FATE: A Faithful Audio-Visual Question Answering Benchmark with Modality-Specific Temporal Grounding

Anonymous ACL submission

Abstract

Recent MLLMs have rapidly advanced toward omni-modal reasoning over visual, textual, audio, and speech signals, motivating benchmarks for audio-visual event understanding. However, existing benchmarks mainly rely on whether a model selects the correct answer, leaving open *why* it was produced. As a result, high accuracy does not guarantee faithful audio-visual reasoning: models may exhibit a *Clever Hans effect*, relying on visual priors, language bias, answer-choice artifacts, or incidental cues rather than the modality-specific evidence that supports the answer. In this paper, we introduce **FATE**, a **F**aithful **A**udio-Visual Question Answering Benchmark with Modality-Specific **TE**mporal Grounding. **FATE** follows three principles: *modality-specific evidence grounding*, requiring separate visual and audio evidence windows; *decisive audio reasoning*, where audio provides necessary evidence; and *counterfactual evaluation*, which tests rejection of unsupported conclusions under absent, inconsistent, mismatched, or incorrectly attributed evidence. The benchmark contains 500 videos and 1,235 QA pairs across 7 major categories and 19 subcategories. We further propose a novel metric, *EG-Score*, which jointly evaluates answer correctness and grounding in visual and audio evidence. We demonstrate that most models still struggle to justify their answers with correct modality-specific evidence, even when they achieve high accuracy.

1 Introduction

Recent Multimodal Large Language Models (MLLMs) have advanced from image- and video-centric perception toward omni-modal frameworks (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Lin et al., 2024; Zhang et al., 2023; Cheng et al., 2024; Team et al., 2023) that jointly process visual, textual, audio, and speech signals. This progress has motivated audio-visual and omni-modal benchmarks (Yun et al., 2021; Li et al., 2025b; Gong et al., 2024) that evaluate whether models can understand events from audio-visual inputs, often emphasizing cross-modal alignment (Chen et al., 2023; Wang et al., 2022; Chowdhury et al., 2024), modality complementarity (Yang et al., 2022), or strong audio-video dependency (Chowdhury et al., 2025).

As MLLMs move toward omni-modal reasoning, evaluation must therefore ask not only whether models can process multiple modalities, but whether they use the right evidence from each modality. However, most benchmarks still evaluate models primarily through final answer accuracy. This leaves a critical ambiguity: a correct answer may be produced for the wrong reason. In particular, a model can exploit visual priors, linguistic regularities, answer-choice artifacts, or incidental single-modality cues without grounding its prediction in the evidence that actually supports the answer, and a single global accuracy cannot distinguish them. Such behavior resembles a *Clever Hans effect* (Kauffmann et al., 2025), *i.e.*, the output appears correct, but the underlying reason is unreliable. Thus, to evaluate faithful reasoning, benchmarks need to specify what evidence justifies each answer, separately for vision and audio.

In this work, we introduce **FATE**, a **F**aithful **A**udio-Visual Question Answering Benchmark with Modality-Specific **TE**mporal Grounding. Each sample is designed to connect answer prediction with the evidence that supports it, rather than treating the answer as the only evaluation target. Specifically, FATE is built around three principles that correspond to the limitations in previous audio-visual benchmarks as illustrated in Figure 1.

First, *modality-specific evidence grounding* annotates separate visual and audio evidence windows. This is necessary because the evidence in the two modalities may not appear at the same time or play the same role. As described in the left example, a faithful model should identify both

Figure 1: **Motivation of FATE.** FATE targets three challenges in audio-visual reasoning: modality-specific evidence grounding, decisive audio reasoning, and counterfactual rejection under false premises or missing evidence.

the relevant visual segment and the relevant audio segment for the answer.

Second, *decisive audio reasoning* focuses on cases where audio provides necessary evidence that cannot be replaced by visual context alone. For instance, in the middle example, the visual stream may show a person making an expression or performing an action, but the event outcome is determined by an audio. These examples test whether a model treats audio as evidence that can decide the answer, rather than as optional context appended to visual understanding.

Third, *counterfactual evaluation* tests whether models reject conclusions when the required evidence is absent, mismatched, temporally invalid, or incorrectly attributed. For instance, in the right example of Figure 1, the visual frames may suggest that a woman is shouting, even though the corresponding shouting sound is absent from the audio. In this setting, the correct behavior is not to infer the most plausible event from the image alone, but to reject the unsupported conclusion based on the actual audio-visual evidence.

Further, to evaluate these abilities in a unified way, we suggest *EG-Score*, which measures whether a model not only answers correctly but also grounds the answer in the appropriate visual and audio evidence. This allows us to distinguish correct answers supported by faithful evidence from the answers produced with incorrect or missing grounding. Our experiments show that most models often achieve reasonable answer accuracy, yet struggle to justify their predictions with the modality-specific evidence required by the question.

To summarize, our contributions are threefold:

- We introduce **FATE**, a Faithful Audio-Visual Question Answering Benchmark with Modality-specific Temporal Grounding.
- We construct decisive audio and counterfactual examples to evaluate whether models use necessary audio evidence and reject unsupported audio-visual assumption.
- We propose *EG-Score* and reveal that current models often struggle to provide faithful modality-specific evidence even when their answers are correct.

Benchmark	A+V	Avg. Dur.(s)	# QA	# Cat.	Evidence	Modality-specific Evidence	Decisive Audio Reasoning	Counterfactual Eval.
AVQA (Yang et al., 2022)	✓	10	57.3k	8	✗	✗	✗	✗
MUSIC-AVQA (Li et al., 2022)	✓	60.0	45.8k	9	✗	✗	✗	✗
WorldSense (Hong et al., 2025)	✓	141.1	3,172	26	✗	✗	✗	✗
Daily-Omni (Zhou et al., 2025b)	✓	–	1,197	6	✗	✗	✗	✗
OmniVideoBench (Li et al., 2025a)	✓	384.2	1,000	13	✗	✗	✗	✗
JointAVBench (Chao et al., 2025)	✓	97.2	2,853	15	✗	✗	✗	✗
DAVE (Radevski et al., 2025)	✓	–	2,426	3	✗	✗	✗	✗
TVQA+ (Lei et al., 2020)	✗	–	29.4k	–	✓	✗	✗	✗
CG-Bench (Chen et al., 2025)	✓	22.8	12,129	14	✓	✗	✗	✗
Perception Test (Patraucean et al., 2023)	✓	23	44k	32	✓	✗	✗	✓
COVER (Zhou et al., 2025a)	✗	–	2,923	13	✗	✗	✗	✓
FATE(OURS)	✓	283.2	1,235	19	✓	✓	✓	✓

Table 1: **Comparison with audio-visual reasoning benchmarks.**

2 Related Works

Multimodal and Omni-modal LLMs. Large language models have evolved from text-only systems with strong instruction following and reasoning abilities (Vaswani et al., 2017; Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Achiam et al., 2023) to multimodal models that connect language models with visual encoders for image understanding and reasoning (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Liu et al., 2023). Subsequent work extended this paradigm from images to videos, requiring models to capture temporal dynamics and event-level reasoning (Zhang et al., 2023; Lin et al., 2024; Cheng et al., 2024), and further incorporated audio and speech for audio-language and audio-visual understanding (Girdhar et al., 2023; Chen et al., 2023; Ghosh et al., 2025). More recently, omni-modal models aim to unify text, image, video, audio, and speech within a single interaction framework, enabling integrated reasoning over heterogeneous signals (Team et al., 2023; Xu et al., 2025). Despite this progress, evaluating such models remains challenging: answer-only accuracy cannot reveal whether a model uses the correct modality-specific evidence or relies on shortcuts such as visual priors, language bias, or incidental cues.

Audio-Visual multimodal Benchmarks. With the development of MLLMs, audio-visual and omni-modal benchmarks have evaluated increasingly diverse multimodal reasoning. AVQA (Yang et al., 2022) and MUSIC-AVQA (Li et al., 2022) focus on associating visual events with sounds, while WorldSense (Hong et al., 2025), Daily-Omni (Zhou et al., 2025b), and OmniVideoBench (Li et al., 2025a) extend to real-world omni-modal videos; JointAVBench (Chao et al., 2025) and DAVE (Radevski et al., 2025) further stress audio-video dependency to reduce unimodal shortcuts. However, as summarized in Table 1, these benchmarks largely measure final answers and lack temporal evidence annotations, especially modality-specific windows identifying which audio and visual moments justify an answer.

Other benchmarks add grounding or counterfactual evaluation: TVQA+ (Lei et al., 2020) and CG-Bench (Chen et al., 2025) provide evidence/clue annotations, the Perception Test (Patraucean et al., 2023) includes localization tasks, and COVER (Zhou et al., 2025a) tests counterfactual reasoning. Yet they are not fully suited to faithful audio-visual reasoning, as they focus mainly on visual evidence, provide task-level rather than QA-specific grounding, or fail to disentangle audio and visual evidence. FATE fills this gap by pairing audio-visual QA with separate visual/audio evidence windows, decisive-audio questions, and counterfactual cases to test whether models answer for the right modality-specific reasons.

3 FATE

In this section, we introduce FATE, a diagnostic benchmark for audio-visual reasoning. By separately annotating visual and audio evidence windows, constructing decisive audio categories, and introducing counterfactual categories, FATE evaluates whether models ground their answers in the appropriate modality-specific evidence.

3.1 Benchmark Design

FATE follows a grounding-first objective. The model performance should be judged by both answer correctness and evidence support. We instantiate this objective through three principles: Modality-specific Evidence Grounding, Decisive Audio Reasoning, and Counterfactual Evaluation.

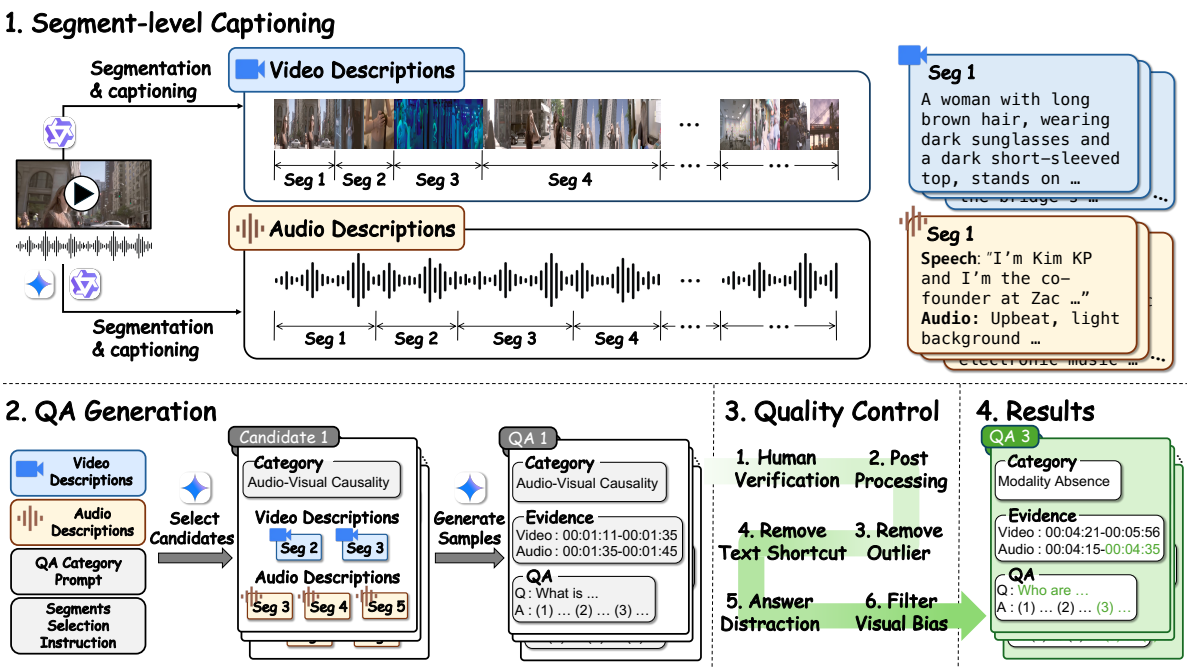


Figure 2: **Overview of FATE construction.** FATE constructs audio-visual QA pairs through segment-level captioning, QA generation, and multi-stage quality control.

Modality-specific Evidence Grounding. Each question is independently annotated with a visual evidence window and an audio evidence window that support the correct answer. These windows cover the modality-specific evidence required by the question and are not required to overlap, reflecting realistic cases where relevant audio and visual evidence appear at different moments.

Decisive Audio Reasoning. Decisive Audio Reasoning reduces single-modality shortcuts by requiring audio evidence to determine the final answer, while visual information provides scene context. We divide this component into three subcategories: State&Material Inference, where sound reveals an object's material or physical state; Fine-grained Action Inference, where audio distinguishes visually similar actions with different intensity or quality; and Unseen Outcome Verification, where audio confirms an outcome not visible in the scene.

Counterfactual Evaluation. Counterfactual Evaluation assesses whether models can recognize unsupported conclusions when the required evidence is missing, contradicted, temporally inconsistent, or incorrectly attributed. To this end, we construct these cases along two axes: Modality Absence and Invalid Assumption. These settings convert answerable questions into cases where the provided evidence does not support one of the standard answer choices. All QA pairs include a "None of the above" option to evaluate whether models reject unsupported answers.

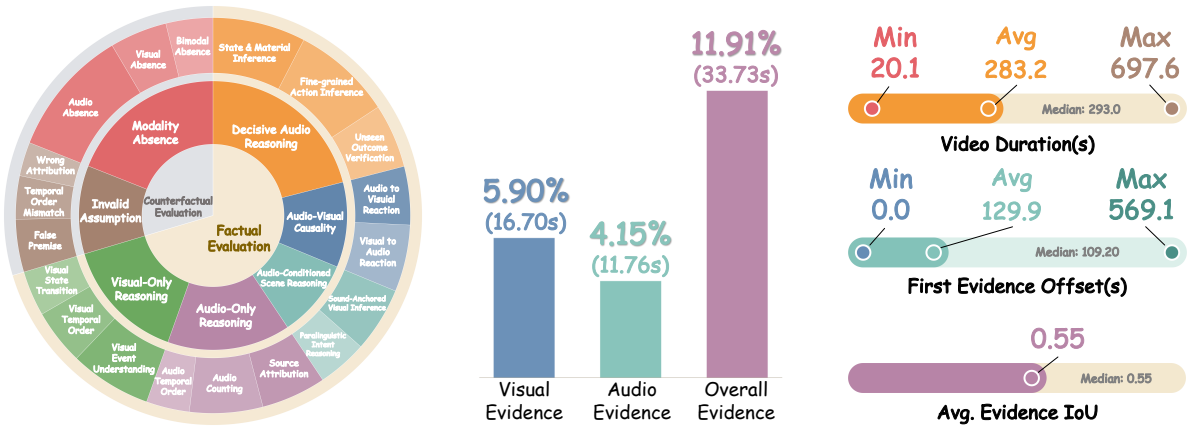
3.2 Benchmark Construction

Figure 2 summarizes the construction pipeline of FATE. Starting from raw videos, we separately caption the visual and audio streams at the segment level, generate QA samples, and conduct multi-stage quality control.

3.2.1 Segment-level Captioning

We first collect 1,000 raw videos from two sources: 750 videos from VideoMME (Fu et al., 2025) and 250 videos from Ego4D (Grauman et al., 2022) to cover diverse source video domains. Then, each video is partitioned into visual and audio segments. We use PySceneDetect (Castellano) to detect visual scene transitions and MFCC (Davis and Mermelstein, 1980) to identify intervals with salient acoustic changes.

Given each segment, we generate modality-specific descriptions. Specifically, Qwen3-VL (Bai et al., 2025) describes visual content, Gemini 3.1 Flash (Team et al., 2023) describes acoustic cues, and Qwen3-ASR (Shi et al., 2026) extracts speech transcripts when speech is present. These outputs serve as structured inputs for evidence selection and QA generation.



(a) QA Category Distribution (b) Evidence Length Ratio (c) Temporal Statistics

Figure 3: **Statistics of FATE.** We report the QA category distribution, evidence length ratio, and temporal statistics, showing that FATE requires models to localize sparse modality-specific evidence in long audio-visual inputs.

3.2.2 QA Generation

Given the segment-level descriptions and transcripts, we prompt Gemini 3.1 Pro to select category-specific candidate evidence segments. For each video and category, the model selects up to five candidate sets, each of which may contain multiple visual and audio segments. The candidate evidence window is defined by aggregating the selected modality-specific segments.

For each candidate evidence window, we generate a multiple-choice QA pair conditioned on the target category and selected evidence. Each question includes one correct answer, three similar distractors, and one "`None of the above`" option.

For Counterfactual Evaluation, samples are constructed in two ways. Some are generated directly from videos where the required evidence is naturally missing or incorrectly attributed. The rest are created by post-processing the previously generated QA pairs: Bimodal Absence constructs the input from non-evidence segments, excluding the annotated visual and audio evidence windows, Temporal Order Mismatch reverses the assumed temporal order, and Wrong Attribution reassigns the referenced source or target.

3.2.3 Quality Control

We apply multi-stage quality control to improve answerability, category consistency, and shortcut resistance. First, trained annotators verify each QA candidate for answer correctness, evidence validity, distractor quality, and category consistency. Minor evidence-boundary errors are corrected, and final approval is determined by majority vote.

We then post-process the approved QA pairs by shuffling answer options while keeping the "`None of the above`" option fixed. Each video is randomly cropped to include the evidence windows, with crop boundaries sampled within a 300-second margin to reduce temporal-position shortcuts.

Finally, we apply shortcut filtering to obtain high-quality samples that require the intended evidence. Specifically, we remove evidence-window outliers above the 99th percentile to exclude imprecisely grounded samples. We also discard text-only shortcut samples answerable from the question and choices alone, and visual-only shortcut samples answerable from visual evidence alone using Gemini 3.1 Pro, except for Visual-Only Reasoning. After this quality-control process, the final benchmark contains 1,235 QA pairs from 500 videos.

3.3 Statistics

Category Taxonomy. The taxonomy of FATE consists of two groups: Grounded Evaluation and Counterfactual Evaluation. Grounded Evaluation assesses whether models can answer questions based on available audio-visual evidence. It includes 1) Decisive Audio Reasoning, where audio provides the decisive cue, 2) Audio-Visual Causality, which tests causal relations between auditory and visual events, and 3) Audio-Conditioned Scene Reasoning, where audio cues must be interpreted with the visual scene. It also includes 4) Audio-Only Reasoning and 5) Visual-Only Reasoning, where the reasoning itself is performed within a single modality, with the other modality rarely contributes beyond minor context. Counterfactual Evaluation assesses whether models can recognize unsupported questions when the required evidence

is absent or invalid. It consists of 6) Modality Absence, where the necessary modality-specific evidence is missing, and 7) Invalid Assumption, where the question relies on a contradicted premise, incorrect temporal order, or wrong attribution. The full category hierarchy and sample distribution are described in Figure 3a and table C.1.

Data Statistics FATE contains 500 videos and 1,235 QA pairs across 7 major categories and 19 subcategories. It includes 868 Grounded Evaluation samples and 367 Counterfactual Evaluation samples, with Decisive Audio Reasoning forming the largest grounded category. Figures 3b and 3c shows that FATE requires temporally grounded reasoning over long videos. The average video duration is 283.2 seconds, while the average overall evidence window spans only 33.73 seconds, corresponding to just 11.91% of the video. This sparse evidence distribution makes evidence localization challenging for models. In addition, the first evidence segment appears at 129.9 seconds on average, ranging from 0 to 569.1 seconds. The average IoU between visual and audio evidence windows is 0.55, showing that evidence from the two modalities is not always temporally aligned. This provides empirical support for the validity of modality-specific grounding in our benchmark.

4 Experiments

4.1 Evaluation Settings

Task Formulation. FATE evaluates models on a joint answer and grounding task that reflects the benchmark's grounding-first objective. Given a video, a question, and five answer choices including `None of the above` option, the model produces three outputs: a predicted answer, a visual evidence window, and an audio evidence window. Either window could be set to `Null`, indicating that the modality is either unnecessary for reasoning or lacks supporting evidence. Similarly, ground-truth windows are defined as either a specific temporal window or `Null`, following Section 3.2.

Evaluation Metrics. We adopt the two complementary metrics. The standard **Answer Accuracy (Acc)** measures the proportion of samples whose predicted answer matches the ground truth:

$$\text{Acc} = \frac{1}{N}\sum_{q=1}^{N} \mathbb{1}_{[\hat{y}_q = y_q]}, \tag{1}$$

where N is the number of samples, $\hat{y}_q$ is the predicted answer for sample q, and y_q is the ground-truth answer. However, Acc cannot reveal whether a correct answer is from correct reasoning or from visual priors and modality shortcuts.

To address this, we further propose **Evidence-Grounded Score (EG-Score)**, which couples answer correctness with the modality-specific localization, as

$$\text{EG-Score} = \tfrac{1}{N}\textstyle\sum_{q=1}^{N} \mathbb{1}_{[\hat{y}_q = y_q]}\, \text{HM}(I_{q,v}, I_{q,a}) \tag{2}$$

$$I_{q,m} = \begin{cases} \text{IoU}(\hat{e}_{q,m}, e_{q,m}), & \hat{e}_{q,m} \neq \text{Null},\ e_{q,m} \neq \text{Null}, \\ 1, & \hat{e}_{q,m} = e_{q,m} = \text{Null}, \\ 0, & \text{otherwise}. \end{cases}, \tag{3}$$

where HM denotes the harmonic mean, q indexes each question, and $m \in \{v, a\}$ denotes the visual or audio modality. For each modality, $I_{q,m}$ measures the match between the predicted evidence $\hat{e}_{q,m}$ and the ground-truth evidence $e_{q,m}$. It is computed as temporal IoU when both are temporal windows, set to 1 when both are `Null`, and set to 0 when only one side is `Null`. Thus, incorrect answers receive zero, while correct answers are scored by the harmonic mean of visual and audio evidence matches, preventing grounding in one modality from compensating for failure in the other. We also report per-category accuracy over the 7 major categories and modality-specific grounding mIoU for the visual and audio modalities.

Baselines. We evaluate a diverse set of MLLMs on FATE, organized into four groups. (i) **Closed-source Omni MLLMs**: Gemini3.1-Pro and Gemini3.1-Flash (Team et al., 2023), which jointly process video and audio inputs. (ii) **Open-source Omni MLLMs**: Qwen3-Omni (Xu et al., 2025), video-SALMONN 2+ (Tang et al., 2025), Phi-4-multimodal (Abdin et al., 2024), Gemma4 (Google DeepMind, 2026), and VideoLLaMA2.1-AV (Cheng et al., 2024). (iii) **Video-only MLLMs**: Qwen3-VL (Bai et al., 2025), VideoLLaMA2, and Qwen3-Omni-Only-Video, and ablated variant of the omni model in which the audio input is removed. (iv) **Audio-only MLLMs.**: Audio Flamingo 3 (Ghosh et al., 2025), which operates solely on audio input, and Qwen3-Omni-Only-Audio, an audio-only ablated variant of the omni model. In addition, we report a **Random Baseline** that samples the answers uniformly.

Method	Param.	# Frames	Factual Eval.					Counterfactual Eval.		mIoU		Overall	
			Decisive Audio Reasoning	Audio-Visual Causality	Audio-Cond. Scene Reasoning	Audio-Only Reasoning	Visual-Only Reasoning	Modality Absence	Invalid Assumption	Visual	Audio	Acc	EG-Score
Random	–	–	19.85	24.60	18.25	21.47	21.51	17.36	17.60	–	–	19.92	–
Closed-source Omni MLLMs													
Gemini3.1-Pro	–	128	**71.54**	75.40	**81.75**	59.89	**78.49**	**53.72**	**51.20**	26.52	**28.91**	**66.72**	**15.76**
Gemini3.1-Flash	–	128	68.91	**81.75**	76.98	**60.45**	74.42	20.66	48.00	**30.46**	25.11	59.03	14.06
Open-source Omni MLLMs													
Qwen3-Omni	30B	128	**58.43**	**59.52**	**70.63**	**45.76**	**63.95**	**31.40**	**37.60**	**5.71**	5.68	**51.34**	**2.67**
video-SALMONN 2+	7B	128	43.82	46.03	54.76	33.33	50.00	0.41	26.40	3.70	3.39	34.25	0.49
Phi-4-multimodal	6B	128	43.07	49.21	49.21	29.38	37.21	4.55	26.40	3.46	5.92	32.31	0.23
Gemma-4-E4B	4B	128	37.45	34.13	50.79	25.99	50.00	**31.40**	32.80	2.03	2.82	36.92	0.50
VideoLLaMA2.1-AV	7B	32	35.58	34.92	38.10	28.25	32.56	19.01	27.20	2.65	**7.42**	30.20	0.41
Video-only MLLMs													
Qwen3-Omni-Only-Video	30B	128	**47.19**	**40.48**	**66.67**	32.77	**65.70**	**33.88**	**37.60**	5.05	–	**45.43**	–
Qwen3-VL	8B	128	41.20	32.54	45.24	**38.42**	60.47	9.50	22.40	**15.14**	–	34.90	–
VideoLLaMA2	7B	32	26.59	16.67	24.60	29.38	22.67	0.00	13.60	3.28	–	18.70	–
Audio-only MLLMs													
Qwen3-Omni-Only-Audio	30B	–	**53.93**	**49.21**	**54.76**	**41.81**	20.93	**58.26**	**31.20**	–	**6.51**	**45.75**	–
Audio Flamingo 3	7B	–	47.57	42.86	46.83	35.03	**25.58**	4.96	20.00	–	3.52	31.01	–

Table 2: **Performance comparison across omni-modal, video-only, and audio-only MLLMs.**

Implementation Details. All open-source models are evaluated on a single node with 4×NVIDIA RTX A6000 (48GB) GPUs, while closed-source models are accessed via their official APIs. All models are provided with 128 uniformly sampled frames as input, except VideoLLAMA2.1-AV and VideoLLaMA2, which follow their default configurations and use 32 frames. For audio-only models, no video frames are provided. All reported results are obtained under the same fixed input sampling and decoding settings.

4.2 Key Observations

Findings 1

EG-Scores are extremely lower than answer accuracies, indicating a substantial gap between correct answering and modality-specific evidence grounding.

Table 2 summarizes performance across omni-modal, video-only, and audio-only MLLMs. Closed-source omni-modal models generally achieve stronger performance, with Gemini3.1-Pro obtaining the highest accuracy of 66.72% and an EG-Score of 15.76. Among open-source omni-modal models, Qwen3-Omni performs best, achieving an overall accuracy of 51.34% and an EG-Score of 2.67. Although unimodal models are competitive in modality-favorable categories, they remain weaker overall.

Method	Factual Eval. (EG-Score)		
	Decisive Audio Reasoning	Audio-Visual Causality	Audio-Cond. Scene Reasoning
Gemini3.1-Pro	13.40	17.01	20.75
Gemini3.1-Flash	17.44	18.43	26.09
Qwen3-Omni	2.27	2.35	5.39

Table 3: **EG-Score comparison on factual categories requiring both modalities.**

Findings 2

MLLMs tend to select an answer even when no valid answer exists, highlighting the need to evaluate whether models can identify unsupported or invalid questions.

Counterfactual evaluation in Table 2 reveals that current MLLMs struggle when evidence is missing or the question premise conflicts with the audio-visual input. Even Gemini3.1-Pro, which achieves the best overall accuracy of 66.72%, drops to 53.72% on Modality Absence and 51.20% on Invalid Assumption. The gap is larger for open-source models: Qwen3-Omni obtains 51.34% overall accuracy but only 31.40% and 37.60% on the two counterfactual categories, while video-SALMONN 2+ and Phi-4-multimodal score only 0.41% and 4.55% on Modality Absence. A more detailed analysis of counterfactual evaluation is provided in the supplementary material.

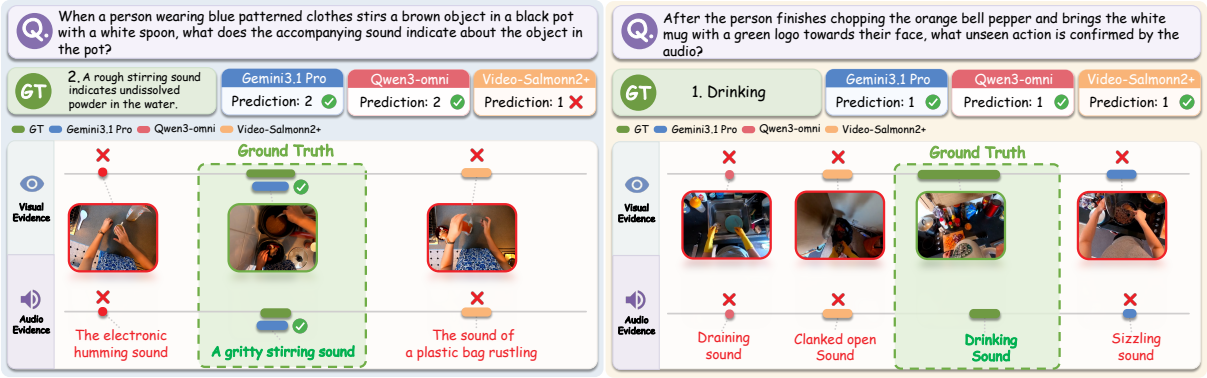


Figure 4: **Qualitative Results.** Representative examples illustrating how the top-performing models answer questions and localize modality-specific evidence. The results highlight differences between correct answer prediction and faithful audio-visual evidence grounding.

Findings 3

MLLMs rely more heavily on modality shortcuts when audio must be decisive.

Table 3 shows that models obtain particularly lower EG-Scores in Decisive Audio Reasoning, where audio provides the necessary evidence for the correct answer. Specifically, Gemini3.1-Pro, Gemini3.1-Flash, and Qwen3-Omni achieve EG-Scores of 13.40, 17.44, and 2.27, respectively, lower than their scores in other categories requiring both modalities. These results indicate that current MLLMs often reach correct answers without grounding them in the decisive audio evidence. In other words, they still struggle to identify when audio is not merely contextual, but instead serves as the key modality required for faithful reasoning.

4.3 Qualitative results

Figure 4 illustrates how **FATE** exposes reasoning failures that are obscured by answer accuracy alone. In the left sample, the question requires the model to infer from a gritty stirring sound that the powder in the pot has not fully dissolved, where a conclusion cannot be reached by the visual stream alone. Gemini 3.1 Pro selects the correct answer and grounds it in the relevant visual and audio evidence. In contrast, Qwen3-Omni also predicts the correct answer, but its evidence windows are temporally unrelated to the ground truth, resulting in zero IoU. video-SALMONN 2+ fails both to answer correctly and to identify the supporting evidence. This case shows that identical answer-level outcomes can reflect very different reasoning paths, and that FATE can distinguish faithful multimodal grounding from shortcut-based prediction.

The second case also demonstrates that answer correctness does not guarantee faithful evidence use. The question asks about an unseen action, drinking, whose occurrence is confirmed only by audio. Although all three models choose the correct option, none localizes the supporting evidence correctly: their predicted windows fall outside the ground-truth regions in both modalities. This suggests that models often reach the right answer while relying on visual priors or incidental cues rather than the true evidence, making such failures invisible under accuracy-only evaluation. Together, these qualitative results show that FATE evaluates not only whether a model answers correctly, but also whether it can justify the answer with the modality-specific evidence that actually supports it.

5 Conclusion

We propose FATE, a diagnostic benchmark for audio-visual reasoning that evaluates not only answer correctness but also whether models ground their reasoning in modality-specific evidence. FATE combines modality-specific evidence grounding, decisive audio scenarios, and counterfactual QA, and introduces EG-Score to jointly measure answer accuracy and evidence consistency. Experiments show that recent omni-modal models often achieve reasonable accuracy but struggle to provide correct visual and audio evidence, especially when audio is decisive. They also frequently fail to reject counterfactual questions with false premises or missing evidence, instead selecting one of the given choices. Overall, FATE reveals limitations hidden by accuracy-only evaluation and provides a diagnostic foundation for developing more reliable audio-visual reasoning models.

6 Limitations

FATE relies on VideoMME and Ego4D, which may not fully cover all real-world audio-visual scenarios. In addition, since Gemini 3.1 Pro is used during the data construction process, the resulting data may inherit biases or artifacts specific to the generator model. To mitigate these concerns, FATE implements several rigorous controls. First, it introduces the EG-Score to evaluate modality-specific temporal grounding, moving beyond simple answer accuracy. Second, human annotators substantially refine the final evidence windows initially selected by the model. Finally, the QA pairs are generated from descriptions produced by both Gemini- and Qwen-based models, reducing dependency on a single generator.

For future work, FATE can be extended to open-ended reasoning, free-form evidence descriptions, broader domains, and richer counterfactual conditions, alongside refining the EG-Score to incorporate semantic similarity.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Brandon Castellano. PySceneDetect.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. 2025. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2025. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *ICLR*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *NeurIPS*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. 2024. Meerkat: Audio-visual large language model for grounding in space and time. In *ECCV*.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE TASSP*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *NeurIPS*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and 1 others. 2024. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Google DeepMind. 2026. Gemma 4. https://ai.google.dev/gemma/docs/core/model_card_4.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Jacob Kauffmann, Jonas Dippel, Lukas Ruff, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. 2025. Explainable ai reveals clever hans effects in unsupervised learning models. *Nature Machine Intelligence*.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, and 1 others. 2025a. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, and 1 others. 2025b. Omnibench: Towards the future of universal omni-language models. *NeurIPS*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*.

Gorjan Radevski, Teodora Popordanoska, Matthew B Blaschko, and Tinne Tuytelaars. 2025. Dave: Diagnostic benchmark for audio visual evaluation. *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

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

Gemini Team and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, and 1 others. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *ACMMM*.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*.

Qiji Zhou, YiFan Gong, Guangsheng Bao, Hongjie Qiu, Jinqiang Li, Xiangrong Zhu, Huajian Zhang, and Yue Zhang. 2025a. Reasoning is all you need for video generalization: A counterfactual benchmark with sub-question evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. 2025b. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

Supplementary Material

FATE: Faithful Audio-Visual Benchmark with Modality-specific Temporal Evidence

A License of Data

Our Benchmark is constructed with videos from VideoMME and Ego4D according to their respective licenses and usage terms.

- **VideoMME.** VideoMME is restricted to academic research use only, and commercial use in any form is prohibited. The copyright of all videos belongs to the original video owners. Without prior approval, VideoMME may not be distributed, published, copied, disseminated, or modified in whole or in part. We therefore use VideoMME only for non-commercial academic research and do not redistribute its original videos.
- **Ego4D.** Ego4D is released under the MIT License. We use Ego4D in accordance with the MIT License.

B Additional Statistics

In this section, we present additional statistics about **FATE**.

B.1 Quantitative comparison between Counterfactual Evaluation and Factual Evaluation across models.

Model	FAR	All Counterfactual		Modality Absence		Invalid Assumption	
		CRR	CRP	CRR	CRP	CRR	CRP
Gemini-3.1-Pro	94.7	**39.0** (−55.7)	75.5	**51.9** (−42.8)	72.9	14.4 (−80.3)	28.1
Gemini-3.1-Flash	99.1	13.1 (−86.0)	85.7	18.3 (−80.8)	84.6	3.2 (−95.9)	33.3
Qwen3-Omni	96.3	21.9 (−74.5)	71.4	30.3 (−66.0)	69.5	5.6 (−90.7)	17.9
video-SALMONN 2+	100.0	0.3 (−99.7)	**100.0**	0.4 (−99.6)	**100.0**	0.0 (−100.0)	–
Phi-4-multimodal	95.7	3.9 (−91.9)	27.5	4.2 (−91.5)	21.3	3.2 (−92.5)	9.8
Gemma-4-E4B	84.2	28.4 (−55.8)	43.2	29.9 (−54.3)	34.4	**25.6** (−58.6)	18.9
VideoLLaMA2.1-AV	89.7	19.1 (−70.7)	44.0	18.2 (−71.6)	33.1	20.8 (−68.9)	22.6
Qwen3-VL	96.2	6.5 (−89.7)	42.1	9.5 (−86.7)	41.1	0.8 (−95.4)	2.9
VideoLLaMA2	100.0	0.0 (−100.0)	–	0.0 (−100.0)	–	0.0 (−100.0)	–
Audio Flamingo 3	99.2	3.8 (−95.3)	66.7	4.2 (−95.0)	58.8	3.2 (−96.0)	**36.4**
Avg.	95.5	13.6 (−81.9)	61.8	16.7 (−78.8)	57.3	7.7 (−87.8)	21.3

Table B.1: **Counterfactual rejection performance (%).** FAR $= \mathrm{TN}/(\mathrm{FP}+\mathrm{TN})$, CRR $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, CRP $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. Red values denote $\mathrm{CRR}-\mathrm{FAR}$, the asymmetry between factual acceptance and counterfactual rejection. "–" marks undefined cells (no reject predictions). Averages exclude undefined cells.

We quantify rejection behavior using FAR(Factual Acceptance Rate), CRR(Counterfactual Rejection Rate), and CRP(Counterfactual Reject Precision). FAR measures the fraction of factual samples answered normally, CRR measures the fraction of counterfactual samples correctly rejected, and CRP measures the fraction of rejected samples that are truly counterfactual. We define counterfactual asymmetry as CRR − FAR, where more negative values indicate a stronger answer-over-reject bias.

As shown in Table B.1, FAR consistently exceeds CRR across all models. The average FAR is 95.5%, whereas the average CRR is only 13.6%, yielding an average CRR–FAR gap of -81.9 percentage points. This large negative asymmetry indicates that current models strongly prefer answering over rejecting counterfactual questions. For example, video-SALMONN 2+ and VideoLLaMA2 achieve FAR scores of 100.0%, but their CRR scores remain only 0.3% and 0.0%, corresponding to CRR–FAR gaps of -99.7 and -100.0 percentage points, respectively. Even Gemini-3.1-Pro, the most balanced model, shows FAR 94.7% versus CRR 39.0%, resulting in a CRR–FAR gap of -55.7 percentage points.

CRP is consistently higher than CRR, suggesting that models are relatively precise when they reject but rarely choose to do so. At the subcategory level, *Modality Absence* is easier than *Invalid Assumption*,

as average CRR drops from 16.7% to 7.7%, while average CRP drops from 57.3% to 21.3%. Overall, current MLLMs can partially detect missing-modality cues, but still struggle to identify contradictions between question premises and video content.

B.2 Modality ablation

Category	Full (A+V)	Only-Video	Only-Audio
Decisive Audio Reasoning	58.43	47.19 (−11.24)	53.93 (−4.50)
Audio-Visual Causality	59.52	40.48 (−19.04)	49.21 (−10.31)
Audio-Cond. Scene Reasoning	70.63	66.67 (−3.96)	54.76 (−15.87)
Audio-Only Reasoning	45.76	32.77 (−12.99)	41.81 (−3.95)
Visual-Only Reasoning	63.95	65.70 (+1.75)	20.93 (−43.02)
Modality Absence	31.40	33.88 (+2.48)	58.26 (+26.86)
Invalid Assumption	37.60	37.60 (+0.00)	31.20 (−6.40)
Overall	51.34	45.43 (−5.91)	45.75 (−5.59)

Table B.2: **Per-category modality ablation on Qwen3-Omni.** Model performance is reported under full audio-visual input, video-only input, and audio-only input. Red and blue in parentheses indicate the performance change relative to the full audio-visual setting.

Table B.2 compares Qwen3-Omni with video-only and audio-only variants. The full omni model reaches 51.34% accuracy, only moderately above its video-only and audio-only variants at 45.43% and 45.75%. Category-level drops further show that Decisive Audio Reasoning depends primarily on audio, while Audio-Conditioned Scene Reasoning depends more strongly on visual grounding. These ablation patterns suggest that FATE *separates distinct modality dependencies while exposing the limited audio-visual integration of current open-source omni models.*

C Benchmark Category Hierarchy

In this section, we provide the complete hierarchical taxonomy of the benchmark categories. Specifically, we list all major categories and their corresponding subcategories, and report the number of QA pairs contained in each category.

Evaluation	Mid Category	Subcategory	#QAs
Factual Eval. (868)	Decisive Audio Reasoning (267)	State & Material Inference	99
		Fine-grained Action Inference	100
		Unseen Outcome Verification	68
	Audio-Visual Causality (126)	Audio to Visual Reaction	58
		Visual to Audio Reaction	68
	Audio-conditioned Scene Reasoning (126)	Sound-Anchored Visual Inference	61
		Paralinguistic Intent Reasoning	65
	Audio-Only Reasoning (177)	Source Attribution	63
		Audio Counting	72
		Audio Temporal Order	42
	Visual-Only Reasoning (172)	Visual Event Understanding	78
		Visual Temporal Order	55
		Visual State Transition	39
Counterfactual Eval. (367)	Modality Absence (242)	Audio Absence	133
		Visual Absence	59
		Bimodal Absence	50
	Invalid Assumption (125)	False Premise	54
		Temporal Order Mismatch	43
		Wrong Attribution	28
Total			**1235**

Table C.1: Category hierarchy with QA counts.

D Prompt for making QAs

In this section, we provide a complete list of the instructions used for the benchmark.

You are a professional video analyst.

Your task is to provide a detailed, objective description of what happens in this short video clip.
This clip corresponds to the following segment of the original video:
- Segment start time: {start_ts}
- Segment end time: {end_ts}

IMPORTANT RULES:
- Describe ONLY what is directly observable from the video. Do NOT infer intentions, causes, or unseen events.
- If something is uncertain, use cautious language such as "appears to", "likely", or "possibly".
- Do NOT add information that cannot be visually confirmed.
- Divide the clip into multiple short timeline intervals and describe each interval separately.
- Use timestamps aligned to the original video timeline, not clip-relative timestamps.
- Every timestamp must stay within the segment range from {start_ts} to {end_ts}.
- The timestamp intervals must be in chronological order and should cover the clip from beginning to end as completely as possible.
- Each line must contain exactly one timestamp interval followed by one description.
- Do not output bullets, numbering, headings, explanations, or a final summary paragraph.

Please describe:
- The main actions and events in temporal order (what happens first, next, and last)
- Characters or people visible and their observable behaviors
- Key visual details: setting, environment, lighting, colors, and notable objects
- Any visible text, signs, or on-screen graphics
- Camera movements or transitions (if noticeable)

Focus on the most salient and important events, avoiding unnecessary minor details.

Output Format:
[START_TIMESTAMP - END_TIMESTAMP] description
[START_TIMESTAMP - END_TIMESTAMP] description
[START_TIMESTAMP - END_TIMESTAMP] description
...
Replace START_TIMESTAMP and END_TIMESTAMP with actual timestamps within {start_ts} to {end_ts}. Do not copy the placeholder text.

Write in English. Be specific and detailed, but concise. Keep each description focused on what is visible within that time interval only.

Figure D.1: Prompt for Video description.

Describe the audio precisely and concisely.

[Task]
For this segment, break the audio into finer-grained time ranges and describe what is heard in each range. Do not provide only one overall segment summary.

[Output Format]
Write 2 or more lines when the audio changes over time, using this exact style:
[HH:MM:SS - HH:MM:SS] description
[HH:MM:SS - HH:MM:SS] description

[Rules]
1. Each line must begin with a timestamp range in square brackets.
2. The timestamps must use HH:MM:SS format and the original video timeline provided in [Segment Time Context], not a 00:00:00-relative timeline.
3. Split the segment into meaningful sub-ranges based on changes in non-speech sounds, music, ambience, or emotion.
4. Do not extract, transcribe, or summarize speech content. If speech is present, mention only non-verbal vocal qualities when needed.
5. If there is no meaningful audible event to describe for an interval, write \Silence\ as the description for that interval.
6. The first timestamp must start at the provided segment start time, and the final timestamp must end at the provided segment end time.
7. Focus on non-speech sounds, background music, ambience, and clearly audible non-verbal emotions only when they are evident.
8. Keep each description concise and specific to that sub-range.

Figure D.2: Prompt for Audio description.

You are selecting candidate segment groups for later multimodal QA generation.

Goal:
- Choose up to {num_groups} high-quality candidate segment groups for the category described below.
- A candidate group may contain multiple video segments and multiple audio/subtitle segments.
- Segment groups should be tight: include only the segments needed to support this category.
- Prefer groups where a later QA would require the intended modality evidence, not generic commonsense.
- If fewer than {num_groups} good groups exist, return fewer.
- If no good group exists, return an empty ‘candidates‘ array.

Category-specific selection prompt:
{category_prompt.strip()}

Return valid JSON only. Do not use markdown.

Required output schema:
{schema}

Hard rules:
- Do not output any metadata fields beyond the required schema.
- Output exactly one JSON object.
- The object must include a ‘candidates‘ array.
- Each candidate must include string ‘reason‘, string ‘evidence_summary‘,
integer array ‘visual_segment_ids‘, and integer array ‘audio_segment_ids‘.
- Use visual ids only from labels that look like V<number>.
- Use audio ids only from labels that look like A<number>.
- Do not invent segment ids.
- Do not repeat the same segment id group.
- Rank candidates strongest to weakest by array order.

Available evidence:
{segment_table}

Figure D.3: Prompt for QA Candidate prompt.

You are selecting candidate segment groups for later visual-only QA generation.

Goal:
- Choose up to {num_groups} high-quality candidate segment groups for the category described below.
- Use only visual descriptions as evidence.
- A candidate group may contain multiple video segments.
- Segment groups should be tight: include only the segments needed to support this category.
- Prefer groups where a later QA would require the intended visual evidence, not generic commonsense.
- If fewer than {num_groups} good groups exist, return fewer.
- If no good group exists, return an empty ‘candidates‘ array.

Category-specific selection prompt:
{category_prompt.strip()}

Return valid JSON only. Do not use markdown.

Required output schema:
{schema}

Hard rules:
- Do not output any metadata fields beyond the required schema.
- Output exactly one JSON object.
- The object must include a ‘candidates‘ array.
- Each candidate must include string ‘reason‘, string ‘evidence_summary‘,
and integer array ‘visual_segment_ids‘.
- Use visual ids only from labels that look like V<number>.
- Do not invent segment ids.
- Do not repeat the same segment id group.
- Do not use audio, subtitles, transcripts, or any information outside the visual descriptions.
- Rank candidates strongest to weakest by array order.

Available evidence:
{segment_table}

Figure D.4: Prompt for QA Candidate visual exclusive prompt.

You are a careful visual QA data generator.

Task:
- Generate up to {num_qas} high-quality multiple-choice QA questions based on candidate evidence.
- Use only visual descriptions as evidence.
- Do not use audio information, subtitles, transcript text, or ‘sub‘ fields even if they are available elsewhere in the data.
- If a candidate does not support a trusted question from visual evidence alone, return ‘is_possible: false‘ and ‘qa_data: []‘.
- Output only valid JSON format. Do not output Markdown format.

QA prompts by category:
{qa_prompt}

Required output schema:
{schema}

Strict rules:
- Do not output candidate_id, video_id, category_name, start_sec, and end_sec. These values are appended in the code.
- ‘is_possible‘ must be a boolean value.
- If ‘is_possible‘ is false, ‘qa_data‘ must be [].
- If ‘is_possible‘ is true, ‘qa_data‘ must contain {num_qas} items.
- Each question item must have exactly 5 different options.
- The last option in ‘choices‘ must be exactly "None of the above".
- ‘answer_idx‘ must be an integer between 0 and 4.
- ‘visual_evidence_segments‘ must use only the visual segment IDs of the provided candidate footage.
- Each valid question must use at least one visual evidence segment.
- Questions must clearly refer to the relevant event or moment within the entire video. They must be designed so that viewers can locate the event without being told the segment ID.
- Do not ask questions about segment IDs, timestamps, descriptions, captions, or metadata.
- Every question must be answerable from visual evidence alone.
- Do not mention audio, speech, narration, subtitles, captions, or unseen sounds.
- Avoid trivial questions such as simply asking what objects are seen without meaningful reasoning.
- Depending on the category, prefer questions regarding cause, reaction, confirmation, inference, source, chronological order, intent, or result.
- Use only the provided visual descriptions as evidence. Do not fabricate events other than the candidate evidence.
- The answer must not be contained within the question. Do not create questions where the correct answer can be determined simply by looking at the question.

Reason for Candidate Selection:
{candidate.reason}

Summary of Candidate Evidence:
{candidate.evidence_summary}

Visual Evidence:
{visual_text}

Figure D.5: Prompt for QA Maker Visual Only prompt.

You are a careful audiovisual QA data generator.

Task:
- Generate up to {num_qas} high-quality multiple-choice QA questions based on candidate evidence.
- All valid questions must require both visual evidence and audio/subtitle evidence.
- If a candidate does not support a trusted question requiring both forms, return ‘is_possible: false‘ and ‘qa_data: []‘.
- Output only valid JSON format. Do not output Markdown format.

QA prompts by category:
{qa_prompt}

Required output schema:
{schema}

Strict rules:
- Do not output candidate_id, video_id, category_name, start_sec, and end_sec. These values are appended in the code.
- ‘is_possible‘ must be a boolean value.
- If ‘is_possible‘ is false, ‘qa_data‘ must be [].
- If ‘is_possible‘ is true, ‘qa_data‘ must contain {num_qas} items.
- Each question item must have exactly 5 different options.
- The last option in ‘choices‘ must be exactly "None of the above".
- ‘answer_idx‘ must be an integer between 0 and 4.
- ‘visual_evidence_segments‘ must use only the visual segment IDs of the provided candidate footage.
- ‘audio_evidence_segments‘ must use only the audio segment IDs of the provided candidate footage.
- Each valid question must use at least one visual evidence segment and at least one audio evidence segment.
- Questions must clearly refer to the relevant event or moment within the entire video. They must be designed so that viewers can locate the event without being told the segment ID.
- Do not ask questions about segment IDs, timestamps, descriptions, captions, or metadata.
- Do not ask questions that can be answered solely by visual evidence.
- Avoid trivial questions such as simply asking what sounds are heard or what objects are seen.
- Depending on the category, prefer questions regarding cause, reaction, confirmation, inference, source, chronological order, intent, or result.
- Use the provided description as evidence. Do not fabricate events other than the candidate evidence.
- The answer must not be contained within the question. Do not create questions where the correct answer can be determined simply by looking at the question.

Reason for Candidate Selection:
{candidate.reason}

Summary of Candidate Evidence:
{candidate.evidence_summary}

Visual Evidence:
{visual_text}

Audio/Subtitle Evidence:
{audio_text}

Figure D.6: Prompt for QA Maker prompt.

Category: 7-2.Temporal Order Mismatch

Definition:
The question presupposes a temporal order between events, but the actual evidence supports the opposite order or a different before/after relation.

Transformation goal:
- Use the source question and source answer to infer the original temporal relationship.
- Rewrite the QA so that the question asks about the reversed order, creating a temporal premise that is not supported by the evidence.
- Because the reversed order does not occur in the evidence window, the transformed question must be unanswerable from the provided evidence.
- Include four plausible non-None distractors, followed by "None of the above".
- The correct answer must be "None of the above".

Good pattern:
Source QA:
Question: "What sound is heard after the scream?"
Answer: "Glass breaking."
Transformed QA:
Question: "What sound is heard after the glass breaking?"
Correct answer: "None of the above."

Rules:
- Prefer questions using before, after, immediately before, immediately after, first, next, or then.
- Do not merely negate the source question.
- Do not create a yes/no question.
- The last choice must be exactly "None of the above".
- answer_idx must be 4.
- answer_text must be exactly "None of the above".
- The first four choices should be plausible but unsupported by the reversed temporal premise.

Figure D.7: Prompt for QA Maker post processing Category Temporal Order Mismatch prompt.

Category: 7-3.Wrong_Attribution

Definition:
The available evidence is present, but the question incorrectly links a sound, action, direction, source, subject, or cause to the wrong entity.

Transformation goal:
Create a hard multiple-choice QA where the question contains an attribution mismatch, but one answer choice identifies the actual source, agent, direction, or cause supported by the evidence.

Allowed mismatch types:
1. Spatial flip:
- left vs right
- front vs back
- inside vs outside
2. Agent swap:
- one visible actor vs another actor
- visible actor vs off-screen actor
3. Causal mismatch:
- the actual cause is A, but the question frames it as B

Examples of desired style:
- "The question asks whether the footsteps come from the right, but the correct option identifies the person moving on the left."
- "The question asks whether the visible dog is barking, but the correct option identifies an off-screen dog."
- "The question asks whether the glass broke because of wind, but the correct option identifies a physical collision."
- "The question asks whether the right person who moves their mouth is speaking, but the correct option identifies the left person."

Rules:
- Do not create a simple yes/no question.
- Make all four non-None options plausible and close to the same semantic field.
- Do not make "None of the above" the correct answer.
- The correct choice must repair the wrong attribution by naming the actual source, agent, direction, or cause.
- Keep the mismatch subtle enough that a model must separate audiovisual evidence from the misleading wording.
- Avoid lexical answer leakage. Do not make only the correct choice contain obvious corrective words such as "actually", "rather than", "instead", "not caused by", "does not cause", or "the real source".
- Prefer writing all four non-None choices in the same grammatical style, with no choice sounding more corrective or explanatory than the others.
- If you use the word "actually" in one non-None choice, use it in all four non-None choices. Better yet, avoid "actually" entirely.
- Bad choices because the correct answer is too obvious:
1. "A metal pot shifting inside the closed cabinet."
2. "The cabinet's metal hinges squeaking as it shuts."
3. "The sound is actually caused by the person pulling a small object from the open drawer."
4. "A dish sliding across the counter."
5. "None of the above"
- Better choices with matched style:
1. "A metal pot shifting inside the closed cabinet."
2. "The cabinet's metal hinges squeaking as it shuts."
3. "A small object being pulled from the open drawer."
4. "A dish sliding across the counter."
5. "None of the above"
- Better choices if using "actually" everywhere:
1. "Actually, a metal pot shifts inside the closed cabinet."
2. "Actually, the cabinet's metal hinges squeak as it shuts."
3. "Actually, a small object is pulled from the open drawer."
4. "Actually, a dish slides across the counter."
5. "None of the above"

Figure D.8: Prompt for QA Maker post processing Category Wrong Attribution prompt.

You are a careful audiovisual multiple-choice QA post-processor.

You will receive one already-approved source QA item from an audiovisual benchmark.
Create one new QA item for the requested negative-query category while preserving the source item's topic, event type, and audiovisual reasoning style as much as possible.

Output only valid JSON with this exact schema:
{
"is_possible": true,
"reason": "Brief explanation of why the transformed QA is valid.",
"question": "The transformed question.",
"choices": [
"A plausible option",
"A plausible option",
"A plausible option",
"A plausible option",
"None of the above"
],
"answer_idx": 0,
"answer_text": "The exact text of choices[answer_idx]"
}

If the source QA cannot support the requested transformation, return:
{
"is_possible": false,
"reason": "Brief reason.",
"question": "",
"choices": [],
"answer_idx": 0,
"answer_text": ""
}

Strict rules:
- Return JSON only. Do not use Markdown.
- Create exactly one QA item.
- The question must be answerable from the same original evidence window, but it must instantiate the requested negative-query category.
- Use the source question and answer as anchors. Do not invent a completely unrelated event.
- Do not mention metadata, timestamps, segment IDs, category names, or that this is a transformed item.
- The choices list must contain exactly 5 distinct strings.
- The last choice must be exactly "None of the above".
- The correct answer should be one of the first four choices unless the category-specific instruction explicitly says otherwise.
- answer_idx must be an integer from 0 to 4.
- answer_text must exactly match choices[answer_idx].
- Keep the wording natural and self-locating, so a viewer can identify the relevant moment without metadata.

Figure D.9: Prompt for QA Maker post processing Common prompt.

You are improving a multiple-choice video/audio benchmark item so it cannot be answered from text alone.

You receive the question, choices, and the correct index. You must preserve the true correct answer semantics and keep the same correct index.

Apply these four methods in order:
1_hard_distractor_regeneration:
Rewrite wrong choices into hard distractors that are as plausible as the correct answer from text alone.
Distractors should share the same modality, scene context, sound/action type, and level of detail.

2_drop_unfixable_text_shortcut:
If the question itself gives away the answer through common sense, world knowledge, lexical clues, or false-premise wording,
and choices alone cannot fix it without changing the benchmark meaning, set action to "drop".

3_specificity_and_length_balancing:
If the correct answer is the only long, detailed, concrete, or benchmark-like option, rebalance all choices.
Make all options similar in length, specificity, grammar, and confidence cues.

4_false_premise_format_normalization:
If this is a false-premise or "actually X, not Y" item, do not let only the correct option contain the obvious correction pattern.
Either make all options share similar correction wording or mark the item as "drop" if that would distort the truth.

Rules:
- Return only valid JSON.
- Keep exactly {len(choices)} choices.
- Keep answer_idx exactly {row.get('answer_idx')}.
- The revised correct choice must remain semantically equivalent to the original answer.
- Do not add media facts that are not already implied by the original item.
- Keep "None of the above" if it exists, but do not make it uniquely attractive.
- If action is "drop", revised_question and revised_choices can equal the original.

Return schema:
{{
"action": "rewrite|drop|keep",
"methods_applied": ["1_hard_distractor_regeneration"],
"shortcut_diagnosis": "brief diagnosis of how text-only model got it right",
"drop_reason": null,
"revised_question": "question text",
"revised_choices": ["choice 0", "choice 1", "choice 2", "choice 3", "choice 4"],
"answer_idx": {row.get('answer_idx')},
"answer_text": "revised correct choice text",
"notes": "brief notes"
}}

Original text-only model rationale, if available:
{rationale}

Question:
{row.get('question', '')}

Choices:
{choice_lines}

Figure D.10: Prompt for Answer distractor prompt.

The clipped video duration is {sample.rel_duration:.3f} seconds. You will receive {modality_description}. All requested timestamps must be relative to the start of this clipped video, in seconds.

Question:
{sample.question}

Choices:
{choices_text(sample)}

Tasks:
1. Choose the single best answer from the choices above.
2. Extract the most relevant video evidence window as [<start_sec>, <end_sec>].
3. Extract the most relevant audio evidence window as [<start_sec>, <end_sec>].

Rules:
- The answer_idx must be a zero-based integer index into the choices.
- Use null for an evidence window if that modality is not needed or no evidence exists.
- Window start and end must satisfy 0 <= start <= end <= {duration_limit}.
- Do not output text outside the JSON object.

Return exactly this JSON schema:
{{"answer_idx": 0, "answer": "choice text", "video_evidence_window": [<start_sec>, <end_sec>], "audio_evidence_window": [<start_sec>, <end_sec>]}}

Figure D.11: Prompt for Inference prompt.

Target category instruction: State & Material Inference

Task definition:
Create questions where visually similar states, materials, or conditions must be distinguished using sound. The visual evidence should narrow the context, while the audio evidence resolves the ambiguity.

The key distinction must be one of:
- Two or more visual interpretations look similar.
- The sound reveals state, material, surface, texture, wetness, hollowness, fullness, hardness, or operating condition.
- The answer cannot be trusted from visual evidence alone.

Design rules:
- Ask about the object's state, material, surface condition, or physical condition.
- Make the audio cue decisive, such as metallic ringing, glass clink, liquid splash, hollow knock, sizzling, scraping, or muffled impact.
- Do not ask a question where the material or state is plainly visible.
- Avoid requiring specialized knowledge beyond common audio-visual reasoning.
- Keep the alternatives plausible from the visual scene.

Good question styles:
- From the sound as the object is placed down, what material is it most likely made of?
- Does the surface seem wet or dry when the item touches it?
- What state of the container is suggested by the sound when it is moved?

Figure D.12: Prompt for category/1-1.State Material Inference.

Target category instruction: Fine-grained Action Inference

Task definition:
Create questions where a visible action looks broadly similar across possibilities, but a subtle impact, contact, operation, or movement sound reveals the quality, accuracy, force, success, or failure of the action.

The key distinction must be one of:
- The visual action alone is not enough to judge quality or success.
- Audio reveals whether contact was clean, weak, missed, forceful, smooth, jammed, successful, or failed.
- The answer depends on matching fine-grained sound quality with the visible action.

Design rules:
- Ask about action quality, success, accuracy, strength, timing, or whether contact occurred.
- Use sound cues such as clean hit, dull thud, scrape, click, misfire, bounce, rattle, or failed mechanism.
- Do not ask only what action is being performed.
- Avoid overclaiming if the sound only weakly supports the distinction.
- Keep the correct answer grounded in the described evidence.

Good question styles:
- Based on the sound of contact, was the hit clean or off-center?
- Does the click indicate that the mechanism successfully locked?
- What does the impact sound suggest about how firmly the object landed?

Figure D.13: Prompt for category/1-2.Fine grained Action Inference.

Target category instruction: Unseen Outcome Verfication

Task definition:
Create questions where an event or result is hidden, occluded, or off-screen, and audio evidence confirms whether the expected outcome actually occurred.

The key distinction must be one of:
- The key result is not directly visible.
- Sound confirms or rejects the outcome.
- The answer depends on audio evidence such as a crash, fall, click, splash, shatter, latch, start-up, or shutdown.

Design rules:
- Ask whether the hidden/off-screen outcome occurred, what result can be confirmed, or what happened out of view.
- The visual evidence should establish the setup or occlusion.
- The audio evidence must be the decisive confirmation.
- Do not create questions where the outcome is directly visible.

Good question styles:
- From the sound behind the box, did the object actually fall?
- What hidden outcome is confirmed by the sound after the person reaches off-screen?
- What result did the object bring about based on the sound heard after it was thrown?

Figure D.14: Prompt for category/1-3.Unseen Outcome Verification.

Target category instruction: Audio -> Visual Reaction

Task definition:
Create questions where an audible event happens first and causes a visible reaction by a person, animal, object, or scene element. The question should require the model to connect the sound with the later visual response.

The key distinction must be one of:
- A sound precedes a visible reaction.
- The visible reaction is best explained by that sound, not by visual context alone.
- The answer depends on recognizing the causal link between audio evidence and visual evidence.

Design rules:
- Ask about why the visible reaction happened, what sound triggered it, or which visual response followed the sound.
- The question must require both the audio event and the visible reaction.
- Do not ask for a generic description of the sound or the object alone.
- Avoid questions where the visual reaction is obvious without hearing the audio.
- Use concrete visual reactions such as turning around, stopping, flinching, looking toward a source, moving away, or changing behavior.

Good question styles:
- After the sudden sound, why does the person turn toward the doorway?
- What visible reaction does the child show after the loud crash is heard?
- Which sound best explains why the animal suddenly looks up?

Figure D.15: Prompt for category/2-1.Audio to Visual Reaction.

Target category instruction: Visual -> Audio Reaction

Task definition:
Create questions where a visible event happens first and produces or triggers a later audible reaction such as speech, screaming, laughter, honking, machine noise, impact sound, or exclamation.

The key distinction must be one of:
- A visual event precedes an audio reaction.
- The sound is a response to what is seen happening on screen.
- The answer depends on identifying the visual cause of the later audio.

Design rules:
- Ask what visible event caused the sound, what the speaker reacted to, or why a later audio reaction occurred.
- The visual cause must be present in the provided visual evidence.
- The audio reaction must be present in the provided audio/subtitle evidence.
- Do not create questions answerable from the audio content alone.
- Avoid vague questions about emotions unless the visual trigger is specific.

Good question styles:
- What does the person see just before they scream?
- Which visible event causes the crowd to cheer?
- Why does the driver honk after the pedestrian moves?

Figure D.16: Prompt for category/2-2.Visual to Audio Reaction.

Target category instruction: Sound-Anchored Visual Inference

Task definition:
Create questions where a sound source may be off-screen or not directly visible, and the answer requires understanding what an on-screen person, animal, object, or scene element is doing in relation to that sound.

The key distinction must be one of:
- A meaningful sound is heard while the relevant source is not directly visible.
- The visible scene shows a simultaneous reaction, behavior, or state conditioned by that sound.
- The answer connects non-visible audio context with in-scene visual evidence.

Design rules:
- Ask what the visible subject is doing while the sound occurs, how they react, or what their state is during the sound.
- The sound should provide temporal or contextual grounding for the visual observation.
- Do not require identifying an unseen source unless source attribution is the category target.
- Avoid questions answerable from visual evidence alone without knowing when the sound occurs.
- Keep the question self-locating within the whole video.

Good question styles:
- While the bird sounds are heard, what is the man doing?
- What happens on screen at the moment the off-screen alarm starts?
- How does the dog behave while the unseen noise continues?

Figure D.17: Prompt for category/3-1.Sound Anchored Visual Inference.

Target category instruction: Paralinguistic / Intent Reasoning

Task definition:
Create questions where the speaker's real intent, emotion, attitude, or stance must be inferred from tone, pitch, rhythm, volume, hesitation, laughter, sighing, or other paralinguistic cues, not just from the literal words.

The key distinction must be one of:
- The spoken words alone are ambiguous.
- Tone or prosody changes the intended meaning.
- The answer depends on audio delivery plus the visible/social context.

Design rules:
- Ask about sincerity, sarcasm, concern, surprise, frustration, confidence, fear, teasing, or reluctance.
- Use the visible context to anchor who is speaking and what situation they are reacting to.
- Do not ask for a transcript-only fact.
- Avoid questions where the emotion is entirely visible without voice cues.
- Make answer choices reflect different plausible intentions or attitudes.

Good question styles:
- From the speaker's tone, is the comment meant as praise or sarcasm?
- What attitude does the person convey while saying the line?
- How does the speaker seem to feel about what just happened?

Figure D.18: Prompt for category/3-2.Paralinguistic Intent Reasoning.

Target category instruction: Source Attribution

Task definition:
Create questions where the model must identify who or what produced a heard sound, using the visible scene or surrounding context to attribute the audio source correctly.

The key distinction must be one of:
- A sound occurs and multiple possible sources are visually or contextually plausible.
- The answer requires mapping the sound to the correct person, object, animal, device, or location.
- Audio is decisive, while visual evidence constrains the candidate sources.

Design rules:
- Ask who or what made the sound, where the sound came from, or which visible object/person is responsible.
- Use clear source cues such as mouth movement, contact timing, instrument use, object operation, or synchronized action.
- Do not ask merely what sound was heard.
- Avoid questions where only one source is possible without reasoning.
- Keep distractors grounded in visible or contextual alternatives.

Good question styles:
- Who produces the sound heard at that moment?
- Which object is responsible for the clicking noise?
- Where does the brief ringing sound appear to come from?

Figure D.19: Prompt for category/4-1.Source Attribution.

Target category instruction: Audio Counting

Task definition:
Create questions where the answer requires counting repeated or separated sound events using audio evidence.

The key distinction must be one of:
- The relevant events are discrete audio occurrences.
- The count cannot be determined reliably from visual evidence alone.
- The answer depends on separating repeated sounds from background noise or continuous sound.

Design rules:
- Ask for the number of clearly countable sounds, such as knocks, claps, beeps, steps, strikes, barks, rings, or spoken repetitions.
- Count only distinct events that are supported by the audio evidence.
- Do not count visual actions unless each is confirmed by sound.
- Avoid ambiguous counts when sounds overlap heavily or are continuous.
- Make numeric choices plausible and include exactly one correct count when possible.

Good question styles:
- How many knocks are heard before the person opens the door?
- How many distinct beeps occur during this moment?
- How many times do you hear the barcode scan sound when paying at the self-checkout counter?

Figure D.20: Prompt for category/4-2.Audio Counting.

Target category instruction: Audio Temporal Order

Task definition:
Create questions where the answer requires determining the before/after order between two or more audio events.

The key distinction must be one of:
- Multiple sounds occur in sequence.
- The relative order is recoverable from audio evidence.
- Visual evidence may provide context but must not be sufficient by itself.

Design rules:
- Ask which sound came first, which came after another, or what sequence best matches the audio.
- Use distinct audio events such as footsteps, speech, door closing, alarm, impact, engine start, whistle, or clap.
- Do not ask about visual event order unless the audio order is the decisive evidence.
- Avoid sequences that are too dense or ambiguous.
- Keep answer choices as ordered pairs or short ordered descriptions.

Good question styles:
- Which is heard first: the footsteps or the door closing?
- What sound occurs immediately after the person speaks?
- In what order do the bell and the crash occur?

Figure D.21: Prompt for category/4-3.Audio Temporal Order.

Target category instruction: Visual Event Understanding

Task definition:
Create questions where the answer is based on understanding a visible action or event without relying on audio.

The key distinction must be one of:
- The relevant evidence is purely visual.
- The question asks who did what, what happened, or what object/action is involved.
- Audio is unnecessary or absent for the correct answer.

Design rules:
- Ask about visible actions, interactions, object use, gestures, movement, or scene events.
- Do not make audio evidence necessary for the answer.
- Avoid asking about sounds, speech, tone, or hidden outcomes.
- Ensure the answer is directly supported by visible evidence.
- Keep distractors visually plausible but incorrect.

Good question styles:
- What does the man pick up from the table?
- Which action does the child perform after entering the room?
- What happens to the object in the person's hand?

Figure D.22: Prompt for category/5-1.Visual Event Understanding.

Target category instruction: Visual Temporal Order

Task definition:
Create questions where the answer requires understanding the order of two or more visible events or actions using visual evidence only.

The key distinction must be one of:
- Multiple visual events occur in a clear sequence.
- The answer depends on before/after reasoning.
- Audio is not needed to determine the order.

Design rules:
- Ask what happened before, after, first, last, or immediately following a visible event.
- Use visible actions or state changes as anchors.
- Do not ask about the order of sounds.
- Avoid sequences where ordering is not clear from the visual evidence.
- Keep answer choices short and event-focused.

Good question styles:
- What does the person do before picking up the cup?
- Which visible action happens immediately after the door opens?
- In what order do the person sit down and reach for the bag?

Figure D.23: Prompt for category/5-2.Visual Temporal Order.

Target category instruction: Visual State Transition

Task definition:
Create questions where the answer requires tracking how an object's or person's visible state, attribute, position, posture, or condition changes over time.

The key distinction must be one of:
- A visible state or attribute changes across the segment.
- The answer requires comparing earlier and later visual evidence.
- Audio is not required to determine the change.

Design rules:
- Ask about transitions such as open/closed, empty/full, on/off visually, clean/dirty, standing/sitting, near/far, held/released, covered/uncovered, or location changes.
- Do not ask about state changes that are only indicated by sound.
- Make the initial and final states visually grounded.
- Avoid single-frame attribute questions that do not require change tracking.
- Keep distractors as plausible alternative state transitions.

Good question styles:
- Does the door change from closed to open?
- How does the person's posture change during the moment?
- Where does the object end up compared with where it started?

Figure D.24: Prompt for category/5-3.Visual State Transition.

Target category instruction: Audio Absence

Task definition:
Create questions where the question appears to require audio evidence, but the selected segment does not contain meaningful audio information needed to answer it. The correct behavior is to recognize that the required audio cue is absent.

The key distinction must be one of:
- The question depends on a voice, sound, tone, count, source, or audio event.
- The relevant audio is absent, silent, too generic, or not meaningful for the asked fact.
- The answer should not assume audio evidence that is not present.

Design rules:
- Ask an audio-dependent question whose required audio cue is missing in the candidate evidence.
- Make the absence of usable audio the reason the answer cannot be determined.
- Do not invent speakers, sounds, counts, or source identities.
- Ensure the visual evidence alone cannot answer the audio-specific question.

Good question styles:
- Whose voice is heard during this moment? (No voice heard)
- What sound causes the person to react in this segment? (Reaction is produced, but no sound was made)
- How many times is the object audibly tapped? (The object is tapped, but no sound is produced)

Figure D.25: Prompt for category/6-1.Audio Absence.

Target category instruction: Visual Absence

Task definition:
Create questions where the question appears to require visual evidence, but the selected segment does not show the necessary visual information. The correct behavior is to recognize that the required visual cue is absent.

The key distinction must be one of:
- The question depends on seeing a person, object, action, location, or visual state.
- The necessary visual evidence is off-screen, occluded, outside the frame, or not shown.
- The answer should not assume visual evidence that is not present.

Design rules:
- Ask a visually grounded question whose required visual cue is missing from the candidate evidence.
- Make the absence of usable visual evidence the reason the answer cannot be determined.
- Do not invent unseen actions, objects, identities, or positions.
- Audio may establish timing, but it must not reveal the missing visual fact.

Good question styles:
- While the applause is heard, what is the person on stage doing? (No person on stage)
- What object falls when the crash is heard off-screen? (No object has ever fallen)
- What facial expression does the speaker show during the line? (No facial expression)

Figure D.26: Prompt for category/6-2.Visual Absence.

Target category instruction: False Premise

Task definition:
Create questions where a key premise in the question conflicts with the actual audio-visual evidence. The model must verify the premise instead of accepting it as true.

The key distinction must be one of:
- The question contains an incorrect assumption about a sound, object, action, person, event, or timing.
- The evidence contradicts the premise.
- The correct answer should reflect that the premise is false rather than reasoning from it.

Design rules:
- Ask a question with a specific but false premise, such as the wrong sound source, wrong object, wrong action, wrong speaker, or wrong timing.
- The contradiction must be clear from the provided evidence.
- Do not make the false premise subtle enough to be unfair.
- Avoid questions where the premise is merely unknown; it should be contradicted.
- Avoid cases where the causative audio is completely absent. Use this only if the causative audio exists.

Good question styles:
- When the guitar sound is heard, what is the person doing? (It is actually the sound of a violin, not a guitar)
- Who is the character looking at after the phone rings? (The reason the character is looking at someone is because the door opened, not because of the phone ringing)
- At the moment the glass breaks, how do the people around react? (it is not the glass that broke, but the sound of a metal object falling.)

Figure D.27: Prompt for category/7-1.False Premise.

E Annotation Tool

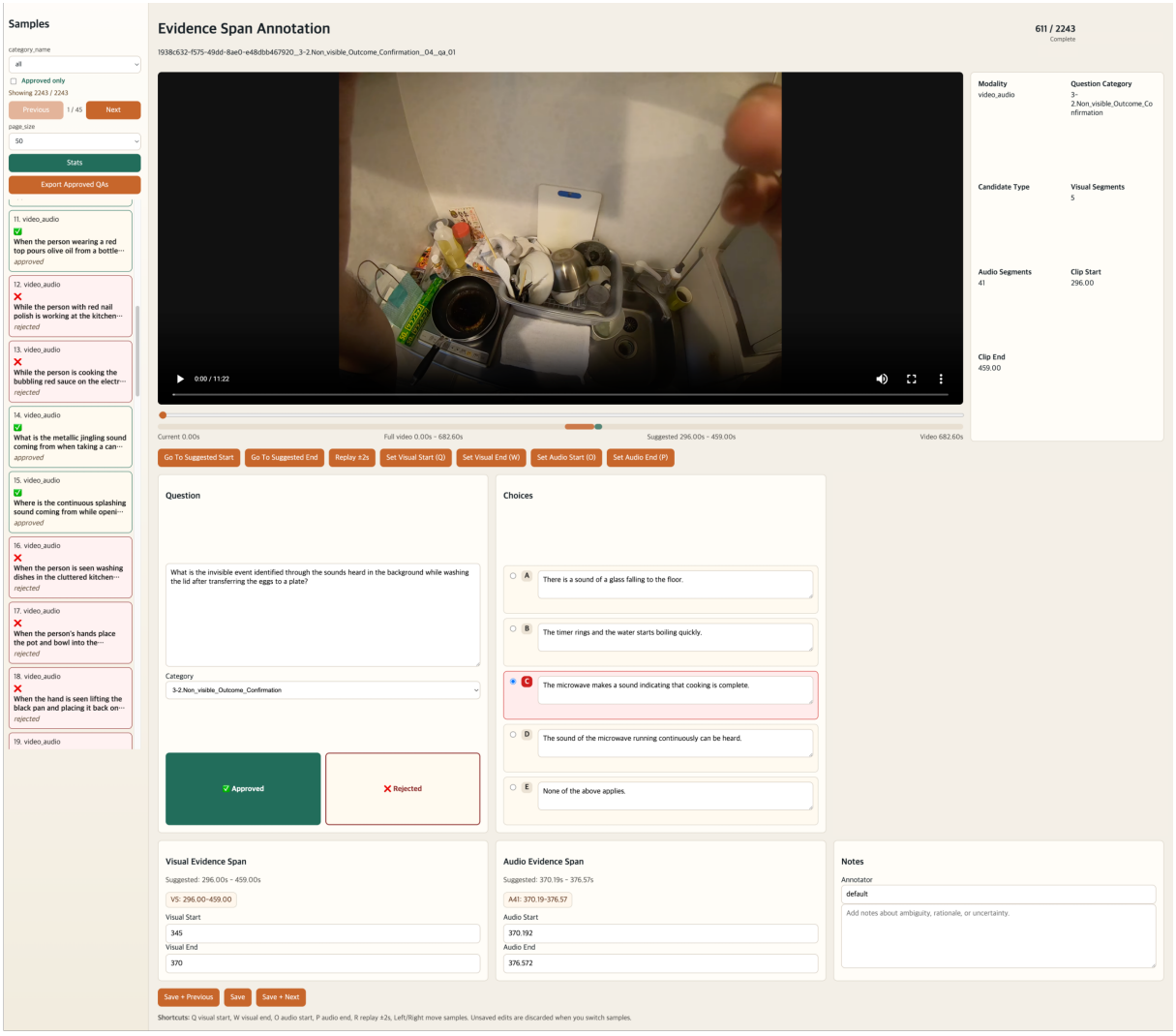


Figure E.1: **Annotation tool used during the human verification step.** The tool is designed to annotate QA pairs for each video. Annotators can directly adjust visual and audio evidence spans while watching the video, and manually refine both questions and answers.