21], which collects categorical emotional responses to video content. As noted by its authors, VCE relies on a relatively controlled crowdworker pool and does not aim to model reactions from a culturally diverse, population-scale audience. In contrast, our work captures naturally occurring, large-scale audience reactions by leveraging social media data, beginning with YouTube as a globally popular platform. This design better reflects the ecological diversity and inherently distributional nature of real-world audience engagement. Furthermore, VCE requires substantial manual annotation effort—reportedly involving 400 annotators—making large-scale expansion and iterative updates costly. Our approach instead employs an agentic annotation pipeline to streamline and automate labeling, enabling faster dataset updates and improved adaptability given the non-static nature of the task.

2.2 Label Distribution Learning

Label Distribution Learning (LDL) was formalized by [10] as a framework to address label ambiguity problems in tasks such as age estimation and sentiment prediction. Unlike traditional classification, which assumes a single ground truth label per instance, LDL represents each instance with a distribution over labels, capturing uncertainty and correlation among neighboring labels.

LDL methods are generally categorized into three families [10]: problem transformation (PT), algorithm adaptation (AA), and specialized algorithms (SA). Further details are included in Section 4 and in Appendix C.

The benchmark we design builds on existing LDL methods, extending them to incorporate foundational Vision-Large Language Models as an additional class of algorithms capable of learning distribution alignment [22].

3 Video2Reaction Dataset

The **Video2Reaction** dataset has been developed to facilitate the prediction of audience reactions from short movie segments. Each instance consists of a movie clip paired with a probability distribution over possible audience reactions (listed in Table 2). To facilitate future research on raw video analysis, we provide YouTube video IDs for all clips. Additionally, we release preprocessed features,

Table 2: Finer-grained reaction categories (21) used in **Video2Reaction** grouped by sentiment.

Sentiment Category	Finer-grained Reaction Categories
Positive	amusement, excitement, joy, caring, admiration, relief, approval
Negative	fear, nervousness, embarrassment, disappointment, sadness, grief, disgust, anger, annoyance, disapproval
Ambiguous	realization, surprise, curiosity, confusion

including state-of-the-art visual embeddings and audio embeddings extracted using pretrained models (more details in Appendix A.2).

3.1 Data Collection

We curate movie clips from the CondensedMovies dataset [1] (licensed with CC BY 4.0), which contains licensed content from the *Movieclips* YouTube channel. Each clip is a channel-curated movie segment capturing a key scene from the film. The use of licensed content improves the longevity of the dataset, as these clips are less likely to be removed from the platform. To ensure meaningful audience engagement, we retain only videos with a minimum of 10,000 views and at least 10 comments. The selected clips, originally uploaded between 2011 and 2019, are downloaded for further processing. Viewer comments on these videos extend through 2025, enabling the dataset to capture evolving audience perspectives and contemporary references, supporting a more comprehensive and generalizable benchmark for reaction prediction. Comments are collected and aggregated into a reaction distribution at the clip level; each clip is further decomposed into key frames for visual feature extraction (Appendix A.2), but both training and prediction remain at the clip level.

3.2 Comment-level Reaction Annotation

Reaction Taxonomy. To represent the complexity of audience reaction, we initially adopt the 28-category emotion taxonomy from GoEmotions [8], which was originally designed for Reddit comments and aligns naturally with our YouTube-based dataset. As a social media platform similar to Reddit, YouTube also features a wide range of audience interactions and emotional responses, making the GoEmotions taxonomy a natural fit. However, we drop 7 of the original categories in GoEmotions due to their significant under-representation in our data. These 7 reactions contribute to less than 0.01% of the distribution mass on average. Our final taxonomy consists of 21 fine-grained emotions (Table 2). The definition for each category is in Appendix A.3.

Two-stage Multi-agent Reaction Annotation Pipeline. Given the volume of raw comments and the non-stationary nature of induced emotions over

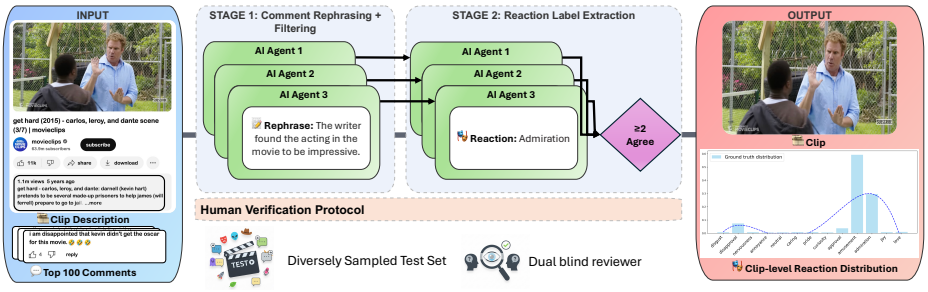


Fig. 2: Overview of Video2Reaction Two-Stage LLM-based Data Annotation Pipeline.
STAGE 1 rephrases comments to explicitly state their reactions towards the clip. It also filters out comments that lack a discernible reaction to the clip. STAGE 2 extracts reaction labels, with majority voting across three LLM agents to ensure consistency and discard ambiguous cases.

time, we designed a scalable and reliable multi-agent annotation pipeline that supports frequent dataset updates, ensuring long-term relevance and impact. Figure 2 outlines our LLM-based pipeline for annotating audience reactions based on user comments. Each comment is processed in two stages. The first stage of rephrasing comments is critical, as many audience comments reflect implicit reactions or off-topic remarks that need contextual interpretation to reveal their emotional intent. For example, a comment like *"I'm so disappointed this actor didn't win an Oscar"* will be mistakenly labeled as *disappointment* without being rephrased as *"the acting in the movie is so impressive"* in the first stage. The second stage then extracts relevant reaction labels from the rephrased comments from the first stage. Prompt details for both stages and qualitative comparison of single-stage versus two-stage annotations are provided in Appendix B.1. Further validation on the Stage 1 filtering step shows that the pipeline achieves 0.83 accuracy with 1.00 sensitivity, confirming it does not discard relevant comments (details in Appendix B.1).

Motivated by prior findings in the use of multi-agent framework in text classification [30], compound systems [4], and chain-of-thought reasoning [5, 31], we employ an ensemble of of three medium-sized multilingual instruction-tuned LLMs [3] and adopt a straightforward majority voting approach for our emotion annotation task. This ensemble design enables inference-time speedup through agent parallelization, in contrast to annotator–critic architectures that require sequential interaction. Our choice to use three medium-sized LLMs with comparable performance aligns with prior observations that ensemble methods are most effective when constituent models exhibit similar strength [30].

Quality of Reaction Annotations. To assess the quality of annotations in **Video2Reaction**, we conduct two complementary evaluations.

[3] LLaMA-3.1-8B-Instruct [12] , Qwen2.5-14B-Instruct [27] , and DeepSeek-R1-Distill-Qwen-7B [7]

Human–LLM annotation alignment. We perform independent human annotation of video–comment pairs with 29 participants and compute both inter-rater correlation and rater–LLM correlation following the protocol of GoEmotions [8]. After filtering ambiguous cases, 233 comments remain with a median of 3 annotators per comments. We compute Spearman correlation per emotion between each rater and the mean of other raters, and between each rater and our LLM-based pipeline. Consistent with prior findings in GoEmotions [8], agreement is moderate and varies substantially across emotions. The mean inter-rater correlation across 21 emotions is 0.428 (std = 0.233), reflecting the subjective and fine-grained nature of induced emotion labeling. The LLM achieves a comparable mean rater-LLM correlation of 0.402 (std = 0.243), closely matching human–human agreement. These results indicate that LLM annotations operate within the same reliability range as human annotators, supporting the validity of our annotation pipeline despite the inherent noise of the task. Further error analysis and breakdown by emotion category can be found in Appendix B.2. Using the same evaluation protocol, an ablated single-stage pipeline achieves a lower rater–LLM correlation of 0.34, compared to 0.40 for our two-stage pipeline and 0.43 for inter-rater agreement, confirming that the rephrasing stage improves annotation reliability (full comparison in Appendix B.1).

Dual-blind human verification. To further account for annotation noise from human raters in this task and to increase the test sample size without incurring significant increase in human evaluation time, we adopt a dual blind human verification protocol. First of all, we randomly sample 100 movie clips with balanced representation across all movie genres. From each clip, 10 comments are randomly selected, yielding a total of 1,000 comments for human evaluation. Then each comment is independently reviewed by two annotators. In cases of disagreement, a third annotator is consulted, and the final label is determined by majority vote. Overall, 86% of the LLM-assigned reaction labels were judged to be correct, 7.8% incorrect, and 6.2% indeterminate (Table 11 in Appendix B.2). Only 0.2% randomly-labeled comments are judged to be correct, showing that human annotators do not exhibit confirmation bias.

We further analyze LLM annotation errors for users' consideration and future improvements. As shown in Table 12 (Appendix B.2), performance remains above 70% across all genres. However, genres such as Comedy and Drama pose greater challenges due to their reliance on subtle cues (e.g., pop culture references), which can obscure emotional tone even for human annotators. Breaking down errors by emotion category instead (Table 13 in Appendix B.2), we find that *embarrassment* and *nervousness* are the most error-prone categories, likely due to their reliance on subtle tonal cues that are easily lost during rephrasing.

Discussion on Alternative Data Construction Approaches. An alternative approach to modeling audience reactions is to recruit participants to watch videos and report their emotional responses, as done in VCE dataset [21]. While this design allows controlled measurement of reaction distributions, it departs from natural viewing conditions and typically relies on relatively small and demographically constrained participant pools. As acknowledged by the cre-

ators of the VCE dataset [21], their annotations are not intended to represent the full diversity of audience emotional responses and are primarily designed for studying whether deep networks can acquire cognitive empathy. Such datasets therefore do not capture large-scale, in-the-wild audience behavior. In addition, participant-based data collection is costly and inherently static, making it difficult to update as cultural context and audience reactions evolve over time.

Another alternative is to manually annotate social media comments. However, fine-grained emotion labeling is time- and cost-intensive (median 40.7 seconds per comment in our human evaluation study), rendering large-scale and continuously updated annotation impractical.

In contrast, leveraging naturally occurring social media reactions enables us to capture large-scale, ecologically valid audience responses. Our LLM-based annotation pipeline makes it feasible to maintain a non-static, updatable benchmark. We do not claim that social-media-derived reactions replace controlled lab studies; rather, they provide a complementary and population-level perspective that would be difficult to obtain through small-scale human annotation alone.

3.3 Dataset Statistics

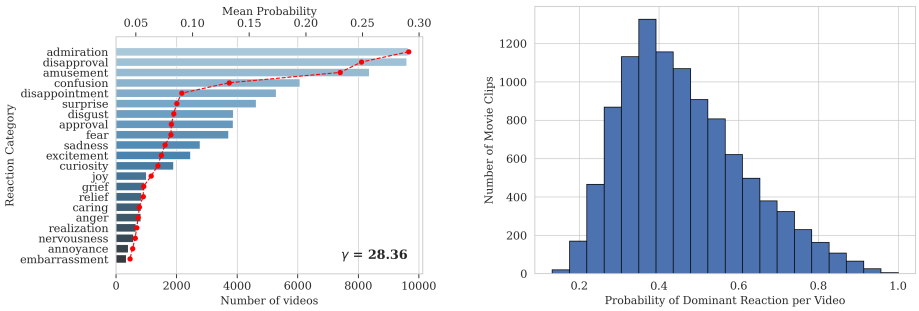
Table 3: Descriptive statistics for video and comment data in **Video2Reaction**.

Category	Total	Min	Mean	Median	Max
Movie-level Statistics					
Inter-segment Chebyshev distance in reaction distribution ([0.0,1.0])	1,545	0.05	0.48	0.48	0.91
Clip-level Statistics					
Clip Duration (sec)	389.81 hrs	23.15	135.61	127.60	367.36
Key Scenes	455,226	16	43.99	39.00	176
Comment-level Statistics					
Raw Comments	771,684	19	74.57	77.00	100
Retained Comments	252,462	10	24.40	22.00	100

Table 3 summarizes **Video2Reaction** at the movie, clip, and comment levels. The dataset spans 389 hours of video from 1,545 movies and includes 10,348 clips. On average, each clip is segmented into 44 scenes and is associated with 24 viewer comments used to construct the reaction distribution—substantially more than prior LDL benchmarks such as *Twitter_LDL* and *Flickr_LDL* [36], which constructs a label distribution from 11 annotators only.

Importantly, **Video2Reaction** captures substantial variation in audience reactions within the same movie, with a median Chebyshev distance of 0.48 (out of 1.0) between clip-level distributions of the same movie. This highlights the need for clip-level rather than movie-level modeling.

As shown in Figure 3a, the dataset is highly imbalanced across the 21 reaction categories (with an imbalance factor of 28.36), a challenge addressed in our evaluation design (Sections 4.2 and 5). Figure 3b further shows that top-1 reaction probability varies considerably across clips (with a median of approximately 0.4), underscoring the importance of modeling full reaction distributions instead of predicting only the dominant label.



(a) Total number of videos and mean video-level probability of each reaction category (γ denotes imbalance factor)

(b) Distribution of dominant reaction probability per video

Fig. 3: Key Statistics on Reaction Outcome in the Video2Reaction dataset.

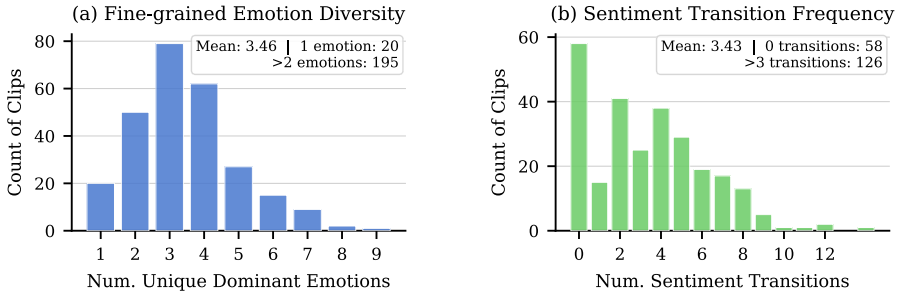


Fig. 4: Distribution of fine-grained emotion and sentiment transitions across movie clips. Sentiment transitions are computed at the monthly level.

3.4 Longitudinal Analysis of Audience Reactions

Further analysis on how audience reactions to the same clip evolve over time shows that audience reactions are non-stationary and can shift meaningfully over time, motivating our scalable annotation pipeline. As shown in Figure 4,

clips exhibit an average of 3.46 distinct dominant emotions over time: only 7.5% maintain a single dominant emotion, while 73.6% exhibit more than three. At the sentiment level, clips show an average of 3.43 sentiment transitions, with 47.5% exhibiting more than 3 transitions. For example, reactions to *There Will Be Blood* shifted from predominantly positive in 2016 to disapproval by 2020.

4 Audience Reaction Forecasting Benchmark

4.1 Problem Formulation

In this work, we frame the problem of audience reaction prediction as a label distribution learning (LDL) task. Unlike traditional classification, where the goal is to assign either a single label or a set of labels to each input instance, LDL seeks to predict a probability distribution over multiple labels, capturing the ambiguity and diversity in audience reactions.

Let x denote an input video clip, which may contain visual, audio, and textual content. The audience reaction to x is represented by a label distribution $\mathbf{d}_x = \{d_{xm}\}_{m=1}^{M}$, where M is the number of affective reaction classes (e.g., amusement, confusion, fear, etc.), and each $d_{xm} \in [0, 1]$ indicates the proportion of annotators or viewers who associated label m with the clip x. The label distribution satisfies the normalization constraint: $\sum_{m=1}^{M} d_{xm} = 1$. This distribution can be interpreted as the conditional probability $p(m|x)$ of observing reaction m given video x. Our objective is to learn a model $f_\theta(x)$ that predicts a distribution $\hat{\mathbf{d}}_x$ approximating the empirical distribution $\mathbf{d}_x$.

Importantly, in our setting, d_{xm} does not represent a soft target for a "correct" label in the traditional sense. Instead, it reflects the proportion of audience that has certain reaction given the video input.

4.2 Evaluation Metrics

Our benchmark is constructed along two complementary axes: (1) **full distribution evaluation**, which assesses how well the predicted distribution over all possible reactions matches the groundtruth distribution; and (2) **dominant reaction evaluation**, which focuses on how accurately the model identifies and estimates the probabilities of the most dominant reactions. The second axis is particularly relevant for real-world applications—such as content recommendation, moderation, or trailer editing—which depend primarily on anticipating the strongest or most likely viewer responses rather than capturing the full reaction spectrum. For each axis, we carefully select a suite of metrics designed to capture different types of errors that are important for our task, allowing future research to prioritize specific evaluation criteria based on downstream use cases. Table 4 summarizes all metrics used in the benchmark, along with their formulas and the error types they capture.

Full Distribution Evaluation. Following [10], we evaluate how closely the predicted reaction distribution matches the ground-truth distribution using a

Table 4: Summary of evaluation metrics used in the **Video2Reaction** benchmark.

Evaluation Category	Metric (Abbr.)	Formula	Error Type Captured
Full Distribution	Chebyshev (Che) ↓	$\max_j \left\|d_j - \hat{d}_j\right\|$	Largest per-class prediction error (worst-case mismatch)
	Clark (Cla) ↓	$\sqrt{\sum_{j=1}^{c} \left(\frac{d_j - \hat{d}_j}{d_j + \hat{d}_j}\right)^2}$	Errors on rare reactions (small denominators amplified)
	Kullback-Leibler (KL) ↓	$\sum_{j=1}^{c} d_j \ln \frac{d_j}{\hat{d}_j}$	Underestimation of true reactions, false-zero sensitivity
	Cumulative Absolute Distance (CAD) ↓	$\sum_j \lvert CDF(d_j) - CDF(\hat{d}_j)\rvert$	Distribution shifts ignoring ordinal relations
	Cosine (Cos) ↑	$\frac{\sum_{j=1}^{c} d_j \hat{d}_j}{\sqrt{\sum_{j=1}^{c} d_j^2}\sqrt{\sum_{j=1}^{c} \hat{d}_j^2}}$	Directional mismatch of prediction vs. groundtruth
	Intersection (Inter) ↑	$\sum_j \min(d_j, \hat{d}_j)$	Non-overlapping reaction prediction error
Dominant Reactions	Mean Reciprocal Rank (MRR) ↑	$\frac{1}{r}$	Incorrect ranking of dominant reaction (r is the predicted rank of the target dominant reaction)
	Top-1 Probability Error (TPE) ↓	$\left\lvert \hat{d}_{j^*} - d_{j^*}\right\rvert, \quad j^* = \arg\max_j d_j$	Probability misestimation of dominant reaction
	Top-k F1 (weighted) ($F1_k$) ↑	$\sum_j \frac{n_j}{N} \cdot \mathrm{F1}_{j,k}$	Both precision and recall-based errors for top-k reactions

suite of statistical distribution distance metrics. Since some metrics (i.e. Canberra and Clark) capture very similar types of errors for our task, we decided to include in the benchmark three distance-based metrics—Chebyshev, Clark, Kullback-Leibler (KL) and two similarity-based metrics—Cosine similarity and Intersection. In addition, inspired by [32], we incorporate an ordinal-aware evaluation by mapping the 21 reaction categories onto a valence-arousal-based emotional space. We then compute the Cumulative Absolute Distance (CAD) between the predicted and true *ordered* distributions. The metric assigns lower penalties to misclassifications involving emotionally similar reactions and higher penalties to those involving emotionally distant categories, encouraging models to make semantically coherent predictions.

Dominant Reaction Evaluation. Unlike traditional emotion classification benchmarks, which often focus on unimodal distributions centered on a single perceived emotion, audience reactions in our task are multimodal, reflecting the diversity of viewer responses (as seen in Figure 3b). Therefore, we evaluate not only the top-1 prediction (single-label) but also the task's performance as a multilabel classification problem. For single-label evaluation, we report the F1 score, Mean Reciprocal Rank (MRR), and Top-1 Probability Error (TPE). For multi-label evaluation, we compute F1 scores based on the top-k emotions from both the ground truth distribution and the model predictions. All F1 metrics are class-weighted to account for label imbalance.

4.3 Comparative Models

Building on the taxonomy of label distribution learning (LDL) algorithms proposed by [10], we organize comparative models into three categories—*Problem Transformation*, *Specialized Algorithms*, and *Algorithm Adaptation*—and extend the framework by introducing a fourth category, *Foundation VLMs*, to evaluate the potential of Video2Reaction as a finetuning dataset to improve the foundation models' capability of forecasting audience reaction distribution.

Problem Transformation (PT). These methods reduce the LDL task to standard single-label learning (SLL) by resampling the training data. Each training instance $(\mathbf{x}_i, \mathbf{d}_i)$ with a label distribution $\mathbf{d}_i$ over c classes is converted into multiple single-label examples $(\mathbf{x}_i, y_j)$ by sampling y_j for class j from $\mathbf{d}_i$. We include PT-Bayes [10] and LDSVR [11] as two representative baselines.

Specialized Algorithms (SA). These methods are explicitly designed for Label Distribution Learning (LDL), typically incorporating task-specific loss functions or optimization techniques. We evaluate three representative algorithms: SA-BFGS [10], which directly optimizes KL divergence using the BFGS algorithm; and LDL-LRR [17] and TLRLDL [20], both of which enhance the training objective by exploiting multi-label correlations.

Algorithm Adaptation (AA). These models were originally designed for related tasks, such as multimodal sentiment analysis, and are adapted here for LDL. We include CubeMLP [26], CTEN [39], and MMIM [13]. These models incorporate modules specifically designed to learn cross-modal features.

Foundation Vision-Language Models (VLMs). Beyond the standard LDL taxonomy [10], we introduce a fourth category to assess the zero-shot performance of large vision-language models. Although not trained for LDL, these models are increasingly used as proxies for human judgment in applications such as agent-based simulations [24] and behavioral research [16,18]. We evaluate one commercial model, Gemini 2.5 Flash [6], and two leading open-source models—LLaVA-Next-Video-7B [38] and Qwen2.5-VL [28]—in a prompted classification setting over fixed reaction labels.

Implementation Details. The PT, SA, and AA algorithms are trained on the dataset using 5 random seeds, and we report the mean of their performance in Tables 5 and 6 (standard deviation is reported in Appendix D.1. For foundation VLMs, we evaluate two variants. First, following [22], we apply temperature scaling on validation data to improve probability calibration. Second, we perform low-rank fine-tuning (LoRA) [15] with rank $r = 8$, $\alpha = 16$, for 3 epochs. During fine-tuning, we incorporate taxonomy random reordering and synonym replacement to improve robustness to label-space variation. Further details on algorithm-specific implementation are listed in Appendix C.

5 Results

5.1 Pretrained VLMs Fail to Predict Audience Reactions from Video Content

Despite large-scale multimodal pretraining, VLMs fail to predict fine-grained audience reaction distributions in a zero-shot setting, even with temperature scaling. As shown in Tab. 5, cosine similarity remains below 0.51 and intersection below 0.38, while Chebyshev distance exceeds 0.34 across models. Performance is similarly poor under dominant reaction evaluation (Tab. 6), with Top-1 F1 below 0.30 for all models. Thus, pretrained VLMs neither recover the overall reaction distribution nor reliably identify the dominant audience response. Although

expected without task-specific supervision, the substantial performance gap despite large-scale video pretraining suggests that audience reaction forecasting is a challenging capability not learned through generic multimodal pretraining, motivating the need for dedicated datasets and benchmarks.

Table 5: Full distribution benchmark results. Both finetuned foundation VLMs and SA-BFGS achieve best and comparable performances. Values in parentheses indicate the improvement of finetuning over the zero-shot temperature-scaled baseline.

Model Name	Cheb ↓	KL ↓	Cla ↓	Cad ↓	Inter ↑	Cos ↑
Foundation VLMs						
Gemini 2.5 Flash	0.3425	4.8102	3.4477	3.4316	0.3787	0.5029
LLaVA-Next-Video-7B						
- *Temperature-scaled*	0.4110	1.5547	3.9710	4.3269	0.2970	0.4185
- *LoRA Finetuned*	**0.1882** (-54.2%)	3.1765 (+104.3%)	3.9433 (-0.7%)	2.1416 (-50.5%)	**0.6861** (+131.0%)	**0.8663** (+107.0%)
Qwen2.5-VL						
- *Temperature-scaled*	0.3985	1.5216	3.9888	4.0333	0.3140	0.4401
- *LoRA Finetuned*	0.2047 (-48.6%)	3.4431 (+126.3%)	3.9446 (-1.1%)	2.2903 (-43.2%)	0.6656 (+111.9%)	0.8427 (+91.5%)
Problem Transformation						
PT_Bayes	0.9668	21.1994	2.7268	6.9506	0.0144	0.0272
LDSVR	0.2584	4.9794	**2.1272**	2.8912	0.6146	0.7872
Specialized Algorithms						
SA_BFGS	0.2306	**0.5976**	3.9147	2.6711	0.6254	0.8089
LDL_LRR	0.3293	2.2569	4.1227	3.5177	0.5242	0.7115
TLRLDL	0.3368	7.9606	3.2362	3.8317	0.4264	0.5968
Algorithm Adaptation						
CubeMLP	0.2738	0.6900	3.9669	3.2122	0.5624	0.7513
CTEN	0.2432	0.6033	3.9542	2.8277	0.6071	0.7977
MMIM	0.2442	0.6076	3.9593	2.8548	0.6019	0.7946

5.2 Finetuning Transforms Foundation VLMs into State-of-the-Art Audience Reaction Predictors

Under full distribution evaluation (Tab. 5), finetuned models reduce Chebyshev distance by roughly 50% and nearly double cosine similarity. Intersection scores alos double, indicating substantially improved structural alignment with ground-truth distributions. Crucially, these gains extend to dominant reaction prediction. As shown in Tab. 6, finetuned VLMs models achieve MRR above 0.75 and Top-3 weighted F1 around 0.77. They consistently outperform all baselines across ranking and classification metrics.

Video2Reaction also shows promise as a scalable pretraining resource for cross-dataset and cross-taxonomy transfer. Fine-tuning Qwen2.5-VL on a mixture of Video2Reaction and 1% of VCE training data achieves 0.46 Top-3 accuracy on the VCE benchmark [21], outperforming training on 10% of in-domain data (0.35) (see Appendix D.2 for details). Although these gains remain far from fully supervised performance, they indicate meaningful cross-dataset transfer and suggest that Video2Reaction with automatic reaction annotation provides complementary supervision signals that can partially substitute expensive human-annotated data.

Table 6: Dominant Reaction Evaluation Benchmark Results. Best performance is shown in bold. Finetuned VLMs significantly outperforms all other baseline models.

Model Name	TPE ↓	MRR ↑	F1 Top 1 ↑	F1 Top 2 ↑	F1 Top 3 ↑
Foundation VLMs					
Gemini 2.5 Flash	0.3026	0.4378	0.2735	0.3332	0.3794
LLaVA-Next-Video-7B					
- *Temperature-scaled*	0.4103	0.1992	0.0143	0.1167	0.1374
- *LoRA Finetuned*	0.1388 (-66.2%)	**0.7833** (+293.2%)	0.6521 (+4460%)	**0.7374** (+531.8%)	0.7672 (+458.3%)
Qwen2.5-VL					
- *Temperature-scaled*	0.3923	0.3088	0.1958	0.2531	0.3037
- *LoRA Finetuned*	0.1516 (-61.3%)	0.7548 (+144.4%)	**0.6577** (+235.9%)	0.7291 (+188.1%)	**0.7725** (+154.4%)
Problem Transformation					
PT-Bayes	0.9661	0.1535	0.0001	0.0031	0.0741
LDSVR	0.1599	0.7054	0.5034	0.5865	0.5696
Specialized Algorithms					
SA-BFGS	0.1882	0.7163	0.5283	0.6075	0.6265
LDL-LRR	0.2496	0.6700	0.4965	0.5684	0.5803
TLRLDL	0.2759	0.5559	0.4516	0.4895	0.4858
Algorithm Adaptation					
CubeMLP	0.2509	0.5996	0.2376	0.4399	0.5587
CTEN	0.2081	0.6939	0.4826	0.5950	0.5109
MMIM	0.2141	0.6749	0.4503	0.5728	0.5775

5.3 Specialized LDL Methods Are Competitive on Distribution Metrics but Inferior on Dominant Reaction Prediction

Classical LDL methods remain competitive under full distribution evaluation. Several specialized approaches achieve strong divergence-based scores, confirming their effectiveness at directly optimizing distributional objectives. However, this advantage does not translate to dominant reaction prediction. While methods such as SA-BFGS and LDSVR attain moderate MRR (∼0.70), their Top-1 F1 scores consistently lag behind finetuned VLMs (∼0.65 vs. ∼0.53). The gap persists across Top-2 and Top-3 metrics. However, LDL methods are substantially more computationally efficient than finetuned VLMs at both training and inference time. Thus, further research on these approaches remain attractive for resource-constrained settings.

Qualitative analysis (Appendix D.3) reveals a consistent pattern: specialized LDL models capture overall distribution shape but underestimate the leading reaction probability, particularly in low-entropy (highly unimodal) samples.

5.4 Text Modality Provides Significant Performance Gain for both LDL and Foundation Models

To assess the contribution of each input modality, we conduct an ablation using two representative models, SA-BFGS and LLaVA-Next, trained on different combinations of visual, audio, and text (clip description) inputs. Table 7 shows that incorporating the text modality leads to substantial performance improvements for both models, while adding the audio modality to SA-BFGS provides only marginal gains over visual and text alone.

Table 7: Reaction prediction performance across methods and input modalities.

Method	Input Modality	Full Distribution		Dominant Reaction	
		Cheb ↓	KL ↓	MRR ↑	$F1_1$ ↑
SA-BFGS	Visual + Audio	0.314	0.907	0.552	0.306
	Visual + Text	0.234	**0.588**	0.709	0.515
	Visual + Audio + Text	**0.231**	0.598	**0.716**	**0.528**
LLaVA-Next	Visual	0.218	4.148	0.731	0.586
	Text	0.202	3.693	0.750	0.605
	Visual + Text	**0.188**	**3.177**	**0.783**	**0.652**

6 Conclusion

This paper introduces **Video2Reaction**, the first dataset and benchmark for predicting fine-grained induced audience reaction distributions towards movie content in the wild. This dataset and algorithms developed using it enables video creators to anticipate audience feedback prior to content release, potentially leading to improved content quality. Our benchmark is also supported by a scalable annotation pipeline that enables iterative updates to adapt to the non-stationary nature of audience engagement. We propose a comprehensive evaluation framework with two complementary axes, full distribution prediction and dominant reaction estimation, and conduct extensive experiments across four categories of algorithms. Our findings demonstrate that this task is challenging yet feasible: low-rank finetuning foundation VLMs can improve the base models' performance significantly on both distribution and classification benchmarks.

Limitations. First, Video2Reaction derives reaction representations solely from YouTube, which may not capture broader audience behavior. Future work will extend the dataset to additional platforms (e.g., TikTok and Bilibili) using our scalable annotation pipeline. Demographic analysis is also not possible due to platform privacy policies. Second, like many fine-grained classification benchmarks, Video2Reaction exhibits a long-tail distribution that remains underexplored. Addressing rare reactions is an important direction for improving prediction performance. Finally, while preliminary results suggest cross-dataset and cross-taxonomy transfer (Appendix D.2), a more systematic study is needed to understand how Video2Reaction can support foundation models that generalize across domains and taxonomies.

Acknowledgements The computational resources for this work are provided by the Unity Research Computing Platform, a multi-institutional cluster led by the University of Massachusetts and the University of Rhode Island. We also thank all the reviewers who provided quality review for our annotation pipeline.

References

1. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings (2020)
2. Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L.: Liris-accede: A video database for affective content analysis. IEEE Transactions on Affective Computing **6**(1), 43–55 (2015)
3. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
4. Chen, L., Davis, J.Q., Hanin, B., Bailis, P., Stoica, I., Zaharia, M.A., Zou, J.Y.: Are more llm calls all you need? towards the scaling properties of compound ai systems. Advances in Neural Information Processing Systems **37**, 45767–45790 (2024)
5. Choi, J., Yun, J., Jin, K., Kim, Y.: Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. In: EMNLP (2024)
6. Comanici, G., Bieber, E., Schaekermann, M., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025), https://arxiv.org/abs/2507.06261
7. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
8. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547 (2020)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
10. Geng, X.: Label Distribution Learning (Apr 2016). https://doi.org/10.48550/arXiv.1408.6027, http://arxiv.org/abs/1408.6027
11. Geng, X., Hou, P.: Pre-release prediction of crowd opinion on movies by label distribution learning. In: IJCAI. pp. 3511–3517. Association for the Advancement of Artificial Intelligence (2015)
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L.,

Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy,

R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783

13. Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4946–4958 (2021)
14. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units (2021), https://arxiv.org/abs/2106.07447
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
16. Hwang, E., Majumder, B.P., Tandon, N.: Aligning language models to user opinions. arXiv preprint arXiv:2305.14929 (2023)
17. Jia, X., Li, Z., Zheng, X., Li, W., Huang, S.J.: Label distribution learning with label correlations on local samples. IEEE Transactions on Knowledge and Data Engineering **33**(4), 1619–1631 (2019)
18. Jiang, H., Beeferman, D., Roy, B., Roy, D.: Communitylm: Probing partisan worldviews from language models. arXiv preprint arXiv:2209.07065 (2022)
19. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing **3**(1), 18–31 (2011)
20. Kou, Z., Wang, J., Jia, Y., Geng, X.: Exploiting multi-label correlation in label distribution learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence (2024)
21. Mazeika, M., Tang, E., Zou, A., Basart, S., Chan, J.S., Song, D., Forsyth, D., Steinhardt, J., Hendrycks, D.: How would the viewer feel? estimating wellbeing from video scenarios. Advances in Neural Information Processing Systems **35**, 18571–18585 (2022)
22. Meister, N., Guestrin, C., Hashimoto, T.: Benchmarking distributional alignment of large language models (2024), https://arxiv.org/abs/2411.05403
23. Muszynski, M., Tian, L., Lai, C., Moore, J.D., Kostoulas, T., Lombardo, P., Pun, T., Chanel, G.: Recognizing induced emotions of movie audiences from multimodal information. IEEE Transactions on Affective Computing **12**(1), 36–52 (2019)
24. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)

25. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)
26. Sun, H., Wang, H., Liu, J., Chen, Y.W., Lin, L.: Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3722–3729 (2022)
27. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/, accessed: March 1, 2026
28. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/, accessed: March 1, 2026
29. Tian, L., Muszynski, M., Lai, C., Moore, J.D., Kostoulas, T., Lombardo, P., Pun, T., Chanel, G.: Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 28–35. IEEE (2017)
30. Trad, F., Chehab, A.: To ensemble or not: Assessing majority voting strategies for phishing detection with large language models. In: International Conference on Intelligent Systems and Pattern Recognition. pp. 158–173. Springer (2024)
31. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2023)
32. Wen, C., Zhang, X., Yao, X., Yang, J.: Ordinal label distribution learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23481–23491 (2023)
33. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision (2020)
34. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation (2022). https://doi.org/10.48550/ARXIV.2211.06687, https://arxiv.org/abs/2211.06687
35. Xu, C., Liu, L., Jin, L., Du, G., Guo, Z., Zhao, Y., Huang, X., Li, R., et al.: Infer induced sentiment of comment response to video: A new task, dataset and baseline. Advances in Neural Information Processing Systems **37**, 103737–103750 (2024)
36. Yang, J., Sun, M., Sun, X.: Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (Feb 2017). https://doi.org/10.1609/aaai.v31i1.10485, https://ojs.aaai.org/index.php/AAAI/article/view/10485
37. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246 (2018)
38. 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/, accessed: March 1, 2026
39. Zhang, Z., Wang, L., Yang, J.: Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18888–18897 (2023)

40. Zlatintsi, A., Koutras, P., Evangelopoulos, G., Malandrakis, N., Efthymiou, N., Pastra, K., Potamianos, A., Maragos, P.: Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. EURASIP Journal on Image and Video Processing **2017**, 1–24 (2017)