LONGCOREF-AV: BENCHMARKING AUDIO-VISUAL IDENTITY COREFERENCE IN LONG-FORM VIDEO

Anonymous authors
Paper under double-blind review

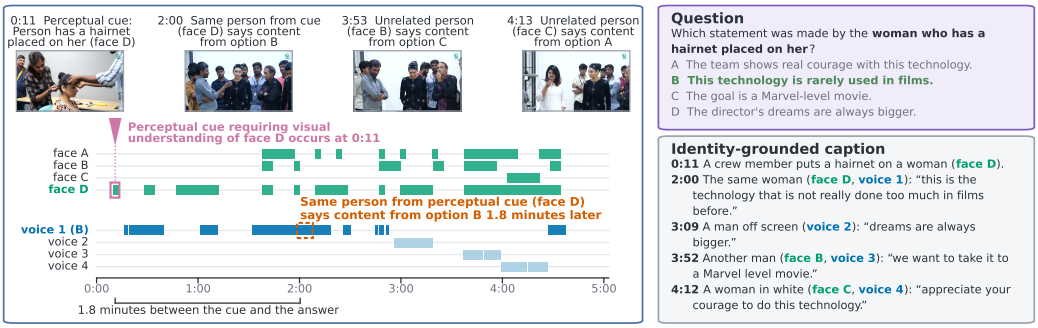


Figure 1: The identity coreference task. Each question identifies a person through a visible or audible cue at one moment, then asks about that same person at a distant moment in the video. Answering correctly requires linking the two distant moments through the person's face or voice, directly testing a model's multimodal perceptual grounding.

ABSTRACT

An unresolved challenge for multimodal models processing long-form video is determining which faces, voices, and actions across distant moments belong to the same person. To measure this capability, we introduce LONGCOREF-AV, a benchmark of 1,126 human-verified multi-choice questions across 678 long-form videos. Questions identify a target individual only through visible or audible cues, such as clothing, gestures, or laughter, and are designed to rely on evidence from moments minutes apart. Each question is labeled with the category of evidence it requires to be answered correctly: voice, visual, or joint. Across 16 systems, closed frontier models far exceed a text-only baseline, open models lag far behind, and voice-dependent questions prove to be the most challenging across all models. When given only the annotated video segments containing the relevant evidence rather than the full video, open models improve far more than frontier models, suggesting that existing open models mainly fail to find evidence, while frontier models are more likely to locate it but often attribute it to the wrong person. In a separate identity-matching test, the two strongest open omni models both incorrectly label more than half of pairs depicting different individuals as the same person, despite their strong performance on multi-choice questions. This discrepancy shows that end-to-end question accuracy does not certify the intermediate identity links: models can select correct options from partial contextual cues without reliably distinguishing individuals across moments. Our findings expose persistent identity attribution as a critical weakness of multimodal LLMs, positioning LONGCOREF-AV to support more reliable attribution of actions to individuals in long-form video. Code and data will be made publicly available.

1 INTRODUCTION

Current multimodal LLMs struggle to distinguish and re-identify individuals across long-form videos involving several people. An individual may appear, speak, leave the scene, and return minutes later, while an off-screen voice must be associated with a previously observed face. Answering a query such as "Which statement was made by the woman who laughs out loud?" requires an *identity chain*: grounding the perceptual cue to an individual, linking that individual across moments and modalities, and attributing the relevant speech or action. We call an association between two

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

appearances of the same individual an *identity link*, such that a *voice-to-person link* specifically associates a voice with its speaker. We term the overall task *identity coreference* (Figure 1). Since errors anywhere in this chain can invalidate claims about who said or did what, direct evaluation is necessary.

Existing benchmarks evaluate this chain only partially (Table 1). Speaker benchmarks use clips lasting seconds to a few minutes (Chowdhury et al., 2026; Kwon et al., 2026; Nguyen et al., 2026) or provide cast lists and pre-segmented utterances at episode scale (Li et al., 2026c); omni suites embed identity questions within broader perception tasks (Li et al., 2026a; Hong et al., 2026; Zhou et al., 2025); and long-video benchmarks remain predominantly vision-centric and exclude audio (Fu et al., 2025; Mangalam et al., 2023; Wang et al., 2025). Most of these benchmarks also leave unclear whether a score reflects the video itself: models can often answer from the question and options alone (Chen et al., 2024a; Shangguan et al., 2025), and accepting audio input does not establish that it is effectively used. Furthermore, few report a text-only baseline, apply option rotation (Sakshi et al., 2025), or report a visual-only condition using sampled frames without audio or an ASR transcript.

Table 1: Positioning against related benchmarks. Top: speaker and identity benchmarks; bottom: general audio-visual suites with an identity subset. *Setting*: **AV**, audio-visual input; **LV**, minutes-long continuous video. *Identity chain*: **VF**, voice-to-face attribution; **NG**, name grounding inside the video; **CT**, cross-time tracking of people who return; **SR**, speaker reasoning beyond attribution. *Protocol*: **EV**, per-question evidence spans; **RE**, required-evidence labels checked by modality ablation; **TB**, text-only baseline; **HV**, human verification of all questions; **PR**, public release (checked Aug. 2026). ● = full, ◐ = partial, ○ = absent; †: unreported, graded conservatively. Rules and per-cell evidence are shown in Appendix D.

	Setting		Identity Chain				Protocol				
Benchmark	AV	LV	VF	NG	CT	SR	EV	RE	TB	HV	PR
M3-SLU (Kwon et al., 2026)	○	◐	○	○	○	◐	○	○	○	●	○
MSU-Bench (Sun et al., 2026)	○	◐	○	◐	◐	●	○	○	○	●	●
AV-SpeakerBench (Nguyen et al., 2026)	●	○	●	○	○	●	○	○	○	●	●
AMusE (Chowdhury et al., 2026)	●	○	●	○	◐	◐	○	○	○	◐	○
HumanOmni-Speaker (Bai et al., 2026)	●	○	●	◐	○	◐	○	◐	○	◐	○
DVD-Bench (Tang et al., 2026)	●	○	●	○	○	◐	○	○	○	●	○
DramaSR (Li et al., 2026c)	●	●	●	◐	●	◐	○	○	○	◐	◐
CrossVideoQA (Meng et al., 2025)	○	○†	○	○	●	○	○	○	○	◐	○
ISYV-Bench (Gao et al., 2026)	○	◐†	○	○	●	○	○	○	○	◐	○
AVUT (Yang et al., 2025b)	●	○	◐	○	○	◐	◐	○	◐	●	●
StreamingBench (Lin et al., 2024)	●	◐	◐	○	○	◐	○	○	○	●	●
JointAVBench (Chao et al., 2026)	●	◐	◐	○	○	◐	○	◐	○	●	◐
EgoLife (Yang et al., 2025a)	●	●	○	○	●	◐	◐	○	○	●	●
LONGCOREF-AV (ours)	●	●	●	●	●	●	●	●	●	●	●

To evaluate this chain directly, we introduce LONGCOREF-AV, comprising 1,126 four-choice questions across 678 long-form videos drawn from open web sources and films. Its three-stage construction pipeline (Figures 1 and 2) produces questions that require resolving an individual's identity across temporally separated observations. First, source filtering retains web videos with at least two speakers separable by voiceprint clustering and film excerpts with at least two recurring faces or released diarization annotations, producing a speaker-attributed transcript and a timeline of face appearances. Second, question generation identifies the target individual exclusively through perceptual cues, including visual attributes, actions, on-screen text, and non-verbal vocalizations, rather than names or discourse roles, thereby limiting transcript-only shortcuts. Each question includes annotated evidence spans, often distributed across moments separated by several minutes, and a required-evidence label: *voice*, *visual*, or *joint*. Third, human reviewers check the candidates against the source video; roughly 70% pass review. Candidate admission is independent of model outputs, avoiding model-conditioned selection bias. LONGCOREF-AV spans six scene families and 91 hours and will be released with the full media and the scoring script.

On LONGCOREF-AV, the leading frontier system, Gemini-3.1-Pro, reaches 86.9% accuracy, compared with a text-only baseline near 60%. The strongest open omni model, Qwen3-Omni-Instruct, trails by 13 percentage points, and voice-dependent questions are the hardest for every video-grounded model. Our annotations localize the bottleneck to the voice-to-person link: open models primarily fail to retrieve evidence, whereas frontier models appear to retrieve it but misattribute the speaker. In a direct identity-link test, Qwen3-Omni-Instruct achieves 74.1% multiple-choice accuracy yet labels 96.6% of different-person pairs as the same person; Qwen3-Omni-Thinking still labels 52.6% as the same. None of the tested interventions repairs this link: denser evidence frames, replacing audio with transcripts, supervised or reinforcement fine-tuning, and voiceprint tools improve evidence retrieval at most, while speaker-attributed captioning exhibits the same failure. These results motivate architectures and training objectives that explicitly maintain person-specific representations and link voices to faces across long contexts.

Our contributions can be summarized as follows:

- **A dedicated benchmark.** We introduce LONGCOREF-AV, the first benchmark for audio-visual identity coreference in long-form video. It contains 1,126 human-verified questions across 678 videos; each refers to its target individual through perceptual cues and is annotated with evidence spans and a required-evidence label.
- **A controlled evaluation of 16 systems.** We evaluate closed models, open models that take video with audio, and open vision-language models given frames and an ASR transcript. Every system is also evaluated on questions and options alone to quantify trivial text-only solvability.
- **Diagnostics of the identity chain.** We introduce a turn-level protocol that varies one input axis at a time, together with evidence-restricted evaluation, direct same-person tests, and speaker-attributed captioning, to distinguish failures of evidence retrieval from failures of identity attribution.

2 RELATED WORK

Speaker and identity benchmarks. Table 1 compares thirteen related benchmarks (Appendix D). Benchmarks for diarization (Xu et al., 2022; Liu et al., 2022), speaker verification (Nagrani et al., 2017), and active-speaker detection (Roth et al., 2020) primarily use short clips or reference identities. Our filtering pipeline builds on these methods, but our questions link individuals across minutes without reference faces or voice samples. Recent benchmarks range from 39-second clips in AMusE (Chowdhury et al., 2026) to audio-only segments of at most 3.5 minutes in M3-SLU (Kwon et al., 2026); others report neither text-only baselines nor evidence spans (Nguyen et al., 2026; Chao et al., 2026). DramaSR (Li et al., 2026c) covers complete episodes but supplies cast lists and pre-segmented utterances, reducing attribution to fixed-list selection. TOC-Bench (Chen et al., 2026) verifies the need for temporal context, but only for visual object consistency.

Long-form video and omni evaluation. We build on long-video and omni benchmarks (Fu et al., 2025; Wang et al., 2025; Chandrasegaran et al., 2024). Whereas EgoSchema (Mangalam et al., 2023) uses temporal certificates to estimate the minimum video content needed to answer a question, we provide evidence spans for every question. Following CG-Bench (Chen et al., 2025a), we evaluate models on these spans alone and rotate answer options as in MMAU (Sakshi et al., 2025). Omni suites (Li et al., 2026a; Hong et al., 2026; Zhou et al., 2025) and captioning work (Ma et al., 2026) cover broad perception but do not isolate identity coreference. Following Chen et al. (2024a) and Shangguan et al. (2025), we report a text-only baseline for every system.

Tool-augmented systems and training. LLMs increasingly use retrieval, ASR, OCR, and zoom tools (Luo et al., 2026; Zhang et al., 2026; Yin et al., 2026; Gao et al., 2025; Ding et al., 2026) or refine diarized speaker labels (Wang et al., 2024). To test whether such tool use improves identity attribution, our text-side study (Appendix Q) lets models select voiceprint queries, performing 1 to 8 points worse than providing all results at once. DramaSR-LRM (Li et al., 2026c) trains tool-using agents with a cast list, whereas LONGCOREF-AV provides none. Post-training methods such as reinforcement learning can improve speaker attribution (Tang et al., 2026), but do not alone resolve identity association: speech-aware LLMs still struggle to identify speakers (Thebaud et al., 2026), and vision-language models to recognize a specific person or object across images (Shi et al., 2026).

3 THE LONGCOREF-AV BENCHMARK

LONGCOREF-AV provides no identity information external to the video: a system receives only the continuous audio-visual input and a four-choice question, without cast lists, character names, or pre-segmented turns. Three properties make the questions challenging. (1) **Temporal separation**: in 69% of questions, the evidence occurs at two or more moments separated by a median of 2.6 minutes, within videos lasting 3–13 minutes. (2) **Speech and voice evidence**: in 53% of questions, the evidence includes speech content or vocal characteristics. (3) **High speaker density**: videos contain 2–13 audible speakers, with a median of 4, requiring the target individual to be distinguished from several others.

3.1 VIDEO SOURCES AND CURATION

Videos are drawn from the openly licensed FineVideo corpus (Farré et al., 2024), covering interviews, vlogs, live events, procedures, and news, and from feature-film excerpts, which contribute

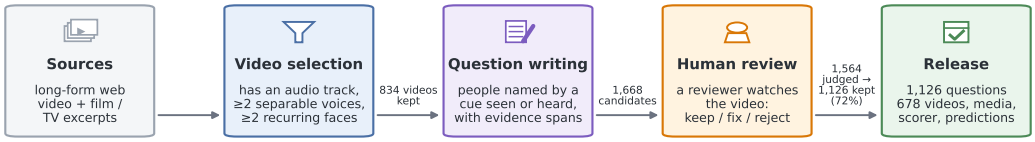
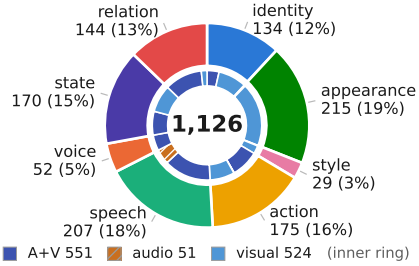


Figure 2: Construction pipeline. A generator writes candidate questions that name people by a perceptual cue, with evidence spans and required-evidence labels, and human reviewers check every question against the video.

Table 2: Benchmark composition. Left: numbers of videos and questions by scene family, with questions grouped by required-evidence label (§3.2); durations are medians. **Right**: distribution across the eight identity dimensions (outer ring), partitioned by evidence modality (inner ring): audio+visual 551, visual-only 524, and audio-only 51.

Scene family	Videos	Questions	Voice	Visual	Joint	Med. dur.
Scripted narrative	227	371	260	73	38	9.5 min
Conversation	141	239	83	69	87	7.8 min
Lifestyle vlog	112	188	91	45	52	8.1 min
Event / Performance	99	162	139	13	10	6.8 min
News / Documentary	56	100	47	30	23	6.9 min
Procedure / Work	43	66	51	6	9	8.1 min
Total	678	1,126	671	236	219	7.9 min



286 questions. Web videos must pass a voice-based screening criterion: diarization must identify at least two speakers with clearly separable voiceprints. Because automatic diarization is unreliable on film soundtracks, clips from CondensedMovies are instead required to contain at least two recurring faces, and segments from AVA-AVD (Xu et al., 2022) enter with their released diarization annotations; Appendix A specifies the thresholds. Filtering produces three resources per video that condition question generation and are released with the benchmark: a speaker-attributed transcript, a face timeline, and a VLM-extracted inventory of distinguishing visual cues. Of 834 eligible videos, 678 yield final questions, totaling 91.3 hours (Table 2).

3.2 Questions anchored on perceptual cues

Every question identifies its target person through a perceptual cue: a visible or audible property other than the content of their speech. Cues include visual attributes, gestures, actions, on-screen text, and non-verbal sounds, drawn from the attribute inventory or face timeline, such as "the man with the neck chain" or "the woman who laughs out loud". Questions exclude dialogue topics, quotes longer than six words, roles such as "the host", and sensitive demographics. Proper names appear in 43 questions, only when spoken or shown in the video. Because these cues are absent from an ASR transcript, the transcript alone cannot identify the target person.

GPT-5.6-sol generates two candidate questions per video: one at *understanding* depth, based on a single moment, and one at *reasoning* depth, combining multiple moments. Questions are distributed across eight identity dimensions (Appendix A). The generator receives the speaker-attributed transcript, face timeline, and attribute inventory and associates each answer and evidence span with timestamps. The prompt prohibits using the cue moment as answer evidence and favors evidence moments at least one minute apart. Each question retains its evidence spans and a required-evidence label, voice, visual, or joint, assigned before review. The label does not always match the evidence annotation: 48.9% of questions have evidence in both audio and visuals, 46.5% in visuals only, and 4.5% in audio only (Table 2, inner ring). The labels are validated in aggregate: replacing audio with noise lowers accuracy on voice and joint questions but not visual questions (Table 5).

3.3 Human verification

Human reviewers evaluated the candidate questions against the video using three criteria: *target necessity*, replacing the target with another plausible person in the video must change the answer; *correctness and uniqueness* within the annotated spans; and *transcript insufficiency*, dialogue alone must not determine the answer. Questions with substantive flaws were rejected, and those whose

reviewer note specified a fix to the cue or option wording were repaired accordingly; roughly 70% of reviewed candidates entered the release (Appendix B).

A question's admission to the benchmark depends only on human review, not model performance. Filtering for questions an agent solves would favor that agent, while filtering for questions a text baseline misses would make difficulty model-dependent. We therefore retain all human-approved questions and report text-only accuracy separately.

3.4 Evaluation protocol

The benchmark release will include transcripts and face timelines, but the standard evaluation does not expose them to the model. A scored system receives only the video with its audio track and the four-choice question. The scorer extracts the selected letter and counts unparsable or truncated responses as incorrect (Appendix I). We report accuracy overall, by required-evidence label, and by scene family, with Wilson 95% intervals in Table 11.

Every system also answers using only the question and options, providing its **text-only** baseline. The 463 questions missed by the GPT-5.5 text baseline form a *text-hard* subset. The tables specify all other input conditions: **frames+ASR** pairs sampled frames with a Whisper (Radford et al., 2023) transcript without speaker labels for models that do not accept audio, while **frames-only** and **audio-only** each omit one channel.

4 Training data, tuned models, and the per-turn protocol

While overall benchmark scores indicate system capability, they confound model scale, pretraining data, and inference interfaces, obscuring whether failures stem from data scarcity or architectural limitations. To isolate these factors, this section introduces: (1) a synthetic training split matching the benchmark's structure, (2) three recipes that fine-tune Qwen3-Omni-Thinking on this split, and (3) a per-turn diagnostic protocol that segments the video into speaker turns and varies one input factor across otherwise matched conditions.

4.1 Training split

We adapt the benchmark pipeline (§3) to synthesize training data from videos outside the benchmark (web video passing quality filters, film clips, and diarized meetings), replacing human review with automated checks. An independent LLM generator writes candidate questions conditioned on attributed transcripts, face timelines, and attribute inventories. A second model filters questions by answering without media, discarding discrepancies, while heuristic filters remove direct quotes and seekable timestamps. Quotas over the required-evidence labels are 50% joint (voice and visual), 25% voice, and 25% visual. The split contains 17,893 training and 1,109 validation questions across 5,445 videos, with at most four per video. With human verification replaced by automated filtering, frozen Qwen answers 63% of questions from transcripts alone, while ablations identify 29% as dependent on the labeled audio or visual input.

Targets use the native reasoning format of Qwen3-Omni-30B-A3B-Thinking. For half the training questions, we distill targets from correct, evidence-citing reasoning chains produced by the frozen model. For the remaining half, a generator produces reasoning chains conditioned on each question's keyed answer and annotated evidence turns, which are then verified by citation checks. Appendix K details the data sources and filtering rules.

4.2 Supervised fine-tuning and reinforcement post-training

We fine-tune Qwen3-Omni-30B-A3B-Thinking, the reasoning variant of the Qwen3-Omni family. It is a mixture-of-experts model with 3B of its 30B parameters active per token. We apply three adaptation strategies for up to two epochs: (1) **LoRA** with rank $r = 128$ targeting attention projections; (2) **Full-parameter** tuning of the reasoning backbone with frozen vision and audio encoders; and (3) **Reinforcement learning** via GRPO atop the full-parameter checkpoint, using training-split inputs with the target reasoning removed and rewarding completions whose final answer matches the option designated as correct in the training data.

Additional checkpoints are trained directly on dense-frame layouts, which retain the uniformly spaced frames and densely sample voiceprint-selected turns at 4 FPS, to disentangle training effects

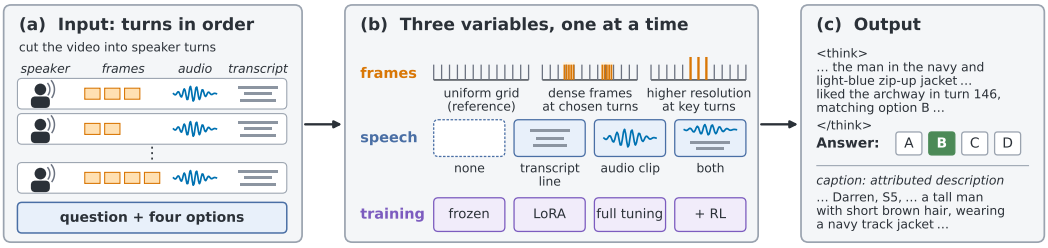


Figure 3: The per-turn diagnostic protocol. (a) Qwen3-Omni-30B-A3B-Thinking receives the video serialized into diarized turns; each turn provides frames, an audio clip, and a transcript line, followed by the question. (b) We ablate a single input factor at a time: frame placement and resolution, speech channel modality, and training targets. (c) The model outputs a reasoning chain ending with the predicted choice.

from inference layouts. We also replicate the frame-placement study on a dense 4B model without an audio encoder, which receives each turn's voice as a single token derived from its voiceprint embedding; see Appendix M for details. All models use the per-turn structure described below; Appendix I provides the optimization hyperparameters.

4.3 The per-turn diagnostic protocol

To isolate the causes of model failure, our diagnostic protocol uses three released annotations: diarized turn segmentations, annotated evidence turns, and required-evidence labels (Figure 3a). As shown in Figure 3, the video is divided into speaker turns (i.e., segments in which one person speaks) and presented chronologically. The model receives uniformly sampled frames, an audio clip, and a transcript line for each turn, then the four-choice question after the final turn.

Holding the turn structure, transcripts, and questions fixed, we ablate one factor at a time (Figure 3b). (1) **Frames**: We compare uniform sampling with denser sampling at turns selected by voiceprint clustering or an annotated-evidence oracle (§5.2). (2) **Speech evidence**: We compare no speech input (vision-only), transcripts, raw audio, and both (§5.3). (3) **Model checkpoints**: We compare frozen, LoRA, full-parameter, and GRPO-adapted models (§5.5). As shown in Figure 3c, the model outputs either a multiple-choice reasoning chain that cites turn indices and ends with a choice letter or an attributed caption describing characters and speech turns (§5.6).

We report paired per-question accuracy differences $\Delta(c, r) = \frac{1}{N}\sum_i(\bar{y}_i^c - \bar{y}_i^r)$ between each condition c and reference r, where $\bar{y}_i$ is correctness averaged over two runs. Strict parsing penalizes budget truncation, whereas lenient parsing extracts any valid choice letter. Differences with $|\Delta| < 2.0\%$ reflect run-to-run sampling variance. Evidence citations are scored by whether a chain's most-cited turns overlap the annotated evidence, distinguishing correct answers with and without supporting evidence. Pairwise identity verification uses balanced accuracy $\mathrm{BA} = \frac{1}{2}(R_{\mathrm{same}} + R_{\mathrm{diff}})$. Appendix I details decoding budgets and prompts, and Appendix L provides our subset analyses.

5 Experiments

5.1 Comparison of existing systems

Table 3 covers closed models, open models with video-audio input, and open VLMs given frames and an ASR transcript. On the 463 *text-hard* questions, Gemini-3.1-Pro scores 80.6% and the best open model scores 60.5% (Appendix E).

Models perform worst on voice-based identity attribution. By required-evidence label, *voice* questions require attributing speech to its speaker, *visual* questions require tracking an on-screen person, and *joint* questions require linking a voice to a face. Joint questions nevertheless score highest for most systems, because their voice is usually heard within seconds of the face it belongs to, while most voice questions must follow a voice with no face to anchor it. Every video-grounded system performs worse on voice than visual questions. Replacing Gemini-3.1-Pro's audio with equal-length white noise lowers accuracy by 14.6 points on joint questions and 9.3 on voice questions, but not on visual questions (Table 5). Together, these results indicate that attributing speech to its speaker is harder for models than tracking a person on screen.

Table 3: Main results on LONGCOREF-AV (1,126 questions). MCQ columns report overall and *text-only* accuracy and accuracy by required-evidence label; unanswered questions count as incorrect. Captioning columns (§5.6) report speaker-attribution accuracy (*attr.*), same-speaker pair F1, and the share of multi-speaker videos assigning $\geq$90% of turns to one label (*coll.*). Bold marks the best entry of each group in each column, the lowest for *coll.* †GPT-5.5 uses a Whisper transcript without speaker labels as its API accepts no audio. ‡VideoLLaMA2.1 captions cover 1.4% of turns, too few to score.

		MCQ Accuracy (%)					Captioning (%)		
System	Input	Overall	Text-Only	Voice	Visual	Joint	*attr.* (↑)	F1 (↑)	*coll.* (↓)
Closed									
Gemini-3.1-Pro (Google DeepMind, 2026)	video + audio	**86.9**	59.6	**83.8**	**89.8**	**93.6**	**76.0**	**72.0**	13
GPT-5.5 (OpenAI, 2026)	32 fr. + ASR†	84.2	58.9	80.5	88.6	90.9	71.4	68.3	**8**
Open (video and audio input)									
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	video + audio	**74.1**	49.0	**70.3**	**76.7**	82.6	68.3	58.5	18
Qwen3-Omni-30B-A3B-Thinking	video + audio	72.4	49.3	67.5	75.0	**84.5**	**69.3**	**61.0**	19
Nemotron-3-Nano-Omni-30B (Deshmukh et al., 2026)	video + audio	70.9	48.0	67.8	73.3	77.6	63.0	57.1	48
video-SALMONN-2+-7B (Tang et al., 2025)	video + audio	64.3	43.4	61.3	63.6	74.4	59.5	53.1	19
AV-Flamingo-9B (Ghosh et al., 2026)	video + audio	63.9	49.1	61.0	63.1	74.0	57.9	55.6	42
MiniCPM-o-4.5 (Cui et al., 2026)	video + audio	61.7	51.8	55.6	72.9	68.5	50.4	46.8	**2**
OmniVinci-9B (Ye et al., 2026)	video + audio	60.8	42.9	57.5	60.6	71.2	57.5	54.9	28
Qwen2.5-Omni-7B (Xu et al., 2025a)	video + audio	56.2	47.2	53.5	58.9	61.6	33.0	26.8	3
Gemma-4-12B (Team et al., 2026)	frames + audio	55.8	49.0	50.2	64.0	63.9	48.2	45.7	10
Inkling-Small (Thinking Machines Lab, 2026)	6×(8 fr. + 30 s audio)	50.4	49.3	44.4	59.3	58.9	62.8	49.2	15
VideoLLaMA2.1-7B-AV (Cheng et al., 2024)	video + audio	45.4	40.2	42.2	50.0	50.2	—‡	—	—
Open (frames and ASR substitute replacing audio input)†									
Gemma-4-31B	32 fr./58 s seg. + ASR	**74.2**	52.0	**71.4**	77.1	**79.9**	60.8	57.9	**15**
GLM-4.5V (Hong et al., 2025)	64 fr. + ASR	72.4	49.5	68.0	**78.4**	79.5	69.5	57.5	42
Qwen3-VL-30B (Bai et al., 2025)	768 fr. + ASR	71.9	49.8	67.8	**78.4**	77.6	**74.6**	**58.4**	94

A model given sampled video frames and a speech transcript performs comparably to the strongest open model given video and audio. Gemma-4-31B conditioned on frames+ASR reaches 74.2%, compared with 74.1% for Qwen3-Omni-Instruct, despite receiving no raw audio. This cross-model comparison does not isolate the effect of input modality because the models differ in scale and training; §5.3 therefore compares audio and transcripts within the same model.

Video input improves accuracy beyond strong text-only baselines. The question and options alone score well above the 25% chance level (Table 3). Full-input accuracy should therefore be interpreted relative to each model's text-only baseline. Video adds 15–25 points for the strongest open models. For GPT-5.5, 32 frames add 18 points, whereas one frame adds under 4. Appendices E to H report the full checks.

5.2 EFFECT OF SPEAKER-TURN SELECTION FOR DENSE FRAME SAMPLING

Using frozen and fine-tuned Qwen3-Omni-Thinking checkpoints under the per-turn protocol, we test whether sparse visual sampling limits performance. We retain the uniform frames from every speaker turn and add frames at 4 fps within selected turns, forming a dense-frame layout. We select turns in three ways: (1) a voiceprint method clusters turn-level speaker embeddings and selects speaker changes or uncertain assignments; (2) Gemini-3.1-Pro ranks turns by relevance to each question; and (3) an oracle selects the annotated turns containing the evidence needed to answer each question. Table 4 compares this oracle placement with placement at voiceprint-selected turns.

Table 4 shows that the effect depends on both frame placement and the checkpoint. On the uniform-trained full-parameter model, placing dense frames on annotated evidence turns improves accuracy more than placing them on voiceprint-selected turns. On the frozen model, both placements improve accuracy by less than one point, with voiceprint placement slightly ahead. On the voiceprint-trained model, voiceprint placement helps, whereas evidence placement provides no gain. Thus, evidence-based placement helps most for the uniform-trained full-parameter checkpoint, but no placement consistently improves every checkpoint.

Table 4: Effect of dense-frame placement on Qwen3-Omni-Thinking. *Uniform* is baseline accuracy (%); the other columns report the paired change after adding dense frames at voiceprint-selected or annotated evidence turns.

		Dense-frame Δ	
Checkpoint	Uniform	Voiceprint	Evidence
Frozen	68.9	+0.9	+0.7
Full (uniform)	69.2	+0.8	+2.2
Full (voiceprint)	69.2	+1.2	0.0

Frame layout and resolution have little effect. Mixed resolution, a single timestamped block of frames, and 448 px frames remain within about one point of uniform sampling; 192 px frames

reduce accuracy by 1.0-2.5 points, and removing the uniform grid reduces the dense 4B model by 3.2 points (Appendices L and M). As in the clue-restricted test of §5.4, extra frames help primarily when placed on turns containing relevant evidence, but the automated methods rarely select those turns: the voiceprint method recalls 10% of them, and Gemini-3.1-Pro fewer than 50%.

5.3 Contributions of audio with and without ASR grounding

Without a transcript, models use meaningful audio content, gaining about 8 points. For whole-video models evaluated without ASR, replacing the audio track with equal-length white noise preserves the input structure while removing speech and voice information (Table 5). Accuracy falls for both Gemini-3.1-Pro and Qwen3-Omni-Instruct, primarily on voice and joint questions. For Qwen3-Omni-Instruct, white noise and removing the track produce nearly identical results, confirming that the gain comes from audio content rather than the presence of a track. Likewise, removing Gemini-3.1-Pro's audio or GPT-5.5's transcript causes the largest losses on joint questions, which depend on the removed source of speech information.

Table 5: Within-model input ablations. Strict accuracy by evidence label after removing speech input or replacing audio with equal-length white noise. Each block begins with the Table 3 configuration; bold marks the best entry of each block in each column.

Model	Input	Overall	Voice	Visual	Joint
Gemini-3.1-Pro	video + audio	**86.9**	**83.8**	89.8	**93.6**
	audio → noise	79.2	74.5	**92.8**	79.0
	32 frames only	73.6	69.4	87.7	71.2
GPT-5.5	32 frames + ASR	**84.2**	**80.5**	**88.6**	**90.9**
	32 frames only	77.4	72.9	87.7	79.9
Qwen3-Omni-Inst.	video + audio	**74.1**	**70.3**	76.7	**82.6**
	audio → noise	65.7	63.0	76.3	62.6
	frames only	66.6	64.2	**77.1**	62.6

In our controlled ablation, acoustic information adds little beyond transcripts. Table 6 evaluates the fine-tuned Qwen3-Omni across transcripts, audio, both, or neither while holding all other variables constant. Audio and ASR each add about 10 points over frames alone, primarily on voice and joint questions, corroborating our required-evidence labels. However, once a transcript is present, adding audio does not improve overall accuracy on the tuned checkpoint, nor on the frozen one at the same temperature (Table 15).

Table 6: Effect of speech input on Qwen3-Omni-Thinking. Accuracy with frames, ASR, audio, or their combinations. The first row is frozen at sampling temperature 1.0; the rest are fine-tuned. *trunc.*: responses exceeding the token budget.

Input	Overall	Voice	Visual	Joint	*trunc.*
frozen: frames + ASR + 3 s audio	68.9	64.5	71.2	**80.1**	0.3%
frames only	59.9	54.8	**75.2**	58.7	0.1%
ASR only	48.6	46.9	41.5	61.4	0.6%
frames + ASR	69.6	65.5	73.3	78.1	0.8%
frames + audio (whole turn)	**70.0**	**65.8**	75.0	77.4	1.0%
frames + ASR + audio (3 s)	69.2	65.1	71.6	79.0	0.5%
frames + ASR + audio (whole turn)	68.4	63.6	71.2	**80.1**	0.9%

Qwen3-Omni exploits audio beyond transcripts only in narrow cases. Whole-turn audio training adds 12.5 points on the voice-quality slice, and the audio track also helps the frozen model on the audio-anchored supplement (Appendix N), but these gains do not improve overall accuracy. This characterizes how this model uses audio, not the information audio contains: it uses direct voice-quality cues but does not broadly integrate acoustic information with transcripts and visual context. Similar findings that text outperforms audio for speech content on speech and omni benchmarks suggest that this may reflect a broader pattern (Fu et al., 2025; Li et al., 2026b; Chen et al., 2024b; Wu et al., 2024; Chen et al., 2025b; Pang et al., 2026).

5.4 Separating evidence retrieval from identity attribution

Multiple-choice accuracy does not imply identity linking. The first test asks whether two observations belong to the same person (Table 7). We score voice pairs (two speech segments) and face pairs (two visible faces) separately, using balanced accuracy to weight same-person and different-person cases equally. This removes a shortcut in the pooled set, where pair type predicts the label, and prevents a bias toward same-person predictions from inflating the score (Appendices R and O). Both Qwen3-Omni models remain near chance despite answering most multiple-choice questions correctly; Qwen3-Omni-Instruct predicts the same person for 97.9% of pairs. Gemini-3.1-Pro distinguishes faces more reliably than voices.

Table 7: Tests of identity attribution and evidence retrieval. Identity-link accuracy gives equal weight to same-person and different-person pairs; gain is relevant-clips minus full-video accuracy, in percentage points.

	Identity-Link Test		Question Accuracy		
Model	Voice	Face	Relevant Clips	Full Video	Gain
Gemini-3.1-Pro	**69.5**	**84.5**	88.8	86.9	+1.9
Qwen3-Omni-Think.	55.9	56.2	82.5	72.4	+10.1
Qwen3-Omni-Inst.	51.7	50.2	84.2	74.1	+10.1

Relevant clips help distinguish evidence retrieval from identity attribution. The second test compares the complete video with only the annotated segments containing the answer evidence, extended by 15 s on both sides (Chen et al., 2025a). Improvement with these clips suggests difficulty locating the evidence in the full video; remaining errors may instead reflect identity attribution or subsequent reasoning. Relevant clips improve both Qwen3-Omni models far more than Gemini-3.1-Pro, suggesting that retrieval is a larger bottleneck for the Qwen models. This is not an isolated retrieval intervention: shorter clips are sampled more densely and may omit intermediate identity-link moments (Appendix R).

5.5 Does training improve voice-to-person linking?

Training changes evidence citation more than identity linking. Table 8 evaluates the checkpoints of §4.2. Accuracy stays within two points of the frozen reference for every checkpoint. Full-parameter fine-tuning raises the annotated-evidence citation rate by 6.4 points while changing strict accuracy by only 0.3 (Table 15). On voice pairs under the per-turn protocol, identity-link balanced accuracy changes from 60.7% to 59.3%: different-person recall increases from 54.4% to 60.5%, while same-person recall falls from 67.1% to 58.1%. Thus, training shifts the balance between the two predictions rather than improving both classes. GRPO further increases evidence citation, but changes this accuracy by less than one point relative to the frozen model. Removing the audio track at test time does not reduce accuracy, providing no evidence that this model benefits from the track.

Table 8: Training Qwen3-Omni-Thinking. Strict and lenient accuracy, the rate of completions reaching the output limit ("trunc."), and paired differences from the frozen model. "ep." denotes training epochs.

Setting	Strict	Lenient	*trunc.*	Δ Strict	Δ Lenient
Frozen reference (temperature 1.0)	68.9	69.3	0.3%	—	—
LoRA, rank 128 (2 ep.)	67.4	68.2	3.9%	−1.6	−1.1
Full-parameter tuning (1 ep.)	68.3	68.6	0.9%	−0.6	−0.7
Full-parameter tuning (2 ep.)	69.2	69.4	0.5%	+0.3	+0.1
+ GRPO (1 ep.)	**70.7**	**71.1**	0.7%	+1.8	+1.8
+ GRPO (2 ep.)	69.2	69.4	1.0%	+0.3	+0.1
Greedy decoding	70.1	70.3	1.1%	+1.2	+1.1
No audio track at test	69.6	69.7	0.8%	+0.7	+0.4
Whole-turn audio (train/test)	70.3	70.4	0.5%	+1.4	+1.2

5.6 Identity-grounded captions remain incomplete and error-prone

To test identity tracking beyond multiple choice accuracy, we ask the systems of Table 3 to describe all 678 videos while stating who is visible, who speaks each line, and when an individual reappears. We evaluate speaker attribution on the reference turns the caption covers, the accuracy of a reader given only the caption, and recall of annotated identity relations without contradiction (Appendix P).

Captioning exposes omissions, speaker confusion, and failures across windows. Gemini-3.1-Pro and GPT-5.5 correctly attribute about three quarters of the turns they cover, but recover only half of the annotated identity relations. The strongest open omni models, Qwen3-Omni-Thinking and Qwen3-Omni-Instruct, cover only 42% and 37% of turns. Other open omni models stop early or transcribe little when given the whole video, and fail to preserve identities across 58-second windows. Open VLMs list every turn but often assign them to a single speaker; Qwen3-VL does so in almost every multi-speaker video. Fine-tuning does not improve these results (Table 20).

6 Conclusion

We introduce LongCoref-AV, a human-verified benchmark of 1,126 questions for measuring audio-visual identity coreference across distant moments in long-form video. Each question grounds its target in a perceptual cue and provides evidence spans, a required-evidence label, and text-only baselines. Across 16 systems, voice-dependent questions yield the lowest accuracy, and open models trail frontier systems. When given annotated evidence clips, the two tested Qwen3-Omni models improve far more than Gemini-3.1-Pro, indicating that retrieval is a larger bottleneck for the Qwen models while many Gemini errors persist beyond retrieval. Direct identity matching and full-video captioning further show that question accuracy does not guarantee reliable identity tracking. For Qwen3-Omni-Thinking, denser sampling of evidence turns and supervised or reinforcement fine-tuning do not clearly improve identity linking; across three text controllers, adaptive voiceprint queries also underperform fixed access to the same verdicts. These results motivate architectures with persistent identity representations across long multimodal contexts, objectives that directly supervise cross-time and cross-modal identity links, and diagnostic evaluations that separately measure evidence retrieval, identity attribution, and end-task accuracy.

REFERENCES

Detao Bai, Shimin Yao, Weixuan Chen, Zhiheng Ma, Xihan Wei, and Jingren Zhou. Humanomni-speaker: Identifying who said what and when. *arXiv preprint arXiv:2603.21664*, 2026.

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

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 7124–7128. IEEE, 2020.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Jianghan Chao, Wenhui Tan, Yuchong Sun, Ruihua Song, Liyun Ru, et al. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 84035–84062, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025a.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *ArXiv*, abs/2510.10444, 2025b. URL https://api.semanticscholar.org/CorpusID:282056914.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. Toc-bench: A temporal object consistency benchmark for video large language models. *arXiv preprint arXiv:2605.09904*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Yi-Ming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *ArXiv*, abs/2410.17196, 2024b. URL https://api.semanticscholar.org/CorpusID:273507315.

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

Sanjoy Chowdhury, Karren D Yang, Xudong Liu, Fartash Faghri, Pavan Kumar Anasosalu Vasu, Oncel Tuzel, Dinesh Manocha, Chun-Liang Li, and Raviteja Vemulapalli. Amuse: Audio-visual benchmark and alignment framework for agentic multi-speaker understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22998–23009, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *ArXiv*, abs/2407.10759, 2024. URL https://api.semanticscholar.org/CorpusID:271213498.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Hong Gao, Yiming Bao, Xuezhen Tu, Yutong Xu, Yue Jin, Yiyang Mu, Bin Zhong, Linan Yue, and Min-Ling Zhang. Agentic video intelligence: A flexible framework for advanced video exploration and understanding. *arXiv preprint arXiv:2511.14446*, 2025.

Shibo Gao, Chongxiao Wang, Chenglong Huang, Jie Ma, Haolin Shi, Fei Ding, Jing Li, Qiang Lyu, Yangyang Liu, Yang Liu, Jun Liu, Linlin Huang, and Peipei Yang. I seek you in videos: Identity-conditioned queries for person-centric video reasoning. In *ACM International Conference on Multimedia (ACM MM)*, 2026.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, et al. Audio-visual flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *ArXiv*, abs/2507.08128, 2025. URL `https://api.semanticscholar.org/CorpusID:280293857`.

Google DeepMind. Gemini 3.1 pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Yejin Kwon, Taewoo Kang, Hyunsoo Yoon, and Changouk Kim. M3-slu: Evaluating speaker-attributed reasoning in multimodal large language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pp. 5722–5736, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026a.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026b.

Yuxuan Li, Lingxi Xie, Xinyue Huo, Jihao Qiu, Jiacheng Shao, Pengfei Chen, Jiannan Ge, Kaiwen Duan, and Qi Tian. Reasoning llm improves speaker recognition in long-form tv dramas. *arXiv preprint arXiv:2607.02504*, 2026c.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *ArXiv*, abs/2411.03628, 2024. URL https://api.semanticscholar.org/CorpusID:273850338.

Tao Liu, Shuai Fan, Xu Xiang, Hongbo Song, Shaoxiong Lin, Jiaqi Sun, Tianyuan Han, Siyuan Chen, Binwei Yao, Sen Liu, et al. Msdwild: Multi-modal speaker diarization dataset in the wild. In *Interspeech*, pp. 1476–1480, 2022.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Yiran Meng, Junhong Ye, Wei Zhou, Guanghui Yue, Xudong Mao, Ruomei Wang, and Baoquan Zhao. Videoforest: Person-anchored hierarchical reasoning for cross-video question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 836–845, 2025.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, pp. 2616–2620, 2017. doi: 10.21437/Interspeech.2017-950.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2272–2283, 2026.

OpenAI. GPT-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5, 2026.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. *arXiv preprint arXiv:2605.27772*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. AVA active speaker: An audio-visual dataset for active speaker detection. In *ICASSP*, 2020.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

Liang Shi, Wei Li, Kevin M Beussman, Lin Chen, and Yun Fu. Iir-vlm: In-context instance-level recognition for large vision-language models. *arXiv preprint arXiv:2601.14188*, 2026.

Zhaokai Sun, Shuai Wang, Zhennan Lin, Chengyou Wang, Dehui Gao, Yuang Cao, Chunjiang He, Pan Zhou, and Lei Xie. Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios. *arXiv preprint arXiv:2606.22868*, 2026.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *ArXiv*, abs/2506.15220, 2025. URL https://api.semanticscholar.org/CorpusID:279447296.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing Lyu, and Chao Zhang. D-orca: Dialogue-centric optimization for robust audio-visual captioning. *arXiv preprint arXiv:2602.07960*, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Thomas Thebaud, Yuzhe Wang, Laureano Moro-Velazquez, Jesus Villalba-Lopez, and Najim Dehak. Speaker verification with speech-aware llms: Evaluation and augmentation. *arXiv preprint arXiv:2603.10827*, 2026.

Thinking Machines Lab. Introducing Inkling-Small. https://thinkingmachines.ai/news/inkling-small/, 2026.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. Diarizationlm: Speaker diarization post-processing with large language models. *arXiv preprint arXiv:2401.03506*, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark A. Hasegawa-Johnson, and Mari Ostendorf. Just asr + llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1137–1143, 2024. URL https://api.semanticscholar.org/CorpusID:272524710.

Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. Ava-avd: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3838–3847, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

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

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900, June 2025a.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Conference on Empirical Methods in Natural Language Processing*, 2025b. URL https://api.semanticscholar.org/CorpusID:277321544.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. *arXiv preprint arXiv:2601.13719*, 2026.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026.

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

A Details of the benchmark

Video selection thresholds. Every video has an audio track and is at least 640px wide. Web videos enter through a voice check, which runs pyannote diarization and ECAPA voiceprints and requires two to six diarized speakers, at least 30 utterances, a minimum speaker share of 0.05 with a Gini imbalance of at most 0.60, a speaker-switch rate between 1 and 20 per minute, a within-speaker voiceprint cosine of at least 0.55, and a within-minus-between margin of at least 0.15, so that the aligned transcript used to condition generation is trustworthy. The 116 film-excerpt videos of the CondensedMovies lane, which carry 193 questions, skip the voice check because diarization is unreliable on film soundtracks, and enter through a face check instead. InsightFace detections are clustered at a cosine threshold of 0.40, and the video must contain at least two identities that recur: each has at least three detections, which fall into two or more stretches separated by more than 30 seconds, with the first and last stretch at least 60 seconds apart. Every video, from either source, also passes a recurrence check, in which some speaker or face appears again at least 60 seconds after an earlier appearance, and a montage check that allows at most 20 shot cuts per minute. The speaker counts reported per question come from the speaker-attributed transcript roster and can exceed the six diarized clusters the voice check admits; 145 questions list more than six audible speakers. The visual attribute inventory is produced by a vision-language model from sampled frames; only cues legible at 480p are kept, and each carries a flag for whether it fits exactly one person.

Two-level content taxonomy. Each question is tagged with one of eight identity dimensions: appearance, 215; speech content, 207; action, 175; state, 170; relation, 144; open-set identity/naming, 134; voice quality, 52; and style, 29. This dimension is crossed with a depth flag, with 554 understanding and 572 reasoning questions, and multi-label reasoning operations. The operation counts are temporal 257, compare 239, consistency 201, cross-modal 163, procedural 25, and causal 5. Answer surface types: proposition 639, noun entity 187, adjectival attribute 157, verb action 133, adverb 9, and 1 unlabeled.

Scenes and domains. The six scene families of Table 2 group 18 domain types. The 678 released videos comprise CondensedMovies film clips 116, TV interviews 78, live sports 60, feature-film segments 56, street interviews 41, live events 38, live music 35, short films 34, travel and food 33, variety and comedy 32, esports 30, documentaries 29, studio shows 25, cooking and repair 25, product displays 21, trailer recaps 13, news packages 11, and sports highlights 1. Scene labels are certified by an LLM judge over the footage; the 116 CondensedMovies videos carry 193 questions.

Speaker counts. Audible-speaker counts per question: 2 speakers 195, 3 speakers 342, 4 speakers 203, and 5+ speakers 386. The maximum is 13 and the median is 4.

Referring cues. Cue modality: visual 89.7%, on-screen text 5.8%, non-verbal audio 4.5%. Grammatical form of the cue: verb 383, adjective 342, composite 245, adverb 92, noun 64. The audio-anchored supplement of Appendix N adds 300 questions whose cue is a voice trait or a vocal event.

Generation checks. Automatic checks confirm that every evidence span falls inside its video before human review, and answer letters are rebalanced to near-uniform at that stage; the letter-bias check is in Appendix F.

Evidence structure. A question's evidence length is the total duration of the union of its evidence spans. Its median is 16.0 s, mean 27.5 s, p90 61 s, and maximum 421 s. The per-turn readings of §5.2 that anchor on evidence turns cover the 602 questions whose annotated evidence falls inside diarized speech turns. Evidence separation: in 69.2% of questions the evidence spans lie at different moments, a median of 2.6 minutes between the first and the last span, and 39.9% of questions spread their evidence over more than two minutes. The cue is often closer to the evidence than that. A question's identity links run from the moment of the cue to the evidence, and in 38.0% of questions every moment on these links falls inside an evidence span. In the rest, the farthest one lies a median of 49 s outside the spans, and more than two minutes outside in 17.6% of all questions.

Example questions. Gold answers are bold and options are abbreviated.

[voice / speech / news/documentary; evidence spans 4.5 min apart] Which statement was made by the woman who laughs out loud? *(A) Reaching the highest level requires complete devotion to chess.* ***(B) A prospective professional player should genuinely love chess.*** *(C) Setting smaller goals ultimately produces the bigger picture…(D) Having a parent present creates pressure…*

[visual / appearance / news/documentary; evidence spans 4.6 min apart] What was the person who points his left index finger at the man next to him wearing at the opening? ***(A)*** *A light grey suit and purple patterned tie …*

[joint / speech / conversation; evidence spans 3.8 min apart] Which position did the man with the neck chain who appears first state consistently at two widely separated points in the interview? ***(B)*** *Viewers can enjoy the film without prior familiarity with Warcraft …*

B HUMAN-VERIFICATION PROTOCOL

Reviewers saw the source video, the question, options, keyed answer, and annotated evidence spans, and gave each question a keep, fix, or reject verdict. When a whole source video was unusable, reviewers rejected both of its questions. Of the 1,564 candidate questions reviewed, 1,084 were kept and 42 were admitted after a repair specified in the reviewer's note, such as a corrected cue description or option wording; the repairs are logged in `provenance/`, and questions whose fix would have changed the keyed answer or the person the cue refers to, or required re-checking the video, were rejected. The other 438 were rejected, so 1,126 of the 1,564 questions, 72.0%, entered the release. The release stores per-question human verdicts alongside the questions.

C RELEASE CONTENTS AND SCORING RULES

The release is self-contained. All 678 videos, totaling 21 GB, ship with the questions through release-relative media paths. The release also includes the scorer and the per-question metadata above. Diagnostic resources include the unlabeled transcript, the speaker-attributed transcript, and materialized face-timeline observations. The standard scored protocol gives a system *only* the video with its audio track, the question, and four options; transcripts and observations are reserved for ablation conditions and must not be fed to a scored system. The scorer reports overall accuracy, macro accuracy over the six scene families, and accuracy by scene, identity dimension, depth, and required-evidence label; per-scene slices of the reported systems are in Appendix H. Unanswered or unparsable questions count as wrong and are reported per row.

D GRADING CRITERIA AND PER-CELL EVIDENCE FOR TABLE 1

Operational criteria. Each column of Table 1 is graded by a fixed rule; ◐ is never a judgment call but a named intermediate condition. The LV thresholds of 2 and 5 minutes are anchored to LONGCOREF-AV's evidence structure. Its evidence spans, when they lie at different moments, are a median 2.6 minutes apart, and its median video length is 7.9 minutes. Identity re-appearance needs minutes of separation to exist at all.

Per-cell evidence. Grades follow the source papers' own statements plus repository checks conducted in August 2026. The † symbol marks cells where the source does not report the statistic.

- **M3-SLU**: The benchmark is audio-only, so AV and VF receive ○. Its 1–3.5-minute dialogue segments give LV ◐. Speaker labels come from corpus metadata and there is no in-content naming task, so NG receives ○. Each example is a single continuous segment, giving CT ○. Task 1 is extractive who-said QA and Task 2 is same-speaker matching; attribution is therefore the task itself and SR receives ◐. All instances were inspected manually, giving HV ●. We found no public repository, so PR receives ○.
- **MSU-Bench**: The benchmark is audio-only. Its clips fall into 1–2-minute and 2–5-minute buckets, giving LV ◐. Role/Identity Identification infers identity from dialogue context on the audio side only, giving NG ◐. Speaker Verification is a pairwise same-speaker judgment without narrative re-appearance, giving CT ◐. Tier-2 dialogue reasoning goes beyond attribution, so SR receives ●. Distractor failure-mode tags are not temporally localized evidence, so EV receives ○. Eight researchers reviewed every question, giving HV ●, and the GitHub and Hugging Face releases are live, giving PR ●.
- **AV-SpeakerBench**: Attribution of speech to visible people is the core design, giving VF ●. Clips last 5–30 s, so LV receives ○. There are no persistent names and each question covers a single scene. Several task types reason over attributed content, giving SR ●. Expert quality control is reported, but there are no modality-necessity labels, so RE receives ○. The Hugging Face dataset is live, giving PR ●.

Table 9: Operational grading rules for Table 1.

	●	◐	○
AV	native audio *and* visual input	one modality replaced by ASR, subtitles, or other text, or optional	single modality
LV	typical continuous length $\geq$5 min	2–5 min, or segments cut from longer continuous sources	$<$2 min
VF	speech→on-screen-person attribution is the core task	attribution appears only in a subtask	absent
NG	names must be established from evidence inside the video	names supplied externally through a registration query or cast list	no naming
CT	re-appearances separated within continuous video or across episodes	stitched non-contiguous clips or pairwise verification	within-scene only
SR	reasoning *on top of* attribution through stance, relations, or multi-hop inference	attribution/recognition is itself the task	no speaker attribution
EV	per-question, temporally localized, released with the data	subset-only, non-temporal, or sample-level measurement	none
RE	per-question required-evidence labels, validated by modality ablation	construction-time modality filtering or split-level control	none
TB	text-only baseline reported and used to slice results	text-only filtering applied but no baseline reported	none
HV	every question human-verified	sample-level or coverage unstated	none
PR	all media and annotations downloadable and verified	partial release	not released

- **AMusE**: Association and re-identification make attribution central, giving VF ●. Clips average 38.7 s. Re-identification and narrative linking use stitched non-contiguous clips, giving CT ◐. Summarization and next-speaker prediction are the only inferential families, so SR receives ◐. "Guided mode" supplies oracle metadata as an evaluation condition rather than per-question labels, so EV and RE receive ○. Manual inspection is reported without per-question coverage, giving HV ◐. No repository was found, so PR receives ○.
- **HumanOmni-Speaker**: Visually registered diarization makes the voice-to-face link central, giving VF ●, but an external registration query supplies identities, so NG receives ◐. The benchmark uses short conversation segments, and who-said-what-when is attribution itself, giving SR ◐. The Hard split removes visual biases through a split-level control, so RE receives ◐. Verification covers a sample of about 1.5K questions with $\kappa = 0.82$, giving HV ◐. The GitHub link returns 404, so PR receives ○.
- **DVD-Bench, or D-ORCA**: Speaker-attributed captioning links voices to on-screen character descriptions, giving VF ●and NG ○. Average duration is 84–95 s, so LV receives ○. Captioning is attribution, giving SR ◐. The evaluation set is fully human-annotated, so HV receives ●. The model and code are released but the data are pending, giving PR ○.
- **DramaSR**: The benchmark uses 39-minute episodes with cross-episode recurrence, giving LV and CT ●. Naming is closed-set against an externally provided cast list, so NG receives ◐. The task is recognition rather than QA or reasoning, giving SR ◐. Labels are propagated and then checked, with about 0.5% residual noise, so HV receives ◐. Nine of thirteen dramas are released, giving PR ◐.
- **CrossVideoQA**: The benchmark uses video-only surveillance, so AV and VF receive ○. Durations are not reported, so LV receives ○†. Cross-camera person re-identification is the defining axis, giving CT ●. There is no speech, so SR receives ○. Expert review has unstated coverage, giving HV ◐. The repository contains only a README, so PR receives ○.
- **ISYV-Bench**: The benchmark is visual-only and provides identity as a reference image, so AV, VF, and NG receive ○. It claims long-horizon tracking without extractable length statistics, giving LV ◐†. Cross-domain matching and long-horizon tracking give CT ●. There is no speech, so SR receives ○. Automatic annotation with manual review gives HV ◐, and the unconfirmed release gives PR ○.
- **AVUT**: AVCM, the "who says…" task, and AVDiar are two of eight tasks, giving VF ◐. Videos average 67.8 s and are capped at 2 minutes, so LV receives ○. AVDiar has utterance-level timestamps but the MCQ tasks do not, giving EV ◐. Four-fold option rotation filters text-only-solvable questions, but no text-only baseline is reported, so TB receives ◐. Every question is human-authored or human-reviewed, giving HV ●. The GitHub and Hugging Face releases are live, giving PR ●.
- **StreamingBench**: Source Discrimination, illustrated by "Who just said [quote]?", is one of 18 tasks and contains about 250 questions. VF and SR therefore receive ◐. Source videos reach 24 minutes, but questions target the recent stream, giving LV ◐. Timestamps mark question insertion rather than evidence, so EV receives ○. All pairs are human-verified, giving HV ●, and the GitHub and Hugging Face releases are live, giving PR ●.
- **JointAVBench**: Speaker Spatial Localization is a subtask, giving VF ◐. Videos average 97.2 s, with a full-scene subset longer than 10 minutes, so LV receives ◐. Cross-scene Association links plot elements rather than identities, giving CT ○. CRI and SPER reason about people without requiring who-said attribution, so SR receives ◐. Audio-visual necessity is enforced during generation but not labeled per question, giving RE ◐. Every sample is human verified, giving HV ●. Only a project page is available and the repository is unconfirmed, so PR receives ◐.
- **EgoLife**: Week-long continuous egocentric recording gives LV and CT ●. RelationMap identifies people among six fixed participants and has no voice→face or naming task, so VF and NG receive ○. Interpersonal reasoning uses loose attribution, giving SR ◐. Query timestamps and certification-length annotations give EV ◐. Heavy human filtering covers all released questions, giving HV ●, and the GitHub and Hugging Face releases are live, giving PR ●.
- **LongCoref-AV, ours**: Native audio and video with a median duration of 7.9 minutes give AV and LV ●. Questions link names, voices, and faces from in-video naming events, giving VF and NG ●. Evidence spans at different moments have a median separation of 2.6 minutes, giving CT ●. Eight identity dimensions with reasoning splits give SR ●. Per-question evidence spans and required-evidence labels, validated by the input ablations of Table 5, give EV and RE ●. The main table reports the text-only baseline and text-hard subset, giving TB ●. Every question is human verified, giving HV ●, and all 678 videos ship with the release, giving PR ●.

E TEXT-HARD SUBSET

We define the text-hard subset as the 463 questions the GPT-5.5 text-only baseline answers incorrectly and the text-easy split as the 663 questions it answers correctly. The split comes from a single run of that baseline under the API's default sampling, so a second run would move some questions across the boundary. Chance is 25 on both halves; a real audio-visual signal must hold on the hard half. Table 10 splits every row of Table 3 and the input ablations of Table 5 this way. Among the open models, GLM-4.5V is highest on the text-hard subset at 60.5, just above Qwen3-Omni-30B-A3B-Instruct and Gemma-4-31B at 60.3.

Table 10: Accuracy on the 463-question text-hard subset, the 663-question text-easy complement, and the full benchmark, grouped as in Table 3.

System	Input	text-hard	text-easy	full
Closed				
Gemini-3.1-Pro	native video+audio	**80.6**	**91.4**	**86.9**
Gemini-3.1-Pro	audio → white noise	68.3	86.9	79.2
Gemini-3.1-Pro	32 frames only	60.9	82.5	73.6
GPT-5.5	32 frames + ASR	77.3	89.0	84.2
GPT-5.5	32 frames only	64.1	86.6	77.4
GPT-Audio	audio only	37.8	60.2	51.0
Open, video+audio input				
Qwen3-Omni-30B-A3B-Instruct	video+audio	60.3	83.7	74.1
Qwen3-Omni-30B-A3B-Instruct	audio → white noise	49.2	77.2	65.7
Qwen3-Omni-30B-A3B-Instruct	frames only	50.5	77.8	66.6
Qwen3-Omni-30B-A3B-Instruct	audio only	41.7	68.6	57.5
Qwen3-Omni-30B-A3B-Thinking	video+audio	58.7	81.9	72.4
Nemotron-3-Nano-Omni-30B	video+audio	57.2	80.4	70.9
video-SALMONN-2+-7B	video+audio	49.7	74.5	64.3
AV-Flamingo-9B	video+audio	48.8	74.5	63.9
MiniCPM-o-4.5	video+audio	49.2	70.4	61.7
OmniVinci-9B	video+audio	44.9	71.9	60.8
Qwen2.5-Omni-7B	video+audio	41.3	66.7	56.2
Gemma-4-12B	frames + audio	40.0	66.8	55.8
Inkling-Small	6 windows of 8 frames + 30 s audio	35.0	61.1	50.4
VideoLLaMA2.1-7B-AV	video+audio	32.8	54.1	45.4
Open, frames+ASR substitute				
Gemma-4-31B	32 frames per 58 s segment + ASR	60.3	84.0	74.2
GLM-4.5V	64 frames + ASR	60.5	80.7	72.4
Qwen3-VL-30B	768 frames + ASR	59.6	80.5	71.9
Gemma-4-12B	32 frames per 58 s segment + ASR	48.6	75.3	64.3
Open, audio only				
Audio-Flamingo-3	audio, 10-minute cap	35.2	56.0	47.4
Qwen2-Audio-7B	audio, 8 chunks of 30 s	31.1	45.7	39.7

F ROBUSTNESS CHECKS

Option rotation under the five-permutation MMAU protocol. GPT-5.5 with frames+ASR scores 83.5 / 84.1 / 83.8 / 83.9 / 83.7. The range is 0.6 points, majority-vote accuracy is 84.3, and all five rotations agree on 83.3% of questions. Qwen3-Omni with AV scores 74.1 / 74.2 / 75.0 / 74.3 / 73.2. The range is 1.8 points, majority-vote accuracy is 75.0, and all rotations agree on 71.2% of questions. On the 300-question Gemini-3.1-Pro AV subset, two rotations differ by 2.3 points.

Frame-budget curve for Qwen3-VL-30B with frames+ASR. The model scores 64.9 with 16 frames, 69.8 with 64, 72.0 with 256, and 71.9 with 768, showing saturation at 256. A separate one-frame control on GPT-5.5 bounds static-frame shortcuts: given a single frame without the transcript, GPT-5.5 scores 62.5, 3.6 points above its text-only baseline of 58.9, against 77.4 with 32 frames.

Subsample stability. We use 200 draws of video-level 80% subsamples and report the 5th–95th percentile range. Gemini-3.1-Pro AV runs from 86.2 to 87.8, GPT-5.5 frames+ASR from 83.3 to 85.1, Qwen3-Omni AV from 73.1 to 75.3, and Gemma-4-31B frames+ASR from 73.1 to 75.5.

Letter-bias check. Gold-letter counts are near-uniform at 277/275/284/290, and the most frequent letter accounts for 25.8% of questions. Letter-preference χ^2 with 3 degrees of freedom is 0.1–2.5 for the frontier systems and higher for some open models, including 11.1 for Qwen3-Omni and 26.7 for MiniCPM-o. Under the rotation protocol above, GPT-5.5 and Qwen3-Omni move by at most 1.8 points; MiniCPM-o was not rotated.

G DIFFICULTY STRUCTURE

Solve-count distribution. The analysis covers the 16 systems of Table 3, each in its main configuration. Five questions, or 0.4%, are solved by no system; 87 questions, or 7.7%, are solved by three or fewer; the median question is solved by 12; and 112 questions, or 9.9%, are solved by all 16.

Composition of the hardest questions. We define the hardest questions as the 87 questions answered by at most 3 of the 16 systems. Speech questions make up 32.2% of this set versus 18.4% of the full benchmark, a difference of $+13.8$ points. Voice-evidence questions make up 69.0% versus 59.6%, a difference of $+9.4$ points.

H CONFIDENCE INTERVALS AND PER-SLICE ACCURACY

Table 11 gives the Wilson 95% interval of every overall accuracy in Tables 3 and 5; the tables after it split accuracy by scene family and by identity dimension.

Table 11: Wilson 95% intervals for the overall accuracies of Tables 3 and 5. *Unanswered*: questions left unanswered or unparsable, scored as wrong.

System	Input	Accuracy	Lower	Upper	Unanswered
Gemini-3.1-Pro	native video+audio	86.9	84.9	88.8	1
Gemini-3.1-Pro	audio → white noise	79.2	76.8	81.5	3
Gemini-3.1-Pro	32 frames only	73.6	71.0	76.1	4
GPT-5.5	32 frames + ASR	84.2	81.9	86.2	8
GPT-5.5	32 frames only	77.4	74.8	79.7	16
GPT-5.5	1 frame only	62.5	59.7	65.3	16
GPT-Audio	audio only	51.0	48.1	53.9	40
Nemotron-3-Nano-Omni-30B	video+audio	70.9	68.1	73.4	0
Qwen3-Omni-30B-A3B-Thinking	video+audio	72.4	69.7	74.9	9
Qwen3-Omni-30B-A3B-Instruct	video+audio	74.1	71.4	76.5	0
Qwen3-Omni-30B-A3B-Instruct	audio → white noise	65.7	62.9	68.4	0
Qwen3-Omni-30B-A3B-Instruct	frames only	66.6	63.8	69.3	0
video-SALMONN-2+-7B	video+audio	64.3	61.5	67.0	9
AV-Flamingo-9B	video+audio	63.9	61.1	66.7	0
MiniCPM-o-4.5	video+audio	61.7	58.8	64.5	0
OmniVinci-9B	video+audio	60.8	58.0	63.6	0
Qwen2.5-Omni-7B	video+audio	56.2	53.3	59.1	0
Gemma-4-12B	frames+audio	55.8	52.9	58.6	19
Inkling-Small	6 windows of 8 frames + 30 s audio	50.4	47.4	53.3	155
VideoLLaMA2.1-7B-AV	video+audio	45.4	42.5	48.3	0
Audio-Flamingo-3	audio, 10-minute cap	47.4	44.5	50.3	0
Qwen2-Audio-7B	audio, 8 chunks of 30 s	39.7	36.9	42.6	22
Gemma-4-31B	32 frames per 58 s segment + ASR	74.2	71.6	76.7	10
GLM-4.5V	64 frames + ASR	72.4	69.7	74.9	63
Qwen3-VL-30B	768 frames + ASR	71.9	69.2	74.5	0
Gemma-4-12B	32 frames per 58 s segment + ASR	64.3	61.5	67.0	9

Flatness checks for Gemini-3.1-Pro AV. The five 20% evidence-position bins score 86.9 / 89.6 / 86.0 / 82.5 / 88.0. Speaker-count buckets for 2 / 3 / 4 / 5+ speakers score 88.2 / 85.4 / 85.2 / 88.6. Duration buckets of $\leq$5 / 5–8 / 8–10 / $>$10 minutes score 86.7 / 86.3 / 89.7 / 86.1. Buckets of the

Table 12: Accuracy by scene family. Column abbreviations are G-AV for Gemini-3.1 native AV, G5-FA for GPT-5.5 frames+ASR, G-F for Gemini 32 frames only, Q-AV / Q-F / Q-A for Qwen3-Omni AV / frames / audio, and Text for GPT-5.5 text-only question+options.

Scene	n	G-AV	G5-FA	G-F	Q-AV	Q-F	Q-A	Text
conversation	239	92.1	90.0	80.3	76.2	67.8	55.2	62.3
event/performance	162	78.4	78.4	62.3	67.3	61.1	52.5	49.4
lifestyle vlog	188	90.4	83.0	75.0	76.1	68.1	66.5	62.2
news/documentary	100	90.0	87.0	71.0	82.0	74.0	66.0	71.0
procedure/work	66	75.8	84.8	68.2	68.2	62.1	51.5	50.0
scripted narrative	371	86.8	82.7	75.2	73.6	66.3	55.5	57.4

Table 13: Accuracy by identity dimension, using the same column key. The speech-content dimension shows the largest native-AV advantage over 32 frames without audio. Gemini scores 80.7 versus 55.1, consistent with the voice-to-person link being the difficulty.

Dimension	n	G-AV	G5-FA	G-F	Q-AV	Q-F	Q-A	Text
action	175	86.9	78.3	70.9	70.3	60.0	51.4	55.4
appearance	215	87.9	82.3	86.5	79.1	80.0	47.9	52.6
identity/naming	134	85.1	87.3	73.9	67.2	67.2	31.3	55.2
relation	144	93.8	93.1	83.3	82.6	72.9	79.2	72.9
speech content	207	80.7	79.2	55.1	68.1	51.7	67.1	50.2
state	170	86.5	88.2	70.6	73.5	66.5	60.0	64.7
style	29	96.6	89.7	86.2	82.8	79.3	65.5	79.3
voice quality	52	90.4	82.7	78.8	80.8	67.3	75.0	71.2

time between the first and last evidence span, 0 / $\leq$1 / 1–5 / $>$5 minutes, score 89.0 / 82.0 / 87.4 / 86.4.

I Evaluation and training settings

The prompts of every setting are given verbatim in Appendix J. Every reply is scored by the letter it chooses; a reply that names no option, runs out of budget, or refuses counts as wrong.

I.1 Whole-video systems

Gemini-3.1-Pro, video+audio. The full video with its audio track is transcoded to fit inline limits at 480 px wide, stepping down to 384 or 320 px only when needed, with AAC 64k mono and a maximum size of 18 MB, and passed as a video part; 1 question is unanswered.

GPT-5.5. The protocol uniformly samples 32 frames at 448 px. The API accepts images but not audio, which we measured directly. The ASR condition adds the benchmark's unlabeled Whisper transcript, and the completion cap is 8192 tokens.

GPT-Audio. The full audio track as a 64k mono MP3; the API accepts audio but not images.

Text-only baselines. Every model receives the question and its four options with the same text-only prompt and no media. Open models decode greedily and run through vLLM where the architecture is supported and through their own transformers code path otherwise (AV-Flamingo, OmniVinci, video-SALMONN-2+, VideoLLaMA2.1); GPT-5.5 and Gemini-3.1-Pro run once each under their API's default sampling. Reasoning models receive an 8k budget and MiniCPM-o-4.5 a 4k budget.

Nemotron-3-Nano-Omni-30B. vLLM serving, 2 fps up to 128 frames; videos over its 10-minute limit are split into parts and the part notes combined.

Qwen3-Omni-30B-A3B-Instruct. The official transformers inference path, 2 fps capped at 256 frames, with the audio interleaved into the video; greedy decoding. The audio-only input scores 57.5 with every question answered.

Qwen3-Omni-30B-A3B-Thinking. The same whole-video protocol as the Instruct row: the video with its audio track, 2 fps capped at 256 frames, audio interleaved, the same prompts, no transcript, and no turn segmentation. It is served with vLLM 0.27.1 under the released sampling config and a 32k budget. Answers are parsed from the text after the reasoning block, and a chain that exhausts the budget counts as unanswered (9 questions). The track is extracted with ffmpeg to 16 kHz before being interleaved, and the 15 questions whose prompt exceeds the model's 65,536 positions (up to 77,899 tokens) run with the position limit lifted, as they do in the Instruct row. Its caption cells in Table 3 come from the same path with the caption prompt (Appendix P).

Qwen2.5-Omni-7B. The vLLM path of the Thinking row with greedy decoding: 2 fps capped at 256 frames, audio interleaved, the Qwen prompts.

AV-Flamingo-9B. 128 frames and the full-length audio through the released transformers fork; non-commercial license, evaluation only.

MiniCPM-o-4.5. The official omni video path, which keeps at most 64 one-second frame-and-audio blocks, 8–17% of a track. Given every second of the track with 128 frames instead, the model starts its reply with a letter on 29% of the questions and scores 25.2 under the lenient parser, the chance level, so the table reports the official path.

Gemma-4-12B and Gemma-4-31B. 58-second segments with 32 frames each under both frames+audio and frames+ASR. The two conditions differ in more than the form of speech: the audio encoder accepts at most 30 s, so frames+audio hears only the first 30 s of each segment, while frames+ASR reads the transcript of the whole video with every segment and again when the segment notes are combined. Gemma-4-31B ships no audio encoder; its model card specifies text+image, so it runs frames+ASR only.

GLM-4.5V. A thinking-mode model. Answers are read from the final `ANSWER:` tag after the reasoning block and otherwise from the last standalone letter near the end of the reply. The main table uses a 64-frame budget; the 16-frame run scores 72.3.

Qwen3-VL-30B. The processor's default 2 fps and a 768-frame cap in the main table; Appendix F reports the frame curve.

Inkling-Small. The model accepts text, images, and audio but not video. Each question uses 6 windows with 8 frames and 30 s of audio, followed by aggregation. This window-limited protocol leaves 155 questions unanswered, which count as wrong.

OmniVinci-9B. The official model-repository recipe with remote code, 128 frames, and an audio window of `max_3600` s.

video-SALMONN-2+-7B. The official minimal-inference path on the released full checkpoint, `max_frames` 768 at `max_pixels` 61250 as in the paper setting, with flash-attention-2. Nine questions fail inside the model's audio windowing with a shape error and count as wrong.

VideoLLaMA2.1-7B-AV. The `audio_visual` branch in its native 16-frame configuration with the joint audio branch enabled, and sdpa attention, because fp16 eager attention overflows to degenerate outputs.

Open audio-only models. These are not in Table 3. Audio-Flamingo-3 (Goel et al., 2025) scores 47.4 overall against a text-only baseline of 42.6, and Qwen2-Audio-7B (Chu et al., 2024) 39.7 against 36.2, three and a half to five points from the track; their text-hard and caption rows are in Appendix E and Table 20. Audio-Flamingo-3 uses the transformers-native build, and its inputs are truncated at the model's 10-minute cap; the median video lasts 7.9 minutes, so only the tail of videos up to 13 minutes is cut. The model has a non-commercial license and is used for evaluation only. Qwen2-Audio-7B runs through native transformers inference; its 30-second encoder cap is bridged by 8 uniform non-overlapping 30 s chunks presented as a temporally ordered multi-audio message. In 22 questions its reply stops before naming an option, and these count as unanswered.

I.2 PER-TURN PROTOCOL AND DECODING

Input ladder (Table 6). Qwen3-Omni-30B-A3B-Thinking is served with vLLM 0.27.1 under the per-turn protocol. Turns come from the benchmark diarization; frames are sampled on a 0.4 fps grid and attached to the turn they fall in; each turn carries its own audio clip and its Whisper transcript line, and the system prompt names only the channels present in a given condition. Rows labeled 3 s cap each clip at 3 s per turn, the training default; whole-turn rows use the uncapped turn (median 2.0 s, 90th percentile 6.0 s), which adds about 1.7k input tokens per question over the cap. Prompts never exceed 28k tokens against a 64k context, so the 32k generation budget is never shortened by the input; every truncated question stops at exactly 32,768 generated tokens. Decoding follows the released generation config (temperature 0.6, top-p 0.95, top-k 20, 32,768 new tokens), which is also the vendor's own benchmark setting for the Thinking variant; answers are parsed only from a final `Answer:` line, and questions that exhaust the budget count as wrong. Seeds 0 and 1 are averaged per question. The ladder is scored on the full-parameter 2-epoch checkpoint of Table 8, trained on frames+ASR+3 s clips, with no further training. It is run on the tuned checkpoint because the frozen model, at the released temperature, exhausts the budget on up to 18% of questions and under greedy decoding on 25%, and each added channel raises that rate, so its contrasts would track truncation rather than content. The frozen row of Table 6 is therefore taken at temperature 1.0, where 0.3% of its chains run out of budget, and serves as a reference.

Decoding temperature (Table 14). The released temperature is swept on the tuned checkpoint and, for the tuned−frozen contrast of Table 8, on the frozen model, holding the 32k budget, top-p, and top-k fixed; greedy decoding drops top-p and top-k. Each cell averages two independent runs. On the tuned model temperature acts only as sampling noise: on the questions both settings answer within budget, temperature 1.0 costs -2.4 and greedy decoding gains $+1.4$ relative to 0.6, and truncation stays at or below 1%. On the frozen model temperature also sets the loop rate, from 25% of chains under greedy decoding to 0.3% at 1.0, so a tuned−frozen contrast taken at a shared temperature runs from $+13.1$ to -2.2 across the sweep while the tuned model's own score moves by 3.4 points; Table 8 therefore takes the frozen model at its best temperature as the reference row.

Table 14: Decoding-temperature sweep on the frames+ASR+3 s input, two runs per cell, strict and lenient parsers, and the tuned−frozen strict contrast at that temperature.

Temperature	tuned strict	tuned lenient	tuned trunc.	frozen trunc.	tuned − frozen
0 (greedy)	70.1	70.3	1.1%	25.0%	+13.1
0.6 (released)	69.2	69.4	0.5%	14.7%	+6.5
1.0	66.7	66.8	0.1%	0.3%	−2.2

I.3 EVIDENCE CITATION

Every question carries annotated evidence spans, and chains under the per-turn protocol cite turns by index, so the reasoning can be scored against the annotation without a judge (Table 15). A chain's evidence claim is the three turns it cites most often; the claim *hits* when one of them lies within one turn of an annotated evidence turn, where evidence turns are the annotated ids plus every turn overlapping an annotated span (median 6 per question). *Precision* is the share of the three that hit, *right for the right reason* is a correct answer with a hit, and *correct without evidence* a correct answer without one. Each setting averages two runs, and deltas are paired per question against the frozen reference. The top-3 rule keeps chains that cite every turn from hitting trivially; counting any cited turn, the frozen and tuned hit rates are 69.2 and 69.7. Turn ranges are capped at 12 turns and out-of-range indices dropped.

I.4 TRAINING

Fine-tuning (Table 8). The training split is the one of Appendix K. Inputs use the per-turn layout of the ladder with the 3 s audio clip, a maximum length of 26,624 tokens, gradient accumulation 4, two epochs, and 5% warm-up. LoRA uses rank 128 and α 128 at learning rate 10^{-4} on all linear layers, which under transformers 5 resolves to the attention projections of this architecture. Full-parameter tuning trains the thinker at 10^{-5} with the vision and audio encoders frozen, using

Table 15: Evidence citation under the per-turn protocol: accuracy, hit rate and precision of the three most-cited turns against the annotated evidence, right for the right reason (RRR), correct without evidence (CwoE), and paired deltas of hit rate and RRR against the frozen reference.

Setting	acc	hit	prec.	RRR	CwoE	Δ hit	Δ RRR
frozen, T=1.0	68.9	53.4	39.8	40.1	28.8	reference	
frozen, T=1.0, no track	69.8	56.8	39.7	43.2	26.6	+3.4	+3.1
LoRA r128, 2 ep	67.4	46.8	39.3	34.5	32.9	−6.6	−5.7
full 2 ep	69.2	59.9	41.7	45.5	23.7	+6.4	+5.4
full 2 ep + GRPO 1 ep	70.7	61.5	42.6	46.7	24.0	+8.0	+6.5
full 2 ep + GRPO 2 ep	69.2	63.3	43.6	47.3	21.9	+9.9	+7.2
full 2 ep, greedy	70.1	61.4	43.0	47.6	22.5	+7.9	+7.5
full 2 ep, whole-turn audio	70.3	62.5	42.5	47.9	22.5	+9.1	+7.7
full 2 ep, no track at test	69.6	61.3	43.4	45.9	23.7	+7.9	+5.7
full 2 ep, frames only at test	59.9	38.6	27.7	24.1	35.8	−14.8	−16.1
full 2 ep, ASR only at test	48.6	45.8	35.8	28.6	20.0	−7.6	−11.5

FSDP with CPU offload. The whole-turn row keeps every setting and replaces each 3 s clip with the uncapped turn in training and scoring; on the 52 voice-quality questions it scores 12.5 points above the 3 s checkpoint scored without the track. The lenient parser is lmms-eval's multiple-choice extractor applied after stripping the reasoning block.

Frame layouts. The dense-frame layout adds 4 Hz frames over whole turns, taken in the order of a voiceprint selector, with a 300-frame cap. The selector embeds each turn's three-second clip with ECAPA (Desplanques et al., 2020), clusters the turns by average-linkage agglomerative clustering at a cosine-distance threshold of 0.85, chosen on 300 training windows, and scores each turn as $0.5\,b + 0.5\,(1 - \hat{m})$, where b is 1 when the turn's cluster differs from the previous turn's and $\hat{m}$ is the normalized margin between the turn's cosine similarity to its own cluster centroid and to the nearest other centroid; the ranking does not depend on the question. The mixed-resolution layout keeps key turns at 448 px and the rest at 192 px.

Identity-link test under the per-turn protocol. Every different-person pair of the first set compares two voices (Appendix R), so we score its 2,195 voice pairs. Over two runs, the frozen model reaches 67.1 same-person and 54.4 different-person recall on them, balanced 60.7; the full-parameter 2-epoch checkpoint reaches 58.1 and 60.5, balanced 59.3; and GRPO after one and two epochs reaches balanced 60.5 and 60.8. On the 1,443 same-person pairs that involve a face or an on-screen moment, same-person recall is 73.1, 72.9, 71.7, and 72.5 for the same four models.

GRPO. The 2-epoch full-parameter checkpoint is continued with rank-128 adapters on the same attention projections, with the fine-tuned checkpoint itself as the reference model and the original GRPO objective: group-normalized advantages, clipping 0.2, KL coefficient 0.04, and one policy update per rollout batch. The reward is 0.9 for a strict gold answer line plus 0.1 for any parseable answer line. Each step draws sixteen prompts with eight samples each at temperature 1.0 and an 8k completion cap, at learning rate 10^{-5}, for two epochs (134 steps), and the end-of-epoch checkpoints are reported. The prompts are the 1,087 of 3,000 randomly drawn training-split prompts whose four samples from the fine-tuned checkpoint were neither all right nor all wrong, so every group carries a non-zero advantage without online resampling. Rollouts come from a co-located vLLM engine; a mean chain is about 450 tokens, and the KL to the fine-tuned checkpoint ends at 0.007.

J PROMPTS

This appendix gives every prompt verbatim, with long lines wrapped. Names in braces are filled for each question: {`q`} is the question; {`opts`} is the four options, one per line, each starting with its letter, as in `A: I still remember when`; {`plain`} is the benchmark transcript without speaker labels, one utterance per line; and {`numbered`} is the same transcript with every line written as `[k] (t=SECONDSs) words`.

J.1 WHOLE-VIDEO MULTIPLE CHOICE

Gemini-3.1-Pro, Nemotron-3-Nano-Omni-30B, and AV-Flamingo-9B. These systems receive the video with its audio track and this prompt. The white-noise row of Gemini-3.1-Pro uses it unchanged.

```
You are given a video -- watch AND listen carefully; the audio track (speech, voices) matters
as much as the visuals. Answer the multiple-choice question about the people in it.

Question: {q}
Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.
```

Nemotron-3-Nano-Omni-30B splits a video longer than its 10-minute limit into parts. Each part receives the prompt below, where {q} is the full prompt above, and a text-only call combines the part notes.

```
This is part {k}/{K} ({s0}-{s1}s) of a longer video.
{q}
First give one sentence of evidence from this part, then your best guess letter.

{K} consecutive parts of one video were examined separately. Notes:
[{s0}-{s1}s] {part note}
...

{q}
```

GPT-5.5 and the 32-frame rows of Gemini-3.1-Pro. These rows receive thirty-two frames sampled uniformly from the video. {n} is 32 and {asr} is empty for the frames-only rows.

```
You are given {n} frames uniformly sampled, in order, from one video. There is NO audio.{asr}
Answer the multiple-choice question about the people in the video.

Question: {q}
Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.
```

For the frames+ASR row of GPT-5.5, {asr} is the sentence below, and the message ends with the transcript block after the frames.

```
 A transcript of the speech (WITHOUT speaker labels) is provided after the frames.

TRANSCRIPT (no speaker labels):
{plain}
```

GPT-Audio. The audio track as a 64k mono MP3 with this prompt.

```
You are given the AUDIO TRACK of a video (speech and voices matter). There is no visual input.
Answer the multiple-choice question about the people in the recording.

Question: {q}
Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.
```

Qwen3-Omni-30B-A3B-Instruct, Qwen3-Omni-30B-A3B-Thinking, and Qwen2.5-Omni-7B. The system message below, then the video and the user text. The frames-only, white-noise, clue-restricted, and option-rotation rows use the same text with the media or the option order changed; the audio-only row describes its input as an audio recording.

```
You are Qwen, a multimodal assistant that can see and hear.

You are given a video (watch and listen carefully). Answer the multiple-choice question about
the people/objects in it.
```

```
Question: {question}
Options:
{options}

Reply with ONLY the letter (A, B, C or D).
```

OmniVinci-9B, video-SALMONN-2+-7B, and VideoLLaMA2.1-7B-AV. The same user text as the Qwen rows, with no system message of ours.

Audio-Flamingo-3 and Qwen2-Audio-7B. The audio with the prompt below. Qwen2-Audio-7B receives the sentence "These are consecutive audio excerpts sampled uniformly from one long recording, in temporal order.", then its eight 30 s chunks, then the prompt.

```
You are given an audio recording (listen carefully). Answer the multiple-choice question about
the people in it.

Question: {q}
Options:
{opts}

Reply with ONLY the letter (A, B, C or D).
```

MiniCPM-o-4.5. The model's own omni-mode system message, the interleaved one-second frame and audio blocks, then this prompt.

```
Answer the multiple-choice question about the people in the video you just watched and heard.

Question: {q}
Options:
{opts}

Reply with ONLY the letter (A, B, C or D).
```

GLM-4.5V and Qwen3-VL-30B. The frames with this prompt. Qwen3-VL-30B receives the same text with the last line replaced by `Reply with ONLY the letter (A, B, C or D).`

```
You are given a video (silent to you -- you cannot hear it) and the transcript of its speech
WITHOUT speaker labels. Answer the multiple-choice question about the people in the video.

TRANSCRIPT (no speaker labels):
{plain}

Question: {q}
Options:
{opts}

Finish your reply with exactly this line: ANSWER: <letter>
```

Gemma-4-31B and Gemma-4-12B with frames+ASR. Each 58 s segment receives its 32 frames and the first prompt below. A text-only call then combines the segment notes, one per line as `[{s0}-{s1}s] note`, with the second prompt.

```
This is segment {k}/{n} (seconds {a}-{b}) of a longer video. The transcript of THE WHOLE
video's speech (no speaker labels) is:
{plain}

Question about the whole video: {q}
Options:
{opts}

From THIS segment alone, briefly note any visual evidence relevant to the question (who is
visible, what they wear, what happens), then give your best current guess. Format: EVIDENCE:
<one or two sentences> GUESS: <letter>


Different segments of one video were examined separately. The per-segment notes are below. The
full transcript (no speaker labels):
{plain}

Question: {q}
```

```
Options:
{opts}

SEGMENT NOTES:
{notes}

Combining all segment notes, reply with ONLY the letter (A, B, C or D).
```

Gemma-4-12B with frames+audio. Each segment receives its frames and audio with the prompt below; the combining prompt is the one above, with the transcript line reading `(audio was heard directly per segment)`.

```
This is segment {k}/{n} (seconds {a}-{b}) of a longer video, WITH ITS AUDIO.
Question about the whole video: {q}
Options:
{opts}

From THIS segment alone, briefly note any evidence relevant to the question (who is visible,
who speaks, what they say), then give your best current guess. Format: EVIDENCE: <one or two
sentences> GUESS: <letter>
```

Inkling-Small. Each of the six windows receives its eight frames and 30 s of audio with the first prompt; a text-only call combines the window notes with the second. A reply without a letter gets one more short call that asks for the letter alone.

```
This is window {k}/{n} ({a}-{b}s) of a longer video: frames from this window plus its AUDIO.

Question about the whole video: {q}
Options:
{opts}

From THIS window alone, note any evidence (who is visible, who speaks, what is said), then a
best guess. Format: EVIDENCE: <1-2 sentences> GUESS: <letter>


Windows of one video were examined separately; notes below.

Question: {q}
Options:
{opts}

WINDOW NOTES:
{notes}

Combining all notes, finish with exactly: ANSWER: <letter>
```

J.2 TEXT-ONLY BASELINES

Every system receives one user message with no media and no system message: the first prompt for the text-only baseline and the second for the options-only reading of Appendix R.

```
Answer this multiple-choice question. You are NOT given the video, audio or transcript it
refers to -- answer with your best guess from the question and options alone.

Question: {q}
Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.


Below are four answer options from a multiple-choice question about a video. The question
itself is NOT shown. Pick the option most likely to be the intended correct answer.

Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.
```

J.3 IDENTITY-LINK TEST

Gemini-3.1-Pro and both Qwen3-Omni models receive the video and this prompt; the Qwen models also receive the system message of Appendix J.1.

```
Watch and listen to this video, paying attention to the time positions mentioned below.

Consider: (1) {a}; (2) {b}.

Are (1) and (2) the SAME person, or DIFFERENT people?
Reply with ONLY one JSON: {"answer":"SAME"} or {"answer":"DIFFERENT"}.
```

Each moment is written as `the person SPEAKING at around {t} seconds` for a voice, `the person whose face is visible at around {t} seconds` for a face pair, and `the person visible at around {t} seconds` for the other visual moments; a moment longer than two seconds reads `from {t0}s to {t1}s` instead.

J.4 PER-TURN PROTOCOL

The system message names the channels present. Frames+ASR+audio, with 3 s clips or whole turns, and the mixed-resolution and resolution variants use the first prompt below.

```
You are analyzing a conversation transcript. Speaker names are not given. Each turn may start
with video frames sampled around that moment showing the scene, and has the audio recording of
that turn -- the voice of its speaker: turns spoken by the same person sound alike and often
show the same person on screen. First group the turns by speaker using the voices and the
faces in the frames, then identify the evidence turns, then think step by step about the
question, then give the final answer.
```

Frames+audio without the transcript:

```
You are analyzing a conversation. Speaker names and the transcript are not given. Each turn
may start with video frames sampled around that moment showing the scene, and consists of the
audio recording of that turn -- what is said and the voice of its speaker: turns spoken by the
same person sound alike and often show the same person on screen. First listen to each turn
and group the turns by speaker using the voices and the faces in the frames, then identify the
evidence turns, then think step by step about the question, then give the final answer.
```

Frames+ASR without audio:

```
You are analyzing a conversation transcript. Speaker names are not given. Each turn may start
with video frames sampled around that moment showing the scene: turns spoken by the same
person often show the same person on screen. First group the turns by speaker using the faces
in the frames and the content of what is said, then identify the evidence turns, then think
step by step about the question, then give the final answer.
```

Frames only:

```
You are analyzing a conversation. Speaker names, the transcript and the audio are not given:
each turn is represented only by video frames sampled around that moment showing the scene;
turns spoken by the same person often show the same person on screen. First group the turns by
speaker using the faces in the frames, then identify the evidence turns, then think step by
step about the question, then give the final answer.
```

ASR only:

```
You are analyzing a conversation transcript. Speaker names are not given, and no video frames
or audio are provided -- only the transcript text of each turn. First group the turns by
speaker using the content and style of what is said, then identify the evidence turns, then
think step by step about the question, then give the final answer.
```

The dense-frame layouts, at selector turns or at the annotated evidence turns:

```
You are analyzing a conversation transcript. Speaker names are not given. Each turn may start
with video frames showing the scene (moments that matter most for telling speakers apart are
sampled more densely), and has the audio recording of that turn -- the voice of its speaker:
turns spoken by the same person sound alike and often show the same person on screen. First
group the turns by speaker using the voices and the faces in the frames, then identify the
evidence turns, then think step by step about the question, then give the final answer.
```

The user message lists the turns in order. A turn carries one `<image>` per grid frame that falls inside it and one `<audio>` for its clip, each replaced by the image or audio itself; a condition without a transcript starts the block with `Turns:` and leaves out the words. The question comes last. The beginning of one question under frames+ASR+audio:

```
Transcript:
Turn 1 <image> <image> <image> <image> <image> <image> <image> <image> <audio>: Well here we
are again.
Turn 2 <image> <image> <audio>: Hold on, hold on one sec.
Turn 3 <image> <image> <audio>: Hold on one sec.
...

Question: Which lyric is sung by the performer who later wears large, ornate dangling
earrings?
A: I still remember when
B: We keep coming back
C: It all comes down to me
D: I guess it must be fate
Answer with the option letter only after your reasoning, as a final line "Answer: X".
```

The per-turn identity-link test uses the frames+ASR+audio system message and ends the same block with the question below, where each moment reads, for example, `the person SPEAKING in Turn 12` or `the person visible in the frames of Turn 76`.

```
<per-turn block>

Question: Consider (1) {A} and (2) {B}. Are (1) and (2) the same person, or different people?
A: the same person
B: different people
Answer with the option letter only after your reasoning, as a final line "Answer: X".
```

The dense 4B model of Appendix M receives no transcript and a voiceprint token in place of each turn's audio, with this system message for the uniform layout:

```
You are analyzing a conversation. Speaker names are not given and the speech is not
transcribed -- the words are not available. Each turn may start with video frames sampled
around that moment showing the scene, and has a voiceprint token that encodes the voice of its
speaker: turns spoken by the same person have similar voiceprints and often show the same
person on screen. Use the voiceprints and the faces and visible details in the frames to work
out who is who, find the evidence turns, think step by step about the question, then give the
final answer.
```

J.5 IDENTITY-GROUNDED CAPTIONS

Whole video with its audio track. Gemini-3.1-Pro, both Qwen3-Omni models, and Nemotron-3-Nano-Omni-30B.

```
Watch AND listen to this video, then write an identity-grounded description of it in three
labeled sections.

Dialogue: transcribe every spoken utterance in order, one per line, as '[m:ss] Speaker X:
words', with consistent speaker labels (Speaker A, Speaker B, ... in order of first
appearance). The same voice must always get the same label, even when that person is off
screen.
Names: for each speaker whose name the video itself supports (said aloud or shown on screen),
one line 'Speaker A = Name'; write 'none' if no name is supported.
Narrative: describe the video in temporal order with explicit speaker attribution: what each
speaker looks like and wears, what they say and do, when each person enters and leaves, and
whether people seen or heard at different times are the same person. Do not speculate beyond
what you can see and hear.
```

GPT-5.5 with 32 frames and the indexed transcript. The prompt with {`n`} set to 32, the frames, then `TRANSCRIPT (no speaker labels):` followed by {`numbered`}.

```
You are given {n} frames uniformly sampled, in order, from one video, and a transcript of its
speech WITHOUT speaker labels, one utterance per line as '[k] (t=SECONDSs) words'. There is NO
audio. Write an identity-grounded description in three labeled sections.

Speakers: assign EVERY transcript line to a speaker, one line per speaker, as 'S1: [0], [1],
[5], ...' using the bracketed line indices (each index exactly once; S1, S2, ... in order of
first appearance).
Names: for each speaker whose name the video itself supports (in the transcript or on screen),
one line 'S1 = Name'; write 'none' if no name is supported.
Narrative: describe the video in temporal order with explicit speaker attribution: what each
speaker looks like and wears, what they say and do, when each person enters and leaves, and
whether people seen at different times are the same person. Do not speculate beyond what you
can see and read.
```

Qwen3-VL-30B and GLM-4.5V. Sixty-four frames with this prompt.

```
You are given a video (silent to you, you cannot hear it) and the transcript of its speech
WITHOUT speaker labels, one utterance per line as '[k] (t=SECONDSs) words'. Write an
identity-grounded description in three labeled sections.

Speakers: assign EVERY transcript line to a speaker, one line per speaker, as 'S1: [0], [1],
[5], ...' using the bracketed line indices (each index exactly once; S1, S2, ... in order of
first appearance).
Names: for each speaker whose name the video itself supports (in the transcript or on screen),
one line 'S1 = Name'; write 'none' if no name is supported.
Narrative: describe the video in temporal order with explicit speaker attribution: what each
speaker looks like and wears, what they say and do, when each person enters and leaves, and
whether people seen at different times are the same person. Do not speculate beyond what you
can see and read.

TRANSCRIPT (no speaker labels):
{numbered}
```

Gemma-4-31B and Gemma-4-12B. Each segment receives its frames, its audio for the 12B model, and the first prompt, where {avnote} reads , WITH ITS AUDIO for the 12B model and is empty for the 31B model; a text-only call merges the segments with the second.

```
This is segment {k}/{n} (seconds {a}-{b}) of a longer video{avnote}. The transcript lines that
fall in this segment (no speaker labels, '[k] (t=SECONDSs) words') are:
{lines}

From THIS segment, assign each of these lines to a speaker using LOCAL labels (P1, P2, ...)
and describe each speaker's appearance in one clause so the segments can be linked later.
Format, one line per speaker: 'P1 (<appearance, voice>): [k], [k], ...'

Segments of one video were described separately with local speaker labels. Using the
appearance/voice clauses to decide which local labels across segments are the same person,
produce ONE identity-grounded description in three labeled sections.
Speakers: global labels S1, S2, ... in order of first appearance, each line 'S1: [k], [k],
...' listing every transcript index exactly once.
Names: 'S1 = Name' where the transcript supports it, else 'none'.
Narrative: the video in temporal order with explicit speaker attribution.

FULL TRANSCRIPT (no speaker labels):
{numbered}

SEGMENT NOTES:
{notes}
```

GPT-Audio. The audio track with this prompt.

```
Transcribe every spoken utterance in this video in order, one per line, exactly in the form
'[m:ss] Speaker X: words'. Use consistent speaker labels (Speaker A, Speaker B, ... in order
of first appearance): the same voice must always get the same label. Output only these lines,
nothing else.
```

Frames-only captions. Gemini-3.1-Pro and GPT-5.5 with 32 frames, and GPT-5.5 with one frame; {dur} states the video length and the time of each frame.

```
You are given {n} frames uniformly sampled, in order, from one video.{dur} There is NO audio
and NO transcript. Write an identity-grounded description of the video in three labeled
sections.

Dialogue: transcribe every utterance you can recover from the frames themselves (on-screen
captions, subtitles, or other legible text), in order, one per line, as '[m:ss] Speaker X:
words', with consistent speaker labels (Speaker A, Speaker B, ... in order of first
appearance); the same person must always get the same label. Write 'none' if no speech is
recoverable.
Names: for each speaker whose name the video itself supports (shown on screen), one line
'Speaker A = Name'; write 'none' if no name is supported.
Narrative: describe the video in temporal order with explicit speaker attribution: what each
person looks like and wears, what they appear to say and do, when each person enters and
leaves, and whether people seen at different times are the same person. Do not speculate
beyond what you can see.
```

Windowed rows. Each 58 s window of MiniCPM-o-4.5, OmniVinci-9B, Qwen2.5-Omni-7B, video-SALMONN-2+-7B, and VideoLLaMA2.1-7B-AV receives the first prompt, AV-Flamingo-9B the second, and each 30 s window of Audio-Flamingo-3 and Qwen2-Audio-7B the third.

```
Transcribe this clip as a dialogue: one line per utterance, in the form 'Speaker X: words'
(Speaker A, Speaker B, ... in order of first appearance; the same voice keeps the same label).
Output only these lines.

Who is speaking in this video and what do they say? Give the dialogue with speaker labels and
the time of each line.

Transcribe this audio clip as a dialogue: one line per utterance, in the form 'Speaker X:
words' (Speaker A, Speaker B, ... in order of first appearance; the same voice keeps the same
label). Output only these lines.
```

Inkling-Small. Each window receives its frames and audio with the first prompt; a text-only call merges the windows with the second.

```
This is window {k}/{n} ({a}-{b}s) of a longer video: frames from this window plus its AUDIO.
Transcribe every utterance you hear in this window, one per line, as '[m:ss] P1: words' with
the clock time measured from the START OF THE VIDEO (this window starts at {a}s) and LOCAL
speaker labels P1, P2, ... Then, one line per local speaker, describe that person's voice and,
if visible, appearance: 'P1: <description>'.

Windows of one video were transcribed separately with LOCAL speaker labels (P1, P2 restart in
every window); each window also describes its speakers. Using the voice/appearance
descriptions to decide which local labels across windows are the same person, write ONE
identity-grounded description in three labeled sections.
Dialogue: every utterance in order as '[m:ss] Speaker X: words' with GLOBAL labels (Speaker A,
Speaker B, ... in order of first appearance), keeping the given clock times.
Names: 'Speaker A = Name' where the video supports it, else 'none'.
Narrative: the video in temporal order with explicit speaker attribution.

Output ONLY the three sections, starting with the line 'Dialogue:'; no preamble, no reasoning,
no notes.

WINDOW NOTES:
{notes}
```

Per-turn rows. The system message below, then the per-turn block of Appendix J.4 without the question, followed by `Produce the annotation with the three sections Speakers, Names, Narrative.`

```
You are analyzing a conversation transcript. Speaker names are not given. Each turn may start
with video frames sampled around that moment showing the scene, and has the audio recording of
that turn, the voice of its speaker: turns spoken by the same person sound alike and often
show the same person on screen. Produce an identity-grounded annotation of the whole video in
three labeled sections:
Speakers: group EVERY turn by speaker, one line per speaker, as 'S1: Turn 1, Turn 4, ...'
(each turn assigned to exactly one speaker; use S1, S2, ... in order of first appearance).
Names: for each speaker whose name the video itself supports (said aloud, on screen, or in the
transcript), one line 'S1 = Name'; write 'none' if no name is supported.
Narrative: a detailed description of the video with explicit speaker attribution: who each
speaker is (appearance, role), what they say and do, and how the people relate, in the order
things happen.
```

J.6 CAPTION READERS

The answerability reader receives each caption with each of its video's questions:

```
Below is a description of a video, written by someone who watched it. Using ONLY this
description, answer the multiple-choice question about the video.

VIDEO DESCRIPTION:
{cap}

Question: {q}
Options:
{opts}

Reply with ONLY one JSON: {"answer":"X"} where X is A, B, C, or D.
```

The caption-level identity-link test gives the reader the caption and the two moments:

```
Below is a description of a video written by someone who watched it. Using ONLY this
description, decide whether two moments of the video involve the SAME person or DIFFERENT
people.
```

```
VIDEO DESCRIPTION:
{cap}

Moment A: {ma}
Moment B: {mb}

Is the person in Moment A the same person as in Moment B? If the description does not let you
tell, give your best guess. Reply with ONLY one JSON: {"answer": "SAME"} or {"answer":
"DIFFERENT"}.
```

Each moment is described as follows:

```
voice:  the person SPEAKING at {m:ss}-{m:ss} (transcript line [{k}], Turn {k+1}), whose words
there are: "{words}"
visual: the person VISIBLE ON SCREEN at {m:ss} (around transcript line [{k}], Turn {k+1},
where the words spoken are: "{text}")
```

K TRAINING SPLIT CONSTRUCTION

The training split reuses the benchmark's collection funnel and generator design on sources outside the benchmark, with every human step replaced by a model check.

Sources. Four pools contribute: web video from the benchmark's collection funnel, about 72% of the questions; concatenated CondensedMovies clips, 19%; AMI meetings, 5%; and AVA film segments with gold diarization, 3%. The web videos are not held to the benchmark's voice and face checks. They need an English diarized transcript with at least 40 turns, a length of 4 to 12 minutes, and at least two speakers, none of whom has more than 92% of the turns. No source appears in the benchmark, asserted at every merge. At most four questions are kept per video.

Cue inventories. For each source Gemini-3.1-Pro lists from sampled frames the visible attributes that could single out a person, and GPT-5.5 re-checks every entry against the frames. Gemini also listens to up to three clips of 3 to 6 s per speaker and lists non-semantic voice traits such as timbre, pace, and laughter, and GPT-Audio re-checks them. The generator is shown only the verified entries.

Generation. Gemini-3.1-Pro, a different family from the benchmark's generator, writes up to five candidate questions per source from the speaker-attributed transcript, the face timeline, and the verified inventories. Each candidate is assigned a cell that fixes its cue class, joint, voice, or visual, its identity dimension, and its depth, which puts the pool near the 50/25/25 class quota; the exact shares come from trimming at assembly. A voice-class question must be anchored on a voice trait and contain no visual attribute word.

Filters. GPT-5.6-sol answers every candidate question from the same annotations with no media, and a candidate it answers differently is dropped. Rule-based filters remove quoted speech, seekable time expressions, and questions whose structural cue class disagrees with the assigned cell. The generator is instructed not to use sensitive attributes. Answer letters are rebalanced. A cleanliness filter drops every source with even one turn among its first 260 that Whisper filled with ellipses, repetition loops, or gibberish.

Difficulty and necessity. The frozen Qwen3-Omni-30B-A3B-Thinking answers every question twice under its official decoding on the per-turn input. Two correct runs mark a question `easy`, one `relaxed`, and none `hard`, which gives 68, 14, and 18% of the split. Three further runs, without frames, without audio, and with the transcript alone, record whether the question needs its labeled channel and whether the transcript answers it. The transcript alone answers 63% of the questions and 29% are certified as needing their labeled channel; both flags ship with every row.

Targets. Every target is a `<think>` block followed by an answer line. When one of the two frozen runs answers correctly, finishes within budget, stays under 16k characters, and cites the evidence location and the channel its class requires, that run's reasoning is the target; this covers half of the rows. For the other half Gemini-3.1-Pro writes the reasoning from the gold answer, the evidence

turns and spans, the required identity edges, the verified inventories, the face timeline, and the labeled transcript. The prompt asks it to write in the first person and present tense in 150 to 450 words, cite turns and moments, state every required identity link in a fixed form, rule out the distractors, end with the answer line, and never mention speaker or face labels or that it was given anything. A check verifies the answer line, the evidence citations, the class-specific citations, the absence of speaker and face labels, and the length, and a target that falls into a repetition loop is discarded. A target from the frozen model is preferred whenever it passes.

Assembly. Rows use the per-turn layout of Table 6: the base frame grid at 0.4 fps assigned to turns, the first 3 s of each turn as audio, the transcript line, and the question after the last turn. Classes are trimmed to 50/25/25; scenes are not trimmed. The final split has 17,893 training and 1,109 validation rows over 5,445 sources, with scene shares of scripted narrative 26%, event 21%, conversation 19%, news 14%, lifestyle 12%, and procedure 7%. The GRPO stage draws its prompts from the same rows without the target.

L LAYOUT AND RESOLUTION VARIANTS ON QWEN3-OMNI

Table 16 gives every layout and resolution variant run on the full-parameter checkpoint under the per-turn protocol; Table 17 gives the same matrix on the dense 4B model, and Table 4 the oracle comparison across training settings. The reference is the uniform checkpoint re-scored on the same software stack as the variants, because numbers taken across stacks or across evaluation files cannot be paired: with the same model, input, and seed, the correct/incorrect agreement between two runs is 94 to 95% on the same file, 83 to 85% across files, and 78% across seeds, so a mixture-of-experts model served with continuous batching re-samples a question whenever its batch neighbours change. Two same-file controls isolate what training on a layout adds beyond evaluating on it: the checkpoint trained on the dense layout against the uniform checkpoint scored on the identical file reads $+0.4$, and the mixed-resolution pair $+0.2$. In the eval-time oracle row, the 524 questions with no annotated evidence turns have byte-identical inputs in the uniform and oracle files and still moved $+1.9$ between them, so the difference-in-differences for that row is $+0.5$. The uncapped oracle places dense segments on every annotated evidence turn (48.6k dense frames over the benchmark against 22.4k for the budgeted one; a median of 227 frames per question against 214, a maximum of 661 against 301) and scores below the budgeted oracle, so the budget cap is not what limits the gold layout. The resolution rows keep the uniform layout and rescale every frame: 192 px is 28 tokens per frame, the default about 300 px is 66, and 448 px is 144. On the frozen model at temperature 1.0 the same layouts read, against its own uniform reference over all 1,126 questions and two seeds: selector turns $+0.9$, budgeted gold turns $+0.7$, uncapped gold turns $+0.2$, 448 px -0.3, and 192 px -2.5.

How often a layout reaches the evidence. A dense segment spans a selected turn and two seconds on either side. An annotated evidence turn counts as covered when a dense segment overlaps it, and a layout's coverage is the share of a question's evidence turns it covers, averaged over the 602 questions whose evidence falls inside diarized turns. The voiceprint selector covers 10.8% of the evidence turns, Gemini-3.1-Pro ranking the turns for each question covers 46.1%, and the budgeted gold layout covers 59.8%; the uncapped gold layout covers all of them by construction. The voiceprint selector reaches at least one evidence turn in 130 of the 602 questions, and the budgeted gold layout in 416.

Where the gain comes from. On the full-parameter checkpoint, the questions whose evidence turns receive dense frames gain about 3 points under either placement, and the remaining questions gain less. With the voiceprint selector, the 130 questions it reaches gain $+3.1$ strict ($+2.7$ lenient), the other 472 questions with evidence turns $+0.9$ ($+0.6$), and the 524 questions without annotated evidence turns $+0.1$ (0.0). With the budgeted gold layout, the 416 reached questions gain $+3.0$ ($+2.9$), the other 186 $+1.1$ ($+1.1$), and the 524 without evidence turns $+1.9$ ($+1.7$). Each group is paired per question against the uniform reference and averaged over two seeds.

Table 16: Layout and resolution variants under the per-turn protocol on the full-parameter 2-epoch checkpoint: strict accuracy and paired deltas against the same checkpoint on the uniform layout, two seeds; lenient-parser deltas in the last column. *Read* rows score the uniform-trained checkpoint on the layout; *trained* rows train and score a checkpoint on it.

Layout	strict	Δ strict	Δ lenient
uniform 0.4 fps, reference	69.2	reference	
dense at voiceprint-selector turns, read	69.9	+0.8	+0.6
dense at voiceprint-selector turns, trained	70.3	+1.2	+1.0
dense at gold evidence turns, budgeted (coverage 59.8%), read	**71.4**	+2.2	+2.0
dense at gold evidence turns, uncapped (coverage 100%), read	69.8	+0.6	+0.4
selector-trained checkpoint, read at gold evidence turns	69.2	0.0	−0.1
mixed resolution, 448 px key turns and 192 px rest, read	69.9	+0.7	+0.7
mixed resolution, trained	70.1	+0.9	+0.8
every frame at 192 px, read	68.2	−1.0	−1.1
every frame at 448 px, read	70.3	+1.2	+1.1

M REPLICATION ON A DENSE 4B MODEL

Qwen3.5-4B has no audio encoder. We give it the frames and, for each turn, one <|SPK|> token whose embedding is a projected ECAPA voiceprint of that turn's three-second clip; the transcript is withheld, so the token is the model's only audio channel. Training uses the same 17,893-question split for two epochs, tuning the language model and the voiceprint projector with the vision encoder and aligner frozen (learning rate 10^{-5}, ZeRO-2, 20,480-token length cap). Evaluation is greedy with thinking enabled and a 16,384-token budget, on the same 1,126 questions. The forced-answer control in this replication truncates a chain at its first repeated 32-token window and asks for the answer; it is applied to every row, and the loops it removes begin after roughly 300 tokens.

Table 17: Full-parameter tuning of Qwen3.5-4B with a voiceprint token in place of audio. Deltas are paired per question against the uniform-frame row. *forced* scores every row after the forced-answer control.

Setting	strict	lenient	forced	trunc.	strict Δ	lenient Δ	forced Δ
frozen	44.4	61.2	67.0	43.2%	−20.8	−4.5	−0.4
full-parameter, uniform frames	65.2	65.7	67.3	3.6%		reference	
voiceprint token removed at test	65.5	66.1	67.5	3.6%	+0.3	+0.4	+0.2
dense frames at gold evidence turns	**66.0**	**66.9**	67.7	3.0%	+0.8	+1.2	+0.4
uniform frames in one timestamped block	65.6	66.7	**67.9**	3.6%	+0.4	+1.0	+0.5
dense frames at Gemini-selected turns	65.4	65.9	67.5	3.6%	+0.2	+0.2	+0.2
mixed resolution	65.2	66.2	66.7	3.6%	0.0	+0.4	−0.6
dense segments, base grid thinned	64.6	65.5	66.2	2.8%	−0.6	−0.3	−1.2
dense frames at voiceprint-clustering turns	64.3	65.5	66.3	4.1%	−0.9	−0.2	−1.1
dense segments only, base grid dropped	62.0	62.9	63.9	2.8%	−3.2	−2.8	−3.5

The pattern of §5.5 repeats. Tuning improves completion: the frozen model loses 43% of questions to the budget and the tuned model 3 to 4%, and once the forced-answer control removes the loops the two are within half a point (67.3 against 67.0). Every layout that keeps the base frame grid moves the score by about a point on any parser: gold-turn placement highest, Gemini selection in between, and the voiceprint-clustering selector lowest, and on Qwen3-Omni gold-turn placement is likewise above the voiceprint selector. Keeping only the dense segments and dropping the grid costs three points on every parser. The tuned model also does not use the voiceprint: removing the token at test time changes the score by a third of a point on every parser.

N AUDIO-ANCHORED SUPPLEMENT

Non-verbal audio cues make up 4.5% of the benchmark, 51 questions, too few to test whether the audio track carries identity information that a transcript cannot. We therefore built a 300-question supplement on the benchmark's own videos in which the referring cue and every identity edge live in the audio channel. The cue is a verified voice trait (252 questions; "the deep, slightly gravelly voice") or a timestamped vocal event (48; a laugh, a cough), the question contains no visual word, and the answer is a checkable fact: what the cued speaker said (speech content, 180 questions), which earlier voice a later voice belongs to (identity, 82), or a role the audio states (relation, 38). Generation reuses

the benchmark's pipeline with Gemini-3.1-Pro as the writer. Gemini builds the per-video inventories of voice traits and vocal events by listening to the audio, and the writer works from the transcript and these inventories; the generation-time checks of §3.2 apply unchanged. A text-side screen by GPT-5.6-sol, reading the transcript without speaker labels, then removes questions a strong reader answers from text alone, questions whose cue is positional or paraphrases transcript words, and questions whose options are not same-kind alternatives. A question is also removed when two GPT-5.5 text conditions, one reading the question and options alone and one reading the transcript without speaker labels, both answer it. Of 2,372 candidate questions 501 pass, and 300 are selected with one question per video, half at reasoning depth, and gold letters 63 / 82 / 75 / 80. Frontier diagnostics, not used for selection: Gemini-3.1-Pro native AV 51.3, audio only 53.0, 32 frames only 32.0, and GPT-Audio audio only 30.0. The two GPT-5.5 text conditions score 16.3 and 22.3; both sit below chance because the screens remove the questions that text answers. The supplement has not been human-verified. It is a diagnostic split, kept separate from the benchmark, and its absolute scores are not comparable with Table 3.

Table 18 repeats the input ladder of Table 6 on this split for the frozen Qwen3-Omni-30B-A3B-Thinking at temperature 1.0, the setting at which it answers within budget (Table 14); the tuned checkpoint has not been scored on it. Scores run from 34 to 40, nine to fifteen points above chance and eleven to seventeen below Gemini's native AV, so the split is hard for every system we have tried. Alone, the two speech channels add the same small amount over frames: $+2.2$ for the transcript and $+1.8$ for the track. Together they add 5.8, more than either alone: with the transcript present the 3 s track adds $+3.7$ overall and $+6.4$ on the speech-content questions, and the whole-turn track adds $+2.7$. On the benchmark the same model at the same temperature loses 0.9 when the 3 s track is added to frames+ASR (Table 15), and the tuned model of Table 6 loses 0.4. The sign does not depend on the decoding temperature: at the released 0.6 the track adds 3.5 points strict and 6.2 lenient. Where the questions are built to require the voice, the track carries content the transcript does not, and the frozen model uses some of it. With three hundred questions, differences of this size are close to the run-to-run variation.

Table 18: Input ladder on the 300-question audio-anchored supplement: frozen Qwen3-Omni-30B-A3B-Thinking under the per-turn protocol of Table 6, temperature 1.0, strict scoring, two seeds. Columns split the questions by dimension; *trunc.* is the share of questions whose reasoning hit the budget.

Input	Overall n=300	Speech content 180	Identity 82	Relation 38	trunc.
frames only	34.2	34.2	**36.0**	30.3	0.0%
ASR only	33.7	34.4	29.9	38.2	0.7%
frames + ASR	36.3	36.9	35.4	35.5	0.0%
frames + audio (whole turn)	36.0	38.9	32.3	30.3	0.3%
frames + ASR + audio (3 s)	**40.0**	**43.3**	**36.0**	32.9	0.0%
frames + ASR + audio (whole turn)	39.0	39.2	**36.0**	**44.7**	0.0%

O IDENTITY-LINK TEST ON FACE PAIRS

Every different-person pair of the first identity-link set, the 3,638 pairs of §5.4, compares two voices. To test faces, we built 2,013 face pairs, 1,085 same-person and 928 different-person, on 456 videos. InsightFace runs at one frame per second; a moment is a stretch of more than five seconds with exactly one face on screen, and the two moments of a pair are at least 30 seconds apart. A pair is labeled different when even its most similar frames have a face-embedding cosine below 0.10, and same when even its least similar frames have a cosine above 0.45, with at most three pairs of each kind per video. We checked 36 pairs by eye, 24 different-person and 12 same-person, and all were labeled correctly. The prompt and scoring are those of the voice pairs, and Qwen3-Omni-Thinking is averaged over two sampled runs. Qwen3-Omni-Instruct answers SAME for 96.9% of the voice pairs, 98.2% of the face pairs, and 97.9% of all 5,651 pairs it was asked about, the 3,638 of the first set and the 2,013 face pairs together.

Table 19: Identity-link test on face pairs and on the voice pairs of Table 7. Balanced accuracy has chance 50.0; *SAME share* is the fraction of all pairs of that kind answered SAME.

	Face pairs (2,013)				Voice pairs (2,195)			
System	SAME rec.	DIFF rec.	balanced	SAME share	SAME rec.	DIFF rec.	balanced	SAME share
Gemini-3.1-Pro	83.5	85.6	**84.5**	51.7	79.5	59.6	**69.5**	43.3
Qwen3-Omni-Thinking	72.2	40.2	56.2	66.4	64.3	47.4	55.9	53.4
Qwen3-Omni-Instruct	98.4	2.0	50.2	98.2	100.0	3.4	51.7	96.9

P IDENTITY-GROUNDED CAPTION TASK

Task. Each of the 678 videos is described once with explicit speaker attribution. The output has three labeled sections: *Speakers*, which assigns every transcript turn to a speaker label (S1, S2, … in order of first appearance); *Names*, one line per label whose name the video supports; and a *Narrative* in temporal order with speaker attribution, appearance, entrances and exits, and whether people seen at different times are the same person. Native audio-visual systems that do not read the transcript instead emit a timed dialogue (`[m:ss] Speaker X: words`), which is aligned to the reference turns by time window and word overlap, with a monotonic text alignment as fallback when timestamps are missing or malformed. Systems that read the transcript receive it unlabeled and indexed by line and attribute by index. The per-turn rows use the protocol of §4.3 with the Table 8 checkpoints and a 16k completion budget; chains that hit the budget yield no caption and count as such. Gemini receives the transcoded video with its track; GPT-5.5 32 frames with the indexed transcript; the open VLMs 64 uniformly sampled frames with the indexed transcript (Gemma-4 in chunks of 32 frames described locally and merged in a text-only pass: 58 s chunks for the 31B model, and 30 s chunks for the 12B model, which also hears each chunk's audio and whose audio encoder accepts at most 30 s); the open omni models their official video inference path with the whole video and its track; the open audio-only models the track in contiguous 30 s windows (their encoders' limit), each window transcribed as a dialogue with local speaker labels and stamped with its clock, since a single pass over the whole track made them summarise rather than transcribe. AV-Flamingo answers every instruction in its dense-caption format (a visual and an audio block every few seconds, with the speech quoted inside the audio block) and exhausts its budget within the first minute of a long video, so it is run in 58 s windows with a direct who-says-what prompt, the block clocks are shifted to the window, and the scorer reads the quoted speech out of the blocks with the speaking subject as the label. Qwen3-Omni-30B-A3B-Thinking is captioned on the path of its Table 3 row (the whole video with its track through vLLM, the released sampling config, a 32k budget); 43 of its 678 chains never close the reasoning block and yield no caption, and 90 more close it but then repeat dialogue lines until the budget, which the scorer's duplicate folding absorbs: capping every caption body at 3,072 tokens moves coverage, attribution, and pair F1 by at most 0.3 points. Qwen3-Omni-30B-A3B-Instruct is captioned through the same vLLM path with the same frames and track, decoding greedily. The frames-only conditions of Table 5 (Gemini-3.1-Pro and GPT-5.5 with 32 frames) and a one-frame control, reported in Table 20 only, are captioned from the same frames with no audio and no transcript and asked to transcribe only speech recoverable from the frames themselves; they write no dialogue for most videos (426 of 676 for Gemini, 536 of 678 for GPT-5.5), so their coverage is 1–2% and their attribution cells are parenthesized. All open-weight systems decode greedily except the Thinking row, which uses its released sampling config. Whole-video completion caps are 2–3k tokens. Nemotron's transcripts ran into a 3k cap, so it is run with 8k. Windowed rows cap each window at 320–1024 tokens.

Reference. The release's speaker-attributed transcript, aligned line by line to the unlabeled transcript, its supported names (35 videos), and its face timeline. The speaker labels come from the diarization of the selection stage, and for the 56 feature-film segments from their released diarization annotations.

Turn-level metrics. Coverage and pair F1 are computed per video and averaged over videos; attribution accuracy is micro-averaged over attributed turns. With N reference turns, $g(k)$ the reference speaker of turn k, C the set of turns the caption attributes, and $p(k)$ the predicted label: coverage $= |C|/N$; attribution accuracy $= \max_\pi \frac{1}{|C|} \sum_{k \in C} \mathbf{1}[\pi(p(k)) = g(k)]$ over one-to-one maps π from predicted labels to reference speakers; pairwise same-speaker F1 over all turn pairs in C, with a pair positive when both turns carry the same label; name recall over the 35 videos with supported names;

and word error rate of the transcribed lines for the timed systems only. Attribution and pair F1 say nothing about the turns a caption omits, and attribution alone rewards putting every turn under one label with the majority speaker's share, which pair F1 only partly penalizes; the tables therefore report coverage, the *collapsed* share (multi-speaker videos in which at least 90% of the attributed turns carry one label), and in Table 20 the mean absolute speaker-count error, next to them. Systems that read the indexed transcript attribute by line index, so their coverage is near 100 by construction; coverage is comparable only among systems that transcribe the track themselves. Rows with coverage under 5% are parenthesized. Table 20 also reports *reach*, the median over videos of the last clock a timed caption mentions as a share of the video length: given the whole video, Gemini and Inkling reach the end, the Thinking variant of Qwen3-Omni reaches 86 with its 32k budget, while Nemotron stops after the opening minutes (reach 56), which is what its low coverage measures; its input is the full video and its whole track, so the stop is the model's. Windowed rows reach the end by construction.

Caption-level metrics. *Answerability*: a fixed GPT-5.5 reader answers each of the 1,126 questions from the caption alone; a missing caption counts as wrong. *Identity-link test*: the 3,638 SAME/DIFFERENT moment pairs of the first identity-link set of §5.4, each moment described to the same reader by clock time, turn number, line index and the words spoken, answered from the caption alone; balanced accuracy over its 2,195 voice pairs, since every different-person pair of the set is a voice pair (Appendix R). Both follow the reference-QA design of description benchmarks such as VDC and AutoDQ, with the questions coming from the annotations rather than from any model's output.

Fact-level metrics. Reference facts are generated deterministically from the reference: the speaker count, each supported name, one utterance cluster per reference speaker (four sampled lines that the same person said, with times), and the on-screen interval of each tracked face. GPT-5.5 judges whether each fact is conveyed (recall), then extracts the caption's own identity claims, at most 25 per video (presence, who said what, names, same-or-different person, speaker count), and marks each SUPPORTED, CONTRADICTED or UNVERIFIABLE against the reference; precision is SUPPORTED over SUPPORTED+CONTRADICTED, and F1 their harmonic mean. The *contradictions* column of Table 20 comes from a separate judge pass that reads the whole caption against the reference and counts every checkable identity claim that contradicts it, with no cap on the count, so it is not bounded by the 25 extracted claims and grows with caption length; appearance and clothing are not checkable and are not counted.

Windowed rows. MiniCPM-o's official video path keeps at most 64 one-second frame-and-audio blocks per video, so on the whole video it hears 8–17% of the track and transcribes almost nothing; in 58 s windows it hears every second and transcribes 64% of the turns (pair F1 46.8). Given the whole video, four further open audio-visual models do not produce a transcript at all: Qwen2.5-Omni-7B writes a narrative for most videos and transcribes the opening seconds of the rest, video-SALMONN-2+ narrates the visuals without dialogue, and OmniVinci and VideoLLaMA2.1 write dialogue that is not the video's (word error rates 94 and 98, at most 5% of the turns aligned in every case). Their rows, like AV-Flamingo's, are therefore the windowed protocol: 58 s chunks from the same cache, a per-chunk dialogue prompt with local speaker labels, and clocks spread over the window; a label assigned in one window has no meaning in the next, which is what the pair F1 of every windowed row measures. Windows change the picture unevenly: OmniVinci then transcribes 43% of the turns, Qwen2.5-Omni 49% but opens a new speaker label on almost every line (speaker-count error 11.8, pair F1 26.8), video-SALMONN-2+ 19%, and VideoLLaMA2.1 still under 2%: its audio encoder is BEATs, an audio-event model that labels a 16 s clip heard in full as "female speech, woman speaking" when asked to transcribe it, so its windows describe sounds rather than words and its cells stay parenthesized. The audio-only models transcribe two thirds of the turns under the same treatment.

Table 20: Identity-grounded caption task, all systems. (a) Systems that write their own timed dialogue from the video or the audio track. (b) Systems given the indexed transcript, which attribute its lines by index, so their coverage is near 100 by construction. (c) Caption-level scores from the fixed GPT-5.5 reader and judges. *Captions*: videos with a caption, of 678; *reach*: median over videos of the last clock the caption mentions, as a share of the video length; *count err.*: mean absolute speaker-count error; *WER*: word error rate of the transcribed lines; *name R*: name recall on the 35 videos with supported names; *contradictions*: checkable identity contradictions per video; *voice id-link*: identity-link balanced accuracy on the 2,195 voice pairs; *facts R/P/F1*: fact recall, precision, and F1. Parenthesized cells are computed over fewer than 5% of the turns.

(a) Timed dialogue written by the system

System	captions	attribution	pair F1	coverage	reach	collapsed	count err.	WER	name R
Closed									
Gemini-3.1-Pro (video+audio)	678	**76.0**	**72.0**	**84.2**	99	13	1.4	**36.1**	**36**
GPT-Audio (audio only)	678	73.4	64.9	50.8	93	15	1.4	60.4	0
Gemini-3.1-Pro (32 frames only)	676	(70.5)	(44.2)	1.7	89	(69)	(2.7)	75.1	16
GPT-5.5 (32 frames only)	678	(69.5)	(40.4)	1.6	92	(57)	(2.6)	75.4	21
Open, video+audio input									
Qwen3-Omni-30B-A3B-Instruct	678	68.3	58.5	37.2	61	18	2.0	64.5	26
Qwen3-Omni-30B-A3B-Thinking	635	69.3	61.0	41.6	86	19	2.0	60.4	26
Nemotron-3-Nano-Omni-30B	678	63.0	57.1	29.7	56	48	3.2	61.1	7
MiniCPM-o-4.5 (58 s windows)	678	50.4	46.8	63.7	100	2	2.1	46.2	0
Inkling-Small	678	62.8	49.2	19.5	100	15	2.1	73.6	6
AV-Flamingo-9B (58 s windows)	678	57.9	55.6	27.9	67	42	2.3	74.4	0
OmniVinci-9B (58 s windows)	678	57.5	54.9	43.0	99	28	2.0	67.9	0
Qwen2.5-Omni-7B (58 s windows)	678	33.0	26.8	49.2	100	3	11.8	49.1	0
video-SALMONN-2+-7B (58 s windows)	678	59.5	53.1	18.8	97	19	1.9	59.2	0
VideoLLaMA2.1-7B-AV (58 s windows)	678	(68.4)	(30.2)	1.4	77	(71)	(3.0)	93.2	0
Open, audio only, transcribed in 30 s windows									
Audio-Flamingo-3	678	54.1	53.6	66.3	100	19	2.2	41.2	0
Qwen2-Audio-7B	678	48.6	46.5	71.5	100	8	3.2	45.9	0

(b) Indexed transcript given to the system

System	captions	attribution	pair F1	coverage	collapsed	count err.	name R
GPT-5.5 (32 frames + indexed transcript)	651	71.4	68.3	99.9	8	1.5	33
Gemma-4-12B (frames + audio + transcript)	678	48.2	45.7	98.3	10	4.0	10
Qwen3-VL-30B-A3B (frames + transcript)	678	74.6	58.4	97.9	94	3.7	3
Gemma-4-31B (frames + transcript)	678	60.8	57.9	98.9	15	2.7	24
GLM-4.5V (frames + transcript)	678	69.5	57.5	89.1	42	1.9	**37**
Gemma-4-12B (frames + transcript)	678	50.1	48.2	98.8	20	2.4	16
Per-turn protocol, Qwen3-Omni-Thinking (frames + ASR + audio clips)							
frozen, $T=1.0$	635	67.3	59.0	97.6	46	1.9	31
full-parameter 2 ep, 3 s audio	653	68.7	59.4	86.3	53	2.0	25
full-parameter 2 ep, whole-turn audio	649	68.4	59.5	81.5	47	2.0	20

(c) Caption-level scores

System	answerability	contradictions	voice id-link (bal.)	facts R	facts P	facts F1
Gemini-3.1-Pro (video+audio)	**82.7**	24.6	**78.1**	**55.7**	**72.1**	**62.8**
GPT-5.5 (32 frames + indexed transcript)	77.1	27.1	74.2	46.1	70.1	55.6
GPT-Audio (audio only)	72.6	25.1	73.9	21.2	70.4	32.6
Gemini-3.1-Pro (32 frames only)	73.8	2.2	62.9	25.0	81.4	38.3
GPT-5.5 (32 frames only)	73.6	2.6	65.0	28.4	77.7	41.6
GPT-5.5 (1 frame only)	56.4	0.4	60.2	9.6	65.4	16.8
Qwen3-Omni-30B-A3B-Thinking	72.9	65.2	61.6	36.2	64.8	46.5
Nemotron-3-Nano-Omni-30B	74.8	102.3	64.2	29.4	58.5	39.2
MiniCPM-o-4.5 (58 s windows)	71.4	42.6	64.3	8.6	57.2	14.9
Inkling-Small	67.5	14.6	66.4	7.2	60.9	12.9
AV-Flamingo-9B (58 s windows)	75.0	62.2	62.9	11.2	67.3	19.2
OmniVinci-9B (58 s windows)	68.7	50.3	65.8	8.4	60.1	14.7
Qwen2.5-Omni-7B (58 s windows)	70.5	112.5	68.1	2.6	43.8	4.8
video-SALMONN-2+-7B (58 s windows)	62.3	22.1	64.7	3.4	54.7	6.3
VideoLLaMA2.1-7B-AV (58 s windows)	61.4	8.6	63.6	8.2	57.3	14.4
Audio-Flamingo-3 (30 s windows)	69.4	106.4	61.6	9.9	57.2	16.9
Qwen2-Audio-7B (30 s windows)	68.4	38.2	66.1	6.8	52.6	12.1
Gemma-4-12B (frames + audio + transcript)	70.6	66.6	56.3	16.6	48.2	24.7
Qwen3-VL-30B-A3B (frames + transcript)	66.8	155.3	62.1	15.2	61.0	24.3
Gemma-4-31B (frames + transcript)	67.5	38.3	62.4	10.7	60.3	18.1
GLM-4.5V (frames + transcript)	69.9	62.5	60.8	16.2	68.2	26.2
Gemma-4-12B (frames + transcript)	70.7	65.1	56.9	19.9	49.5	28.4
Per-turn, frozen, $T=1.0$	67.0	52.9	60.1	15.1	66.7	24.7
Per-turn, full-parameter 2 ep, 3 s audio	63.4	51.3	59.6	13.7	68.3	22.8
Per-turn, full-parameter 2 ep, whole-turn audio	65.9	50.5	61.2	13.2	68.3	22.1

Q TEXT-SIDE STUDY WITH VOICEPRINT TOOLS

We ask whether letting a model decide which identity checks to run, a common response to the voice gap, helps when the benchmark does not preselect for it. The evidence is fixed: per-turn ECAPA voiceprints (Desplanques et al., 2020) with per-scene calibrated thresholds over the benchmark's own turn segmentation. What varies is who decides what to verify, and every condition is text-only. In the first three conditions the controller starts from the question and options and calls tools until it answers. **Transcript** offers the unlabeled transcript. **Speaker labels** also offers the transcript with automatic pyannote (Bredin et al., 2020) speaker labels. **Agent** offers the numbered transcript and `same_speaker`(i, j), which returns the voiceprint verdict for any two turns the controller names, up to 12 calls. The other two conditions are single prompts without tools. **All verdicts** gives the numbered transcript with every pairwise verdict, as the full table for windows up to 40 turns and as the resulting speaker clusters for longer windows, so it holds everything the agent could ask for. **Router** gives the numbered transcript with the verdicts for the pairs a fixed rule selects, the turns that share the most words with the question. Tool loops cap each completion at 4,096 tokens and single prompts at 8,192. Table 21 reports three controllers.

Table 21: Text-side study, overall accuracy. Every condition is text-only: audio and video reach the controller only as transcripts and voiceprint SAME/DIFFERENT verdicts computed offline, and agent, all verdicts, and router draw on the same verdicts. *Agent − best fixed* is the agent minus the better of all verdicts and router. The last column is each controller's own result from Table 3, frames+ASR for GPT-5.5 (†) and video+audio for Gemini and Qwen, shown for reference; it is a different input condition. ‡Adding the face-timeline text to GPT-5.5's all-verdicts prompt lowers it to 73.4. ¶The best of five pre-registered agent prompts raises GPT-5.5's agent score to 70.4.

Controller	transcript	+speaker labels	agent	all verdicts	router	agent − best fixed	*Table 3*
GPT-5.5	62.0	58.2	67.8¶	**75.3**‡	73.3	−7.5	*84.2*†
Gemini-3.1-Pro	74.4	76.4	74.2	75.3	**76.1**	−1.9	*86.9*
Qwen3-Omni-Instruct	54.2	55.7	48.8	52.8	**53.2**	−4.4	*74.1*

Adaptive tool use does not outperform fixed access to the same voiceprint evidence. The agent never has the highest score. For every controller it is two to eight points below the better fixed use of the same verdicts. All verdicts gives the controller every verdict the agent could request, so querying on demand can only recover a subset, and choosing the right subset requires knowing which turns matter, which is most of the question. The Qwen3-Omni controller shows the extreme case: it called a tool on only 1 of the 1,126 questions, so its agent row answers from the question and options alone and scores 48.8, level with its text-only baseline of 49.0.

For GPT-5.5 we also tried to improve the agent before running it on the full set. A pre-registered sweep over five agent prompts, all fixed before any result was seen and run on 50 questions per scene family, 300 in total, compared the original prompt at 67.7 macro accuracy with strategy guidance at 72.7, a think-then-act loop at 70.0, guidance with a doubled 24-call budget at 71.7, and plan-then-verify with 24 calls at 65.7. The best variant, strategy guidance, scores 70.4 overall on the full set, below the 75.3 of all verdicts.

Derived text can also hurt. GPT-5.5 scores 58.2 with pyannote speaker labels against 62.0 with the plain transcript, and adding face-timeline text to its all-verdicts prompt lowers it from 75.3 to 73.4, because the controller trusts incorrect who-is-who text. Every controller's best text-side condition is 9 to 18 points below its own result with the video, so transcript and voiceprint tools add to looking and listening rather than replacing them.

R LIMITATIONS

Residual guessability. A text-only reader given the question and its options scores 58.9 (GPT-5.5) and 59.6 (Gemini-3.1-Pro) against a chance level of 25.0, and the open models of Table 3 score 40.2 to 52.0 under the same prompt, and the two open audio-only models 42.6 and 36.2. The options alone carry a shared signal: GPT-5.5 scores 40.9 from the options without the question, the 16 open models 28.8 to 40.6, a majority vote of the 16 open models 41.7, and 167 questions, or 14.8%, are answered correctly from the options by at least 13 of these 17 readers. A longest-option rule scores 35.3. Perceptual-cue anchoring closes the transcript route to the referent, but four written options still carry plausibility signal, so absolute accuracies should be read against these measured text-only baselines rather than against chance. The text-hard subset is defined by the questions one text-only reader, GPT-5.5, answers incorrectly in a single sampled run, which makes it a conditional diagnostic rather than a model-independent split.

What human review certifies. Reviewers judged every question on the three criteria of §3.3 with the keyed answer in view. This certifies the questions; it does not measure human performance. A row in which people answer the questions themselves, with and without audio, is left for future work. Required-evidence labels are assigned before review and retained through it, and they are validated in aggregate by modality ablation (§5.1) rather than certified question by question.

Identity-link test. Every different-person pair of the first identity-link set of §5.4 compares two voices, whereas 1,443 of its 1,604 same-person pairs involve a face or an on-screen moment. The kinds of the two moments therefore predict the label: a rule that answers DIFFERENT exactly when both moments are voices, without the video, reaches 95.0% balanced accuracy over all 3,638 pairs. Table 7 therefore reports voice pairs and face pairs separately, §5.5 and the caption-level test of Appendix P report voice pairs, and neither kind of pair carries this shortcut. Different-person pairs are checked by voiceprint, and same-person pairs come from the identity links annotated with each question.

Clue-restricted inputs. The clue-restricted condition of Table 7 keeps the annotated evidence spans with a 15 s margin. The annotation covers the answer's evidence, not every moment the identity link passes through, so for 519 of the 1,126 questions the clip leaves out such a moment, most often the cue. A clip is also much shorter than its video and is therefore sampled more densely. The difference between clue and full input mixes finding the evidence with both changes.

Generator and evaluated systems. Every question comes from one generator, GPT-5.6-sol. One evaluated system, GPT-5.5, belongs to the same family as the generator, although it does not top the table.

Scope of the diagnostics. The per-turn protocol and the training runs cover one open model family, one training split, supervised fine-tuning, and one reinforcement-learning recipe with an answer-only reward. Whether other training recipes close the voice-to-person gap is the comparison the benchmark is built to support.

Speech-centric audio. Every audio cue in the benchmark is a property of a speaker's voice or a non-verbal vocalization, and no question turns on a non-speech sound source. That is one reason the transcript substitutes so well for the track. Sound-event questions are out of scope for this release.

S DISCLOSURE OF AI USE

Large language models are part of the data construction, and these uses are described with the method. GPT-5.6-sol writes the candidate questions of the benchmark, and human reviewers check every released question against its video (Sections 3.2 and 3.3). LLMs also write and filter the synthetic training split (Appendix K) and the audio-anchored supplement (Appendix N), and GPT-5.5 scores the caption task (Appendix P). In addition, we used LLM assistants to help write and debug experiment code and to edit the text for readability. The authors checked all AI-assisted code and text, and take full responsibility for the content of this paper.