FinCap: Topic-Aligned Captions for Short-Form Financial YouTube Videos

Siddhant Sukhani[1,2,*] ✉ Yash Bhardwaj[2*] Riya Bhadani[2] Veer Kejriwal[2]
Michael Galarnyk[2] Sudheer Chava[2]
[1] Stanford University
[2] Georgia Institute of Technology
✉ Corresponding author: sukhani@stanford.edu

Abstract

We evaluate multimodal large language models (MLLMs) for topic-aligned captioning in financial short-form videos (SVs) by testing joint reasoning over transcripts (T), audio (A), and video (V). Using 624 annotated YouTube SVs, we assess all seven modality combinations (T, A, V, TA, TV, AV, TAV) across five topics: main recommendation, sentiment analysis, video purpose, visual analysis, and financial entity recognition. Video alone performs strongly on four of five topics, underscoring its value for capturing visual context and effective cues such as emotions, gestures, and body language. Selective pairs such as TV or AV often surpass TAV, implying that too many modalities may introduce noise. These results establish the first baselines for financial short-form video captioning and illustrate the potential and challenges of grounding complex visual cues in this domain. All code and data can be found on our Github[1] under the CC-BY-NC-SA 4.0 license.

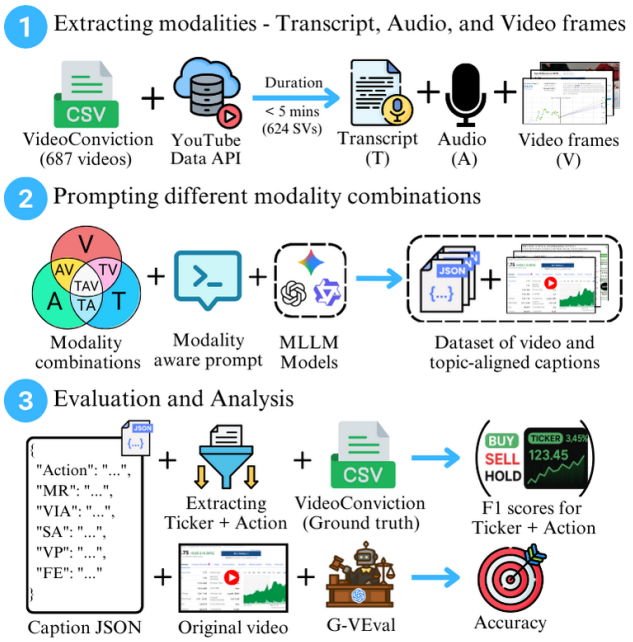

Figure 1. End-to-end pipeline: (1) Extract text, audio, and video frames from VideoConviction [11], leading to 624 SVs; (2) prompt MLLMs with all non-empty modality pairs to generate topic-aligned captions; (3) evaluate captions with G-VEval for accuracy and weighted F1 score on ticker–action pairs.

1. Introduction

Short-form videos (SVs) have become a dominant medium for information exchange, condensing dense multimodal signals into seconds [13, 60]. Platforms like YouTube host billions of videos [57], where rapid visual transitions, layered audio, and textual overlays are tightly interwoven [19, 39]. Capturing structured meaning from such content goes well beyond generic video captioning, demanding models that integrate visual, auditory, and textual cues with domain-specific reasoning. While multimodal large language models (MLLMs) have advanced video understanding [40, 45, 46, 51, 54], SVs remain largely unexplored [58]. Financial SVs present unique challenges: (i) overlapping on-screen elements such as charts, stock tickers (e.g., Microsoft's ticker is MSFT), logos, and annotations that demand robust object, scene, and multi-concept recognition; (ii) dense financial terminology and abbreviations that require precise grounding; and (iii) cues dispersed across non-contiguous segments, where stocks' tickers, spoken rationale, and affective signals (e.g., gestures, tone) may only align when considered together [12, 55].

We address this gap by establishing baselines for financial SV understanding through topic-aligned captioning. Using the VideoConviction dataset [11], we systematically test all single-, dual-, and tri-modality combinations of transcripts (T), audio (A), and video (V) across five tasks: main recommendation, sentiment analysis, video purpose, visual

[0] * These authors contributed equally.
[1] https://github.com/gtfintechlab/FinCap

analysis, and financial entity recognition (Figure 1). Captions are evaluated for both fidelity (capturing correct stock ticker–action pairs) and accuracy, using G-VEval [43]. Results show consistent gains from multimodal integration over single-modality baselines, with selective subsets sometimes surpassing full tri-modal fusion [27]. These findings establish the first benchmarks for financial SV understanding and provide a foundation for analyzing domain-specific SV content.

2. Related Work

Captioning has progressed from static image descriptions to temporally coherent video understanding [6, 25]. Attention mechanisms [47, 53], transformer architectures [63], and large-scale vision–language pretraining [21] have enabled topic-aligned and zero-shot captioning. However, most benchmarks emphasize generic or long-form domains, leaving short, domain-specific formats underexplored [50].

In finance, early work focused almost exclusively on text, including earnings calls, regulatory filings, and market news [24, 26, 36, 41]. Sentiment analysis in particular often relied on lexicons such as Loughran–McDonald [24], before newer corpora captured finer-grained stance and subjectivity [10, 30]. More recently, multimodal approaches have incorporated visual data such as financial charts [38], but short-form financial video remains unstudied. Prior work has not aligned finance-specific cues (e.g., stock tickers, charts, prosody, financial metrics) across modalities. We close this gap by establishing the first baselines for topic-aligned multimodal captioning of financial SVs, integrating transcripts, audio, and video in a unified evaluation.

3. Experiments & Data

We use the VideoConviction dataset [11], which contains YouTube videos from financial influencers issuing explicit stock recommendations. Each full-length video (avg. 9 minutes) is manually segmented into shorter clips, each containing a single actionable recommendation of buy, sell, or hold. We filter the 687 recommendation clips for segments under 5 minutes to align with short-form videos [15], yielding 624 SVs (avg. 4 minutes). Compared to ultra-short formats such as YouTube Shorts [48], video segments contain richer visual and auditory cues [3]. Human annotators labeled the stock ticker and corresponding action for every segment, enabling quantitative evaluation.

The resulting segments provide three aligned modalities:

1. **Video frames (V)** sampled at 0.25 fps following HourVideo [5], consisting of various on-screen elements such as charts, overlays, and gestures.
2. **Audio (A)** preserving prosody and tonal cues.
3. **Transcript (T)** aligned to segment timestamps for precise mapping between speech and visuals.

Together these modalities yield $2^3 - 1 = 7$ possible configurations: T, A, V, TA, TV, AV, and TAV.

Content Captioning For each modality combination, we prompt MLLMs to generate concise, topic-aligned captions. This task requires chart interpretation, spatiotemporal reasoning, and multimodal grounding. Captions are evaluated across five topics: **Main Recommendation (MR)**, which captures the core financial advice with stock ticker, action, and rationale; **Sentiment Analysis (SA)**, which measures tone and confidence from multimodal cues; **Video Purpose (VP)**, which identifies communicative intent and target audience; **Visual Analysis (VIA)**, which describes charts, indicators, overlays, and gestures to test OCR and spatiotemporal reasoning; and **Financial Entities (FE)**, which extracts company names, tickers, and numerical values within the SV. Together these topics target core multimodal challenges, including visual grounding, prosody, domain-specific language, and entity recognition, providing a focused benchmark for MLLMs in financial video understanding [2, 7, 56].

4. Evaluation

We evaluate all seven modality combinations (T, A, V, TA, TV, AV, TAV) across five multimodal large language models (MLLMs): Qwen2.5-Omni-7B [52], Gemini 2.0 Flash [8], Gemini 2.5 Flash [9], GPT-4o mini [32], and GPT-5 mini [28]. Since the GPT models do not support audio inputs, their evaluation is restricted to T, V, and TV combinations [28, 32]

Performance is measured using two complementary metrics: (i) F1-score for ticker–action extraction, based on the stock ticker and action (buy, sell, and hold) labels annotated in the VideoConviction dataset [11], and (ii) G-VEval [42], which scores accuracy on the remaining topics (MR, SA, VP, VIA, FE).

4.1. Structured Stock Ticker + Action Extraction

Table 1 reports F1-scores across modality combinations. Gemini 2.0 Flash performs best with 0.60 under AV, while V alone is consistently weakest. T, A, and TA provide moderate gains, and performance peaks when V is fused with either T or A. This pattern highlights stock ticker–action extraction as an intrinsically multimodal task: text and audio supply the primary signal, while visuals add grounding through ticker placement, chart overlays, and on-screen prompts [33]. However, fusion is not always beneficial as TAV seldom achieves the best results [17, 49].

4.2. Topic-aligned Captioning

G-VEval Metric Evaluating captions in finance is challenging because small semantic shifts can change the meaning of a recommendation. Standard n-gram metrics (BLEU

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	28	28	9	32	29	34	36
Gemini 2.0 Flash	46	55	41	49	59	**60**	59
Gemini 2.5 Flash	38	41	31	36	50	46	46
GPT-4o-mini	41	–	26	–	50	–	–
GPT-5-mini	40	–	32	–	57	–	–

Table 1. F1-scores for stock ticker and action extraction across T, A, and V, with best, worst, and overall highest in **bold**.

[29], ROUGE [22], METEOR [1], CIDEr [44]) penalize valid paraphrases and often miss domain-specific nuances [23]. Hence, we use G-VEval, which leverages GPT-4o with chain-of-thought prompting to assess the accuracy of generated captions. Unlike CLIPScore [16], EMScore [37], and PAC-Score [34], which rely on zero-shot prompting and limited token windows, G-VEval [42] uses structured reasoning to capture fine-grained, domain-specific semantics.

We organize our results by topic (MR, SA, VP, VIA, FE) to illustrate how different modality combinations contribute to each task. The following sections highlight the strengths and limitations of each model, revealing where each modality is most critical for short-form video understanding.

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	69	69	30	73	73	72	74
Gemini 2.0 Flash	80	83	74	81	83	**85**	84
Gemini 2.5 Flash	73	71	45	71	75	80	76
GPT-4o mini	78	–	77	–	82	–	–
GPT-5 mini	62	–	28	–	76	–	–

Table 2. G-VEval accuracy on topic **MR**, with best, worst, and overall highest in **bold**.

Main Recommendation (MR) MR, which captures the core financial advice, shows strong dependence on multiple modalities. As reported in Table 2, Gemini 2.0 Flash delivers the best overall performance, reaching an accuracy of 85 with the AV modality, while GPT-4o-mini follows closely with 82 under TV. These results underscore the importance of visual grounding—stock tickers, prices, and overlays—often cannot be recovered from transcripts alone. Qwen2.5 attains moderate results with TA and TV but fails with visual-only inputs, further suggesting that MR requires alignment between T, A, and V. Overall, the findings indicate that while text and audio contribute to capturing rationale and time frame, reliable extraction of the main recommendation depends critically on integrating multiple modalities.

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	61	55	48	57	66	70	68
Gemini 2.0 Flash	70	69	70	70	74	75	73
Gemini 2.5 Flash	73	74	79	74	77	76	76
GPT-4o mini	69	–	67	–	71	–	–
GPT-5 mini	71	–	**82**	–	76	–	–

Table 3. G-VEval accuracy on topic **SA**, with best, worst, and overall highest in **bold**.

Sentiment Analysis (SA) Table 3 illustrates the difficulty of modeling tone and confidence in short-form financial videos. Visual input emerges as the strongest signal, consistent with prior findings that gestures, expressions, and delivery style are key for sentiment recognition [35]. GPT-5 mini achieves the highest score (82) with video alone, underscoring the importance of visual cues. Gemini models show a similar vision-first pattern, though they maintain balanced performance across all modality combinations. In contrast, Qwen2.5 performs poorly with video alone (48) but improves substantially when audio is incorporated (70), suggesting reliance on audio cues. Overall, these results suggest that while sentiment is anchored in visual signals, complementary gains come from fusing audio and text, aligning with prior multimodal sentiment studies [4, 31].

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	61	44	73	47	79	75	78
Gemini 2.0 Flash	83	84	87	84	87	87	87
Gemini 2.5 Flash	85	86	**89**	86	88	87	87
GPT-4o mini	78	–	81	–	81	–	–
GPT-5 mini	74	–	84	–	79	–	–

Table 4. G-VEval accuracy on topic **VP**, with best, worst, and overall highest in **bold**.

Video Purpose (VP) VP targets the communicative intent and intended audience of a video. Table 4 shows that performance is consistently vision-dominant, with Gemini 2.5 Flash reaching the highest accuracy (89) under the V modality. Scene composition, sponsor tags, and instructional overlays provide strong cues for purpose classification, and across models V alone often matches or exceeds multimodal fusion. These results suggest that VP is determined primarily by how content is packaged visually, with transcripts and audio contributing little additional signal.

Visual Analysis (VIA) VIA evaluates OCR and spatiotemporal reasoning by requiring models to interpret charts, indicators, overlays, and gestures. Table 5 shows

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	20	16	85	19	83	83	83
Gemini 2.0 Flash	20	27	87	18	87	87	87
Gemini 2.5 Flash	14	9	90	9	90	90	89
GPT-4o mini	51	–	80	–	81	–	–
GPT-5 mini	24	–	**92**	–	91	–	–

Table 5. G-VEval accuracy on topic **VIA**, with best , worst , and overall highest in **bold**.

that performance is overwhelmingly vision-dominant. Accuracy collapses with text or audio alone (e.g., Gemini 2.5 Flash at 14 for T and 9 for A, GPT-5 mini at 24 for T) but rises to 90 and 92 with visual inputs alone for Gemini 2.5 Flash and GPT-5 mini, respectively. Adding transcripts or audio (TV, AV, TAV) offers no major benefit and sometimes reduces performance, indicating that non-visual modalities introduce redundancy or noise rather than complementary signal. These findings confirm VIA as a fundamentally video-driven task, where robust chart and overlay interpretation cannot be approximated through speech or text alone.

Model	T	A	V	TA	TV	AV	TAV
Qwen 2.5	65	68	79	66	78	78	78
Gemini 2.0 Flash	78	78	85	77	83	83	82
Gemini 2.5 Flash	76	80	86	80	86	86	86
GPT-4o mini	76	–	83	–	81	–	–
GPT-5 mini	78	–	**87**	–	**87**	–	–

Table 6. G-VEval accuracy on topic **FE**, with best , worst , and overall highest in **bold**.

Financial Entities (FE) FE targets company names, stock tickers, and numerical values. Table 6 shows a clear reliance on vision, with V yielding the strongest results across models and GPT-5 mini achieving peak accuracy of 87 under both V and TV. Unlike VIA, where visuals dominate due to charts and layouts, FE depends on structured identifiers that are typically displayed on-screen. Text and audio provide only partial cues and often add noise, offering little benefit over visual input alone. Overall, FE is a vision-dominant task, distinguished from VIA by its focus on precise finance-specific entities rather than broader visual context.

5. Discussion

Our findings reveal two consistent patterns. First, full tri-modal fusion does not guarantee superior performance over selective modality combinations [62]. In several cases, particularly VIA, FE, and VP, adding modalities to V introduced contradictory signals that reduced accuracy or induced hallucination [20, 61]. This suggests that careful modality selection, rather than indiscriminate fusion, is essential for robust multimodal reasoning.

Second, the role of V varies sharply across task types. For explicitly visual tasks such as FE and VIA, models succeed at extracting precise financial details (e.g., "the stock price of $49.39"). By contrast, for tasks where the financial signal is not directly visible, such as MR, visual input often defaults to scene-level descriptors (e.g., speaker appearance) instead of structured financial content. For SA and VP, contextual visual cues such as camera setup, gestures, and delivery style proved critical for interpreting tone and audience targeting. These trends point to a division of labor across modalities: vision should serve as the primary anchor, with audio and text acting as task-sensitive supplements. Effective prompting strategies are critical to ensure models attend to the right visual features.

Finally, our study is scoped to financial short videos drawn from the VideoConviction dataset. While this controlled setting enables systematic benchmarking, it also limits generalizability. Other short-form video domains may exhibit different multimodal dynamics, making cross-domain extensions a promising direction for future work.

6. Conclusion

We presented, to the best of our knowledge, the first benchmark for topic-aligned captioning of financial SVs. Our evaluation spans five tasks: MR, SA, and VP, which capture core financial advice, sentiment, and communicative intent; VIA, which probes visual grounding; and FE and ticker–action extraction, which test entity recognition and multimodal reasoning. Together, these tasks provide a structured foundation for assessing MLLMs in domains where dense jargon and layered visual cues make comprehension uniquely demanding [59].

Our results demonstrate that selective modality fusion can outperform full tri-modal integration, underscoring the need for careful input design in multimodal reasoning [14, 20]. Given the market influence of financial video recommendations, accurate topic-aligned captions are critical both for research and real-world applications. Looking ahead, we expect these benchmarks to guide future development of multimodal models in finance and to serve as a stepping stone for captioning in other high-stakes, domain-specific settings [18].

Acknowledgment We appreciate the generous support and funding provided by the Georgia Institute of Technology Financial Services Innovation Lab and the Center for Finance and Technology.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Metrics for MT and/or Summarization*, page 65–72, 2005. 3

[2] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022. 2

[3] Alejandro Bernales, Marcela Valenzuela, and Ilknur Zer. Effects of information overload on financial markets: How much is too much? 2023. 2

[4] Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter J. Scheirer. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities, 2018. 3

[5] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding, 2024. 2

[6] Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation, 2025. 2

[7] Hyundong Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T. Chavez, and Jonathan May. Can vision language models understand mimed actions?, 2025. 2

[8] Google DeepMind. Gemini 2.0 flash, 2025. 2

[9] Google DeepMind. Gemini 2.5 flash, 2025. 2

[10] Quanqi Du, Loic De Langhe, Els Lefever, and Veronique Hoste. Sentiment analysis on video transcripts: Comparing the value of textual and multimodal annotations. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 10–15, Albuquerque, New Mexico, USA, 2025. Association for Computational Linguistics. 2

[11] Michael Galarnyk, Veer Kejriwal, Agam Shah, Yash Bhardwaj, Nicholas Meyer, Anand Krishnan, and Sudheer Chava. Videoconviction: A multimodal benchmark for human conviction and stock market recommendations. *Available at SSRN 5315526*, 2025. 1, 2

[12] Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. 2024. Describes challenges of interpreting dense multimodal signals—charts, financial jargon, annotations—relevant to compressed finance video formats. 1

[13] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts, 2025. 1

[14] Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[15] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21064–21075, 2022. 2

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3

[17] C. Hori et al. Attention-based multimodal fusion for video description. In *ICCV*, 2017. 2

[18] Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[19] Sarah Lee and Qing Zhao. Youtube as a financial news source: Credibility and market impact. *IEEE Transactions on Multimedia*, 25:789–801, 2023. 1

[20] Yujian Lee, Peng Gao, Yongqi Xu, and Wentao Fan. How do optical flow and textual prompts collaborate to assist in audio-visual semantic segmentation? In *Proceedings of ICCV*, 2025. 4

[21] Juntao Li, Xintian Zhu, Wei-Lun Chao, and Zhi-Hua Yu. Clip-it: Language-guided video summarization via multimodal embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, ACL*, page 74–81, 2004. 3

[23] Jiaxin Liu, Yang Yi, and Kar Yan Tam. Beyond surface similarity: Detecting subtle semantic shifts in financial narratives. *arXiv preprint arXiv:2403.14341*, 2024. 3

[24] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011. 2

[25] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, Xiaopeng Hong, Yongjian Wu, and Rongrong Ji. Image captioning via dynamic path customization, 2024. 2

[26] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, page 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. 2

[27] Sahid Hossain Mustakim, S M Jishanul Islam, Ummay Maria Muna, Montasir Chowdhury, Mohammed Jawwadul Islam, Sadia Ahmmed, Tashfia Sikder, Syed Tasdid Azam Dhrubo, and Swakkhar Shatabda. Watch, listen, understand, mislead: Tri-modal adversarial attacks on short videos for content appropriateness evaluation, 2025. 2

[28] OpenAI. GPT-5 Mini — openai api documentation, 2025. Accessed: 2025-08-13. 2

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311–318, 2002. 3

[30] Huzaifa Pardawala, Siddhant Sukhani, Agam Shah, Veer Kejriwal, Abhishek Pillai, Rohan Bhasin, Andrew DiBiasio, Tarun Mandapati, Dhruv Adha, and Sudheer Chava. Subjective-qa: Measuring subjectivity in earnings call transcripts' qa through six-dimensional feature analysis, 2024. 2

[31] Moisés H. R. Pereira, Flávio L. C. Pádua, Adriano C. M. Pereira, Fabrício Benevenuto, and Daniel H. Dalip. Fusing audio, textual and visual features for sentiment analysis of news videos, 2016. 3

[32] Alec Radford, Joon W. Kim, Christina Hallacy, et al. Gpt-4o: Openai's multimodal transformer. Technical report, OpenAI, 2024. Technical Report. 2

[33] Mouli Rastogi, Syed Afshan Ali, Mrinal Rawat, Lovekesh Vig, Puneet Agarwal, Gautam Shroff, and Ashwin Srinivasan. Information extraction from document images via fca-based template detection and knowledge graph rule induction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[34] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[35] Zoya Shafique, Haiyan Wang, and Yingli Tian. Nonverbal communication cue recognition: A pathway to more accessible communication. In *CVPR Workshops*, 2023. 3

[36] Agam Shah, Siddhant Sukhani, Huzaifa Pardawala, Saketh Budideti, Riya Bhadani, Rudra Gopal, Siddhartha Somani, Michael Galarnyk, Soungmin Lee, Arnav Hiray, Akshar Ravichandran, Eric Kim, Pranav Aluru, Joshua Zhang, Sebastian Jaskowski, Veer Guda, Meghaj Tarte, Liqin Ye, Spencer Gosden, Rutwik Routu, Rachel Yuh, Sloka Chava, Sahasra Chava, Dylan Patrick Kelly, Aiden Chiang, Harsit Mittal, and Sudheer Chava. Words that unite the world: A unified framework for deciphering central bank communications globally, 2025. 2

[37] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *CVPR 2022*, 2021. 3

[38] Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. Finchartbench: Benchmarking financial chart comprehension in vision-language models, 2025. 2

[39] John Smith and Arjun Kumar. The rise of finfluencers: Impact and influence on retail trading behavior. *Journal of Financial Communication*, 10(3):45–60, 2022. 1

[40] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zujun Ma, and Chao Zhang. Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization. *arXiv preprint arXiv:2410.06682*, 2024. 1

[41] Nikita Tatarinov, Siddhant Sukhani, Agam Shah, and Sudheer Chava. Language modeling for the future of finance: A quantitative survey into metrics, tasks, and data opportunities, 2025. 2

[42] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. *arXiv preprint arXiv:2412.13647*, 2024. 2, 3

[43] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o, 2024. 2

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 4566–4575, 2015. 3

[45] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Chapter-llama: Efficient chaptering in hour-long videos with llms, 2025. 1

[46] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence—video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542, 2015. 1

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 2

[48] Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. Shorts vs. regular videos on youtube: a comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference*, pages 213–223, 2024. 2

[49] Wenwu Xia et al. On modality bias in multimodal learning. *arXiv preprint arXiv:2302.10912*, 2023. 2

[50] Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *Advances in Neural Information Processing Systems*, 37:16623–16644, 2024. 2

[51] Jin Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video-description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 1

[52] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni: An end-to-end multimodal model, 2025. Accessed: 2025-08-20. 2

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 2

[54] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale, 2023. 1

[55] Xiao-Yang Liu Yanglet, Yupeng Cao, and Li Deng. Multimodal financial foundation models (mffms): Progress, prospects, and challenges, 2025. 1

[56] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. 2

[57] YouTube Press. Youtube celebrates more than twenty billion videos uploaded. *YouTube Press Blog*, 2025. As of April 2025, over 20 billion videos have been uploaded; over 20 million videos are uploaded daily. 1

[58] Ting Yu, Kunhao Fu, Shuhui Wang, Qingming Huang, and Jun Yu. Prompting video-language foundation models with domain-specific fine-grained heuristics for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[59] Liwei Zhang, Rishabh Kumar, Yifan Liu, Yuan Li, and Jing Xu. Mm-vidqa: Multimodal video question answering on short-form videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[60] Zhipeng Zhang and Liyi Zhang. Most significant impact on consumer engagement: An analytical framework for the multimodal content of short video advertisements. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(2):54, 2025. 1

[61] Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict, 2025. 4

[62] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025. 4

[63] Yubo Zhu, Yichen Duan, Yue Zhao, Zhiwei Zhang, Hui Xiong, and Zhi-Hua Lin. Topic-aware video summarization using multimodal transformer. *Pattern Recognition*, 140: 109506, 2023. 2