FAVE: A Structured Benchmark for Fine-Grained Audio-Visual Temporal Evaluation in Multimodal LLMs — Supplementary Material

Weiheng Lu[1*] An Yu[2*] Jian Li[3,4†] Zhenfei Zhang[2] Felix X.-F. Ye[2] Ming-Ching Chang[2]
[1]Peking University [2]State University of New York at Albany [3]Tencent YouTu Lab [4]Nanjing University
[*]Equal contribution [†]Corresponding author: swordli@tencent.com

In this supplementary material, we provide additional analysis and results omitted from the main paper due to space constraints. This document includes:

- Complete data construction and evaluation protocol, including all GPT prompt templates (§1).
- GPT scoring validation: consistency and error-category analysis (§2).
- Discussion of model selection and audio data choices (§3).
- Future research directions (§4).
- Data licenses (§5).

1. Data and Evaluation Protocol

We present the complete data construction and evaluation protocols for FAVE, including all GPT prompt templates used at each stage. Runtime placeholder values are shown in {braces}. Crucially, the GPT pipelines for data annotation and model evaluation are fully isolated: they use distinct model instances, prompt seeds, and share no context, preventing circular evaluation or data leakage.

1.1. Data Construction Protocol

The FAVE pipeline processes raw QVHighlights videos through four sequential stages, as illustrated in Figure 2 of the main paper.

Stage 1: Shot Segmentation and Material Extraction. TransNetV2 segments each video into shots. For each shot, LongVA and InternVL2.5 generate visual captions using a standard prompt instructing the model to *describe in detail the visual content appearing in the video clip*. Whisper transcribes speech and 3D-Speaker assigns speaker identities. The outputs (visual caption, speech transcript) are passed to Stage 2.

Stage 2: Event Identification and Annotation (Prompt P1). GPT-4 receives the structured shot-level material from Stage 1 and identifies meaningful temporal events, each with a generated multimodal annotation. An event must be specific and concrete, have clear boundaries from other segments, last longer than 5 seconds, and require both visual and audio content to be fully understood. Events where audio or visual content is trivial, ambiguous, or redundant are excluded.

Stage 3: Multi-Level Task Construction. The event annotations from Stage 2 are used to construct the three subsets, each targeting a different level of temporal reasoning.

FAVE-align. For each event, we construct two cross-modal QA items. In the Visual-to-Audio (V2A) direction, the visual caption is presented to the model along with a question about the audio content occurring at the same moment; the audio information is withheld from the input. The question is designed so that it cannot be answered by audio information alone; the model must ground the visual description in the video to identify the concurrent audio. The Audio-to-Visual (A2V) direction is constructed symmetrically: the audio caption is provided, the visual content is withheld, and the model is asked about the visual scene at that moment. In both cases, the ground-truth answer is drawn from the withheld modality in the annotation.

FAVE-low. For each item, two events are sampled from the same video. Their temporal relationship is derived directly from their timestamps: relative order (which event occurs first), temporal proximity (whether the two events are adjacent with no other events in between), and event position (whether an event falls in the earlier or later half of the video). Each relationship yields a multiple-choice question (A/B or Yes/No), with the correct answer determined from the timestamp metadata.

FAVE-high. Each item presents the model with a time range (in both absolute seconds and relative percentage of total video length) and asks it to generate a detailed description of the audio-visual content within that interval. The answer is the event annotation caption produced in Stage 2.

Stage 4: Quality Control (Prompt P2). Quality control proceeds in two passes: an automated Gemini-based filtering pass followed by two-round human verification by 15 trained annotators (inter-annotator agreement $\approx$ 85%). The automated pass uses a single structured prompt (P2) that evaluates each sample against a checklist covering data

quality and task validity; items failing any criterion are discarded before human review.

After automated filtering, all remaining samples undergo two-round human verification. Each item is independently reviewed by two annotators and accepted only upon consensus; disagreements are resolved by a third annotator.

P1 — Event Identification and Annotation

The following contains structured descriptions of segments from a video. `caption` is the visual description of each segment; `audio` contains the speech and dialogue information.

`{structured_shot_material}`

Identify one or more specific, clear, and logically coherent events from this video. Each event must satisfy all of the following requirements:

1. The event must be definite and concrete, with clear temporal boundaries distinguishing it from other segments. Duration must exceed 5 seconds.
2. The event must incorporate both visual and audio content, with each modality playing an equally essential role. Neither can be omitted without losing the event's meaning.
3. The event description must be written in fluent narrative prose that integrates visual and auditory information naturally. Do not simply repeat the audio transcript verbatim.
4. Visual information should be described comprehensively. Audio information should capture the speakers' specific viewpoints, attitudes, and expressed content in detail, integrated into the narrative. Do not use generic role labels (e.g., "the speaker"); reference content and context instead. Ignore trivial audio such as interjections or filler words.
5. The description must be certain and accurate. Do not include speculation, hedging, or uncertain language. If no clear event can be identified, produce no output.

For each event, output on a new line in the following format:

`time: [start_time, end_time]; caption: {event description in English}; caption_zh: {event description in Chinese}`
`raw_visual: {relevant raw visual captions}`
`raw_audio: {relevant raw audio/speech content}`

If an event recurs at multiple time intervals, list all intervals separated by commas within the `time` field. Times are in seconds.

P2 — Quality Control Checklist (Gemini)

You are reviewing a video QA sample for the FAVE benchmark. The sample contains the following fields... Evaluate the sample against each criterion below. Output `PASS` or `FAIL` for each, followed by a brief reason if `FAIL`.

Quality Criteria:

1. *Perceptual clarity.* Is the video segment visually clear (no severe blur, artifacts, or disruptive cuts)? Is the audio intelligible without heavy noise?
2. *QA clarity.* Is the question clearly and unambiguously stated, with no implicit assumptions or undefined references?
3. *Answer accuracy.* Is the ground-truth answer factually correct and supported by the video content, with no extrapolation beyond observable content?
4. *No hallucination.* Does the caption and QA contain only information present in the video? (Reject if it includes inferred emotions, unnamed persons, or fabricated details.)

Task Validity Criteria:

5. *Multimodal necessity.* Does answering the question require both visual AND audio understanding? If either modality alone is sufficient, output FAIL.
6. *Answer determinism.* Is there exactly one correct and unambiguous answer? Reject if the answer is subjective, context-dependent, or admits multiple reasonable responses.
7. *Temporal boundary integrity.* Does the time range correspond to a complete, self-contained segment? Reject if the boundary cuts mid-action or mid-sentence.
8. *Semantic independence (FAVE-low only).* Do the two events lack an obvious semantic or causal relationship that would allow temporal order to be guessed from text alone (e.g., "preparing food" then "eating")? The correct answer must require actual localization in the video.
9. *Text-bias check (FAVE-low only).* Based solely on reading the two event descriptions, could a reasonable person guess the correct temporal order without watching the video? If yes, output FAIL.

Output a JSON object: `{"1": "PASS/FAIL", "2": ..., ..., "9": "PASS/FAIL (N/A if not FAVE-low)", "overall": "PASS/FAIL"}`
The sample passes overall only if all applicable criteria pass.

1.2. Evaluation Protocol

Benchmark Card. Figure 1 illustrates a concrete case showing how the three FAVE subsets are constructed from the same annotated events. Table 1 provides a concise specification of each subset: exact model inputs, expected output

format, and evaluation metric.

Model Inference Prompts. The following prompts are provided verbatim to evaluated models. Runtime values are in {braces}.

Inference Prompt: FAVE-align (V2A)

Please first locate the time segment in the video where the following description occurs: {visual_caption}. Based on the content happening during this time segment, answer the following question: {audio_question}

Inference Prompt: FAVE-align (A2V)

Please first locate the time segment in the video where the following description occurs: {audio_caption}. Based on the content happening during this time segment, answer the following question: {visual_question}

Inference Prompt: FAVE-high

Please describe the content of the video between {start}s ({start_pct}%) and {end}s ({end_pct}%). Your description should cover both what is visually happening and what is being said or heard during this time period.

Inference Prompt: FAVE-low (Relative Order)

Read the following content and answer A or B. These are the descriptions of two events in the video.
A. {caption_A}
B. {caption_B}
Please select the event that occurred earlier in time. You should answer A or B directly.

Inference Prompt: FAVE-low (Temporal Proximity)

This is a description of two events in the video.
1. {caption_A}
2. {caption_B}
Are these two events adjacent to each other? Please answer Yes or No.

Inference Prompt: FAVE-low (Event Position)

Please determine whether this description occurs in the earlier part or the later part of the video: {caption}. Choose from the following options: B. the latter part of the video, A. the early part of the video.

GPT Evaluation Prompts (Prompt P3). All automated evaluation uses a single prompt template instantiated per task type, running on a GPT instance fully isolated from the annotation pipeline (distinct seeds and context).

P3 — Evaluation Prompts (all tasks)

FAVE-low: Choice Extraction (A/B questions)

This is the question we input to the model: {question}
This is the answer of the model: {pred}.
Please determine whether the model's answer leans more towards "A" or "B", and respond with A or B only. If unable to judge, output -1 with no additional output.

FAVE-low: Choice Extraction (Yes/No questions)

This is the answer of the model: {pred}.
Please determine whether the model's answer leans more towards "Yes" or "No", and respond with Yes or No only. If unable to judge, output -1 with no additional output.

FAVE-align and FAVE-high: Open-Ended Scoring

This is the question we input to the model: {question}.
This is the ground truth: {truth}.
This is the answer of the model: {pred}.
Please rate the model's response based on the correct answer.
The 6-point grading criteria are as follows:

- **5**: Fundamentally correct and clear; addresses all core aspects accurately.
- **4**: Mostly correct with good quality; minor details missing or slightly different perspective.
- **3**: Reasonable; does not directly match the expected response but provides solid value from a different angle.
- **2**: Partially correct but contains incorrect or misleading content.
- **1**: Completely irrelevant or provides no useful information.
- **-1**: Rejected answer or invalid content (e.g., refusal, garbled output).

Note: Ignore differences between Chinese and English; focus on semantic content only. Respond with the score directly, no additional output.

2. GPT Scoring Validation

We validate the reliability of our GPT-based evaluation through two complementary analyses. For FAVE-align and FAVE-high, where responses are open-ended, we measure *cross-model agreement*: we re-score a random sample of

Subset	Model Input	Expected Output	Metric	Notes
FAVE-align (V2A)	Video clip + visual description of a target moment	Open-ended description of audio at that moment	GPT 5-pt score	Audio channel enabled
FAVE-align (A2V)	Video clip + audio/speech description of a target moment	Open-ended description of visual scene at that moment	GPT 5-pt score	Audio channel enabled
FAVE-low (Order)	Video clip + two event captions (A & B)	"A" or "B" (which occurred earlier/later)	Accuracy (%)	GPT parses free-form outputs
FAVE-low (Proximity)	Video clip + two event captions (A & B)	"Yes" or "No" (consecutive?)	Accuracy (%)	GPT parses free-form outputs
FAVE-low (Position)	Video clip + one event caption	"A" (early half) or "B" (late half)	Accuracy (%)	GPT parses free-form outputs
FAVE-high	Video clip + time range in seconds and relative %	Open-ended audio-visual scene description	GPT 5-pt (Vision / Audio / Overall)	Relative % provided for models lacking timestamp support

Table 1. FAVE benchmark card: inputs, outputs, and metrics for each task subset. All video clips include synchronized audio unless noted. Vision-only models (TimeChat, VTimeLLM) receive muted audio and are evaluated on FAVE-high only.

50 model responses per task using both GPT-4o and Gemini as independent second raters, then compare against the original GPT-3.5 scores. For FAVE-low, where evaluation reduces to choice extraction (A/B or Yes/No parsing), we report *parse accuracy*: the fraction of model responses that can be unambiguously resolved to a valid choice.

Table 2 reports inter-rater agreement across all three FAVE subsets using GPT-4o as an independent second rater on 50 held-out samples per task. For FAVE-align and FAVE-high, which use open-ended scoring on a 6-point scale, we report *within-$\pm$1 agreement*: the fraction of pairs where the two raters' scores differ by at most one point. For FAVE-low, which reduces to choice extraction (A/B or Yes/No), we report *parse agreement*: the fraction of cases where GPT-4o's extracted choice matches the original GPT-3.5 extraction.

Task	Subtask / Dimension	Agreement (%)
FAVE-align	V2A (visual $\rightarrow$ audio)	86.0
	A2V (audio $\rightarrow$ visual)	88.0
FAVE-high	Visual scoring	86.0
	Audio scoring	80.0
	Overall scoring	86.0
FAVE-low	Relative Order	100.0
	Temporal Proximity	96.0
	Event Position	92.0

Table 2. Inter-rater agreement between GPT-3.5 (original scorer) and GPT-4o (independent second rater) on 50 randomly sampled responses per task. For FAVE-align and FAVE-high, "Agreement" is within-$\pm$1 on a 6-point scale. For FAVE-low, "Agreement" is exact parse match (A/B or Yes/No).

Agreement exceeds 86% for FAVE-align and most FAVE-high dimensions. Audio scoring is slightly lower (80%), reflecting the inherent subjectivity of evaluating spoken content descriptions. FAVE-low parse agreement is near-perfect across all three subtasks (92–100%), confirming that choice extraction is highly deterministic.

Error Attribution. To understand failure modes, we collected incorrect responses per task (filtering out timeout and incomplete outputs) and classify each into one of six error categories using GPT-4o. Tables 3 and 4 report the results.

Error Category	FAVE-align (%)	FAVE-high (%)
Temporal mislocalization	35.5	37.0
Factual error	32.0	39.0
Question misunderstanding	16.0	0.0
Incomplete / insuff. detail	14.0	3.0
Hallucination	2.0	21.0
Modality confusion	0.5	0.0

Table 3. Error attribution for FAVE-align (V2A + A2V merged, $n = 200$) and FAVE-high ($n = 100$) wrong answers (score ≤ 2), classified by GPT-4o. "Temporal mislocal." = correct content but references the wrong time segment.

For FAVE-align, temporal mislocalization (35.5%) and factual errors (32.0%) together account for over two thirds of failures, reflecting that current models struggle both to retrieve correct content and to anchor it to the right temporal window. Question misunderstanding is notable in FAVE-align (16.0%) but absent in FAVE-high, consistent with the more complex cross-modal retrieval instructions of the former. In FAVE-high, hallucination rises sharply to 21.0%, as the open-ended captioning format gives models more freedom to generate plausible-sounding but fabricated details. For FAVE-low, temporal reasoning errors dominate (70.5%), confirming that the benchmark successfully probes temporal understanding rather than surface-

Figure 1. A concrete example illustrating how each FAVE subset is constructed from the same pair of annotated events (A and B). The top section shows the video frames, audio waveform, and event captions. FAVE-align asks the model to retrieve the counterpart modality within the same temporal window (V2A or A2V). FAVE-low derives multiple-choice questions about temporal relationships between the two events (relative order, proximity, and position). FAVE-high presents the model with a time range and requires a detailed audio-visual description.

Error Category	FAVE-low (%)
Temporal reasoning error	70.5
Localization failure	16.5
Semantic prior error	5.0
Question misunderstanding	5.0
Option mapping error	3.0

Table 4. Error attribution for FAVE-low wrong answers (parsed choice $\neq$ ground truth, $n = 200$ sampled from all three subtasks merged). "Localization failure" = model cannot find the relevant event in the video; "Semantic prior error" = model relies on commonsense expectations that contradict actual video content.

level matching. Localization failure (16.5%) indicates that a significant fraction of errors occur even before temporal reasoning, when the model cannot ground the described event in the video.

3. Addressing Model Selection and Audio Data Choices

In this section, we clarify key decisions regarding model selection and audio data processing in our study.

3.1. Exclusion of Certain Multimodal Models

While models like GPT-4o [3], XinstructBlip [4], and Vast [1] support multiple modalities, they were not included in our evaluation due to accessibility constraints and limited alignment with the objectives of our study. Specifically, GPT-4o was excluded as its API supporting simultaneous image, video, and audio processing is not publicly available. Other GPT-4o variants process either image-text or

audio-text but do not jointly handle vision and audio, which is essential for audio-visual temporal perception. Additionally, certain open-source AVLLMs were excluded after preliminary testing due to either poor instruction-following capabilities or the lack of available model checkpoints.

3.2. Use of Speech Captioner for Audio Processing

During data generation, we prioritized speech-based audio descriptions by utilizing a Speech Captioner, resulting in limited coverage of environmental sounds. This choice was based on the current limitations of large audio-language models (ALLMs), which remain less reliable than speech recognition models and are prone to hallucinations. Given the availability of established datasets like VALOR, our study focuses on the joint temporal perception of speech and vision. Since speech carries significantly higher information density than other audio elements in everyday videos, FAVE primarily emphasizes speech-based audio information. For non-speech sounds (*e.g.*, waves, rain, dog barking), we manually annotated the data to supplement the dataset.

4. Future Research Directions

Despite recent advancements in audio-visual large language models (AVLLMs), our evaluation highlights significant challenges in temporal perception, multimodal alignment, and fine-grained event localization. Existing models struggle with precise temporal integration, particularly in moment-to-caption tasks, where they fail to jointly process audio and visual information over extended timeframes. Additionally, token compression techniques, while improving efficiency, often lead to loss of absolute temporal cues, impairing the models' ability to detect event boundaries and temporal relationships.

To address these limitations, future work should explore enhanced feature fusion mechanisms that retain fine-grained temporal information while reducing computational overhead. Adaptive token compression strategies could help preserve event structures without compromising efficiency. Moreover, current models lack explicit timestamp encoding, making frame- and moment-level reasoning challenging. Future AVLLMs could benefit from timestamp-aware training or contrastive learning techniques to improve temporal precision.

Another critical direction is improving multimodal alignment, particularly in audio-to-visual (A2V) tasks, where models perform significantly worse than in visual-to-audio (V2A) tasks. Augmenting training datasets with balanced, high-quality audio-visual pairs and leveraging self-supervised pretraining on audio-visual synchronization tasks may enhance models' understanding of cross-modal dependencies. Additionally, while some models like video-SALMONN perform better in audio descriptions, they still fail to fully synthesize visual and auditory elements. Future research should advance joint representation learning to better capture contextual interactions across modalities.

Lastly, benchmark expansion is necessary to push model capabilities further. Future versions of FAVE could include longer-duration videos, more complex temporal reasoning tasks, and human evaluation metrics to assess the semantic coherence of generated outputs. By addressing these challenges, future AVLLMs can move closer to human-level temporal reasoning, enabling more reliable audio-visual understanding in real-world applications.

5. Data Licenses

We obtained our data from the open-source database QVHighlights [2], available under the MIT License. Additionally, the annotations for our FAVE dataset will be publicly released under the CC BY-NC-SA 4.0 license. We aim for our dataset to serve as a key benchmark in advancing comprehensive multimodal video understanding. [1] [2]

References

[1] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *NIPS*, 36: 72842–72866, 2023. 5

[2] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 6

[3] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[4] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2024. 5

[1] https://opensource.org/license/mit

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/