GEEZ: GROUNDED, EFFICIENT, AND EFFECTIVE VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing video benchmarks can contain unanswerable or subjective questions, or questions solvable without the intended audio-visual evidence. We present GEEK, a benchmark of 1,000 multiple-choice questions on web videos lasting 15 s to 5 min: 800 answerable from the video alone and 200 linking video evidence to verified external knowledge. We use a multimodal large language model to build EntityScript, a timestamped record of entities, events, and audio cues. An optimizer model iteratively refines prompts for generating questions from EntityScript, while web-search agents construct questions involving external knowledge. Candidate screening combines adversarial filtering with manual review. We also present GEEZ, a video question-answering framework that augments EntityScript with speech transcripts and on-screen text extracted offline. A planner prioritizes the transcripts for spoken wording and the extracted text for on-screen content, while querying an audio-visual recognizer for visual details. With Qwen3-Omni-30B as the recognizer, GEEZ achieves 67.4% accuracy on GEEK. Compared with the EntityScript-based agent without these additional text sources (ES-AGENT), it improves accuracy by 8.5 percentage points while using 24% fewer query-time recognizer calls. It also outperforms a planner using only EntityScript, speech transcripts, and on-screen text by 17.9 points. Across recognizers, gains over ES-AGENT range from 4.1 to 5.1 percentage points on OmniVideoBench and Video-MME-v2. We will release the benchmark, construction records, video annotations, and framework code.

1 INTRODUCTION

Video benchmarks assess the audio-visual understanding of multimodal large language models (MLLMs) (Fu et al., 2025; Li et al., 2026). Reliable evaluation, however, faces two risks. First, questions intended to test joint audio-visual reasoning may be answerable from visual input or audio alone. Second, automatically generated questions may be unanswerable or subjective, or have incorrect reference answers. These questions have dimished value in evaluating perception capabilities of the MLLMs. Benchmark construction should therefore assess evidence dependence and answer validity, rather than difficulty alone.

To address these risks, we combine agent-based question generation with adversarial filtering and manual review. The pipeline uses EntityScript, a structured video representation that records entities, events, and audio cues with timestamps. An MLLM processes segments obtained with PySceneDetect (Castellano, 2025) in chronological order. It updates an entity registry with stable, programmatically assigned identifiers and links events to segment intervals. Related video systems also use entity-centered representations (Cai et al., 2026; Tencent Hunyuan Team, 2026; Tian et al., 2026). Built independently of any question, EntityScript supports inspection and evidence tracing to the source video.

GEEK: agent-guided benchmark construction. We build GEEK (Grounded Evaluation with Entity Knowledge), 1,000 multiple-choice questions on 15 s–5 min web videos: 800 answerable from the video alone and 200 linking the video to verified external media. QAGP-ES generates them across 13 reasoning types with an optimizer that refines generator prompts against a fixed validation set, while QAGP-A^2 uses a tutor agent to guide web-browsing generators toward externally grounded questions; adversarial filtering and manual review screen candidates (§3.2, §3.4).

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.

GEEZ: answering with explicit evidence-source priorities. The base answering framework, ES-AGENT, has a planner read EntityScript and query an audio-visual recognizer, but paraphrased speech and omitted on-screen text lose needed wording, numbers, and names. GEEZ adds offline timestamped ASR transcripts (ByteDance Seed, 2026) and OCR text (Bai et al., 2025), exposed to the planner as local `transcript` and `read_text` tools that need no model calls.

With Qwen3-Omni-30B as the recognizer, GEEZ achieves 67.4% accuracy on GEEK. Compared with ES-AGENT, it improves accuracy by 8.5 percentage points while using 24% fewer query-time recognizer calls. It also outperforms a planner using only EntityScript, ASR transcripts, and OCR output by 17.9 points. Across recognizers, gains over ES-AGENT range from 4.1 to 5.1 percentage points on OmniVideoBench (Li et al., 2026) and Video-MME-v2 (Video-MME Team, 2026).

Contributions. First, we introduce GEEK, a 1,000-question video benchmark with adversarial filtering, manual review, and per-item screening records. Second, we develop QAGP-ES for feedback-driven prompt refinement and QAGP-A^2 for tutor-guided generation of externally grounded questions. Third, we present GEEZ, which integrates offline ASR/OCR evidence, source-priority rules, and local retrieval to improve accuracy with fewer query-time recognizer calls. We will release the benchmark, construction records, video annotations, and code for question generation and answering.

2 RELATED WORK

Video understanding benchmarks. Video benchmarks probe complementary temporal and multimodal skills: MVBench covers short-clip temporal understanding (Li et al., 2024); Video-MME evaluates diverse domains and durations with audio and subtitles (Fu et al., 2025); LongVideoBench, EgoSchema, and HourVideo study extended temporal context (Wu et al., 2024; Mangalam et al., 2023; Chandrasegaran et al., 2024); Video-MME-v2 tests visual aggregation, temporal modeling, and answer consistency through grouped questions (Video-MME Team, 2026); OmniVideoBench targets joint audio-visual reasoning with manually annotated evidence chains (Li et al., 2026); and KnowIT VQA requires knowledge of a television series beyond the current clip (Garcia et al., 2020).

Benchmark construction and quality control. Adversarial filtering (HellaSwag, AFLite) and text-only screening (MMStar, MMMU-Pro) are established defenses against exploitable dataset biases and language-only shortcuts, and OmniVideoBench combines single-modality and text-only filtering with manual refinement (Le Bras et al., 2020; Zellers et al., 2019; Chen et al., 2024; Yue et al., 2024; Li et al., 2026).

Structured video representations and question generation. Structured representations link video content to entities, events, and time: AGQA builds compositional questions from spatio-temporal scene graphs (Grunde-McLaughlin et al., 2021); OmniVideo-100K uses entity-anchored scripts with timestamped audio-visual descriptions (Cai et al., 2026); Script-a-Video links caption streams through entity and temporal references (Tencent Hunyuan Team, 2026); and prompt-optimization methods such as OPRO refine prompts from evaluation feedback (Yang et al., 2024).

Agentic video understanding. Agentic frameworks gather evidence over multiple steps: VideoAgent and TraveLER plan iteratively (Wang et al., 2024; Shang et al., 2024); LVAgent and AOP-Agent coordinate multiple agents (Chen et al., 2025; Xu et al., 2026); video-tree, DRVideo, and memory-based systems organize hierarchical, evolving, or persistent representations (Wang et al., 2025; Ma et al., 2025; Fan et al., 2024); EGAgent and CAM use entity-based graphs (Rege et al., 2026; Tian et al., 2026); Video-RAG retrieves speech transcripts, on-screen text, and object information (Luo et al., 2024); and VideoAgent2 weighs tool evidence by uncertainty (Zhi et al., 2025).

3 THE GEEK BENCHMARK

GEEK contains 1,000 multiple-choice questions: 800 GEEK-ES questions generated from EntityScript and answerable from the video alone, and 200 GEEK-A^2 questions that connect video evidence to verified external images or facts.

3.1 CORPUS AND GROUNDED SUBSTRATE

We collect web videos covering news, vlogs, tutorials, shopping, sports, and music, retaining their original audio. For each video, we generate EntityScript with Gemini 3.1 Pro Preview. The script contains an entity registry with names, appearances, and visual descriptors, together with timestamped events. Each event belongs to one of four types: actions, quoted speech turns, sounds, and interactions. A segment may also include a caption describing musical events. Entity references and events can be traced to their source segments, allowing evidence cited in generated questions to be checked against the video.

3.2 QAGP-ES: OPTIMIZER-EVOLVED GENERATORS

Generation step. Drawing inspiration from existing benchmarks (Li et al., 2026; Hu et al., 2026), we curate 13 reasoning types, referred to as nodes. For each node, a prompt refiner M_{pr} produces a generation prompt P_1^i, where i indexes refinement iterations. A question generator M_q reads the prompt and the rendered EntityScript to produce an open-ended or multiple-choice question, an answer, and supporting evidence:

$$M_q\big(v_j, \text{EntityScript}(v_j), P_1^i\big) \to (Q_j, A_j, E_j), \tag{1}$$

where v_j is the j-th video, Q_j includes the options for a multiple-choice question, A_j is the authored answer, and E_j contains supporting EntityScript entries and their timestamps. The benchmark retains only multiple-choice questions.

QA filtering. We use three MLLMs to assess candidate questions. A baseline model M_l receives both audio and video to measure question difficulty. A modality-restricted model M_m answers with audio removed, video removed, or both removed. The last condition provides only the question and options. Success under these conditions identifies questions that this model can answer without the omitted evidence; it does not by itself establish answer leakage. A strong reference model M_h provides an additional check of answer validity and response consistency.

Each model-input condition uses up to three attempts, with independently shuffled option orders for multiple-choice questions. A question passes a condition only when a strict majority of completed attempts agree with the authored answer (§3.4). For open-ended responses, a judge model M_j assesses agreement with that answer.

Prompt evolution. The refiner M_{pr} uses `view`, `revise`, `revert`, and `run_test` tools to inspect and edit each node's prompt. We first select a fixed validation set sampled uniformly across video-duration buckets. At iteration i, we generate questions with P_1^i, evaluate them on this set, and return the quality metrics defined in §3.4 to M_{pr} to guide P_1^{i+1}. The refiner can then retain, revise, or revert the prompt. Refinement ends when M_{pr} judges further edits unlikely to improve the reported metrics. Testing uses a separate set of videos sampled across the same duration buckets; individual test cases are not exposed to M_{pr}. For the five reasoning types with detailed validation results in Appendix A.2, refinement takes 12–17 iterations.

3.3 QAGP-A^2: TUTOR-GUIDED WEB AGENTS

Unlike GEEK-ES, GEEK-A^2 connects an entity or event in the video to external images or facts. Generating these questions requires more than retrieving relevant information: the external evidence must support the answer, and the question must remain grounded in the video. Otherwise, a generator may invent an unsupported connection or produce a generic lookup question that does not require watching the video.

We therefore use a tutor agent M_t to supervise question-generating agents. The tutor examines records $\mathcal{T}$ of trial generator sessions and converts observed failures into generation and verification instructions P_2. Using M_q to denote the web-browsing generator role, the two stages are

$$\begin{gathered} M_t(\mathcal{T}) \to P_2, \\ M_q\big(v_j, \text{EntityScript}(v_j), P_2\big) \to (Q_j, A_j, E_j) \text{ or } \varnothing. \end{gathered} \tag{2}$$

Here, E_j includes timestamped video evidence and external source citations, and $\varnothing$ denotes abstention. The tutor thus supervises how generators construct and verify questions, rather than only checking their final answers.

The instructions specify four question templates: matching an on-screen item to retrieved images, linking a specific video moment to an external fact, verifying an attribute or price of an on-screen product or venue, and combining a video fact with external knowledge. They also specify how to combine two verified reasoning steps into one question. Each generator uses a separate browsing session with image search, web search, page inspection, and screenshot tools, producing at most one candidate. An external claim must be confirmed by two independent sources; otherwise, the generator abstains. Time-sensitive claims, such as prices, cite dated source snapshots. The full process, including verification and shared candidate screening, yields 200 questions from 648 sessions (31%). Appendix B gives the instructions and verification protocol. Web access is used for construction, but is not available to the answering systems evaluated in §5.

3.4 Metrics and Feedback Loop

Question-level scoring. For a candidate question Q and one of the five model-input conditions above, indexed by c, let $n_c(Q) \in \{1, 2, 3\}$ be the number of completed attempts and $k_c(Q)$ the number that agree with the authored answer. We define the pass indicator and aggregate pass rate as

$$b_c(Q) = \mathbf{1}\left[k_c(Q) > \frac{n_c(Q)}{2}\right], \qquad r_c = \frac{1}{|\mathcal{C}_c|} \sum_{Q \in \mathcal{C}_c} b_c(Q), \tag{3}$$

where $\mathbf{1}[\cdot]$ is an indicator and $\mathcal{C}_c$ contains the candidates with completed responses in condition c. Missing conditions are not counted as failures. The reported screening rates therefore measure question-level majority success, not single-attempt accuracy.

Feedback to generation. On the fixed validation set, we track the number of generated candidates, the pass rates of M_l and M_m, the agreement and consistency checks from M_h, and lexical diversity. These measurements jointly assess generation yield, difficulty, potential answer shortcuts, answer validity, and diversity. In QAGP-ES, M_{pr} uses these aggregate metrics to refine P_1^i. In QAGP-A^2, feedback instead comes from generator session records, which M_t uses to formulate the verification and abstention instructions in P_2. Both pipelines apply the same model-based screening protocol to the resulting candidates.

Candidate selection. For GEEK-ES, we pool candidates across generation runs and remove malformed questions, questions outside the 4–10 option range, and questions missing any required screening condition or the reference-model consistency check. We exclude questions that M_m passes with text-only input. Candidates for which M_h consistently disagrees with the authored answer while M_l also fails are withheld from the benchmark for manual review. Manual quality review also precedes final sampling. To limit redundancy, we remove near-duplicates with trigram Jaccard similarity at least 0.5 and retain at most three questions per node and video. We then sample approximately evenly across the 13 nodes using a fixed seed. Each selected question retains the option order used in the first screening attempt, together with its per-condition pass decisions and completed-attempt counts. Appendix A.5 specifies the screening protocol, and Appendix A.3 gives the final composition.

Benchmark statistics. Table 1 summarizes the selected questions. Within GEEK-ES, the 303 questions on which M_m fails both the no-audio and no-video conditions form the GEEK-AV core, covering all 13 reasoning types and 210 videos. Membership is defined by the recorded screening outcomes; the reference answers are additionally audited by human annotators (Appendix A.5).

We would like to clarify that the AV core subset is not an absolute proof that both modalities are logically necessary since we only consider the rollout from one model. Likewise, the 0% text-only pass rate follows from the selection rule and does not imply that other text-only models cannot answer these questions. Because the screening models participate in selection, their pass rates are not independent benchmark evaluations. Comparisons between the full-media baseline and modality-restricted model also change the model, so their differences do not isolate modality effects. Evaluation results and model overlap are discussed in §5.

Table 1: GEEK composition and construction-stage screening results. Except for the first row, statistics describe the 800-question GEEK-ES subset. Screening rates are strict-majority pass rates, not single-attempt answer accuracies. Results for GEEK-A^2 appear in Appendix B; the screening protocol and human answer review are described in Appendix A.5.

Composition		**Screening pass rates (%)**	
Questions (GEEK-ES / GEEK-A^2)	800 / 200	Audio + video (M_l)	61.9
Videos	408	No audio (M_m)	47.5
Reasoning types	13	No video (M_m)	32.3
AV core (questions / videos)	303 / 210	Text only (M_m)	0.0
Duration range (median)	17.6–299.9 s (117 s)	Audio + video (M_h)	74.0
Median stem length	68 words	AV core: M_l	41.3
Options per question	4–9 (mean 5.0)	AV core: M_h	57.4
English / Chinese / other	462 / 245 / 93		

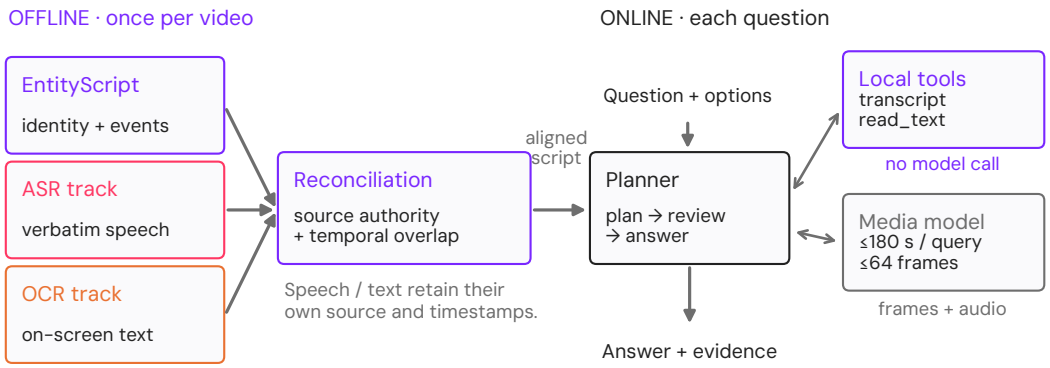

Figure 1: The GEEZ VQA harness. Offline, speech transcripts and on-screen text are extracted and aligned with EntityScript. At question time, the planner consults the augmented script and retrieves text for relevant intervals. It can also query an audio-visual recognizer for further evidence before selecting an answer.

4 GEEZ: A GROUNDED, EFFICIENT AND EFFECTIVE VQA HARNESS

GEEZ is a video question-answering (VQA) harness that extends ES-AGENT with speech transcripts and on-screen text extracted offline (Figure 1). These sources are aligned with EntityScript and reused across questions. The planner prioritizes them for spoken wording and displayed text, while querying an audio-visual recognizer for further evidence.

4.1 BASE HARNESS: ES-AGENT

ES-AGENT uses EntityScript as its initial video description. A language-model planner receives the question, answer options, and script, but does not process audio or video directly. Instead, it asks an audio-visual recognizer to inspect selected intervals. The recognizer is a multimodal model that answers a focused question from sampled frames and the corresponding audio, with supporting timestamps. The planner reviews these responses and may request further evidence before selecting an answer with cited evidence.

4.2 EXTRACTING SPEECH AND ON-SCREEN TEXT

To capture details that may be omitted or paraphrased in EntityScript, we extract speech and on-screen text separately.

Speech transcripts. We apply automatic speech recognition (ASR) to each EntityScript segment. The model is instructed to transcribe speech verbatim, with timestamps and speaker labels. To handle speech that spans segment boundaries, we provide the end of the preceding transcript as context and merge continuing utterances afterward. All timestamps are mapped to the full video.

On-screen text. We apply optical character recognition (OCR) to frames already sampled for EntityScript construction. Before extraction, we remove near-duplicate consecutive frames to avoid reading static content repeatedly. We then merge similar readings across successive sampled frames and record their time intervals. The result is a timestamped record of subtitles, captions, and other on-screen text.

4.3 Aligning the Text with EntityScript

We align transcript utterances with EntityScript speech events using timestamps and text similarity. When an utterance matches an existing event, we replace the event's quoted speech with the transcribed wording while retaining its entity attribution. Unmatched utterances are added as speech events under their transcript speaker labels. We attach extracted on-screen text to overlapping segments without changing the visual descriptions. The planner receives the augmented script, with transcripts and on-screen text listed separately for each segment.

4.4 Using the Evidence to Answer Questions

The planner can retrieve transcripts or on-screen text for a selected interval. The planner can also request further evidence from the audio-visual recognizer using the same procedure as ES-Agent.

When sources disagree, the planner is instructed to prioritize transcripts for spoken wording and utterance timing, and extracted on-screen text for displayed content. For example, a spoken number is checked against the transcript, while a displayed number is checked against the extracted text. For visual details, the recognizer's response takes precedence over the script's description. These priorities guide the planner's use of evidence, although errors in transcription or text extraction can still affect the answer.

5 Experiments

5.1 Setup

Evaluated models. We evaluate five audio-visual recognizers, which answer queries from audio and video frames: Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), Seed 2.0 Lite (ByteDance Seed, 2026), Gemini 3.6 Flash (Google DeepMind, 2026b), Gemini 3.1 Pro Preview (Google DeepMind, 2026a), and GPT-5.6 Sol (OpenAI, 2026). All agentic methods use GPT-5.6 Sol with high reasoning effort as the planner. The planner selects evidence queries and produces the final answer without directly accessing audio or frames.

Construction and extraction models. For the benchmark-construction roles defined in §3, Kimi-K3 serves as the prompt refiner M_{pr} and question generator M_q in QAGP-ES, and as the tutor M_t in QAGP-A^2. DeepSeek V4.1 Flash serves as the web-browsing generator M_q in QAGP-A^2 and the judge M_j for open-ended answer agreement. Thus, M_q denotes the generator role within each pipeline, not a shared model. Screening uses Qwen3-Omni-30B-A3B-Instruct as the full-media baseline M_l, MiMo-V2.5 as the modality-restricted model M_m, and Gemini 3.1 Pro Preview as the strong reference model M_h. MiMo-V2.5 is used in all three restricted-input conditions: no audio, no video, and question-and-options only. Gemini 3.1 Pro Preview also produces EntityScript; Seed 2.0 Lite produces the automatic speech recognition (ASR) transcripts, and Qwen2.5-VL-7B-Instruct extracts on-screen text by optical character recognition (OCR). Qwen3-Omni-30B-A3B-Instruct and Gemini 3.1 Pro Preview are therefore used in both screening and evaluation. GPT-5.6 Sol is not used for question generation or screening, although it is the shared planner in all agentic methods.

Methods. *No-Agent* answers in a single recognizer call using 64 uniformly sampled frames and the full audio on Geek; on the transfer benchmarks, its audio input is limited to the first 600 s. ES-Agent uses EntityScript to guide recognizer queries (§4.1). Geez adds the offline ASR/OCR tracks and local retrieval tools of §4.2–4.4. Its source-priority rules give speech transcripts priority for spoken wording, OCR text for on-screen content, and recognizer responses for visual details. On the transfer benchmarks, we also include *ShotScript*, which gives the same planner flat timestamped shot captions in place of the entity registry.

Table 2: Accuracy (%) on the 1,000-question GEEK benchmark and mean query-time recognizer calls per question. All agentic methods use the same GPT-5.6 Sol planner; GEEZ adds offline ASR/OCR tracks, source-priority rules, and local retrieval tools.

	Accuracy (%)			Recognizer calls	
Recognizer	No-Agent	ES-AGENT	GEEZ	ES-AGENT	GEEZ
Seed 2.0 Lite	49.3	61.8	69.6	5.9	4.4
Qwen3-Omni-30B-A3B-Instruct	44.6	58.9	67.4	5.4	4.1
Gemini 3.6 Flash	52.7	64.3	71.9	6.2	4.8
Gemini 3.1 Pro Preview	56.2	67.1	74.8	6.6	5.1
GPT-5.6 Sol	55.4	66.2	73.6	6.0	4.6

Table 3: ASR/OCR track ablations and controls on GEEK with Qwen3-Omni-30B-A3B-Instruct as the recognizer where used. Calls, queries, and frames are query-time means per question. Local queries retrieve stored text without invoking a model.

Method	Acc. (%)	Recognizer calls	Local queries	Frames
No-Agent + tracks	53.8	1	–	64
Planner-only + tracks	49.5	0	4.1	0
ES-AGENT	58.9	5.4	–	148
+ ASR track	64.7	4.6	2.4	127
+ OCR track	61.8	5.0	1.9	136
GEEZ (both)	67.4	4.1	3.2	113

Evaluation protocol. We evaluate on the 1,000-question GEEK benchmark described in §3: 800 GEEK-ES questions and 200 GEEK-A^2 questions. All answering evaluations use the released reference answers. The evaluated systems have no web access, including when answering GEEK-A^2 questions. Accuracy is the fraction of predicted option letters that match the reference answers. This differs from the construction-stage screening pass rates, which use a strict majority of up to three completed attempts per question and model-input condition (§3.4). We report mean query-time recognizer calls and frames processed per question; local queries retrieve stored text without additional model calls and are counted separately. For paired-bootstrap confidence intervals, we resample videos with replacement and compare methods on the same sampled videos (Appendix F).

5.2 MAIN RESULTS ON GEEK

Table 2 shows accuracy gains across all five recognizers: No-Agent ranges from 44.6% to 56.2%, ES-AGENT improves by 10.8–14.3 points, and GEEZ adds a further 7.4–8.5 points while using 23–25% fewer query-time recognizer calls. With Qwen3-Omni-30B-A3B-Instruct, GEEZ reaches 67.4% versus 44.6% (No-Agent) and 58.9% (ES-AGENT), a gain over ES-AGENT with paired-bootstrap 95% confidence interval [5.6, 11.3] (Appendix F); the same three methods score 31.7%, 47.5%, and 58.1% on the 303-question GEEK-AV core. Appendix F also reports a No-Agent control using video input directly rather than 64 sampled frames: Gemini 3.1 Pro Preview reaches 58.8%, versus 74.8% for GEEZ.

Using only questions and options, without media or EntityScript, GPT-5.6 Sol, Kimi-K3, and Llama-4-Maverick (Meta AI, 2025) score 21.9%, 23.6%, and 22.8% on the 800-question GEEK-ES subset—close to the 20.7% uniform-choice chance rate and the 22.5% always-A baseline. These checks suggest limited answer predictability without video-derived evidence but do not establish the absence of leakage.

5.3 ABLATING THE ASR/OCR TRACKS

Table 3 evaluates the ASR and OCR tracks with Qwen3-Omni-30B-A3B-Instruct. Adding ASR or OCR alone improves accuracy by 5.8 and 2.9 points over ES-AGENT, and using both yields an 8.5-point gain while reducing query-time frames from 148 to 113 (24%). Two controls assess the complete system: No-Agent + tracks supplies both tracks in a single recognizer call and reaches 53.8%, while Planner-only + tracks uses the GPT-5.6 Sol planner with EntityScript and both tracks, but no recognizer calls, and reaches 49.5% overall and 31.4% on the AV core. GEEZ outperforms these controls by 13.6 and 17.9 points overall, respectively. These comparisons support combining

Table 4: Accuracy (%) on the full test sets of OmniVideoBench and Video-MME-v2. Parentheses give mean query-time recognizer calls per question for agentic methods, all of which use the GPT-5.6 Sol planner. Appendix E specifies the evaluation protocol, including the No-Agent audio limit.

Recognizer	No-Agent	ShotScript	ES-AGENT	GEEZ
OmniVideoBench				
Seed 2.0 Lite	42.7	55.9 (6.4)	56.4 (5.9)	61.2 (4.5)
Qwen3-Omni-30B-A3B-Instruct	39.2	54.6 (5.7)	54.5 (5.2)	59.6 (4.0)
Gemini 3.6 Flash	51.8	61.0 (7.1)	62.3 (6.9)	66.7 (5.3)
Gemini 3.1 Pro Preview	57.0	65.8 (8.2)	66.2 (8.0)	70.4 (6.3)
Video-MME-v2				
Seed 2.0 Lite	40.3	58.7 (6.8)	60.9 (6.2)	65.6 (4.8)
Qwen3-Omni-30B-A3B-Instruct	34.4	59.8 (6.2)	61.5 (6.1)	66.3 (4.7)
Gemini 3.6 Flash	49.5	63.2 (7.3)	64.8 (7.0)	69.0 (5.5)
Gemini 3.1 Pro Preview	62.4	68.9 (8.4)	70.1 (8.1)	74.2 (6.4)

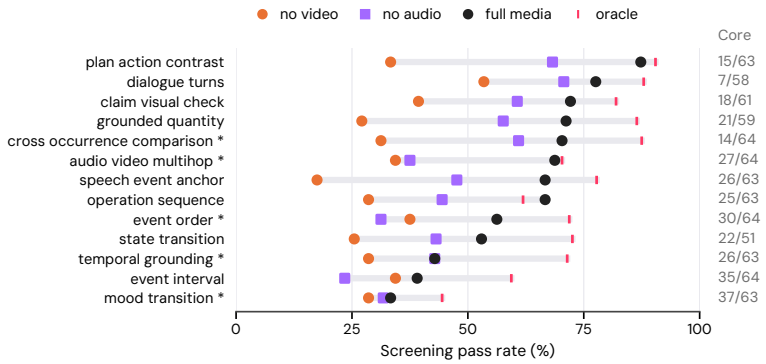

Figure 2: Screening pass rates by reasoning type on the 800-question GEEK-ES subset (M_l: full-media; M_m: single-modality; M_h: reference). Stars mark the five types with prompt-refinement results in Appendix A.2; the right column gives AV-core/total counts.

offline text evidence with iterative recognizer queries, but do not isolate the effect of source-priority rules.

Within GEEK-ES, the largest gains over ES-AGENT are on speech–event anchoring (+14.3 points; a net increase of 9 correct answers among 63 questions), dialogue turns (+12.1; 7 among 58), and grounded quantities (+11.9; 7 among 59). Mood transition improves by only 1.6 points (a net increase of 1 correct answer among 63 questions), with a paired-bootstrap 95% confidence interval of $[-4.9, 8.1]$ points that includes zero (Appendix F).

5.4 TRANSFER TO EXTERNAL BENCHMARKS

We apply the same extraction and answering pipeline to the full test sets of OmniVideoBench and Video-MME-v2 (Appendix E). Across the four recognizers in Table 4, GEEZ improves accuracy over ES-AGENT by 4.2–5.1 percentage points on OmniVideoBench and 4.1–4.8 points on Video-MME-v2, with fewer query-time recognizer calls in every case. The ES-AGENT–ShotScript accuracy differences are smaller: -0.1 to $+1.3$ points on OmniVideoBench and $+1.2$ to $+2.2$ points on Video-MME-v2. Thus, the gains from adding ASR/OCR tracks and local tools extend beyond GEEK, while the advantage of entity-structured input over flat shot captions varies across benchmarks.

5.5 DIFFICULTY ANALYSIS

Figure 2 shows variation across reasoning types in the construction-stage screening results on GEEK-ES. The full-media baseline M_l spans 33.3% on mood transition to 87.3% on plan–action contrast; for temporal grounding, both M_l and M_m without audio pass at 42.9%, versus 28.6% for M_m without video. Because M_l and M_m are different models, these comparisons do not isolate modality contributions. Figure 4 shows that text-only screening removes the most candidates; the baseline

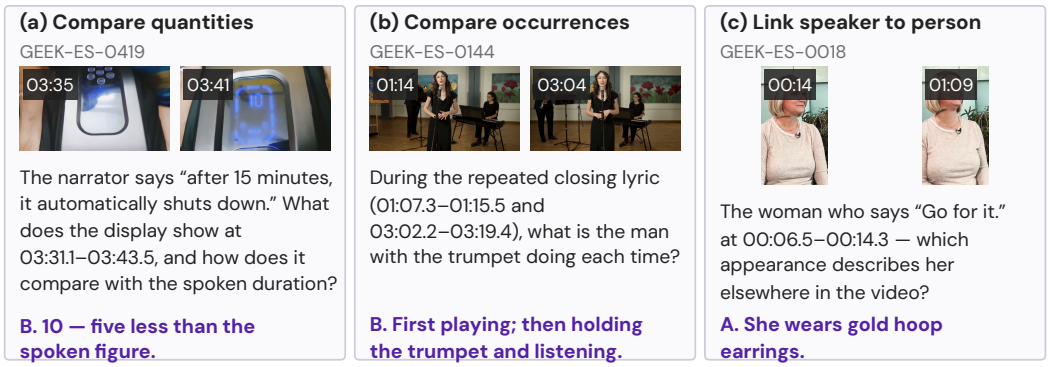

Figure 3: Source video frames paired with three benchmark questions and reference answers. Frames are unmodified extracts at the displayed timestamps; question wording is shortened for space. Answer letters follow the released option order. Full questions and options appear in Appendix C.

pass rate is lower on 3–5 min than on 15–30 s clips (56.7% vs. 71.4%), but reasoning-type and content differences prevent attributing this to duration alone. These are construction-stage screening results, not independent answer-accuracy evaluations.

5.6 Qualitative Examples and Cost

Figure 3 illustrates three evidence requirements: comparing a displayed value with a duration quoted in the question, comparing a trumpeter's actions across two intervals, and linking a referenced speaker to her appearance in a later shot. The examples illustrate on-screen text reading, temporal comparison, and entity linking; they show reference-answer evidence rather than recorded agent trajectories. Appendix C provides the full questions and options.

Cost accounting. With Qwen3-Omni-30B-A3B-Instruct on GEEK, GEEZ averages 3.2 local retrieval queries per question and cuts query-time recognizer calls from 5.4 to 4.1 and frames from 148 to 113 relative to ES-AGENT, at 11% more planner tokens and 102 s to 86 s query-time latency (excluding offline extraction). For the median 117 s video, extraction requires 9 ASR and 64 OCR requests. Under the pricing model in Appendix F, both methods cost $41 per 1,000 questions amortized over 2.0 questions per video, with GEEZ cheaper under further reuse; both totals exclude the shared EntityScript extraction cost.

6 Discussion

Effectiveness of adversarial filtering. Our filtering screen removed 40% of panel-complete candidates answerable by a text-only model from option priors, stem leakage, or world knowledge, which give weak evidence of multimodal understanding. The AV core records failure of two single-modality M_m responses, not proof that every item needs both modalities.

Limitations. GEEK clips are capped at five minutes, so the duration gradient may not generalize to longer videos; some components also depend on a single model deployment.

7 Conclusion

We showed that agent-tutored generation over grounded EntityScript state, with adversarial filtering and interleaved manual review, yields GEEK: text-only-screened by construction yet hard for strong omni models. We further showed with GEEZ that combining dedicated ASR/OCR tracks, reconciliation, and local tools adds +7.4–+8.5 points over the ES-AGENT with fewer recognizer calls, and transfers beyond the benchmark.

REPRODUCIBILITY STATEMENT

We will open-source the complete construction and evaluation stack upon acceptance (kept anonymous during review): the GEEK release file with per-item reasoning type, duration bucket, options in scored presentation order, and key; the assembly script with its fixed sampling seed and every filter threshold of §3.4; the QAGP-ES node prompts with their optimizer-iteration histories and frozen split definitions; the QAGP-A^2 doctrine and generator tool contracts; and the GEEZ harness (ASR/OCR track extraction with caching, alignment thresholds of §4.3, planner schemas and instructions, and the evaluation runner). The EntityScripts and ASR/OCR tracks for GEEK and both transfer benchmarks ship as artifacts, and the serializer deployment is named at camera-ready. All harness hyperparameters appear in §4.1 and Appendix D: plan and review budgets, span cap, frame budgets, dHash and text-similarity thresholds, and stitching rules. Evaluation uses temperature-0 recognizers and structured planner outputs. We report sampling uncertainty via video-resampled bootstrap intervals, characterize run-to-run provider variance by repeated runs (Appendix F), and include the full observation trace in per-item result files, so every reported cell can be audited from released artifacts.

ETHICS STATEMENT

GEEK is built from publicly accessible web video; clips are used for evaluation only and ship as identifiers plus derived annotations rather than redistributed media. Questions may reference visible or audible people. The construction pipeline targets on-screen evidence (appearance, actions, speech, on-screen text) and does not solicit inferences about protected attributes or off-screen identity beyond what QAGP-A^2 can verify from cited public sources. Manual review screens contested keys and removes questions relying on private or sensitive information. ASR and OCR tracks transcribe only content already present in the media. The benchmark inherits the topical and linguistic distribution of web video (§6); scores on GEEK should not be read as unbiased estimates of capability over other domains or languages.

AI USE STATEMENT

Generative AI tools are part of the reported benchmark-construction and evaluation workflow: they were used to implement and run software, serialize videos, propose and revise candidate questions, and assist with qualitative analysis and interpretation of model outputs. They were also used for recommended authoring tasks including reference lookup and formatting, drafting and editing prose for readability, and limited figure-authoring assistance. We did not use generative AI to invent experimental measurements, alter reported results, or replace the authors' verification of keys, citations, code, figures, and claims. The authors reviewed all AI-assisted work and take responsibility for the final paper and released artifacts.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

ByteDance Seed. Seed2.0 Lite model card. Model card, 2026. URL https://seed.bytedance.com/en/seed2/.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. OmniVideo-100K: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026. URL https://arxiv.org/abs/2606.14702.

Brandon Castellano. PySceneDetect: Python library and command-line tool for scene cut/transition detection. Computer software, 2025. URL https://www.scenedetect.com/.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1684. URL https://doi.org/10.52202/079017-1684.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025. doi: 10.1109/ICCV51701.2025.01882. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_LVAgent_Long_Video_Understanding_by_Multi-Round_Dynamical_Collaboration_of_MLLM_ICCV_2025_paper.html.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 75–92, 2024. doi: 10.1007/978-3-031-72670-5_5. URL https://eccv.ecva.net/virtual/2024/poster/1474.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10826–10834, 2020. doi: 10.1609/aaai.v34i07.6713. URL https://ojs.aaai.org/index.php/AAAI/article/view/6713.

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

Google DeepMind. Gemini 3.6 Flash model card. Model card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-6-flash.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11282–11292, 2021. doi: 10.1109/CVPR46437.2021.01113. URL https://openaccess.thecvf.com/content/CVPR2021/html/Grunde-McLaughlin_AGQA_A_Benchmark_for_Compositional_Spatio-Temporal_Reasoning_CVPR_2021_paper.html.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1078–1088, 2020. URL https://proceedings.mlr.press/v119/bras20a.html.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, et al. OmniVideoBench: Towards audio-visual understanding evaluation for omni MLLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ItRYEe8E61.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024. URL https://arxiv.org/abs/2411.13093.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. doi: 10.1109/CVPR52734.2025.01764. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal intelligence. Model release, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. GPT-5.6 Sol model. OpenAI API model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 46575–46602, 2026. doi: 10.18653/v1/2026.acl-long.2161. URL https://aclanthology.org/2026.acl-long.2161/.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. TraveLER: A modular multi-LMM agent framework for video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9740–9766, 2024. doi: 10.18653/v1/2024.emnlp-main.544. URL https://aclanthology.org/2024.emnlp-main.544/.

Tencent Hunyuan Team. Script-a-video: Deep structured audio-visual captions via factorized streams and relational grounding. *arXiv preprint arXiv:2604.11244*, 2026. URL https://arxiv.org/abs/2604.11244.

Yizhou Tian, Zizhe Chen, Shiyuan Deng, Garry Yang, Zijie Dai, Luohao Pan, Hao Lin, Peiqi Yin, Xiao Yan, and James Cheng. CAM: Question answering on entity-centric videos with continuous extraction and adaptive querying. *arXiv preprint arXiv:2609.06504*, 2026. URL https://arxiv.org/abs/2609.06504.

Video-MME Team. Video-MME-v2: Extending comprehensive video evaluation to omni-modal reasoning. Benchmark release, 2026. URL https://video-mme.github.io/.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision*, pp. 58–76, 2024. doi: 10.1007/978-3-031-72989-8_4. URL https://eccv.ecva.net/virtual/2024/poster/1090.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3282, 2025. doi: 10.1109/CVPR52734.2025.00311. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VideoTree_Adaptive_Tree-based_Video_Representation_for_LLM_Reasoning_on_Long_CVPR_2025_paper.html.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

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*, 2025. URL https://arxiv.org/abs/2509.17765.

Ke Xu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Agentic active omni-modal perception for multi-hop audio-visual reasoning. *arXiv preprint arXiv:2605.28192*, 2026. URL https://arxiv.org/abs/2605.28192.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. URL https://arxiv.org/abs/2409.02813.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zhuo Zhi, Qiangqiang Wu, Minghe shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. VideoAgent2: Enhancing the LLM-based agent system for long-form video understanding by uncertainty-aware CoT. *arXiv preprint arXiv:2504.04471*, 2025. URL https://arxiv.org/abs/2504.04471.