FinEngage: Benchmarking Multimodal Language Models on Content-Conditioned Human Feedback Inference

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have advanced rapidly in video understanding, yet existing benchmarks mainly test visual recognition, factual question answering, and event reasoning. We study a different problem: whether models can infer real user engagement from domain-specific persuasive videos. We introduce FINENGAGE, a benchmark for content-conditioned human feedback inference in financial recommendation short videos. It contains 10,505 Chinese stock recommendation videos with titles, transcripts, stock entities, visual inputs, and platform engagement signals. From likes, comments, favorites, and shares, we construct continuous log-engagement targets and five-level ordinal engagement labels. Under a leakage-free protocol, we evaluate representative MLLMs using classification and ordinal agreement metrics. Results show that current models capture only weak engagement signals: the best model achieves 0.290 Accuracy and 0.274 Spearman correlation. Further analysis shows that models can identify persuasive cues such as profit claims, chart evidence, emotional language, and calls to action, but these cues do not reliably translate into real engagement. FINENGAGE reveals content-conditioned human feedback inference as a distinct challenge beyond conventional video understanding.

1 Introduction

Multimodal Large Language Models (MLLMs) perform well on video question answering, summarization, event recognition, and temporal reasoning benchmarks (Li et al., 2024b; Fu et al., 2025; Wang et al., 2023). However, understanding videos on real-world social platforms requires more than identifying what appears in a video; it also requires reasoning about how audiences will respond to their content. A model may correctly recognize entities, events, and visual evidence yet still fail to predict whether the video will trigger likes, comments, favorites, or shares. Somewhat surprisingly, existing video understanding evaluations mainly assess content-side semantics, leaving audience response inference from multimodal content underexplored.

A related but distinct line of work studies social media popularity prediction. Existing approaches typically combine content features with platform features, user features, posting time, early engagement, or social diffusion signals to predict eventual popularity (Szabo and Huberman, 2010; Bandari et al., 2012; Tatar et al., 2014; Khosla et al., 2014). These features are effective for practical forecasting, but they make the task closer to platform diffusion, user behavior, or recommender systems. They are therefore more suitable for popularity extrapolation than for evaluating whether an MLLM understands content-side signals associated with audience response.

In this paper, we study a more constrained problem: **content-only engagement inference**. In plain terms, the task asks whether a model can predict aggregate viewer engagement (likes, comments, favorites, and shares) using only pre-publication content-side signals, with no access to *post-publication interaction counts, account history, user profiles, social network structure, or diffusion paths*. We do not claim engagement is fully determined by content. Instead, we use this leakage-free setting to establish a lower bound on content-conditioned predictability and to test whether current MLLMs capture predictable structure in content that is associated with real audience engagement.

We ground this task in the domain of Chinese financial recommendation short videos. Financial social media has emerged as a crucial channel for capturing investor attention, gauging market sentiment, and disseminating investment advice (Da et al., 2011; Kakhbod et al., 2023; Cookson et al., 2025). Unlike entertainment content, where engagement is driven primarily by aesthetic preference, interac-

tion with stock-recommendation videos hinges on whether viewers find an investment claim credible and actionable. This dynamic tightly couples audience response to content semantics. Consequently, models must visually ground charts and trading screenshots, parse domain-specific language, and evaluate persuasive claims under real financial stakes. Because this engagement actively shapes investment attention and either amplifies or corrects prevailing market narratives, these videos provide a uniquely demanding and highly interpretable testbed for content-only engagement inference.

We construct FINENGAGE, a multimodal benchmark of Chinese financial recommendation short videos covering the A-share market. The benchmark provides content-side signals (titles, ASR transcripts or subtitles, stock entities, posting time, and video frames) as model inputs, and platform engagement signals (likes, comments, favorites, shares) as supervision targets. Following prior work on popularity prediction, we treat these engagement signals as proxies for aggregate audience response (Bandari et al., 2012; Tatar et al., 2014; Khosla et al., 2014). We release raw interaction counts, log-transformed continuous targets, and five-level ordinal engagement labels constructed by equal-frequency binning of log-engagement. To prevent label leakage, model inputs additionally exclude gold labels, rank information, popularity scores, and platform feedback metadata. Models must predict from content-side signals only.

We formulate FINENGAGE evaluation as an ordinal prediction task. Given a video and its content-side inputs, a model predicts one of five engagement tiers from A to E, where A denotes the highest tier and E the lowest. Because the tiers are equal-frequency, a uniform-random baseline scores 0.20 accuracy. Since labels are ordered, we evaluate exact classification with Accuracy and Macro-F1, and ordinal consistency with Quadratic Weighted Kappa, Rank MAE, 1-off Accuracy, and Spearman correlation. Such ordinal and rank-consistency metrics are widely used to distinguish adjacent errors from long-distance errors (Gutiérrez et al., 2016; de la Torre et al., 2018; Vargas et al., 2023).

Under a leakage-free evaluation protocol, current MLLMs exhibit only weak alignment with real-world engagement. The top-performing model achieves an accuracy of just 0.290 (against a random baseline of 0.20) and a Spearman correlation of 0.274. This indicates that while MLLMs capture marginal content-conditioned engagement signals, they remain far from providing reliable predictions. Further analysis reveals that although these models successfully identify persuasive cues, such as return promises, chart evidence, emotional expressions, controversial topics, and calls to action, they struggle to map these features to realized engagement metrics. Ultimately, our results suggest that audience response inference is not a natural byproduct of general video understanding, thereby motivating the need for dedicated evaluation benchmarks. Inferring engagement requires synthesizing multimodal content, domain-specific persuasion tactics, and aggregate audience behavior under strict, leakage-free conditions; a complex mapping that current MLLMs have yet to master. The framework of the paper is shown in the Figure 1

Our contributions are as follows:

- We propose and formalize **content-only engagement inference**, a leakage-free evaluation setting that tests whether MLLMs can infer aggregate viewer engagement from domain-specific persuasive videos without post-publication information.
- We construct FINENGAGE, a multimodal benchmark of **10,505** Chinese financial recommendation short videos covering the A-share market, with raw interaction counts, log-transformed continuous targets, and five-level ordinal engagement labels.
- We design a strict leakage-free evaluation protocol and evaluate representative MLLMs using classification and ordinal-consistency metrics. The best leakage-free model reaches only 0.290 accuracy (random = 0.20); error analysis reveals a capability gap in which models detect persuasive cues but fail to predict whether those cues translate into aggregate audience engagement.

2 Related Work

Video Understanding and Contextualized Multimodal Reasoning. The proliferation of video–language models has driven the development of benchmarks evaluating *multimodal reasoning over videos*. Tasks like VideoQA (Xiao et al., 2025), TVQA (Lei et al., 2019), and ToolQA (Zhuang et al., 2023) test the integration of visual, auditory, and textual cues to answer content-grounded queries. Broader benchmarks, such as MVBench (Li et al., 2024b) and Video-MME (Fu

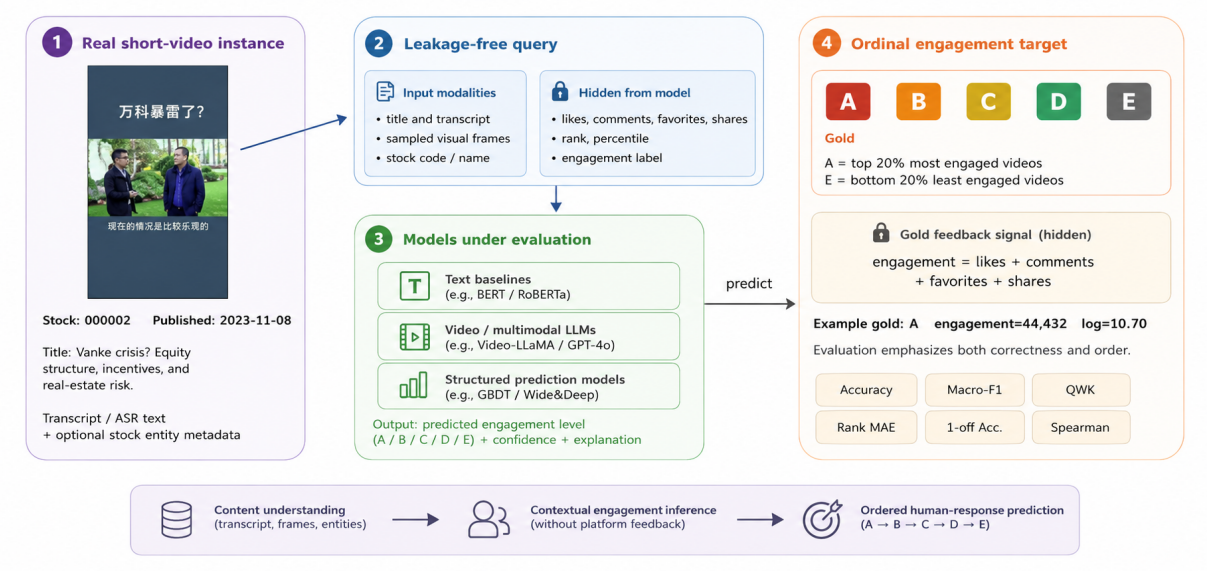

Figure 1: The framework of the paper

et al., 2025), employ multi-task frameworks to assess action recognition, causal reasoning, and instruction following. Recently, long-video benchmarks (e.g., LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025), MLVU (Zhou et al., 2025), and Video-XL (Shu et al., 2024)) have extended these efforts to target long-context coherence and cross-segment reasoning. However, these benchmarks predominantly rely on intra-video information, rarely demanding reasoning over external, time-varying world states.

Recent work has begun exploring this *external context* in short-video understanding. For instance, VideoConviction (Galarnyk et al., 2025) shows that verifying financial video claims necessitates external market data, while SMTPD (Xu et al., 2025) integrates contextual signals, such as publication time and social feedback, to predict video popularity. These studies highlight the critical role of non-visual information in modeling audience response and validating real-world claims.

Concurrently, research has incorporated *external knowledge sources* into video QA and reasoning. Representative approaches (Mondal et al., 2024; Jang et al., 2017; Gupta and Gupta, 2022) augment video representations with knowledge graphs, timestamps, or retrieved textual evidence. Yet, these methods typically treat external information as static background knowledge or auxiliary features, rather than as a dynamically evolving world state that requires explicit, joint reasoning alongside the video content.

Diverging from prior work, our focus shifts from factual retrieval and predefined QA toward evaluating whether models can interpret the *reasoning underlying viewpoint expression* in real-world videos. The limitations of existing static benchmarks are especially pronounced in stock recommendation videos, where viewpoints are inextricably linked to time-sensitive financial conditions and rapidly fluctuating market environments.

Video Popularity Prediction Understanding and forecasting the popularity of online videos has attracted broad attention due to its importance in recommendation, advertising, and content delivery systems. Deep learning advanced this domain by harnessing multimodal signals. While early models (Zhao and et al., 2017) applied recurrent neural networks strictly to visual cues, subsequent work (Li et al., 2020) introduced interpretable, DNN-based systems to predict YouTube popularity. More recently, benchmarks like SMTPD (Xu et al., 2025) and AMPS (Chen et al., 2024) have targeted short-form videos, fusing video, audio, and text to forecast engagement. Ultimately, existing literature focuses on predicting popularity directly from content features or interaction dynamics. In contrast, our objective is not popularity forecasting. Instead, we leverage engagement signals as a proxy for human response, evaluating whether models can discern how varying viewpoint expressions drive audience behavior under comparable external conditions.

Early works on video understanding and video popularity prediction are discussed in Appendix A.

3 Method

3.1 Task Definition

Our work focuses on whether models can infer real users' engagement responses from financial recommendation videos. We formalize this task as *content-conditioned human feedback inference*: given the content-side and context-side inputs of a financial recommendation video, the model predicts an ordinal engagement level derived from real user feedback.

We represent each sample as:

$$x_i = \{v_i, t_i, a_i, m_i, c_i\}, \tag{1}$$

where v_i denotes the video content, t_i denotes the video title, a_i denotes the transcript or ASR text, m_i denotes stock entity metadata such as stock code and stock name, and c_i denotes market context available at or before the video publication time.

The observed user feedback signal is defined as the aggregate engagement count:

$$e_i = L_i + C_i + F_i + S_i, \tag{2}$$

where L_i, C_i, F_i, and S_i denote likes, comments, favorites, and shares, respectively. In our dataset, this aggregate value is provided by the platform as the social attribute field. Following prior work on social media popularity prediction, we use observable engagement signals as a behavioral proxy for aggregate user response (Szabó and Huberman, 2008; Bandari et al., 2012; Tatar et al., 2014; Khosla et al., 2014).

The main prediction target is an ordinal engagement label:

$$y_i \in \{A, B, C, D, E\}, \tag{3}$$

where higher levels (in alphabetical order) indicate stronger aggregate user engagement. A model f_θ predicts:

$$f_\theta(x_i) \rightarrow \hat{y}_i, \tag{4}$$

where $\hat{y}_i$ is the predicted engagement level. A model may additionally output a continuous engagement score or a natural language explanation, but the main benchmark evaluates the predicted ordinal label.

We note that our task differs from recommender-system popularity prediction as our goal is not to optimize platform ranking or click-through rate.

3.2 Dataset Construction

FINENGAGE is built from real-world Chinese financial recommendation short videos, The videos used in our study were collected from Douyin, the Chinese counterpart of TikTok. Specifically, we gathered stock-recommendation videos related to the Chinese A-share market from the platform, covering content associated with individual stock recommendations, investment advice, stock analysis, and market commentary. After obtaining the initial video collection, we further conducted a second-stage screening using both ChatGPT and Gemini to improve the relevance and quality of the dataset. Only videos that were identified by both models as relevant to A-share stock recommendations were retained and included in our final raw dataset.

To mitigate noise and off-topic content within the raw collection, we apply a conservative relevance filtering procedure. A video is retained only if it is independently classified as stock-related by two model-based annotators. This filtering step eliminates non-financial and tangentially related content, thereby ensuring the semantic consistency of the benchmark. Following filtering and preprocessing, the final dataset comprises 10,505 video records across 10,422 unique video IDs, with publication dates spanning from January 11, 2018, to January 23, 2024. For each retained sample, we construct three benchmark fields:

- `engagement_raw`: the raw aggregate engagement value, equal to the platform social attribute;
- `log_engagement`: the log-transformed engagement value, $\log(1 + \texttt{engagement_raw})$;
- `engagement_level`: the five-level ordinal engagement label A–E.

Dataset statistics can be found in Appendix B.

3.3 Engagement Label Construction

Raw engagement counts in social media data are highly skewed, a property repeatedly observed in online content popularity and social media popularity prediction (Szabó and Huberman, 2008; Tatar et al., 2014; Khosla et al., 2014). To reduce the influence of extreme values, we first construct a continuous engagement target:

$$y_i^{reg} = \log(1 + e_i). \tag{5}$$

Log transformation is commonly used when modeling long-tailed view or engagement counts.; for

Level	Rank Percentile	Meaning
A	0–20% from top	Very high engagement
B	20–40% from top	High engagement
C	40–60% from top	Medium engagement
D	60–80% from top	Low engagement
E	80–100% from top	Very low engagement

Table 1: Def. of the five ordinal engagement levels.

example, log-normalized view counts are used to define image popularity (Khosla et al., 2014).

For the main benchmark task, we convert engagement counts into five ordinal levels:

$$y_i^{ord} \in \{A, B, C, D, E\}. \tag{6}$$

Specifically, all videos are sorted by their global engagement counts and divided into five equal-frequency groups. Level A denotes the top 20% most engaged videos, while level E denotes the bottom 20%, cf. Table 1. This design follows prior popularity prediction work formulating popularity estimation as range classification rather than exact count prediction (Tatar et al., 2014).

3.4 Input Protocol

FINENGAGE is designed as a video-centered multimodal benchmark. Each instance contains the raw video content, textual metadata, stock entity information, and optional market context. These fields jointly define the available input sources for model evaluation. Different MLLMs support different video input interfaces. Thus, we define each benchmark sample as a video-level instance and adapt the input representation to each model's supported interface. Details can be found in Appendix C.

In addition to the standard text and video inputs, FINENGAGE defines an optional market-aware setting.Details of the market-context construction are provided in Appendix D. Details of the evaluation setting can be found in Appendix E.

4 Experiments

Evaluated MLLMs and Metrics. We evaluate five representative multimodal large language models, including Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), LLaVA-OneVision variants (Li et al., 2024a), and Gemini 2.5 Flash (Comanici et al., 2025). Accuracy and Macro-F1 measure exact classification performance, while Quadratic Weighted Kappa (QWK), Rank MAE, 1-off Accuracy, and Spearman correlation evaluate whether predictions preserve the ordered structure of the A–E engagement labels. Details of the evaluation metrics can be found in Appendix F.

4.1 Main Results

Table 2 presents the main results on ordinal engagement-level prediction. Overall, MLLMs show limited agreement with real user engagement feedback. Gemini 2.5 Flash achieves the highest Accuracy 0.290 and Spearman correlation with 0.274, suggesting relatively stronger alignment with the overall engagement ordering. Qwen3-VL obtains the best Macro-F1 with 0.212, indicating more balanced performance across the five engagement levels. LLaVA-Video achieves the lowest Rank MAE with 1.307 and the highest 1-off Accuracy 0.620, suggesting that although its exact predictions are often incorrect, they tend to remain close to the gold ordinal level. OneVision performs substantially worse than the other models, with the lowest QWK for 0.004 and Spearman correlation 0.104. Further inspection shows that this is mainly caused by prediction collapse: the model predicts most evaluated samples as high-engagement level A, while many gold labels belong to lower engagement levels. This mismatch reverses the ordinal relationship for many examples and leads to poor ranking consistency.

These findings indicate that even advanced MLLMs extract only marginal engagement-related signals, suggesting that predicting aggregate user feedback necessitates capabilities extending well beyond standard video understanding.

Ordinal Error Analysis. Since the A–E labels are ordinal, exact accuracy alone does not fully capture the severity of model errors. For example, predicting an A-level video as B is less severe than predicting it as E. We therefore compute the ordinal error distance for each sample:

$$d_i = |\mathrm{rank}(y_i) - \mathrm{rank}(\hat{y}_i)| \tag{7}$$

where y_i and $\hat{y}_i$ denote the gold and predicted engagement levels, respectively, and $\mathrm{rank}(\cdot)$ maps A–E to ordered numerical indices.

Figure 2 shows the distribution of ordinal prediction errors across models. The results indicate that model errors are not entirely random: several models make a substantial portion of errors within a small ordinal distance, which is consistent with their relatively high 1-off Accuracy. However, severe errors still occur, suggesting that current MLLMs lack reliable fine-grained calibration over engagement levels.

Model	Acc.	Macro-F1	QWK	Rank MAE	1-off Acc.	Spearman
Qwen2.5-VL (Bai et al., 2025b)	0.235	0.189	0.104	1.438	0.571	0.141
Qwen3-VL (Bai et al., 2025a)	0.250	**0.212**	0.167	1.377	0.600	0.228
LLaVA-Video (Zhang et al., 2024)	0.219	0.149	0.065	**1.307**	**0.620**	0.116
LLaVA-OneVision (Li et al., 2024a)	0.212	0.169	0.004	1.910	0.411	0.104
Gemini 2.5 Flash (Comanici et al., 2025)	**0.290**	0.200	**0.167**	1.361	0.608	**0.274**

Table 2: Main results on ordinal engagement-level prediction. Metrics are computed on valid parsed outputs after status filtering.

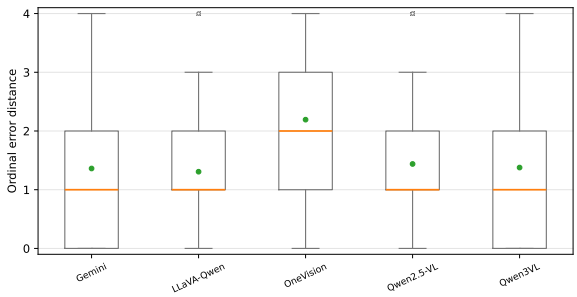

Figure 2: Distribution of ordinal prediction errors across models. The error distance is defined as $|\mathrm{rank}(y) - \mathrm{rank}(\hat{y})|$. Orange lines and green dots denote median and mean values.

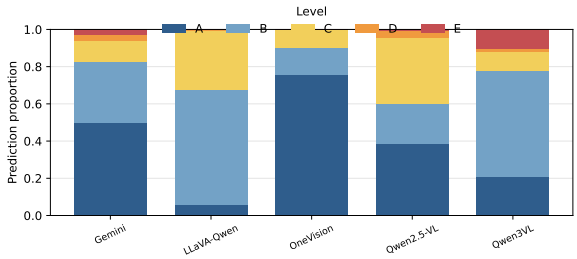

Figure 3: Distribution of predicted engagement levels across models. The figure illustrates whether models are biased toward specific engagement levels.

Prediction Bias Analysis. To better understand model failures, we further analyze the distribution of predicted engagement levels. Figure 3 shows that different models exhibit clear label biases. Some models over-predict high-engagement levels, while others collapse toward middle levels. This suggests that performance is affected not only by content understanding, but also by calibration over the ordinal label space. Figure 4 presents the normalized confusion matrix of Gemini. The model does not make errors uniformly across the label space; instead, many errors occur between nearby engagement levels. This further supports our observation that current MLLMs can capture weak coarse-grained engagement signals but remain unreliable for fine-grained ordinal prediction.

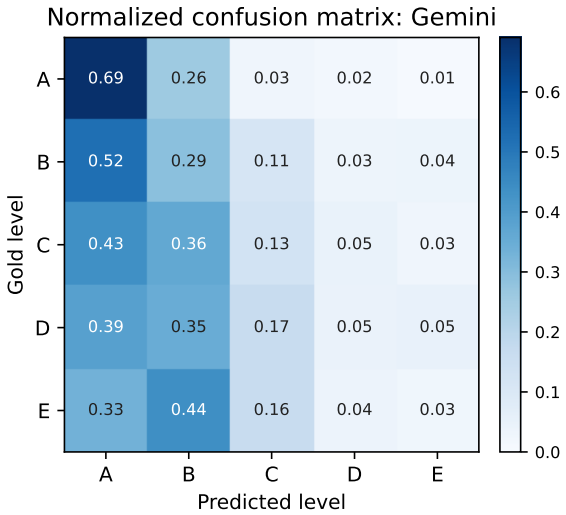

Figure 4: Normalized confusion matrix of Gemini on engagement-level prediction. Rows denote gold labels and columns denote predicted labels.

Collectively, the results and error analyses yield three key observations. First, current MLLMs struggle to reliably infer user engagement in financial videos. Second, errors possess underlying structure: high 1-off Accuracy indicates models can identify broad engagement regions. Third, models exhibit prediction bias and poor calibration, particularly when differentiating adjacent engagement levels.

4.2 Ablation study

To isolate how much engagement signal can be inferred from textual content alone, we evaluate several prompted LLMs using only the video title, transcript text, and stock entity metadata. This setting removes visual frames from the input while keeping the same ordinal engagement label space. The goal is not to train a task-specific classifier, but to test whether instruction-tuned language models can infer realized user engagement from the linguistic and entity-level cues that appear in financial recommendation videos.

Table 3 reports the text-only LLM results.

Overall, the models achieve only modest performance. DeepSeek-Chat obtains the highest accuracy (0.202), one-off accuracy (0.635), and Spearman correlation (0.152), while Qwen-Plus achieves the highest Macro-F1 (0.193) and QWK (0.106). However, all models remain weak in ordinal agreement: QWK is close to zero for every model, and exact accuracy ranges only from 0.184 to 0.202. These results suggest that prompted LLMs can identify some plausible engagement-related textual cues, such as salient stock entities, urgent language, profit-oriented claims, and calls to action, but they are not well calibrated to the actual distribution of platform feedback.

The ablation provides two takeaways. First, textual information contains some signal for engagement inference, but this signal is difficult for prompted LLMs to map into calibrated ordinal labels. The relatively higher one-off accuracy compared with exact accuracy indicates that models sometimes place a video near the correct engagement region, yet still struggle to distinguish adjacent levels. Second, the weak QWK and Spearman scores show that textual reasoning alone does not recover the global ranking structure of realized engagement. Thus, text-only prompting is useful as a diagnostic setting, but it does not solve the core FINENGAGE task.

4.3 Case Study

To analyze model behavior beyond aggregate metrics, we conduct a qualitative case study using Gemini 2.5 Flash, the top-performing leakage-free model by Spearman correlation. We select four representative examples: correct high- and low-engagement predictions, a near-miss, and a severe error. This allows us to examine the content cues the model utilizes and why they often diverge from actual aggregate engagement.

Table 4 summarizes these cases. The correct predictions demonstrate the model's ability to identify broad engagement signals. For the high-engagement video, it highlights controversy, humor, and a shareable narrative regarding a "demon stock." Conversely, for the low-engagement video, it correctly flags dry reporting, a weak narrative structure, and absent investment opinions. These successes suggest MLLMs can recognize features typically associated with social-media engagement. However, the failure cases reveal limitations in this cue-based reasoning. In the near-miss, the model overestimates a personal investment update by overweighting transparency, portfolio screenshots, and aspirational goals. In the severe error, it predicts high engagement for a lecture-style explanation based on timely themes, authoritative presentation, and technical analysis. In both instances, the model conflates plausible persuasive cues with realized audience response.

In Appendix G, we present two additional failure modes. Overall, this case study confirms that persuasive cues lack a stable correspondence with aggregate engagement. Predicting actual user feedback demands more than recognizing emotionally salient or authoritative content; it necessitates modeling audience preferences, platform dynamics, and contextual salience.

5 Discussion

Unlike conventional video understanding tasks, FINENGAGE does not merely require models to identify the stocks, charts, people, or events appearing in a video. Instead, it further requires models to judge whether such content expressions will trigger user engagement on real platforms.The results above lead to three observations.

Prompted large language models capture only weak text-side engagement signals. The text-only LLM results show that models can extract some engagement cues from titles, transcripts, and stock entities, but their ordinal agreement remains limited. This suggests that instruction-following models can identify plausible persuasive cues in financial text, but they are not well calibrated to the realized engagement distribution.

Adding multimodal input does not by itself solve engagement inference. The pilot MLLM results are also modest despite access to visual frames in addition to textual content. These patterns indicate that generic video understanding is not sufficient for calibrated human-feedback inference: models can recognize salient financial content, but this recognition does not reliably translate into the engagement level that the platform audience actually produces.

The main failure is ordinal calibration, not only semantic understanding. Across models, one-off accuracy is consistently much higher than exact accuracy, which suggests that predictions are not entirely random. However, QWK and Spearman remain low. This indicates that models may capture coarse high-versus-low engagement cues while failing to preserve the global ranking structure of real

Model	Acc.	Macro-F1	QWK	MAE	1-off	Spearman
GPT-4o-mini	0.193	0.135	0.074	1.315	0.599	0.116
DeepSeek-Chat	0.202	0.137	0.100	**1.241**	**0.635**	**0.152**
Qwen-Plus	0.195	**0.193**	**0.106**	1.433	0.577	0.140
Qwen-Turbo	0.184	0.139	0.030	1.449	0.540	0.050
Gemini-2.5-Flash-Lite	**0.209**	0.147	0.080	1.221	0.625	0.089

Table 3: Text-only LLM ablation results. Models receive only title, transcript/ASR text, and stock entity metadata, without video frames or any feedback-derived variables.

Case Type	**Gold / Pred.**	**Video Topic**	**Model-Identified Cues**	**Interpretation**
Correct high-engagement	A / A	A controversial "demon stock" with repeated name changes and speculative history.	Controversy, reveal-style narrative, humor, memes, and shareable social-media presentation.	The model identifies cues aligned with actual high engagement.
Correct low-engagement	E / E	A factual annual-report presentation for a listed company.	Dry numerical reporting, weak narrative, lack of explicit investment opinion, and limited audience scope.	The model recognizes the lack of engagement-oriented affordances.
Near miss	C / B	A personal investment update showing daily loss while holding a stock toward a portfolio goal.	Transparency, portfolio screenshots, concrete profit/loss numbers, personal storytelling, and an aspirational goal.	The model detects plausible engagement cues but overestimates their effect.
Severe error	E / B	A lecture-style video introducing several types of stocks during an epidemic-related market context.	Timely market theme, authoritative speaking style, technical analysis, and concrete stock examples.	The model mistakes plausible persuasive cues for realized engagement.

Table 4: Qualitative case study using Gemini 2.5 Flash. The examples show that the model can identify plausible persuasive and presentation cues, but these cues do not always translate into accurate prediction of real aggregate engagement.

feedback. For an A–E label space, this calibration problem is central: a useful model must not only decide whether a video looks appealing, but also place that appeal at the correct relative level among other videos in the same behavioral distribution.

Overall, these findings clarify the role of FINENGAGE. The benchmark does not test whether a model can describe a financial video, identify a stock, or recognize a chart. It tests whether those content signals can be mapped to realized aggregate user behavior under a leakage-free protocol. Current results suggest that this mapping remains weakly learned.

6 Conclusion

This paper introduced FINENGAGE, a multimodal benchmark for content-conditioned human feedback inference in financial recommendation short videos. Using real platform interaction signals, we construct continuous engagement scores and five-level ordinal engagement labels, and evaluate whether current large language and multimodal models can infer realized user engagement from video, text, and optional contextual inputs. Our results show that existing models capture only weak engagement-related signals: they can often recognize persuasive cues such as profit claims, emotional language, chart-based evidence, and calls to action, but struggle to calibrate these cues against actual user feedback, leading to low ordinal agreement and prediction bias. These findings suggest that human engagement inference is not a natural byproduct of general video understanding, but a distinct capability requiring dedicated evaluation. FINENGAGE provides an initial testbed for this problem, while future work should extend it across languages, platforms, markets, and video genres, and further study how non-content factors such as creator influence, recommendation exposure, posting time, and market events shape observed engagement.

Limitations

FINENGAGE is currently limited to Chinese financial recommendation short videos, and observed engagement is influenced by platform algorithms, creator popularity, posting time, and market events. The five-level ordinal labels also abstract away finer-grained variation in engagement intensity and interaction type.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, and 1 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6.

Xiaoyu Chen, Tao Li, and 1 others. 2024. Amps: Multimodal short-form video popularity prediction. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Justin Cheng, Lada A. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. 2014. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

J Anthony Cookson, Runjing Lu, William Mullins, and Marina Niessner. 2025. Market signals from social media. *Available at SSRN*.

Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653.

Zhi Da, Joseph Engelberg, and Pengjie Gao. 2011. In search of attention. *The journal of finance*, 66(5):1461–1499.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Michael Galarnyk, Veer Kejriwal, Agam Shah, Yash Bhardwaj, Nicholas Meyer, Anand Krishnan, and Sudheer Chava. 2025. Videoconviction: A multimodal benchmark for human conviction and stock market recommendations. *Preprint*, arXiv:2507.08104.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. *Preprint*, arXiv:1706.04261.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. *Preprint*, arXiv:2110.07058.

Pranay Gupta and Manish Gupta. 2022. Newskvqa: Knowledge-aware news video question answering. *Preprint*, arXiv:2202.04015.

Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro, and César Hervás-Martínez. 2016. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. *Preprint*, arXiv:2007.10937.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Preprint*, arXiv:1704.04497.

Xia Jin and et al. 2016. Predicting emotion distributions in social media via visual and textual features. In *Proceedings of the 24th ACM International Conference on Multimedia*.

Ali Kakhbod, Seyed Mohammad Kazempour, Dmitry Livdan, and Norman Schuerhoff. 2023. Finfluencers. *Available at SSRN*, 4428232.

Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 867–876.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering. *Preprint*, arXiv:1809.01696.

Bo Li, Yuanhan Zhang, Dong Guo, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jing Li and et al. 2015. Learning from emotions in online social networks: A dual sentiment hawkes process model. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Wei Li, Jing Zhang, and 1 others. 2020. Precwd: A deployable framework for video popularity prediction and explanation. *IEEE Transactions on Multimedia*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. 2016. Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. *Preprint*, arXiv:2401.12863.

Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. 2017. A hawkes process model for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. *Preprint*, arXiv:2104.00990.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *Preprint*, arXiv:2409.14485.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. *Preprint*, arXiv:1604.01753.

Gábor Szabó and Bernardo A. Huberman. 2008. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88.

Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88.

Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):8.

Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. 2023. Soft labelling based on triangular distributions for ordinal classification. *Information Fusion*, 93:258–267.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. *Preprint*, arXiv:2303.16727.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2025. Lvbench: An extreme long video understanding benchmark. *Preprint*, arXiv:2406.08035.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Preprint*, arXiv:2407.15754.

Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, and Angela Yao. 2025. Videoqa in the era of llms: An empirical study. *Preprint*, arXiv:2408.04223.

Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. 2025. Smtpd: A new benchmark for temporal prediction of social media popularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18847–18857.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Kai Zhao and et al. 2017. Using recurrent neural networks for online video popularity prediction. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 210–215.

Wei Zhao, Tong Xu, and Xiaofeng Wu. 2022. Popularity prediction over arbitrary time horizons. *Proceedings of the VLDB Endowment*, 15(12):3290–3302.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. *Preprint*, arXiv:2406.04264.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Preprint*, arXiv:2306.13304.

A Related Work

Video Understanding and Contextualized Multimodal Reasoning. Early research on video understanding has largely focused on recognizing events and actions from visual sequences. Representative datasets such as ActivityNet (Heilbron et al., 2015), Something-Something (Goyal et al., 2017), and Charades (Sigurdsson et al., 2016) evaluate models on action classification and human–object interaction recognition, emphasizing whether models can identify what actions occur and when they happen. These benchmarks have played a foundational role in advancing video representation learning, but they primarily treat understanding as recovering predefined action categories or event labels.

Subsequent work has moved toward more structured representations of video content. VidSitu (Sadhu et al., 2021) extends action recognition by introducing event-centric semantic role annotations, enabling models to reason about agents, objects, and their roles within events. At a larger narrative scale, MovieNet (Huang et al., 2020) evaluates a model's ability to understand plot development, character relationships, and story progression across entire movies. Similarly, Ego4D (Grauman et al., 2022) emphasizes goal-oriented reasoning, long-horizon activity understanding, and anticipation in continuous egocentric videos. Despite these advances, these benchmarks still assume that video semantics can be inferred primarily from information internal to the video itself.

Video Popularity Prediction Early studies mainly focused on exploiting early view patterns to predict long-term popularity. For instance, some works modeled view dynamics from early access signals and external shocks (Szabó and Huberman, 2008; Crane and Sornette, 2008). Later works incorporated temporal and structural patterns in social diffusion processes (Cheng et al., 2014; Mishra et al., 2016), showing that the more observations available, the higher the predictability of viral cascades. With the rise of social platforms, several models have adopted point processes to capture self-exciting dynamics. A Hawkes process is introduced to distinguish exogenous promotion from endogenous user reactions (Rizoiu et al., 2017), while subsequent studies extended this framework to arbitrary horizons (Zhao et al., 2022). Recent variants also integrated emotional or textual cues to model content virality (Li and et al., 2015; Jin and et al., 2016).

B Dataset Statistics

Table 5 presents the statistics of the FinEngage dataset.

Statistic	Value
Total records	10,505
Unique video IDs	10,422
Time range	2018-01-11 to 2024-01-23
Engagement signal	Likes + Comments + Favorites + Shares
Engagement levels	A/B/C/D/E
Videos per level	2,101 each

Table 5: Basic statistics of FinEngage. Engagement levels are constructed from global engagement quintiles.

C Input Protocol

Different MLLMs support different video input interfaces. Thus, we define each benchmark sample as a video-level instance and adapt the input representation to each model's supported interface. For models that support direct video input, we provide the raw video through the model's native video interface. For models that do not support native video input, we follow common video-MLLM evaluation practice and convert videos into sampled frame sequences (Maaz et al., 2024; Li et al., 2024b; Fu et al., 2025).

Textual metadata includes the title, transcript or ASR text, stock code, and stock name. Market context is treated as optional external context associated with the stock and publication time. Whether textual metadata or market context is included depends on the specific experimental setting. We further analyze these input sources through ablation experiments.

To prevent label leakage, inputs never include raw engagement counts, gold engagement levels, rank percentiles, popularity scores, or any platform feedback metadata. Since the gold labels are directly derived from user engagement, exposing any such information would invalidate the evaluation. Models are required to infer engagement only from content-side and context-side signals available before evaluation.

D Optional Market Context Construction

Financial recommendation videos are often interpreted in relation to the contemporaneous performance of the featured stock and the broader market. For this reason, FinEngage includes an optional market-aware evaluation setting in addition to the

standard text and video settings. In this setting, each video is linked to the stock mentioned in the video, and the model is provided with market information that was available at or before the video publication time.

For each video, we collect recent stock-level indicators such as price movements, trading volume, percentage change, and other available trading statistics around the publication date. We also include broader market index information when available. The resulting market context is serialized into a textual or structured representation and appended to the model input only in the market-aware setting. The standard FINENGAGE setting does not include these variables.

To avoid look-ahead bias, we restrict all market variables to information that would have been publicly available no later than the video's publication timestamp. In particular, the model is not allowed to access future stock returns, future volume changes, post-publication market movements, realized engagement counts, engagement levels, rank percentiles, or any other feedback-derived variables. This constraint ensures that market-aware evaluation tests whether contemporaneous financial context provides complementary signal, rather than allowing the model to infer the label from future or feedback-side information.

This optional setting is intended as a diagnostic extension rather than the main benchmark protocol. It allows us to test whether external financial context helps models distinguish videos that are similar in content but released under different market conditions. If space is limited, all implementation details and market-aware results can be reported in the appendix, while the main paper focuses on the standard leakage-free text and multimodal settings.

E Evaluation Setting

FINENGAGE is model-agnostic: any model capable of producing an engagement-level prediction can be evaluated using our standardized label space and metrics. Our work focuses on instruction-following multimodal language models, as they represent the predominant paradigm for deploying general-purpose multimodal systems. For each benchmark instance, the model processes the inputs defined by the selected setting and predicts one engagement level from $\{A, B, C, D, E\}$.

To ensure rigorous comparability, all input configurations are evaluated across an identical set of labeled benchmark instances. The precise input formatting is tailored to the specific model's interface: video-native architectures process raw video directly, whereas frame-based models operate on uniformly sampled frames. Furthermore, supplementary textual metadata and serialized market context are conditionally appended based on the designated experimental setup.

F Evaluation Metrics

The main task is five-class ordinal engagement prediction. We report both standard classification metrics and ordinal agreement metrics.

For classification correctness, we report Accuracy and Macro-F1. Macro-F1 gives equal weight to each class and is suitable for multi-class prediction, especially when avoiding dominance by particular classes. Prior surveys on popularity prediction also note that accuracy alone may be insufficient under long-tailed or imbalanced settings (Tatar et al., 2014).

Since A–E labels are ordinal, we additionally report Quadratic Weighted Kappa (QWK), rank MAE, one-off accuracy, and Spearman rank correlation. QWK penalizes distant ordinal errors more heavily than adjacent errors and is commonly used for ordinal classification; ordinal classification studies also commonly use kappa and MAE to evaluate distance-aware label prediction (Gutiérrez et al., 2016; de la Torre et al., 2018; Vargas et al., 2023). Rank MAE maps A–E to ordinal indices and measures the average absolute distance between predicted and gold labels. One-off accuracy counts a prediction as correct if it falls within one adjacent level of the gold label, which is appropriate for noisy behavioral labels. Spearman correlation measures whether the predicted labels preserve the overall engagement ordering.

For the auxiliary continuous prediction setting, models predict $\log(1 + e_i)$ and are evaluated using MAE, RMSE, and Spearman correlation. Unless otherwise stated, our main results focus on ordinal engagement-level prediction.

G Case Study

Figure 5 details two additional failure modes. The first demonstrates surface-cue overestimation: the model heavily weights a limit-up chart and claimed prior success, despite the video receiving minimal actual engagement. The second illustrates fine-

grained calibration failure: the model correctly identifies high engagement potential in a dramatized Vanke-related video but predicts B instead of the gold label A. Corroborating our quantitative results, MLLMs capture coarse signals but lack calibration against true user feedback.

Figure 5: Representative failure cases in FINENGAGE. Highlighted cues denote signals that models rely on when predicting engagement. The first case shows surface-cue overestimation, where the model overweights limit-up charts and private-traffic solicitation. The second case shows coarse engagement recognition but fine-grained calibration failure.