

See, Hear, and Understand: Benchmarking Audiovisual Human Speech Understanding in Multimodal Large Language Models

Supplementary Material

A. Annotation Details

This section outlines the details of our annotation process. We first describe the interface used to collect audio–visual annotations, followed by the task guidelines provided to annotators to ensure consistency across examples. We then outline the criteria used to select experienced annotators for this work. Finally, we include a trivially solvable case to illustrate the intended task clarity and establish a lower bound on annotation difficulty.

A.1. Annotation Interface

Figure 5 shows the web-based interface used to collect multiple-choice audio–visual annotations. Each example is presented as a self-contained card with a video player at the top and all annotation fields displayed below. The card includes immutable metadata (e.g., `video_id`, `video_type`, `category`, `sub_category`, `task_id`) and an approved flag, followed by the natural-language question and four answer options (A–D). Annotators watch the clip, specify the temporal span used for reasoning via the `start_time` and `end_time` fields, select the correct answer, and provide a short explanation in the “reason for the answer” field, typically referencing visual or auditory evidence. The researchers responsible for review will examine the reason together with the question and answer choices to determine whether a question should be refined or approved.

A.2. Annotation Guideline

A.2.1. General Guidelines

We provide annotators with a set of general guidelines to ensure consistent and high-quality audio–visual question creation. Annotators work from a pool of English-language YouTube videos and follow a standardized workflow: each clip is added to the interface, a 5-to-30-second segment is selected, and a single multiple-choice question is authored according to the assigned task type. Annotators record the answer, and supply a brief justification.

Video Selection and Clip Duration. We provide the following constraints to maintain data quality:

- **Captions.** Select videos without persistent on-screen subtitles or captions that would trivially reveal the spoken content.
- **Content safety.** Exclude clips containing extreme or explicit violence or gore. Mild or non-gratuitous physical

conflict is allowed, but highly graphic or disturbing content should be avoided.

- **Language.** All selected videos must be in English.
- **Identity visibility.** Ensure that the queried identity is clearly visible in the foreground and is visually distinguishable from other people in the scene.
- **Discriminative attributes.** When designing questions about a specific person, choose attributes (e.g., clothing, position, actions) that clearly distinguish that person from others in the clip.
- **Contextual distractors.** Construct all answer options from entities, actions, or speech events within the same clip. When there are not enough distinct candidates, form distractors by recombining visible or auditory attributes (e.g., clothes, positions, or phrases) so that each option remains contextually plausible and balanced.
- **Clip length.** Localize each question to a contiguous 5–30 s segment, depending on the task’s requirements, trimming longer spans to the minimal window that still supports correct reasoning.

A.2.2. Task-specific Guideline

We provide task-specific requirements in Table 6. Annotators apply these templates when constructing multiple-choice questions. The annotators will add a few questions for us to provide feedback before they scale up the data annotation.

A.3. Annotator Selection

To ensure high-quality labels for tasks requiring fine-grained audio–visual reasoning, we restrict annotation to experienced researchers rather than crowd workers. Our annotator team has prior research experience in video understanding, speech processing, or multimodal large language models, enabling them to identify and select clips that are meaningful and informative for the tasks. This careful selection process is crucial for producing labels of sufficient quality and reliability, which would be difficult to achieve with crowd-sourced annotators.

In addition, researchers are trained on the full task taxonomy (see Section A.2) before annotation. This ensures that annotators understand the distinction between trivially solvable cases and those requiring joint audio–visual inference, and could apply task definitions precisely.

A.4. Eliminating Trivially Solvable Questions

A key step in our quality-control pipeline is identifying and removing *trivially solvable* questions—items whose an-

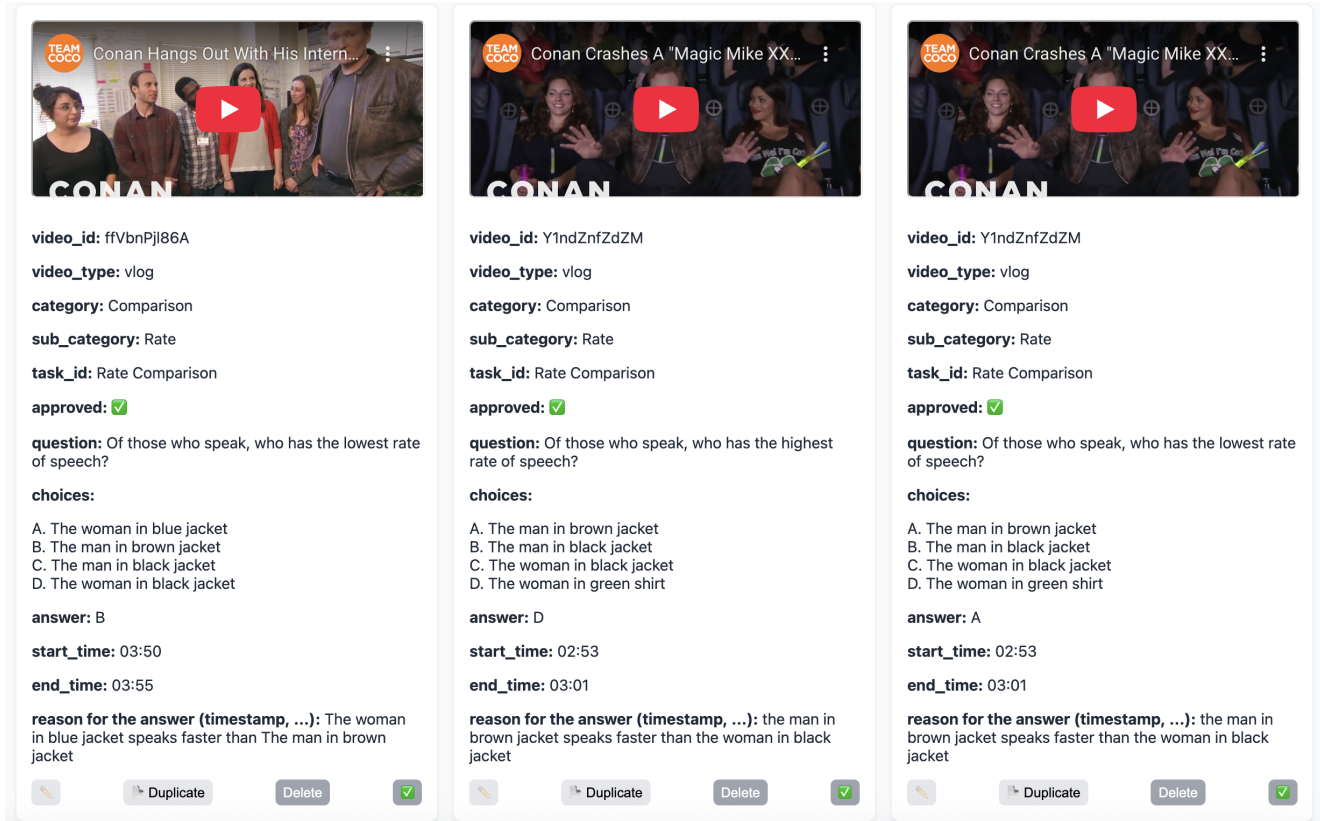


Figure 5. **Annotation interface for rate-comparison tasks.** The interface presents annotators with the video clip, metadata (video ID, category, task type), the question, all answer choices, and the selected response. Annotators also specify the temporal window used for judgment and provide a brief justification. The examples shown correspond to (left) lowest rate of speech, (middle) highest rate of speech, and (right) lowest rate of speech for a different time span within the same video. These examples illustrate how annotators validate temporal reasoning by explicitly grounding answers in the video timeline.

swers can be inferred without reference to specific temporal or spatial cues in the clip. These questions arise when the structure of the video or the phrasing of the question makes the correct answer globally obvious, regardless of any localized event. Typical examples include questions where the target event occurs throughout the video or is otherwise easily predictable, providing little meaningful challenge for the fine-grained audio-visual reasoning our tasks are designed to assess.

- **Persistent speaker patterns:** e.g., one person speaks continuously throughout the clip, making questions such as “Does the man speak when event X happens?” solvable without checking the requested moment.
- **Constant scene composition:** e.g., the number of visible people remains unchanged across the entire segment, allowing “How many people are visible when event X occurs?” to be answered without performing moment-specific counting.
- **Globally obvious answers:** cases where the question refers to a specific moment or segment, but the answer

can be determined from a property that holds across the entire clip rather than from the localized event.

- **Always-on video captions:** some videos contain burned-in subtitles throughout the entire clip. In such cases, questions involving spoken content (e.g., “What does the man say after event X ?”) become answerable by simply reading the captions rather than performing speech recognition or audio-visual alignment. Therefore, these questions fail to evaluate the intended modality.

Figure 6 shows two representative examples. In the first, only a single person is visible across the entire segment, so the answer to a moment-specific visibility question is globally obvious. In the second, the spoken line appears verbatim as burned-in captions on the screen, allowing the correct answer to be selected without listening to or timing the utterance. Such items do not evaluate the intended audio-visual capabilities and are therefore removed.



Q: At the moment the man in black suit says, "someone is making a ton of money", how many people are visible, even partially?
A. One B. Two C. Three D. Four



Q: What does the man in grey sweater says before the man in black sweaters points to him?
A. "Steven Vegali?"
B. "Steven Vegetabli"
C. "What would the plumber say to the singer?"
D. "If Steven Seagal removed all animal products from his diet, what would we call him?"

Figure 6. **Examples of trivially solvable questions removed during filtering.** (Top) A moment-specific visibility question becomes trivial because only one person is visible throughout the entire clip, making the answer recoverable without grounding to the referenced utterance. (Bottom) A speech-content question becomes trivial because the spoken line appears as burned-in captions, allowing the answer to be selected without performing audio-based reasoning.

B. Detailed Evaluation

This section provides additional details on our evaluation protocol beyond what is reported in the main paper. We first specify the exact prompts and answer format used when querying multimodal language models on our multiple-choice audio-visual questions (Sec B.1). We then describe our procedure for measuring human performance, including annotator recruitment, instructions, and aggregation of responses (Sec B.2), which serves as a reference ceiling for model accuracy on our benchmark.

B.1. Evaluation Prompt

At test time, we query models with a single unified prompt for all tasks. Figure 7 illustrates the format. Each instance consists of (i) the video input (uniformly sampled frames plus the full audio track), (ii) a fixed instruction, and (iii) a multiple-choice question with four options labeled A–D.

The instruction reads:

“Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.”

The question and answer block then specifies the task (e.g., speech intensity, speaker counting) and lists four candidate answers in natural language. We parse the model’s prediction by mapping it to the corresponding option among A, B, C, or D. Any response that does not contain a valid option is counted as incorrect. This strict, letter-only



Instruction
 Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.

Question and Answers
 Among the people who speak, who speaks the most quietly overall?
A. The woman in the gray sweater
B. The man in the dark brown shirt
C. The man in the striped shirt
D. The woman in the brown jacket
The best answer is:

Figure 7. Evaluation prompt used for multimodal LLMs. For each example we show a short video (represented here by keyframes and the waveform), a fixed natural-language instruction, and a multiple-choice question with four options (A–D). Models must answer by outputting only the letter of the correct option.

format helps create deterministic results through standardized instructions across all sub-benchmarks, following prior work [14].

B.2. Human Performance

To obtain a reliable human upper bound, we conduct a separate human evaluation that is fully decoupled from the dataset annotation workflow. None of the human evaluators participated in authoring or reviewing the dataset questions. The evaluation team consists of ten undergraduate and master’s students, all with strong English proficiency. Each evaluator answered roughly 300 questions, providing broad, non-overlapping coverage of the test set. This design follows common practice in recent multimodal benchmarks and ensures that the reported human accuracy reflects genuine task difficulty rather than annotator familiarity.

Evaluation Interface and Protocol. Figure 8 shows the interface used for human evaluation. The interface presents the video clip, accompanied by playback controls (*Play, Pause, Replay*), followed by the multiple-choice question exactly as posed to the model. Evaluators select their answer from four options (A–D) without access to transcripts, subtitles, or any auxiliary textual resources. To keep track on the number of speakers in each question, the interface also includes a control prompt asking evaluators to report the *total number of people visible in the video*.

Each evaluator completes their assigned questions through the interface in random order. After submission, the interface immediately advances to the next item, preventing revisiting or revision. Evaluators do not have access

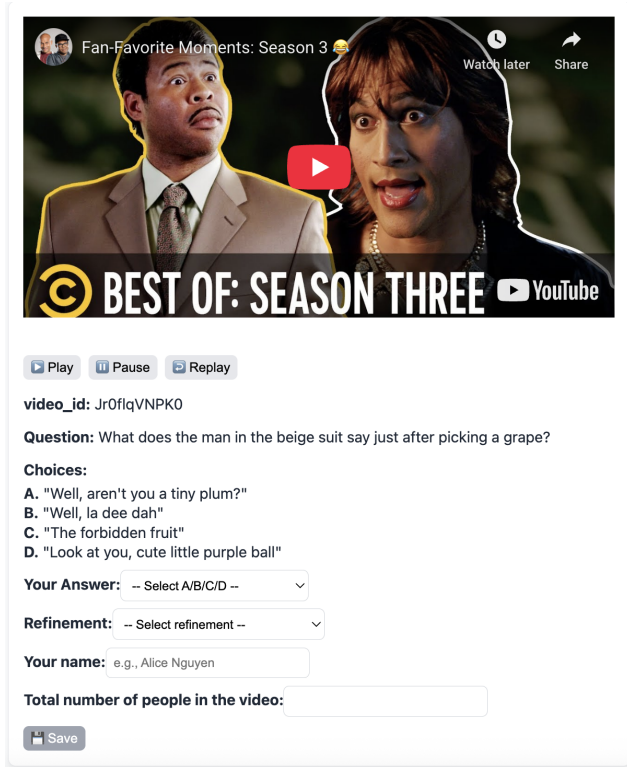


Figure 8. **Human evaluation interface.** Evaluators watch the video clip, then answer the corresponding multiple-choice question (A–D). No transcript or subtitle is provided. The interface also includes an optional refinement tag and a control question asking for the total number of people visible in the video. This setup ensures that human performance is independent of annotation and directly comparable to model outputs.

to ground-truth labels or the responses of other participants. The interface additionally provides an optional refinement tag (e.g., *trivial error*) that allows evaluators to flag questions with clear issues, such as mismatches between the question and answer options, or between the video and the question. These tags are used for data sanity checks and are not shown to models or used in computing human accuracy. In practice, no refinement tags were submitted during the human evaluation process.

This protocol ensures that human performance is (i) independent of the annotation process, (ii) based solely on audio–visual information, and (iii) directly comparable to model predictions under the same multiple-choice setting.

C. Qualitative Analysis by Error Pattern

Figure 9 presents qualitative examples of Gemini 2.5 Pro’s reasoning traces on AV-SpeakerBench, organized by four representative error patterns. In all cases, the model has access to the full video and audio segment, yet its thinking reveals where the audiovisual reasoning breaks down.

(a) Cross-modality attribution. In Figure 9a, the question asks how many unique people speak after the host waves his hand toward the screen. The video shows that only the host and a woman respond verbally. However, the model attributes the utterance to the man standing on the left, and further hallucinates an additional off-screen male speaker. This error reflects a failure to align the acoustic evidence (voice timbre and timing) with the correct visual identity, leading to incorrect speaker attribution and over-counting of speakers.

(b) Audio and visual perception. The example in Figure 9b targets the person with the lowest voice pitch among several women. The model first misperceives the visual scene, describing a *non-existent* “woman in a dark green off-the-shoulder top” instead of the actual woman in a blue dress. It then claims that the woman in a khaki sweater has the lowest pitch, even though her pitch is not the lowest in the clip. This case jointly exposes incorrect visual (wrong clothing description) and incorrect audio perception (wrong comparative pitch), showing that errors in one modality can reinforce errors in the other.

(c) Temporal grounding. The example in Figure 9c probes whether the model can restrict its reasoning to a short temporal window. The question asks how many people are visible, even partially, when the man with a cup in his hand speaks at the specific frame that the man starts saying the phrase. The model correctly identifies this time span as the anchor, but then counts people who appear before (a person walking in the background) and after (a person farther back near the cars). The answer is therefore based on people outside the annotated window, illustrating a temporal grounding error: the anchor is found, but the visual evidence is aggregated over a broader interval than the question specifies.

(d) Temporal localization. In Figure 9d, the example focuses on the number of times “I remember it” is mentioned by all speakers from the beginning of the clip until a specific action: the man in the blue sweater putting on the headphones. In the video, the relevant action occurs at approximately 00:08, and the correct count is determined by utterances up to that moment. The model, however, localizes the anchor around 00:10 in its reasoning trace, shifting the cut-off point later in time. This mis-localization changes which utterances are counted and leads to an incorrect answer, highlighting that even small temporal offsets can cause large reasoning errors in speech-counting tasks.



After the man in the middle waves his hand toward the screen, until the end of the video, **how many unique people speak, out of all the people visible on screen, even partially?**



... First, I need to pinpoint the start time, which is after the host waves his hand toward the screen. I found that moment around the 0:08 mark. I've identified: ****Speaker 1:**** the host, is clearly speaking. ****Speaker 2:**** A man on the left says something celebratory. ****Speaker 3:**** Another man on the left also speaks.

... This aligns perfectly with option D, which is "3 out of 6". I am now confident in my selection. ❌

(a) Cross-modality Attribution



Among those who speak, **who has the lowest pitch?**



... The task is to identify the woman with the lowest voice pitch. The first is a woman in a dark green, off-the-shoulder top her voice is in the mid-range. Then, a woman in a khaki or beige sweater speaks, and her voice is noticeably lower. There are other women in the video but they're not speaking.

... Only the woman in the khaki sweater, actually speaks with the lower pitch. Option D is clearly the answer. ❌

(b) Audio and Visual Perception



At the moment the man with a cup in his hand starts saying, "Yay, I'm a great student," **how many people are visible, even partially?**



... Let's see, the timestamp is around 00:12 - 00:13. Right, got it. We've got the obvious ones: the guy filming, pretty clear. Then the guy with the cup who's talking, definitely visible. Now, let's look closer. I spot someone walking in the background, and there's another person further back near the cars

... The answer is obviously B. Four. Confirmed. ❌

(c) Temporal Grounding



From the start of the video until just before the man in a blue sweatshirt is seen wearing the headset, **how many times is "I remember it" mentioned by all speakers?**



... The question states "from the beginning" to "just before" the headphones are fully on. I'll need to scrutinize the dialogue between 0:00 and roughly 0:10. Right before he's putting on the headphones at 00:10, he says, "Yeah, I remember it. I remember it." Yes, at 00:09, he definitely says, "I remember it," *twice!

... The answer is, without a doubt, A. Two. ❌

(d) Temporal Localization

Figure 9. **Qualitative examples of Gemini 2.5 Pro** reasoning traces on **AV-SpeakerBench**. Green and red highlight colors indicate the model's correct and incorrect reasoning, respectively. The figure above contains representative failure cases spanning four key error patterns: (a) cross-modality attribution, (b) audio and visual perception, (c) temporal grounding, and (d) temporal localization. Detailed analyses are provided in the subsection.

D. Question Examples

To make the task definitions concrete, we visualize representative multiple-choice question-answer pairs from AV-SpeakerBench. Each example is shown as a panel containing (from top to bottom) a strip of video frames, the corresponding audio waveform, a natural-language question, and four answer options. Across all tasks, the model is evaluated by its accuracy in selecting the correct option, given the paired audio-visual clip.

Figure 11 groups the *speaker-centric* tasks: Speaker Detection, Speaker Recognition, and Speaker Counting. In these tasks, the evaluation focuses on whether the model can correctly associate speech segments with the corresponding visible speaker(s), distinguishing who is speaking, when they speak, and how many distinct speakers are active.

Figure 12 illustrates *speaker-visual* tasks such as Activity Recognition and visually grounded counting. Here, the model must reason jointly about a speaking person and their on-screen actions or surrounding context (e.g., who performs an action while or around speaking, or how often a visually specified person speaks), coupling appearance, motion, and speech.

Figure 13 showcases *speech reasoning* tasks: Speech Recognition, Speech Counting, and Speech Duration. The evaluation in these tasks requires recovering short spoken phrases, counting phrases, and compare the duration of speech.

Finally, Figure 14 presents *speech attribute* tasks, including and comparative speech questions on rate, pitch, and intensity. These examples emphasize fine-grained acoustic reasoning: the model must use the audio waveform to com-

Task Type	Speaker-centric			Visual-centric			Audio-centric						Overall
	Detection	Recognition	Counting	Attribute Recognition	Activity Recognition	Counting	Recognition	Duration	Pitch	Rate	Intensity	Counting	
Human Performance	96.02	93.13	94.28	93.14	93.20	94.15	96.52	90.68	93.20	91.39	94.17	93.40	93.74
Gemini 2.5 Pro Thinking [8]	81.73	74.15	74.13	72.55	73.30	62.93	77.11	78.81	67.48	69.86	71.84	63.89	73.04
Gemini 3 Pro Thinking	85.95	79.86	73.13	80.39	79.13	71.71	87.06	75.85	70.87	78.47	75.24	70.14	77.62

Table 4. Evaluation Results of Gemini 3 Pro (Thinking) on AV-SpeakerBench.

Model (Size)	8f	16f	1fps
Gemini 2.5P (-)	70.4	72.0	73.0
Phi4-MM (5.6B)	38.5	39.2	39.5
VITA 1.5 (7B)	36.5	36.8	36.3
Qwen2.5-O (7B)	41.0	41.3	42.3
Qwen3-O (30B)	51.3	53.1	54.1

Table 5. Performance on different sampling strategies.

pare how different people speak (e.g., how fast, how high or loud).

For compactness, each figure arranges three examples per row and one row per task group, while maintaining a consistent interface across all panels. Together, these four figures highlight the diversity of natural-language formulations and reasoning skills targeted by AV-SpeakerBench, spanning identity, actions, counting, appearance, speech, and attributes within a unified audio-visual QA framework.

E. Performance of Gemini 3 Pro on AV-SpeakerBench

We evaluate the recently released Gemini 3 Pro (Thinking)¹ and compare it with Gemini 2.5 Pro on AV-SpeakerBench.

As shown in Table 4, Gemini 3 Pro attains an average accuracy of **77.62%**, outperforming Gemini 2.5 Pro (73.04%) by **+4.6** points. The improvements are broad and most pronounced in tasks requiring fine-grained audio-visual grounding, including speaker-centric understanding, visual reasoning, speech recognition, and prosody-related perception. However, several categories—most notably counting and speech-attribute reasoning—remain far from solved, and a substantial gap persists relative to human performance (93.74%).

Overall, Gemini 3 Pro represents meaningful progress but still falls short of demonstrating robust and generalizable audiovisual reasoning.

F. Evaluation beyond default frame sampling.

We conducted additional experiments with models achieving non-trivial accuracy on AV-SpeakerBench under different frame sampling densities. Notably, the strongest models

¹Released after the main paper deadline.

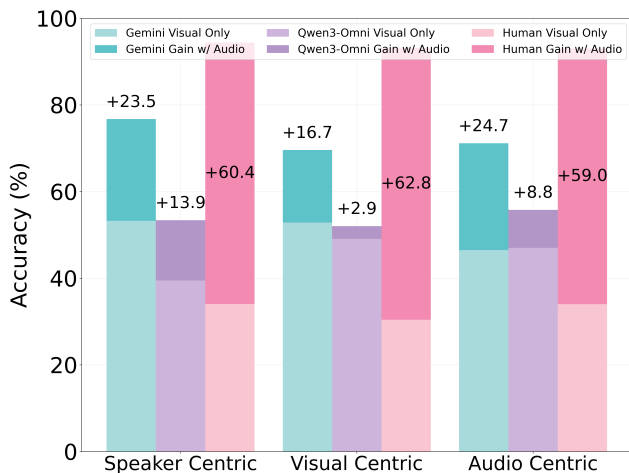


Figure 10. Performance under visual-only (V) and audiovisual (A+V) settings.

already adopt a similar default policy (1 fps with a frame cap). We report results with 8 and 16 frames since these 2 numbers are widely adopted in audiovisual large language model. Results show that these fixed number of frames consistently leads to slightly *lower* accuracy, while preserving the same model ranking (Table 5).

G. Audio necessity vs. vision-only solutions.

AV-SpeakerBench enforces fusion through question design, grounding speech to visible identities and temporal events. Human evaluation shows that the benchmark is largely *not solvable* from vision alone, while audiovisual input yields near-ceiling performance (Figure 10). In fact, we observe the same trend across all 12 tasks. Some items may still be inferred from natural visual conversational cues (e.g., expressions or gestures), but overall audiovisual fusion provides the most reliable and general solution. We further label questions that humans cannot answer under vision-only as an *audio-required* subset; Gemini 2.5 Pro achieves 72.7% on this subset versus 73.04% overall.

H. Ethical Statement

We define our dataset as a collection of publicly available YouTube videos paired with precise timestamps and human-written annotations. The benchmark is used solely for evaluation, not model training, and thus poses min-

imal risk of bias amplification or unintended memorization. The dataset will be released under the CC BY-NC-SA 4.0 license, which restricts use to non-commercial research and prohibits applications involving facial recognition, surveillance, or biometric identification. Individuals featured in referenced videos may request removal, and we will promptly withdraw the corresponding segments from future releases.

Speaker Detection



- Q:** Does the man wearing glasses speak **before** the man in the blue suit starts saying "Good, uh?"
- A. No, he only shows an awkward smile
 - B. Yes, he gives in after refusing at first**
 - C. Yes, he chews and replies in an unfriendly tone
 - D. No, he only looks at the other man



- Q:** Does the woman with blonde hair speak in the video?
- A. No, she stands up and walks away
 - B. No, she only nods and smiles
 - C. Yes, she is listing a few things she knows
 - D. Yes, she acknowledges the man's point**



- Q:** Does the man in the khaki coat speak **after** the man in the dark green sweater stands up?
- A. No, he only listens to the argument between the other two men
 - B. Yes, he refutes the request from the man in the dark green sweater**
 - C. Yes, he argues with the man in the red clothes
 - D. No, he only yells at the other men

Speaker Recognition



- Q:** Who speaks **right after** the man in the red shirt and the man in the gray T-shirt do a fist bump?
- A. The man in a gray shirt says "great to meet you!"
 - B. The man wearing glasses says "what's your name?"**
 - C. The man wearing a backpack says "what's your name?"
 - D. The man in red shirt says "It's never too late"



- Q:** Who speaks **immediately before** the woman in the black shirt says, "Tarzan. Well, you say because he just can't compete with me"?
- A. The woman with blonde hair**
 - B. The man with blonde hair
 - C. The man with black hair
 - D. The woman with black hair



- Q:** Who says, "Oh, I see what's going on"?
- A. The man in the blue shirt
 - B. The man in the striped jacket
 - C. The man in the white sweater
 - D. The man in the black suit**

Speaker Counting



- Q:** **After** the man in the red patterned shirt finishes saying "No actually, I think JD would be the one", and until the end of the video, how many unique people speak, out of all the people visibly present on screen, even partially?
- A. 2 out of 4
 - B. 2 out of 6
 - C. 3 out of 5**
 - D. 1 out of 4



- Q:** From the start of the video until **just before** the man in the checkered shirt starts saying "oh yeah wait", how many unique people speak, out of all the people visibly present on screen, even partially?
- A. 5 out of 7
 - B. 3 out of 5
 - C. 4 out of 6
 - D. 5 out of 6**



- Q:** **After** the man without a beard closes the door, until the end of the video, how many unique people speak, out of all the people visible on screen, even partially?
- A. 2 out of 3
 - B. 2 out of 2**
 - C. 3 out of 3
 - D. 3 out of 4

Visual Reference Audio Reference

Figure 11. Visualization of speaker tasks. Top: Speaker Detection. Middle: Speaker Recognition. Bottom: Speaker Counting.

Activity Recognition



- Q:** When does the woman in the black skirt say, "That would be so much fun"?
- A. After the man in the black shirt walks to the window
 - B. After the woman in the black choker sits on the couch
 - C. As the man in the black shirt points at the window
 - D. As she walks across the room



- Q:** What does the man in the green shirt do right before the man in the white t-shirt says, "Long time, no see"?
- A. Touches the lips of the man in the white t-shirt
 - B. Claps the hand of the man in the white t-shirt
 - C. Pats the back of the man in the white t-shirt
 - D. Hugs the man in the white t-shirt



- Q:** What does the man with the mustache do right after he says, "The lord hates a coward"?
- A. Brings his hands together in prayer
 - B. Shakes the hand of the man next to him
 - C. Taps the man next to him on his arm
 - D. Puts a locket away in his pocket

Visual Counting



- Q:** At the moment the man in a black suit sitting in a chair finishes saying, "This is what I meant," how many unique people are visible, even partially?
- A. Five
 - B. Three
 - C. One
 - D. Four

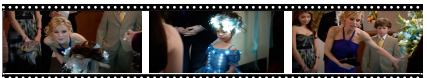


- Q:** After the woman with blonde hair says, "I'm a history major," until the end of the video, how many unique people are visible, even partially?
- A. One
 - B. Two
 - C. Three
 - D. Zero



- Q:** From the start of the video until just before the man driving the car starts saying, "Uh, what's reverse?", how many unique people are visible, even partially?
- A. One
 - B. Two
 - C. Three
 - D. Four

Attribute Recognition



- Q:** What does the person who speaks immediately before the person who says, "that was in the car" wear?
- A. Green dress
 - B. Burgundy jacket
 - C. Black suit
 - D. Blue dress



- Q:** What is the person wearing who speaks immediately after the line "I thought this was the bathroom" is spoken?
- A. Blue tank top
 - B. Gray turtleneck
 - C. Black crop top
 - D. Silver earrings

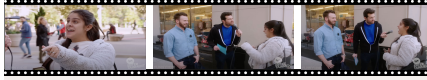


- Q:** What does the person who speaks immediately before the person who says, "that was in the car" wear?
- A. Green dress
 - B. Burgundy jacket
 - C. Black suit
 - D. Blue dress

Visual Reference Audio Reference

Figure 12. Visualization of speaker-visual tasks. Top: Activity Recognition. Middle: Visual Counting. Bottom: Attribute Recognition.

Speech Counting



Q: How many times the "15th anniversary" is mentioned by the man wearing blue button up?
A. Zero
 B. One
 C. Two
 D. Three



Q: From the start of the video until **just before** the man with the grey scarf receives the money, how many times is "Britney Spears" mentioned?
 A. Zero
B. Three
 C. Five
 D. Seven



Q: **After** the woman with white earrings turns away, until the end of the video, how many times is "never that name" mentioned?
 A. One
 B. Two
 C. Zero
D. Three

Speech Duration



Q: From the start of the video until **just before** a person holding a tray appears, who speaks the least among those who speak?
 A. The person in a dark green T-shirt
B. The person in a dark red T-shirt
 C. The person in a light yellow T-shirt
 D. The person in a light blue T-shirt



Q: **After** the person enters the door, until the end of the video, who speaks the most among those who speak?
A. The man in the green uniform
 B. The man in the white shirt
 C. The man in the black vest
 D. The man in the purple shirt



Q: **After** the smartphone screen appears, until the end of the video, who speaks the most among those who speak?
 A. The man with black hair
 B. The man with blonde hair
 C. The woman with black hair
D. The woman with blonde hair

Speech Recognition



Q: What does the man in the black suit say **just after** he stands up?
A. "What do you have in your mouth?"
 B. "Qualification?"
 C. "Wait a moment."
 D. "Nothing, mayhem."



Q: What does the woman sitting on the porch swing say **just before** she drinks from her glass?
 A. "The voice just now."
B. "We don't hear anything."
 C. "What is it?"
 D. "What is that?"



Q: What does the man in the white shirt say **while** pointing at the girl?
 A. "That's how it is."
 B. "Do you mean Jacquelin?"
 C. "You all want to play."
D. "I have my eyes on you."

Visual Reference Audio Reference

Figure 13. Visualization of speech tasks. **Top:** Speech Counting. **Middle:** Speech Duration. **Bottom:** Speech Recognition.

Speech Intensity



Q: Among the people who speak, who speaks the most quietly overall?

- A. The woman in the gray sweater
- B. The man in the dark brown shirt
- C. The man in the striped shirt
- D. The woman in the brown jacket



Q: Of those who speak, whose voice gets the loudest at any moment in the video?

- A. The man with gray shirt
- B. The woman with purple blazer
- C. The woman with pink dress
- D. The man with navy shirt



Q: Among the people who speak, who speaks the most quietly overall?

- A. The woman in green shirt
- B. The woman in purple sweater
- C. The man in blue gilet
- D. The man in black gilet

Speech Pitch



Q: Among those who speak, who has the lowest pitch?

- A. The man with a mustache
- B. The man with gray hair
- C. The man in a green t-shirt
- D. The man in a green sweater



Q: Among those who speak, who has the highest pitch?

- A. The man in glasses
- B. The woman with gray hair
- C. The woman with blonde hair
- D. The man with gray hair



Q: Among those who speak, who has the lowest pitch?

- A. The man with no hair
- B. The man with pink hair
- C. The woman in brown shoes
- D. The woman in black shoes

Speech Rate



Q: Of those who speak, who has the highest rate of speech?

- A. The man in brown jacket
- B. The man in black jacket
- C. The woman in black jacket
- D. The woman in green shirt



Q: Of those who speak, who has the lowest rate of speech?

- A. The woman in the black top
- B. The woman in the pink top
- C. The woman in the blonde top
- D. The woman in the blue top



Q: Of those who speak, who has the lowest rate of speech?

- A. The woman in black dress
- B. The woman in golden dress
- C. The man in black shirt
- D. The man in white shirt

Visual Reference

Audio Reference

Figure 14. Visualization of speech attribute tasks. **Top:** Speech Intensity. **Middle:** Speech Pitch. **Bottom:** Speech Rate.

Table 6. Task-specific goals and requirements for all annotation tasks.

Task Name	Requirement
Speaker Detection	<p>Goal: Test whether models can detect whether a person is currently speaking among multiple visible people.</p> <ul style="list-style-type: none"> • Select a 5–30 s clip with at least two visible individuals, where only one person is clearly speaking during a short moment. • Choose a target person (e.g. the man in black suit), and a short utterance as an audio anchor or a short event as a visual anchor. • Formulate a question such as: “Does the [person] speak” or “Does the [person] speak before/after/when [utterance/event] occurs?” • Provide four answer options, with 2 No and 2 Yes. For each No option, put a visual reference (e.g. No, he just stares at the opposite person). For each Yes option, put an audio reference (e.g. Yes, he talks about his plan). • If there is not enough audio/visual reference, please create plausible distractors that fit the scene context.
Speaker Recognition	<p>Goal: Evaluate whether models can link an utterance to the correct visible speaker.</p> <ul style="list-style-type: none"> • Select a 5–30 s clip with multiple visible speakers. • Choose a short utterance as an audio anchor or a short event as a visual anchor. • Formulate a question using: “Who speaks right after/before/when [person] says [utterance]?”, “Who responds to [person] saying [utterance]?”, or “Who says [utterance]?” • Provide four answer options referring to different individuals (e.g., “the man in the black suit”, “the woman in the red dress”). • If fewer than three people appear, create plausible distractors that fit the scene.
Speaker Counting	<p>Goal: Assess whether models can count how many distinct people speak in a segment.</p> <ul style="list-style-type: none"> • Select a 10–30 s clip with multiple visible people and several distinct spoken turns. • Choose a short utterance as an audio anchor or a short event as a visual anchor. • Formulate a question such as: “From the start of the video until [utterance/event], how many people speak, out of the people in the video?” or “After [utterance/event] until the end of the video, how many people speak, out of the people in the video?” • Provide four numerical answer options in the format “n out of m” (e.g., “1 out of 2”, “2 out of 3”); the first number must be the true count of speakers in the specified interval, and the second number must be the number of visible people in that segment. • Ensure that speakers are visually distinguishable and separated by clear turn-taking; create plausible nearby counts as distractors.
Attribute Recognition	<p>Goal: Test whether models can connect a speaker’s appearance to their speech right before or after an anchor.</p> <ul style="list-style-type: none"> • Select a 10–30 s clip with multiple visible speakers. • Choose a short utterance as an audio anchor. • Formulate a question such as: “What is the appearance of the person who says [immediately before/after/when] the person says [utterance]?” or “What does the person who says [utterance] wear?” • Ask about attributes like clothes, clothing shape, clothing color, hair, hairstyle, or hair color. • Provide four answer options describing different plausible appearances; exactly one must match the correct speaker. • If fewer than four distinct appearances are visible, create plausible distractors consistent with the scene.

Task Name	Requirement
Activity Recognition	<p>Goal: Measure whether models can associate a speaker with actions that happen before/after/while they speak.</p> <ul style="list-style-type: none"> • Select a 10–30 s clip with multiple visible speakers and at least one clear action (e.g., standing up, waving, pointing). • Choose a short utterance as an audio anchor. • Formulate a question such as: “What does the person do [immediately before/after/when] the person says [utterance]?” or “When does the person say [utterance]?” • For the first question type, encode the activity (e.g., “stands up”, “raises a hand”) into the four answer choices. • For the second question type, encode [before/after/when] + [activity] inside the four answer choices. • If fewer than three distinct activities appear, create plausible distractors consistent with the scene.
Visual Counting	<p>Goal: Assess whether models can count visible entities conditioned on an audio anchor.</p> <ul style="list-style-type: none"> • Select a 10–30 s clip with multiple visible people. • Choose a short utterance as the reference time. • Formulate a question such as: “After/Before/When the [utterance] occurs, how many people are visible, even partly?” • Provide four numerical answer options (e.g., 1, 2, 3, 4) or phrases that clearly encode the count. • Ensure all counted entities are clearly visible in the frame; create plausible distractors using nearby counts.
Speech Recognition	<ul style="list-style-type: none"> • Goal: Evaluate whether models can recognize the spoken content in the clip. • Select a 10–30 s clip with multiple visible speakers. • Choose a short visual event as an anchor. • Formulate a question such as: “What does the person say [just before/after/when] the [event] occurs?” • Provide four answer options corresponding to different possible transcripts; exactly one must match the spoken phrase (allow light paraphrasing only for the correct option). • Ensure distractor phrases are grammatical and plausible in the context of the video.
Speech Duration	<p>Goal: Check whether models can reason about relative speaking time.</p> <ul style="list-style-type: none"> • Select a 5–20 s clip where at least two people each speak for clearly different durations. • Choose an event as a visual anchor. • Formulate a question such as: “Of those who speak, who speaks the most/least?”, “From the start of the video, until [event], who speaks the most/least?”, or “After [event], until the end of the video, who speaks the most/least?” • Provide four answer options referring to different individuals (e.g., “the man in the black suit”, “the woman in the red dress”). • Ensure the duration difference is perceptible (at least 1–2 s); construct plausible but incorrect visual options as distractors.
Speech Pitch	<p>Goal: Test whether models can distinguish speakers by relative pitch.</p> <ul style="list-style-type: none"> • Select a 5-10 s clip where at least two speakers have clearly different pitch ranges. • Formulate a question such as: “Among those who speak, who has the highest/lowest pitch?” • Provide four answer options referring to different individuals in the scene (e.g., “the man in the black suit”, “the woman in the red dress”). • If fewer than three speakers are present, create plausible distractors based on other visible characters that could plausibly speak.

Task Name	Requirement
Speech Rate	<p>Goal: Measure whether models can compare how fast different speakers talk.</p> <ul style="list-style-type: none"> • Select a 5–10 s clip with at least two speakers whose speaking rates differ noticeably. • Formulate a question such as: “Among those who speak, who has the highest/lowest rate of speech?” • Provide four answer options referring to different speakers (e.g., by clothing or position). • Ensure the difference in speaking rate is clear (e.g., more syllables in similar time), and create plausible distractors when needed.
Speech Intensity	<p>Goal: Assess whether models can reason about speech loudness.</p> <ul style="list-style-type: none"> • Select a 5–10 s clip where at least two speakers talk with clearly different loudness levels (e.g., one shouting, one speaking softly). • Formulate a question such as: “Among those who speak, who has the highest/lowest voice?” • Provide four answer options referring to different individuals (e.g., “the man in the black suit”, “the woman in the red dress”). • Avoid clips where microphone distance alone explains loudness, unless this is visually clear; create plausible distractors as needed.
Speech Counting	<p>Goal: Evaluate whether models can count number of phrases or keywords.</p> <ul style="list-style-type: none"> • Select a 5–30 s clip with multiple short utterances or turns of speech. • Choose an event as a visual anchor. • Formulate a question such as: “From the start of the video, until just before [event], how many times [phrase] is mentioned by [person/everyone]?” or “After [event], until the end of the video, how many times [phrase] is mentioned by [person/everyone]?” • Provide four numerical answer options; exactly one must match the true count. • Ensure utterances are separated enough to be countable (clear pauses or turn-taking), and use nearby counts (e.g., $n - 1$, n, $n + 1$, $n + 2$) as distractors.