AVTEXTBENCH: BENCHMARKING CROSS-MODAL ON-SCREEN TEXT GROUNDING IN AUDIO-VISUAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-screen text has become a widely applied tool to convey video information more effectively, such as subtitles. Despite the success of audio-visual large language models (LLMs), it remains unclear whether they can ground on-screen text in the surrounding audio-visual context. In this work, we show that audio-visual LLMs perform poorly in understanding the relationships between on-screen text and audio-visual streams. To systematically study this problem, we introduce a benchmark for probing these capabilities. Specifically, we introduce a controlled framework covering three fundamental relationships: **corresponding**, **complementary**, and **independent**. This framework provides a comprehensive evaluation of whether audio-visual LLMs can correctly **match**, **integrate**, and **segregate** on-screen textual and audio-visual information in real-world videos. We evaluate our benchmark on state-of-the-art audio-visual LLMs, including both open-source models and commercial models. Experimental results show that existing open-source audio-visual LLMs fail to ground on-screen text in context, with performance near random guessing in multiple tasks. On the other hand, the Gemini-3.8-Flash model achieves strong accuracy, highlighting a large gap between open-source and commercial models. Our work uncovers an unexplored limitation in current audio-visual LLMs and provides new insights for future improvement.

1 INTRODUCTION

With the wide application of online platforms (e.g., YouTube), videos have become an increasingly popular medium for communication. To convey information more effectively, people frequently augment the captured audiovisual stream with on-screen text. For example, subtitles transcribe spoken content into text, making speech easier to follow (Figure 1(c1)), while textual annotations provide critical and complementary information (Figure 1(c2)). Recently, audio-visual large-language models (LLMs) (Sun et al., 2024; Xu et al., 2025a;b; Tang et al., 2025) have demonstrated great success in video understanding (Zhao et al., 2022; Zhou et al., 2025b; Sung-Bin et al., 2025; Hong et al., 2026). This raises a fundamental question: *could modern audio-visual LLMs ground on-screen text in audio-visual context?* Crucially, grounding these texts not only requires **recognizing what the on-screen text says**, but also **understanding the relationship between on-screen text and audiovisual content**. Failing to capture this relationship could lead to incorrect video interpretation, which is particularly risky in safety-critical applications such as emergency news understanding and content moderation (Jahagirdar et al., 2023).

Recent video-OCR benchmarks (Fei et al., 2025; Tom et al., 2023; Zhao et al., 2022) have extended video text understanding to diverse scenarios, such as news content (Jahagirdar et al., 2023) and egocentric scenes (Zhou et al., 2025a). Despite this progress, these benchmarks largely formulate text understanding as a vision modality-only problem, where audio is not included for answering the questions (Figure 1(a)). As a result, they can assess what the text content is but overlook the prevalent scenarios that require cross-modality understanding. On the other hand, recent audio-visual benchmarks (Hong et al., 2026; Zhou et al., 2025b; Sung-Bin et al., 2025) incorporate audio and evaluate the models' ability to integrate information across modalities (Figure 1(b)). However, despite the wide application of on-screen text in modern videos, how audio-visual LLMs ground on-screen text in audio-visual context remains largely unexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

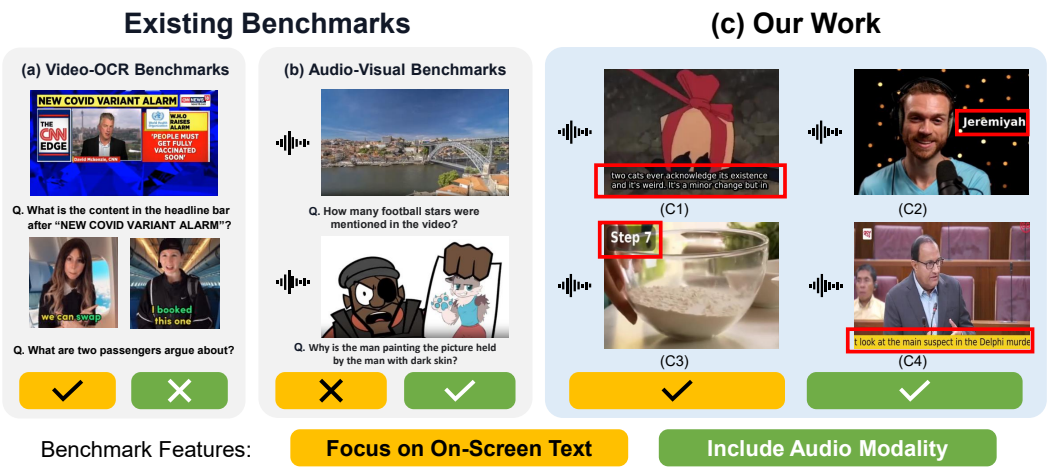

Figure 1: **Comparison with existing benchmarks.** Recent video-OCR benchmarks studied on-screen text understanding across diverse scenarios, but treat it as a vision-only problem without including audio. On the other hand, audio-visual benchmarks evaluate multimodal understanding but rarely focus on the role of on-screen text. Our work bridges this gap by investigating a fundamental yet underexplored question: **how do audio-visual LLMs ground on-screen text in audio-visual content**, i.e., **understand its relationship with the surrounding audio and visual information**.

In this work, we investigate a fundamental yet unexplored question for real-world understanding: how do audio-visual LLMs ground on-screen text in broader audio-visual context? To systematically study this problem, we propose a comprehensive benchmark, AVTextBench, which identifies three fundamental and cross-modal relationships between on-screen text and audio-visual content: **corresponding**, **complementary**, and **independent**. Corresponding tasks examine whether models understand how on-screen text matches the audio-visual content and whether they can recognize mismatched information if it exists. Complementary tasks require the model to integrate non-redundant information conveyed by on-screen text and audio-visual content to obtain a complete interpretation. Independent tasks evaluate whether models can identify the source where specific information comes from (e.g., on-screen text or audios). Together, these three relationships provide a comprehensive study of whether audio-visual LLMs can correctly **match**, **integrate** and **segregate** on-screen textual and context information in real-world videos.

To examine the cross-modal grounding ability of on-screen text in audio-visual LLMs, we perform a comprehensive evaluation of state-of-the-art models on AVTextBench. Specifically, we evaluate both open-source models (e.g., Qwen3-Omni (Xu et al., 2025b)) and commercial models (e.g.,Gemini 3.8 Flash (Google, 2026)). Experiment results show that open-source models perform poorly on all tasks of our AVTextBench. Specifically, even the advanced Qwen3-Omni (Xu et al., 2025b) achieves accuracy near random guess in correspondence tasks. In contrast, the Gemini 3.8 Flash (Google, 2026) achieves strong performance higher than 80% for all tasks. These results highlight the high quality of our benchmark, while also shows a large gap between open-source and commercial models. To understand the failures of open-source models, we further analyze the models' abilities from multiple aspects. We find that models exhibit different text and speech recognition performance across different tasks. Specifically, models often confuse spoken content with on-screen text in the independent tasks. These findings reveal a largely underexplored yet fundamental capability of audio-visual LLMs in grounding on-screen text in audio-visual context, and provide new insights for improving future models. We will release our dataset and code upon acceptance.

2 AVTextBench: Cross-modal On-Screen Text Grounding in Audio-visual LLMs

In this section, we start by outlining the taxonomy of our AVTextBench in Section 2.1. We then identify three fundamental and cross-modal relationships between on-screen text and audio-visual content from three perspectives: **corresponding**, **complementary**, and **independent** in Sections 2.2,

2.3, and 2.4, respectively. We then introduce how the dataset is created in Section 2.5, followed by an overview of dataset statistics in Section 2.6.

2.1 TASK TAXONOMY

Overview. Our work proposes AVTextBench, the first comprehensive benchmark that systematically investigates the cross-modal on-screen text grounding in audio-visual LLMs. Given a video containing an audiovisual stream C with on-screen text T, we study whether audio-visual LLMs can correctly identify the cross-modal relationship between on-screen text T in their surrounding audiovisual context C. Specifically, we identify three fundamental relationships as follows: (1) **Corresponding**: whether audio-visual LLMs understand if on-screen text matches the audio and recognize the mismatched information if exists. (2) **Complementary:** whether audio-visual LLMs can integrate information only available from on-screen text together with audio-visual content for comprehensive reasoning. (3) **Independent:** whether audio-visual LLMs can identify the source of a statement (e.g., on-screen text and audio). Together, our AVTextBench provides a systematic framework for studying how audio-visual LLMs can **match**, **integrate** and **segregate** on-screen text and context information in real-world videos. We show examples of our benchmark in Figure 2.

Task formulation. We follow previous audio-visual benchmarks (Fei et al., 2025; Hong et al., 2026; Zhou et al., 2025b) to design text questions for different tasks to evaluate model abilities. We follow to include two-choice and multiple-choice questions for effective and efficient evaluation. We shuffle the options for each question to avoid option position bias for fair evaluation, and introduce the detailed task design in the following subsections.

2.2 CORRESPONDENCE TASKS

Motivation. A canonical example of the *corresponding on-screen text* is subtitles, where the displayed text on screen is expected to correspond to the spoken content in the video. However, such correspondence is not always perfect in real-world videos. For example, mismatches can arise when displayed text substitutes part of the spoken content or when correct text is temporally misplaced. These cases motivate us to examine whether audio-visual LLMs can reliably reason about cross-modal correspondence between on-screen text and spoken content.

Task definition. Given a video containing displayed subtitle text T and speech S, we evaluate whether audio-visual LLMs can correctly determine their correspondence. Specifically, a robust model should not only determine **whether** the two modalities are aligned, but also characterize **what** the mismatched content is and localize **when** the mismatch occurs. Accordingly, we design three coarse-to-fine tasks as follows, with video examples shown in Figure 2(a). We present simplified question pairs as follows, with the completed version in Appendix C.

- ***Alignment verification.*** This task evaluates whether the model can determine if the displayed subtitles consistently match the spoken content, with the following question:

  Do the on-screen subtitles correctly correspond to the spoken content throughout the video? (A) Yes. B. No.

- ***Misalignment type classification.*** When the mismatch between displayed text and speech exists, this task examines if the model can identify the types of mismatch errors, such as substitution and permutation.

  What type of mismatch exists between the on-screen subtitles and the spoken content? (A) Substitution: The subtitle content differs in meaning from the corresponding speech. (B) Permutation: The subtitle content belongs to another spoken segment in the same video and is displayed at the wrong time.

- ***Misalignment localization.*** Given a video with mismatched on-screen text and spoken content, this task evaluates whether the model can identify the time interval when the mismatch occurs.

  When does the mismatch between the on-screen subtitles and the spoken content occur? (A) [First third of the video duration] (B) [Second third of the video duration] (C) [Last third of the video duration]

Together, these tasks provide a comprehensive evaluation of whether audio-visual LLMs could understand if the on-screen text **matches** the corresponding video events.

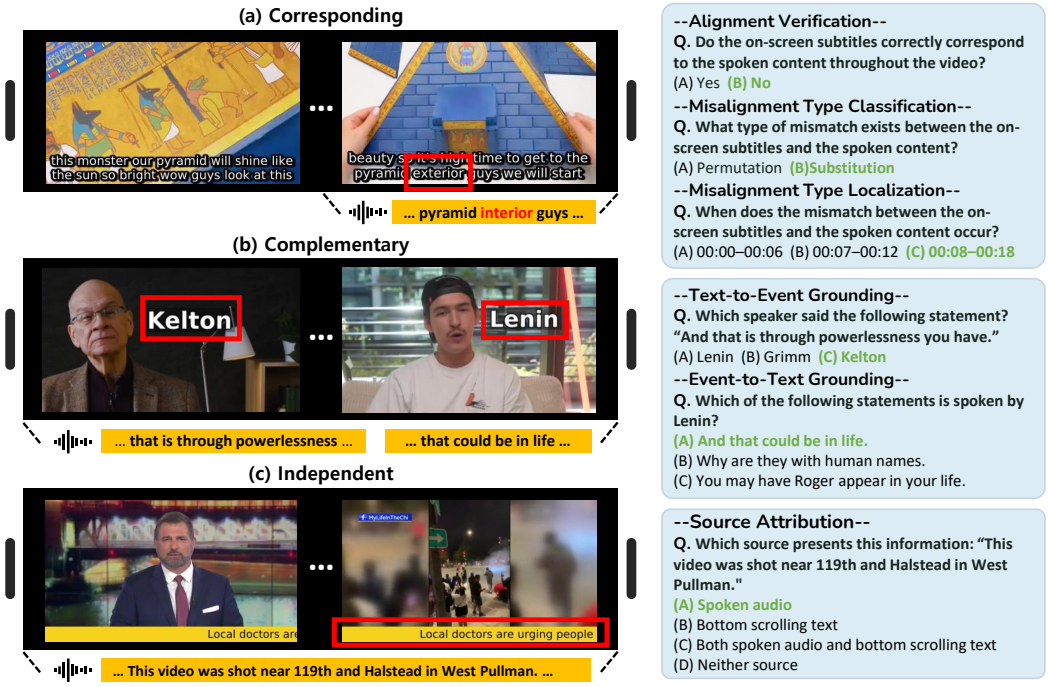

Figure 2: **Cross-modal grounding of on-screen text in audio-visual LLMs.** Our AVTextBench investigates three fundamental relationships between on-screen text and audio-visual streams: **corresponding**, **complementary**, and **independent**. Corresponding tasks evaluate whether audio-visual LLMs can identify the mismatch between on-screen text and recorded video. Complementary tasks require the model to integrate information only conveyed in on-screen text together with audio-visual streams for comprehensive understanding. Independent tasks evaluate whether models can identify the source of a statement (e.g., on-screen text or audio). Our AVTextBench provides a comprehensive evaluation of whether audio-visual LLMs can correctly **match**, **integrate** and **segregate** on-screen text and context information in real-world videos.

2.3 Complementary Tasks

Motivation. *Complementary on-screen text* provides information that is not conveyed by the audiovisual stream, but is necessary for interpreting the whole video. For example, text annotation of speaker name provides identity information which may not be mentioned in the original video (Figure 2(b)-1), while step markers in an instructional video naturally split the video into multiple segments (Figure 2(b)-2). In these cases, the on-screen text serves as a contextual anchor that must be associated with the relevant audiovisual content. Even if these text annotations disappear on screen, a successful model could identify the entity and time duration it refers to. These complementary on-screen text are widely applied in real videos yet remain underexplored in existing benchmarks (Fei et al., 2025; Hong et al., 2026; Zhou et al., 2025b).

Task definition. Given a video with complementary on-screen text T, we systematically evaluate whether audio-visual LLMs integrate this complementary information from on-screen text for cross-modal understanding. Specifically, we design grounding tasks in two directions as follows, with video examples shown in Figure 2(b).

- **Text-to-event grounding.** Given an on-screen text cue, the model must identify which audiovisual event is associated with it. The following is the question example for the speaker name setting:

  Which of the following statements is spoken by [speaker name]?
  (A) [Statement 1] (B) [Statement 2] (C) [Statement 3]

- **Event-to-text grounding.** Given an event in the video, the model must identify which on-screen text cue it is associated with.

Which speaker said the following statement?
(A) [Speaker 1] (B) [Speaker 2] (C) [Speaker 3]

Overall, these bidirectional tasks provide a systematic evaluation of whether audio-visual LLMs could **integrate** the complementary information from on-screen text with original video content.

2.4 INDEPENDENT TASKS

Motivation. In real-world videos, on-screen text does not always convey related information compared to recorded audiovisual content, *i.e.*, *independent on-screen text*. Instead, it may serve as an independent information channel, such as bottom-rolling text widely applied in TV programs, as shown in Figure 2(c). In such cases, it is critical for audio-visual LLMs to preserve the provenance of information across modalities, yet remaining unexplored.

Task definition. Given a video containing on-screen text T and audiovisual content C that convey independent information, we evaluate whether the model can identify the source of a queried statement as follows.

- **Source attribution.** We ask the model to identify whether a queried statement is conveyed by the on-screen text or the audio stream.

  Which source presents this information: [statement]
  (A) Spoken audio. (B) Bottom scrolling text. (C) Both spoken audio and bottom scrolling text. (D) Neither source.

Overall, our work provides an evaluation of whether models can **segregate** information from on-screen text and original recorded videos.

2.5 DATASET CONSTRUCTION PIPELINE

To systematically study how audio-visual LLMs ground on-screen text in audio-visual content, we propose a controlled framework for dataset creation. We will introduce the pipeline for corresponding, complementary and independent tasks, respectively.

Corresponding tasks. We construct corresponding samples from the Shot2Story dataset (Han et al., 2025), which contains diverse video domains, including movies, podcasts, and daily-life videos. We first use the widely applied Whisper large-v3 (Radford et al., 2023) to transcribe the speech and obtain utterance-level timestamps. In the default setting, the resulting subtitles are overlaid on the video at their aligned timestamps. To evaluate whether models can identify misalignment between on-screen text and the audio stream, we further create two representative controlled variants: *substitution* and *permutation*. For substitution cases, we ask GPT-5 model (openAI, 2025) to modify one semantic element of the original matched subtitles so that the semantic meaning could be changed. For permutation, we randomly select two sentences from the subtitle and swap the order when displaying on the screen. This framework allows us to systematically study whether audio-visual LLMs could identify the corresponding relationship between on-screen text and recorded videos.

Complementary tasks. To evaluate whether audio-visual LLMs can integrate complementary information from on-screen text and audio-visual streams, we consider two representative types of on-screen text: *speaker name* and *step markers*. For speaker names, we build on TalkVid (Chen et al., 2026), which contains diverse talking-head videos from YouTube. We concatenate clips from different speakers into a single video and assign each speaker a randomly sampled name consistent with their gender. The name is displayed near the speaker, then disappears. We construct question-answer pairs that require models to associate later speech with the correct speaker, testing whether they can retain and integrate the transient textual identity cue with subsequent audio-visual content. For step markers, we build on the YouCook2 dataset (Zhou et al., 2018), which contains instructional videos with naturally structured procedures and event-level annotations, including both semantic descriptions and temporal boundaries. We overlay a step marker at the beginning of each annotated event and construct questions about the events in the recorded video.

Independent tasks. Independent tasks evaluate whether models can identify the source of a statement, i.e., whether it originates from the on-screen text or the audio stream. We build this task on MultiVENT2.0 (Kriz et al., 2025), where bottom rolling text (i.e., news ticker) can naturally convey

Table 2: **Results on AVTextBench.** We evaluate the state-of-the-art audio-visual LLMs on our benchmark, and report results for each task as follows. Results show that all open-source models (e.g., Qwen3-Omni) perform poorly on AVTextBench, with near-random performance in correspondence tasks. In contrast, Gemini 3.8 Flash achieves strong results, highlighting a gap between open-source and commercial models. Note that Spk. and Step denote accuracy for the speaker and step setting, respectively.

Model	Correspondence					Complementary				Independent
	Verification		Type		Localization	Text→Event		Event→Text		Attribution
	Acc.	F1	Acc.	F1	Acc.	Spk.	Step	Spk.	Step	Acc.
Gemini 3.8 Flash (Google, 2026)	**81.84**	**81.82**	**94.34**	**94.34**	**88.42**	**100.00**	**95.26**	**99.49**	**94.69**	**96.93**
VideoLLaMA 2 (Cheng et al., 2024)	48.58	47.86	50.24	47.83	33.81	36.32	40.10	40.95	36.28	55.19
video-SALMONN (Sun et al., 2024)	46.70	46.50	58.16	53.79	34.04	50.00	35.26	34.12	33.19	0.24
video-SALMONN 2+ (Tang et al., 2025)	45.05	44.71	51.42	51.36	30.26	66.04	75.13	82.94	68.58	46.70
Qwen2.5-Omni-3B (Xu et al., 2025a)	48.11	48.11	52.36	52.34	36.64	59.43	50.76	81.04	78.76	25.71
Qwen2.5-Omni-7B (Xu et al., 2025a)	50.94	50.31	54.72	53.97	30.97	61.32	54.31	80.57	81.42	19.81
Qwen3-Omni-30B (Xu et al., 2025b)	48.35	48.34	49.53	49.50	34.75	70.62	82.69	78.67	78.32	41.75
Random choice	50.00	50.00	50.00	50.00	33.33	33.33	33.33	33.33	33.33	25.00

information independent of the main news event presented in the video and audio. We first select news videos from MultiVENT2.0 and use Whisper large-v3 to transcribe their speech. We then take speech content from a different video and overlay it on the current video in the form of a news ticker, thereby creating an independent textual information source. Finally, we construct questions about statements appearing in the news ticker, requiring models to distinguish the information conveyed by on-screen text from that in the original video.

2.6 Dataset statistics

Our AVTextBench includes 2541 video-question pairs, with 1271 samples for correspondence tasks, 846 samples for complementary tasks and 424 samples for independent tasks. We show the number for each task in Table 1. Our benchmark follows previous studies(Fei et al., 2025; Hong et al., 2026; Zhou et al., 2025b; Sung-Bin et al., 2025) to include two-choice or multi-choice question. Note that within each task, our dataset contains a balanced number of samples for each answer option.

Table 1: Statistics of our benchmark.

Category	Task	Count
Corresponding	Alignment Verification	424
	Misalignment Localization	423
	Misalignment Type Classification	424
Complementary	Text-to-Event Grounding	423
	Event-to-Text Grounding	423
Independent	Source Attribution	424
Total		**2,541**

3 Experiments

In this section, we begin by introducing the evaluation setups in Section 3.1, including state-of-the-art audio-visual LLMs and evaluation metrics. We then discuss the experiment results of those state-of-the-art models on our AVTextBench benchmark in Section 3.2. We further provide a comprehensive analysis of model failures on AVTextBench, providing insights for potential model improvement in Section 3.3.

3.1 Experiment Setup

Models. We evaluate our AVTextBench benchmark with six state-of-the-art audio-visual LLMs. Given a text question, these models take video frames and the audio track as joint input. We evaluate on five popular open-source audio-visual LLMs, including VideoLLaMA 2 (Cheng et al., 2024), video-SALMONN (Sun et al., 2024), video-SALMONN 2+ (Tang et al., 2025), Qwen2.5-Omni (Xu et al., 2025a), and Qwen3-Omni (Xu et al., 2025b). We also evaluate the latest Gemini 3.8 Flash model for the performance of commercial models on our benchmark. All models are evaluated in a zero-shot way, without task-specific training.

Evaluation metrics. Following audio-visual benchmarks (Chowdhury et al., 2025; Radevski et al., 2025; Sung-Bin et al., 2025), every task is posed as a choice between two options or among several, and scored with the metrics those benchmarks use. Two-choice tasks are scored with accuracy and F1, the harmonic mean of precision and recall. For multiple-choice tasks, we report accuracy across all samples.

3.2 RESULTS

We evaluate the ability of six state-of-the-art audio-visual LLMs on our AVTextBench, which include the latest commercial model Gemini 3.8 flash and the open-source model Qwen3-Omni (Xu et al., 2025b). We report the model performance on different tasks in Table 2, where accuracy for random guessing is also reported for fair comparison.

Cross-modal grounding of corresponding on-screen text. As discussed in Section 2.2, corresponding tasks not only evaluate whether audio-visual LLMs could identify the cross-modal mismatch between on-screen text and audio content, but also examine whether models could understand and localize the mismatched content if it exists. As shown in Table 2, all open-source models show severe failure modes in all correspondence tasks, even for the state-of-the-art Qwen3-Omni (Xu et al., 2025b) model in the 30B size. Specifically, Table 2 shows that all open-source audio-visual LLMs achieve poor performance compared to random guessing. For example, Qwen 3-Omni (Xu et al., 2025b) achieves an accuracy of 48.35% for the alignment verification task, which is close to 50% in the random choice setting for two-choice tasks. These results demonstrate that open-source models lack the ability to identify the cross-modal relationship between corresponding on-screen text and audio streams. By contrast, the latest Gemini 3.8 Flash model significantly outperforms open-source models in correspondence tasks, achieving an accuracy of over 80% for all tasks. The results of Gemini 3.8 Flash model show that our correspondence tasks are solvable and highlight the high quality of video-question pairs in our AVTextBench benchmark. Overall, these experiment results show a severe yet previously unexplored failure of open-source audio-visual in cross-modal grounding of corresponding on-screen text, highlighting the necessity for future improvement to reduce the gap with commercial models.

Cross-modal grounding of complementary on-screen text. In complementary tasks, we aim to evaluate whether audio-visual LLMs could identify useful information in on-screen text, and perform integrated cross-modal understanding of text and audio-visual streams. As discussed in Section 2.3, our AVTextBench focuses on two representative complementary on-screen text, speaker names and step markers. Table 2 shows the results for complementary tasks and reveal multiple insightful findings. First, our complementary tasks present a challenge to open-source models, e.g. the 54.31% accuracy of the advanced Qwen2.5-Omni-7B (Xu et al., 2025a) in text-to-event tasks with the step marker setting. Second, Qwen3-Omni model achieves the best performance among all open-source models, while Gemini 3.8 Flash significantly outperforms open-source models on all tasks. Third, different models exhibit distinct trend in different tasks. For example, a group of models achieves better results on event-to-text grounding tasks than text-to-event grounding tasks, including video-SALMONN 2+ (Tang et al., 2025), Qwen2.5-Omni-3B (Xu et al., 2025a), and Qwen2.5-Omni-7B (Xu et al., 2025a). By contrast, other models do not follow the trend, such as Qwen3-Omni (Xu et al., 2025b). Lastly, for Qwen series models, performance generally improves with increasing model size on text-to-event task, e.g., Qwen2.5-Omni-3B (Xu et al., 2025a) and Qwen2.5-Omni-7B (Xu et al., 2025a) achieve an accuracy of 59.43% and 61.32% in the speaker setting. In contrast, this trend that larger model size increases accuracy is not observed on the event-to-text task for Qwen series models. Overall, our complementary tasks in AVTextBench provide a comprehensive evaluation of how the models integrate complementary information from on-screen text with recorded audiovisual streams.

Cross-modal grounding of independent on-screen text. Independent tasks aim to evaluate whether audio-visual LLMs could identify the source of a statement (e.g., on-screen text or audios). Table 2 shows a poor while diverse performance of open-source models in this task, with the best model achieving an accuracy of 55.19%. Interestingly, although Qwen2.5-Omni-3B (Xu et al., 2025a) and Qwen2.5-Omni-7B (Xu et al., 2025a) achieve relatively strong accuracy in other tasks, they obtain a poor accuracy of 25.71% and 19.81% in the source attribution task, revealing a new weakness unexplored by existing benchmarks. On the other hand, Gemini 3.8 Flash model still

achieves superior performance compared to all open-source models, highlighting a improvement direction for future open-source models.

3.3 Analysis

Experiment results and discussions in Section 3.2 have shown that our benchmark poses significant challenges to state-of-the-art open-source audio-visual models, such as the near-random accuracy on correspondence tasks. In this section, we further provide a comprehensive analysis of why these models fail in cross-modal grounding of on-screen text.

Do audio-visual LLMs recognize on-screen text? We first examine whether model errors may stem from failures in reading on-screen text. Specifically, we evaluate the advanced Qwen2.5-Omni-7B (Xu et al., 2025a) on recognizing subtitles for correspondence tasks, speaker names and step markers for complementary tasks, and rolling text for independent tasks. For evaluation metrics, we report WER for long-form text (e.g., subtitles and rolling text) and exact-match accuracy for short-form text (e.g., speaker name and step marker). As shown in Table 3, the model performs well on subtitles and step markers but worse on speaker names and rolling text. We further analyze the failure cases of text recognition and find that speaker-name errors mainly involve omissions, hallucinated non-existent names, and incorrect ordering. On the other hand, the extremely high WER (higher than 100%) in recognizing rolling text in independent tasks is largely caused by transcribing spoken audio as scrolling text, leading to many insertion errors. Note that Gemini 3.8 Flash achieves strong performance in our AVTextBench(Table 2), indicating that the on-screen text in our benchmark is generally recognizable. These results highlight that open-source models still have room for improvement in diverse text recognition.

Table 3: On-screen text recognition performance. *Metrics:* Acc. = Accuracy; WER = Word error rate.

Tasks	Text Type	Metric	
		WER ↓	Acc. ↑
Correspondence	Subtitle	18.24	-
Independent	Rolling text	338.12	-
Complementary	Speaker name	-	66.98
	Step marker	-	93.15

Do audio-visual LLMs recognize speech content in the audio? To answer this question, we investigate the speech recognition ability of audio-visual LLMs by directly asking models to transcribe the speech content in the entire video for correspondence and independent tasks or within a specified temporal interval for complementary tasks. We use the widely used Whisper model (Radford et al., 2023) as the pseudo-reference transcript. We then compute the WER between predictions of Qwen2.5-Omni-7B (Xu et al., 2025a) and the reference transcript.

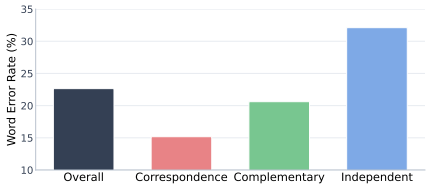

Figure 3: Speech recognition performance across task categories.

Figure 3 shows that the model exhibits substantially different speech recognition capabilities across task types, with particularly pronounced degradation in the independent setting. When on-screen text is unrelated to the speech, models appear more susceptible to interference from visual textual information, indicating that irrelevant on-screen text can mislead the recognition of spoken content. This finding underscores that robust speech recognition in audio-visual LLMs requires more than recognizing speech in isolation: models must maintain reliable recognition across different video contexts while effectively segregating speech from potentially conflicting or irrelevant on-screen text.

Can textualized modality information improve cross-modal understanding? In audio-visual LLMs, on-screen text and speech content are input to the model by visual and audio tokens. Since models can process text tokens more effectively than audio and video tokens, how about input on-screen text and speech transcripts in text format to the model, thus mapping them to text tokens? Specifically, we keep the original audio and

Table 4: Results with additional textual inputs.

Input	Corr.	Comp.	Indep.
Base.	45.54	69.41	19.81
+ Spee.	44.68	67.32	**33.96**
+ On-Sc.	**45.78**	68.51	13.21

video input, and include either speech transcript or on-screen text in text format. Table 4 shows that textualizing either modality brings little improvement on the correspondence and complementary tasks, suggesting that their difficulty of tasks and require more strong cross-modal ability. The independent task behaves differently because it explicitly requires source attribution between spoken audio and on-screen text. Under this setting, converting one source into textual input can bias the model toward that source: adding the speech transcript improves performance, while adding on-screen text degrades it. These results further highlight the necessity to improve models' cross-modal abilities in understanding the relationship between on-screen text and audio-visual streams.

4 Related Work

Multimodal large language models (MLLMs). MLLMs connect a pretrained vision perception encoder to an LLM, and have been used across multiple applications (Lee et al., 2025; Kim et al., 2026). LLaVA (Liu et al., 2023) is an early instance of this design, projecting CLIP (Radford et al., 2021) image features into the token embedding space of an instruction-tuned language model. Audio-visual LLMs add an audio encoder so that the language model receives both vision and audio tokens, as in VideoLLaMA 2 (Cheng et al., 2024), and video-SALMONN 2 (Tang et al., 2025). In this work, we focus on investigating whether audio-visual LLMs could identify the relationship between on-screen text and audio-visual streams.

Benchmarks for video OCR understanding. Recent studies have proposed benchmarks to understand the visual text in diverse scenarios: video text spotting (Wu et al., 2021), question answering that combines visual and textual cues (Zhao et al., 2022; Jahagirdar et al., 2023), driving and egocentric footage (Tom et al., 2023; Zhou et al., 2025a). Specificially, they evaluate diverse abilities, such as spatial and temporal localization and cross-frame integration (Fei et al., 2025; Shi et al., 2025). These benchmarks largely formulate text understanding as a vision-only problem, where audio is not included both as input and for answering the questions. In contrast, our work includes the audio modality and investigates the model's ability to ground on-screen text in audio-visual context, an important yet unexplored setting in previous work.

Benchmarks for audio-visual understanding. Several benchmarks test whether audio-visual LLMs use the audio and the visual input together. (Gong et al., 2024) show that in their AV-Odyssey Bench, these models often struggle to tell which of two sounds is louder or higher in pitch; answering each multiple-choice problem requires clues from both the visual and the audio input. WorldSense (Hong et al., 2026) designs its evaluation tasks so that audio and video are strongly coupled, and DAVE (Radevski et al., 2025) makes both modalities necessary by construction and reports results per atomic subcategory. AVHBench (Sung-Bin et al., 2025) measures cross-modal hallucination between audio and visual signals, and AVTrustBench (Chowdhury et al., 2025) covers adversarial attacks, compositional reasoning, and dependence on a single modality. Daily-Omni Zhou et al. (2025b) focuses on the temporal understanding of audio-visual modalities. In contrast, our work focuses on the cross-modal relationship between on-screen text and audio-visual streams, which is unexplored in previous studies.

5 Conclusion

In this work, we investigate whether audio-visual LLMs can understand the cross-modal relationship between on-screen text and the recorded audio-video stream, an important yet unexplored ability for understanding real-world videos. To systematically evaluate this ablity, we introduce AVTextBench, a benchmark that study the relationship between on-screen text and audio-visual streams from three fundamental perspectives: corresponding, complementary, and independent, which evaluate AVTextBench evaluates whether models can correctly match, integrate, and segregate information across on-screen text and audio-visual content. Experimental results show that even the state-of-the-art open-source audio-visual LLMs perform poorly on our benchmark, while Gemini-3.8-Flash model achieves strong performance. Our work highlights an important yet unexplored limitation of large-language models, and offers new insights for improve future models.

AI USE STATEMENT

In this work, we have not used generative AI tools to develop research ideas or design experiment setups. We have used generative AI tools to assist with paper writing, primarily for polishing purposes, such as improving sentence structure and phrasing, after preparing detailed drafts ourselves. We have also used generative AI tools to assist with coding required for benchmark construction and model evaluation. We have reviewed all AI-assisted work in both writing and coding. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Jianye Hou, Bo Li, Guanbin Li, Ser-Nam Lim, Harry Yang, and Benyou Wang. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3492–3500, June 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. In *ICCV*, pp. 1590–1601, October 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

Yulin Fei, Yuhui Gao, Xingyuan Xian, Xiaojin Zhang, Tao Wu, and Wei Chen. Do current video LLMs have strong OCR abilities? a preliminary study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9860–9876, January 2025. URL https://aclanthology.org/2025.coling-main.659/.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024. URL https://openreview.net/forum?id=xx3DCKbt1T.

Google. Gemini 3.8 flash. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FZv3kPHTtB.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *ICLR*, 2026. URL https://openreview.net/forum?id=YxsfxAvJv4.

Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4441–4450, January 2023.

Kyeong Seon Kim, Baek Seong-Eun, Lee Jung-Mok, and Tae-Hyun Oh. mEOL: Training-free instruction-guided multimodal embedder for vector graphics and image retrieval. In *WACV*, pp. 1191–1200, March 2026.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, et al. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *arXiv preprint arXiv:2410.11619*, 2025.

Jung-Mok Lee, Nam Hyeon-Woo, Moon Ye-Bin, Junhyun Nam, and Tae-Hyun Oh. Automated model discovery via multi-modal & multi-step pipeline. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=qGFvTIMS3W.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

openAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *NeurIPS Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=4ZAX1NT0ms.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yi-Fan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, Xinlong Chen, Bohan Zeng, Sihan Yang, Yuanxing Zhang, Pengfei Wan, Haotian Wang, and Wenjing Yang. Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios, 2025. URL https://arxiv.org/abs/2505.21333.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47198–47217, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sun24l.html.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *ICLR*, 2025. URL https://openreview.net/forum?id=jTEKTdI3K9.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. URL https://arxiv.org/abs/2506.15220.

George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and C. V. Jawahar. Reading between the lanes: Text videoqa on the road. In *Document Analysis and Recognition - ICDAR 2023*, pp. 137–154, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-41731-3.

Weijia Wu, Yuanqiang Cai, Debing Zhang, Sibo Wang, Zhuang Li, Jiahong Li, Yejun Tang, and Hong Zhou. A bilingual, openworld video text dataset and end-to-end video text spotter with transformer. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/b6d767d2f8ed5d21a44b0e5886680cb9-Paper-round2.pdf.

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 technical report. *arXiv preprint arXiv:2503.20215*, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b. URL https://arxiv.org/abs/2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang

Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, Aobotao Dai, and Shuigeng Zhou. Towards video text visual question answering: Benchmark and baseline. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35549–35562. Curran Associates, Inc., 2022. doi: 10.52202/068431-2576. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e726197ffd401df4013cd9f81007b5cf-Paper-Datasets_and_Benchmarks.pdf.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3363–3373, June 2025a.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2025b. URL https://arxiv.org/abs/2505.17862.

A LIMITATION

Our work has two main limitations. First, AVTextBench currently focuses on English-language videos and therefore does not evaluate multilingual or cross-lingual on-screen text grounding. Extending the benchmark to other languages would enable a broader assessment of whether current audio-visual LLMs can reliably ground on-screen text across diverse linguistic settings. Second, our benchmark primarily considers minute-level videos and does not cover hour-scale long-form content. Such hour-scale videos poses additional challenges for cross-modal grounding of on-screen text. Extending AVTextBench to long-form videos would therefore be a valuable direction for future work.

B AUDIO-VISUAL LLMS FOR EVALUATION

VideoLLaMA2. VideoLLaMA2 (Cheng et al., 2024) introduces a Spatial-Temporal Convolution (STC) connector to effectively capture the complex spatial and temporal dynamics of videos. STC connector, together with an MLP, bridges the vision encoder and the LLM, while an additional MLP connects the audio encoder to the LLM. With this structure, the model is trained through a multi-stage training strategy, progressing from vision-centric training to audio-centric training and finally to joint audio-visual training, enabling comprehensive understanding of both modalities. For evaluation, we use VideoLLaMA2 with CLIP (Radford et al., 2021) as the vision encoder, BEATs (Chen et al., 2023) as the audio encoder.

Video-SALMONN. Video-SALMONN (Sun et al., 2024) introduces a Multi-Resolution Causal Q-Former (MRC Q-Former) and a diversity loss to jointly understand general audio, speech, and visual modalities. MRC Q-Former aligns audio-visual features with the text representation space at three different temporal scales, enabling the model to effectively capture temporal causal relationships among events in a video. Meanwhile, the diversity loss used during training encourages the model to capture diverse audio-visual features present in the video. For evaluation, we use Video-SALMONN with Vicuna-v1.5-13B (Chiang et al., 2023) as the LLM, Whisper large-v2 (Radford et al., 2023) as the speech encoder, BEATs as the audio encoder, and the Vision Transformer and Q-Former from InstructBLIP (Dai et al., 2023) as the vision encoder.

Video-SALMONN2+. Video-SALMONN2+ (Tang et al., 2025) introduces Multi-Round Direct Preference Optimization (MrDPO), a training strategy designed to equip existing video LLMs with audio-visual understanding capabilities. Specifically, MrDPO periodically updates the reference policy used in the DPO (Rafailov et al., 2023) objective, enabling continual improvement over multiple rounds of preference optimization. For evaluation, we use the Video-SALMONN2+ series built upon Qwen2.5-VL (Bai et al., 2025b) series, both of which employ the audio encoder from Whisper large-v3 (Radford et al., 2023).

Qwen2.5-Omni. Qwen2.5-Omni (Xu et al., 2025a) adopts a time-interleaving strategy that organizes audio and visual representations according to their temporal order, and further introduces a novel positional encoding approach, Time-aligned Multimodal Rotary Position Embedding (TMRoPE). TMRoPE encodes the spatial and temporal positions of video representations while aligning their temporal positions with those of audio representations, enabling the model to effectively capture temporal relationships across audio and visual modalities. For evaluation, we use Qwen2.5-Omni-3B and Qwen2.5-Omni-7B, whose LLM components are initialized from Qwen2.5 (Qwen et al., 2025), while their vision encoders follow Qwen2.5-VL and their audio encoders are initialized from Whisper large-v3.

Qwen3-Omni-Instruct. Qwen3-Omni-Instruct (Xu et al., 2025b) adopts an Audio Transformer (AuT), trained on large-scale supervised audio data, as its audio encoder to leverage strong audio representations. Similar to Qwen2.5-Omni, it also employs TMRoPE to effectively model audio-visual representations. For evaluation, we use Qwen3-Omni-30B-A3B-Instruct, whose LLM is initialized from Qwen3 (Yang et al., 2025) and paired with the vision encoder from Qwen3-VL (Bai et al., 2025a) and the AuT audio encoder.

C COMPLETE PROMPT FOR EVALUATION

In this section, we provide the prompt templates used to evaluate each task in our benchmark. Each template specifies the task instruction and the corresponding answer options. Figures 4–11 present the prompt templates for the respective tasks. The gray text at the bottom of each template describes the criteria used to select the answer options. For the Text-to-Event and Event-to-Text Grounding tasks, we present separate prompt templates for the step markers and speaker name settings.

Alignment Verification Prompt Template for Correspondence

Do the on-screen subtitles correctly correspond to the spoken content throughout the video? Ignore minor timing offsets and minor wording differences that preserve the original meaning. Mismatch includes cases where subtitle content changes the meaning of the corresponding speech or where subtitle content from another spoken segment is displayed at the wrong time. Choose one option and return only its letter.

A. {option A}

B. {option B}

The two options are randomly assigned to A and B without replacement:

- The subtitles correctly correspond to the spoken content throughout the video.
- There is a mismatch between the subtitles and the spoken content.

Figure 4: Prompt template used to evaluate the Alignment Verification task.

Misalignment Localization Prompt Template for Correspondence

When does the mismatch between the on-screen subtitles and the spoken content occur? Ignore minor timing offsets and minor wording differences that preserve the original meaning. Mismatch includes cases where subtitle content changes the meaning of the corresponding speech or where subtitle content from another spoken segment is displayed at the wrong time. Choose one option and return only its letter.

A. {option A}

B. {option B}

C. {option C}

The incorrect options are selected from intervals containing non-matching speech.

Figure 5: Prompt template used to evaluate the Misalignment Localization task.

Misalignment Type Classification Prompt Template for Correspondence

What type of mismatch exists between the on-screen subtitles and the spoken content? Choose one option and return only its letter.

A. {Option A}

B. {Option B}

The two options are randomly assigned to A and B without replacement:

- Permutation: The subtitle content belongs to another spoken segment in the same video and is displayed at the wrong time.
- Substitution: The subtitle content differs in meaning from the corresponding speech.

Figure 6: Prompt template used to evaluate the Misalignment Type Classification task.

Text-to-Event Grounding Prompt Template for Complementary: Step Markers

Which of the following statements is spoken during {step}? Choose one option and return only its letter.

A. {opion A}

B. {option B}

C. {option C}

The incorrect options are selected from statements of steps containing non-matching statements.

Figure 7: Prompt template used to evaluate the Text-to-Event Grounding task with step markers.

Event-to-Text Grounding Prompt Template for Complementary: Step Markers

Which step contains the following spoken content: {speech}? Choose one option and return only its letter.

A. {opion A}

B. {option B}

C. {option C}

The incorrect options are selected from steps containing non-matching speech.

Figure 8: Prompt template used to evaluate the Event-to-Text Grounding task with step markers.

Text-to-Event Grounding Prompt Template for Complementary: Speaker Name

Which of the following statements is spoken by {name}? Choose one option and return only its letter.

A. {opion A}

B. {option B}

C. {option C}

The incorrect options are selected from statements not spoken by the correct speaker.

Figure 9: Prompt template used to evaluate the Text-to-Event Grounding task with speaker name.

Event-to-Text Grounding Prompt Template for Complementary: Speaker Name

Which speaker said the following statement? {statement} Choose one option and return only its letter.

A. {opion A}

B. {option B}

C. {option C}

The incorrect options are selected from speech without explicitly mentioned speaker names.

Figure 10: Prompt template used to evaluate the Event-to-Text Grounding task with speaker name.

Source Attribution Prompt Template for Independent
Which source presents this information: {information} Choose one option and return only its letter.
A. Spoken audio.
B. Bottom scrolling text.
C. Both spoken audio and bottom scrolling text.
D. Neither source.

Figure 11: Prompt template used to evaluate the Source Attribution task.