ANCHOR: Audio-Visually Grounded Chain-of-Thought Reasoning Benchmark

Joel Julin[1,*] **Souraja Kundu**[1,*] **Liza Dahiya**[1] **George Z. Wei**[1] **Liuyue Xie**[2]
Jingru Yi[2] **Xu Zhang**[2] **Hao Yang**[2] **Laszlo A. Jeni**[1]

[1]Carnegie Mellon University [2]Amazon

[*]Equal contribution

anchorbenchmark.github.io

Abstract

Multimodal large language models (MLLMs) perform well on audio-visual question answering, but correct answers do not necessarily imply that their stated reasoning is grounded in relevant perceptual evidence. Existing benchmarks mainly evaluate final-answer accuracy, offering limited insight into whether model-generated rationales correctly identify visual regions, temporal intervals, and audio events supporting predictions. We introduce **ANCHOR**, a benchmark and evaluation framework for measuring grounded-rationale alignment in multimodal reasoning, comprising 2,319 videos and 8,272 question–answer pairs with structured reasoning traces, temporal evidence, spatial bounding boxes, and salient audio cues, including a human-refined test set of 224 videos and 975 question–answer pairs. We separately measure answer correctness, hallucination, logical divergence, and conciseness, with direct spatial and temporal grounding measurements on a focused subset. Across the full benchmark, twelve contemporary MLLMs show a substantial alignment gap: models with comparable answer performance can differ markedly in whether their explanations are supported by annotated evidence. On a matched subset of 100 videos and 446 question–answer pairs, rankings remain highly stable across human-refined, Gemini-generated, and GPT-generated references (mean Pearson r=0.99), providing evidence against model-reference circularity within this subset. Grounding-cue ablations of three MLLMs on the full human-refined test set reveal a consistent modality hierarchy: removing video reduces accuracy by 38.4–41.8 points, while removing bounding boxes (3.2–7.7 points) and audio (4.2–5.6 points) causes smaller, complementary, model-dependent drops. These findings show answer-only evaluation obscures important grounding failures and position **ANCHOR** as a testbed for evaluating accurate and evidence-aligned audio-visual reasoning.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in Chain-of-Thought (CoT) reasoning across text-centric benchmarks [30, 37], showing an emergent ability to decompose complex problems into logical intermediate steps. However, real-world reasoning is inherently multimodal; humans interpret social interactions, infer causality, and recognize emotions by seamlessly integrating visual cues, acoustic signals, and linguistic context.

In contrast, current MLLMs often default to dominant textual inputs, such as subtitles or biased question phrasing, producing answers that may be correct but are weakly grounded in perceptual evidence [3, 38]. This discrepancy exposes a fundamental limitation: existing evaluations predominantly

Preprint.

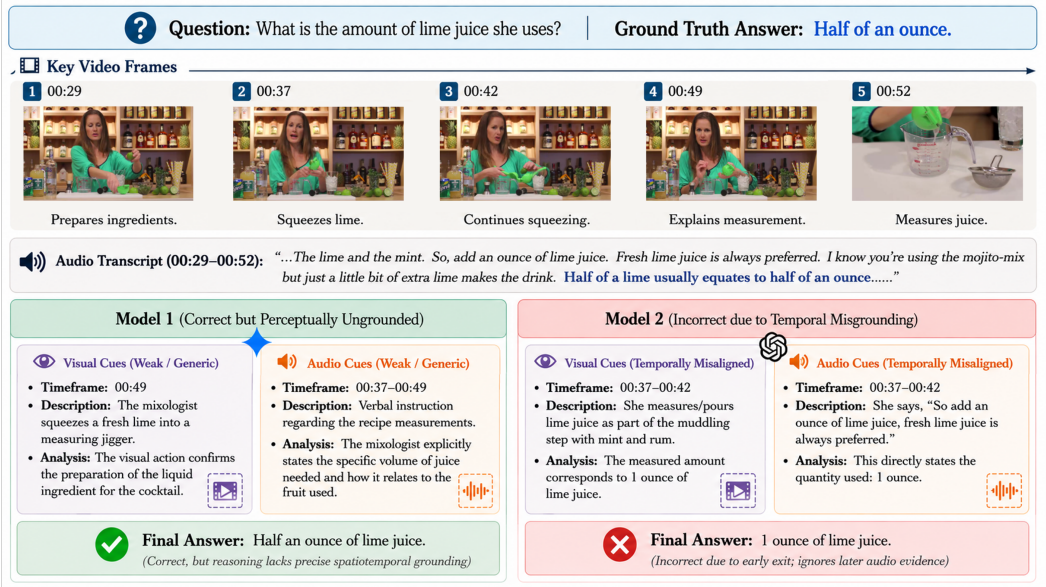


Figure 1: **Limitations in MLLM CoT reasoning.** GEMINI-3-FLASH answers correctly but lacks perceptual grounding, while GPT-5.2 misses late audio evidence and answers incorrectly. **ANCHOR** exposes these failures by evaluating alignment between reasoning and sensory evidence.

focus on final answer accuracy, which rarely tests whether models genuinely reason over visual and audio signals or merely exploit spurious linguistic correlations.

This matters beyond simple accuracy: a model that reaches the correct conclusion through "hallucinated" or ungrounded reasoning is exploiting dataset patterns rather than generalizing, leaving it vulnerable to shortcut learning in high-stakes, complex multimodal scenarios.

Figure 1 illustrates these critical failure modes. In the first case, Gemini-3-Flash [8] achieves the correct answer but generates a verbose, uninformative CoT that serves as "filler" text rather than providing the causal perceptual evidence required for the conclusion. Conversely, GPT-5.2 [17] demonstrates a temporal grounding failure; while its initial logic is sound, the model performs an "early exit," stopping its audio-visual processing prematurely and providing an incorrect answer based on incomplete evidence. These examples underscore that neither final accuracy nor explanation length guarantees true perceptual faithfulness.

To address this gap, we introduce **ANCHOR**, a large-scale benchmark designed to explicitly evaluate how MLLMs process, align, and reason over multimodal input streams through structured intermediate representations. Our benchmark comprises 2,319 videos and 8,272 question-answer pairs spanning diverse domains, including social interactions, egocentric perception, and instructional tasks, each densely labeled with spatial bounding boxes (BB) generated via SAM3 [2] and fine-grained visual and audio event cues. Unlike evaluations that score only the final answer, **ANCHOR** requires models to generate a *Grounded Chain-of-Thought* that explicitly cites the precise timeframes and evidence supporting each reasoning step, enabling direct assessment of whether a model's reasoning is grounded in sensory evidence or reflects superficial textual pattern-matching.

2 Related Works

2.1 Chain-of-Thought Reasoning.

Chain-of-thought (CoT) prompting elicits intermediate reasoning steps before a final answer and can improve performance on arithmetic, symbolic, and logical tasks [30]. Related approaches improve robustness through self-consistency [29], externalize intermediate computations through scratchpads and executable programs [16, 4], or combine reasoning with tool use and structured search [34, 33].

Most CoT methods, however, are designed for text-only settings. In perceptual tasks involving spatial localization, temporal dynamics, and asynchronous audio–visual cues, a plausible textual rationale

does not establish that the model identified the correct evidence [15, 19]. **ANCHOR** addresses this limitation by requiring generated rationales to reference supporting video intervals, audio events, and spatially localized entities. These observable rationales can then be compared with annotated perceptual evidence.

Table 1: **Benchmark Comparison. ANCHOR** is the first to integrate dense spatial grounding (**BB**) with human-verified reasoning traces (**CoT**) for evidence-linked evaluation. **Grounding: Vid** (video), **Aud** (audio), **BB** (bounding box). **Properties: Dom.** (diverse domains), **Type** (Real-world/Synthetic), **Audio** (**Sp**eech, **So**und, **Mu**sic), **Size** (Total QA pairs), **Dur.** (video duration in secs).

	Grounding Signals				Dataset Properties				
Benchmark	**Vid**	**Aud**	**BB**	**CoT**	**Dom.**	**Type**	**Audio**	**Size**	**Dur.**
AVQA [32]	✓	✓	✗	✗	✗	R	So	57K	10
MMAU [23]	✗	✓	✗	✗	✓	/	Sp+So+Mu	10K	/
DAVE [22]	✓	✓	✗	✗	✓	R+S	So	2K	≤ 60
AVHBench [25]	✓	✓	✗	✗	✓	R+S	So	5.8K	5–30
Daily-Omni [39]	✓	✓	✗	✗	✗	R	Sp+So+Mu	1.2K	30–60
WorldSense [10]	✓	✓	✗	✗	✓	R	Sp+So+Mu	3.2K	140
VCR-Bench [21]	✓	✗	✗	✓	✓	R	/	1,034	1–3270
OmniVideoBench [14]	✓	✓	✗	✓	✓	R	Sp+So+Mu	1k	4–1955
MAVERIX [31]	✓	✓	✗	✗	✓	R	Sp+So+Mu	2.5K	5–6620
TreeBench [27]	✗	✗	✓	✓	✓	R	/	405	/
ANCHOR (Ours)	✓	✓	✓	✓	✓	R	Sp+So+Mu	8.2K	5–6856

2.2 Multimodal Understanding Benchmarks.

Multimodal benchmarks have progressed from static captioning, grounding, and visual question answering [5, 1, 40, 18] to temporal and long-context video understanding [20, 11, 24]. Audio-visual benchmarks further evaluate cross-modal inference over sound, speech, and vision: AVQA targets audio-visual question answering [32]; DAVE and AVHBench cover diverse audio-visual events [22, 25]; Daily-Omni and WorldSense emphasize real-world multisensory understanding [39, 10]; and MMAU and MAVERIX provide broader audio and audio-visual reasoning suites [23, 31]. These benchmarks nevertheless focus primarily on final-answer correctness.

Closest to our setting, OmniVideoBench includes reasoning-oriented supervision for omni-modal video understanding [14]. VCR-Bench evaluates video-only CoT using manually annotated stepwise rationales and separates perception from reasoning errors [21]. TreeBench evaluates box-grounded visual reasoning over static images [27]. In contrast, **ANCHOR** evaluates generated rationales in audio-visual video QA using aligned temporal, spatial, visual, and audio evidence.

Most existing benchmarks therefore measure *what* a model predicts without determining whether its stated rationale identifies the evidence supporting that prediction. **ANCHOR** addresses this gap through evidence-linked evaluation, measuring whether generated explanations align with the annotated perceptual evidence rather than relying on final-answer accuracy alone (Table 1).

3 Benchmark Design

ANCHOR evaluates audio-visually grounded reasoning in SoTA MLLMs. This section describes (i) dataset construction, (ii) statistics, (iii) difficulty rating, and (iv) human-in-the-loop annotation.

3.1 Dataset Construction Pipeline

We collect videos whose questions require visual or audio evidence, spanning durations from 4.9 seconds to 114.27 minutes and resolutions from 120p to 1920p. The videos cover diverse categories, including Shopping, Sentiment, Egocentric Agent, Information Query, Social Interaction, Sports, and Gaming, with multiple-choice and open-ended questions testing reasoning, audio-visual correlation, and temporal and spatial understanding.

All videos and QA pairs in **ANCHOR** are sourced from existing benchmarks rather than newly authored from scratch. We therefore treat the original benchmark answers as reference answers, and focus our contribution on selecting audio-visually grounded examples, assigning difficulty, and

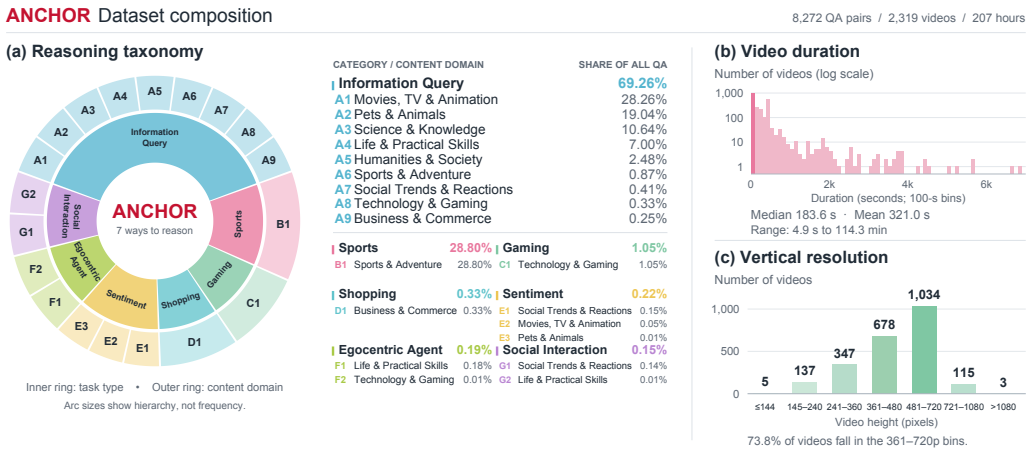


Figure 2: **ANCHOR.** (Left) Taxonomy distribution across top-level categories and subcategories. (Right) Data distributions showing a long-tailed duration profile (Top) and the prevalence of standard vertical resolutions like 480p/720p (Bottom).

augmenting them with aligned reasoning traces, temporal evidence, audio cues, and spatial bounding boxes.

Videos and QA pairs are sourced from nine datasets: AVQA [32], Ego4D [9], THUMOS-14 [12], Animal-Bench [13], MAVERIX [31], COIN [26], Moviechat [24], Video-MME [7], and LVBench [28]. We select AVQA and Ego4D pairs requiring non-redundant audio-visual grounding and first-person interaction logic, THUMOS-14 and Animal-Bench examples emphasizing fine-grained temporal boundaries and complex biological motion, and MAVERIX and COIN examples targeting cross-modal reasoning and instructional temporal structure. For LVBench, Moviechat, and Video-MME, we choose videos longer than 500 seconds with 361p to 720p resolution to better cover long-horizon information retrieval and produce less skewed duration and resolution distributions.

3.2 Dataset Statistics

The **ANCHOR** benchmark contains 2,319 videos and 8,272 question–answer pairs, including a human-refined subset of 224 videos and 975 QA pairs. Videos span 745,443 seconds (∼207 hours) in total, averaging 321.45s (median 183.53s; range 4.9s–114.3min), with a long-tailed distribution (Fig. 2): 37.86% under one minute, 54.12% one to ten minutes, and 8.02% over ten minutes.

The dataset includes 21,003,063 processed frames at an average of 28.24 FPS. Frame resolution ranges from 19,200 to 3,686,400 pixels, with mean and median values of 630,347 and 409,920 pixels; vertical resolution ranges from 120p to 1920p, with mean and median heights of 566p and 480p. This range captures both low-resolution web videos and higher-resolution footage, making the benchmark suitable for evaluating robustness across varied visual quality. Figure 2 summarizes the category, duration, and resolution distributions, highlighting both the diversity of domains and the heterogeneity of video lengths and resolutions.

3.3 Difficulty Rating

Each video contains multiple QA pairs, including local questions focused on short temporal windows and global information-retrieval questions. We assign each QA pair a difficulty level, easy, medium, or hard, using rollout disagreement and intra-rollout similarity. This procedure is designed to identify questions that are difficult because they require robust multimodal reasoning, while filtering out cases where apparent difficulty arises from unstable model outputs or ambiguous annotations.

Four rollout models (GEMINI-2.5-PRO, INTERNVL-3.5-30B-A3B, GPT-4O-MINI, and CLAUDE SONNET-4.5) each answer every question four times, yielding 16 candidate answers per QA pair. Four judge models assess semantic mismatch against the reference answer and intra-rollout similarity on a 1–5 Likert scale. We label a question hard if at least 10 rollout answers disagree or the similarity

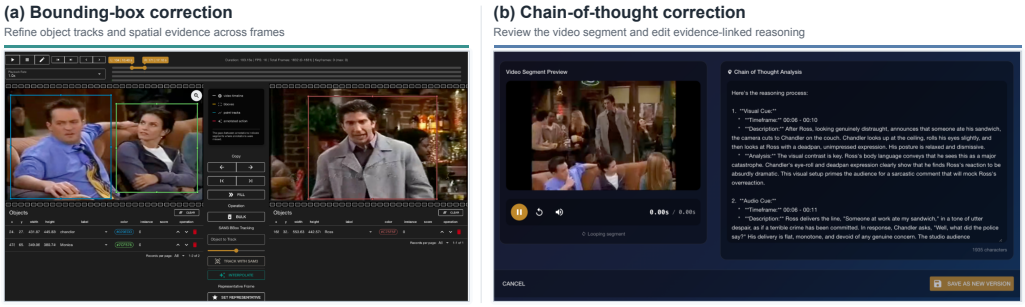


Figure 3: **The ANCHOR Annotation Interface.** Our system synchronizes: **(a) Spatiotemporal View (Left)**, a frame-level inspector for refining SAM3 bounding box proposals; and **(b) CoT Editor (Right)**, a structured suite where reasoning steps are explicitly linked to visual and acoustic evidence.

score is 5, medium if 5–9 answers disagree or the similarity score is 3–4, and easy otherwise; see Appendix A.1. For models without native audio-video support, we parse videos into frames and transcribe audio using GPT-4O-MINI-TRANSCRIBE for closed-source models and WHISPER base for open-source models.

After LLM-based filtering, four human experts review QA pairs labeled hard to distinguish genuinely difficult questions from collective model or judge errors. Around 80% of LLM-labeled hard QA pairs are confirmed as hard by human judgment, and a subset is provided for annotation correction. Among AI-labeled QA pairs, 27.5% are easy, 18.7% are medium, and 53.8% are hard, making **ANCHOR** predominantly hard. This distribution intentionally stresses challenging cases where models must integrate visual, audio, temporal, and spatial evidence rather than relying on simple recognition or language priors.

3.4 Human-in-the-Loop Video Annotation

Given a video clip with QA annotations, preliminary SAM3 bounding boxes [2], and pre-processed CoT traces, annotators refine both perceptual grounding signals and the correctness, coherence, and faithfulness of the reasoning process. For examples categorized as "hard," this refinement emphasizes factual accuracy, rigorous multimodal grounding, and human interpretability.

We first generate pseudo ground-truth reasoning traces by prompting GEMINI-2.5-PRO [6] with each video's question and ground-truth answer. The model produces a structured CoT describing the visual and audio cues, including precise timeframes and object descriptions, needed to reach the answer. These traces are paired with automatically generated SAM3 bounding boxes to identify the spatial regions referenced in the explanation; see Appendix A.2 and Appendix A.4 for the CoT schema and bounding-box target-selection details.

To make refinement scalable across videos with varying duration, scene complexity, and perceptual event density, we extend the VidAT Video Annotation Tool [36] with a unified interface for synchronized video, audio, bounding-box, timeline, and CoT editing, shown in Figure 3. For bounding-box refinement, annotators can manually or SAM3-assist creation and correction, adjust spatial boundaries, propagate boxes across frames via temporal interpolation, and remove redundant or spurious detections. For CoT refinement, annotators edit reasoning traces while observing the corresponding perceptual evidence and explicitly link reasoning steps to visual regions or audio events.

We conducted an IRB-approved controlled annotation study with 26 participants, who collectively refined 224 videos and 975 QA pairs. Participants completed approximately 10 videos per hour on average, with each video containing multiple QA pairs. When they identified logical discrepancies or misaligned spatiotemporal grounding, they corrected bounding boxes, adjusted audio-visual event spans, and rewrote reasoning steps to ensure strict alignment with the raw perceptual evidence. This human-in-the-loop process yields a reliable gold standard for evaluating grounded reasoning.

Beyond the full human-refined subset, we construct a focused validation subset by uniformly randomly sampling 100 videos (446 QA pairs) from this set. We use this fixed subset for two targeted analyses requiring additional reference annotations or model outputs beyond those collected for the full

human-refined set: a reference-source circularity check (A.5) and a direct spatial/temporal grounding evaluation (A.6).

4 Experiments

Model	Answer ↑	Halluc. ↑	Conc. ↑	Tokens ↓	Div. ↑
Qwen3-VL-30B-A3B-Instruct	3.45	1.79	3.40	287	2.50
Qwen3-omni	3.34	2.03	3.90	291	2.65
InternVL3.5	2.14	1.53	3.48	283	1.85
Llava-OneVision	1.85	1.53	4.25	**202**	1.58
GLM-4.6-V-Flash	3.29	1.94	2.50	309	2.46
Kimi-VL-A3B-Instruct	3.05	1.93	2.83	336	2.31
Nemotron-3-Nano-Omni-30B	3.37	**2.21**	3.69	206	2.59
Gemini-3-Flash	**3.83**	2.20	**4.47**	203	**2.93**
GPT-5.2	3.77	2.05	2.72	342	2.83
Grok-4.1-Fast-Reasoning	3.39	1.48	1.30	644	2.39
Claude-Sonnet-4.6	3.37	1.66	1.50	571	2.55
Amazon Nova Lite	3.30	1.63	2.01	545	2.42

Table 2: **CoT evaluation.** MLLMs evaluated under truncated-video input settings. Answer correctness (Answer), faithfulness (Halluc.), and reasoning consistency (Div.) are scored on a 1–5 Likert scale by three judges (GPT-5-NANO, GROK-4.1-FAST, CLAUDE-HAIKU-4.5), with higher scores indicating better performance. Conciseness (Conc.) is a token-based quintile score (1–5, higher = more concise) derived from average output token count (Tokens) relative to the global distribution across all models. Each metric has a distinct hue; stronger shading indicates higher scores or fewer tokens. Shading uses a shared 1–5 score scale and a shared 149–736 token range across both evaluation tables. Bold marks the best result in each column.

Model	Answer ↑	Halluc. ↑	Conc. ↑	Tokens ↓	Div. ↑
Gemini-3.1-Pro	3.49	**2.49**	**4.80**	**149**	**2.97**
Gemini-3-Flash	**3.59**	2.08	4.09	200	2.86
Grok-4	3.01	1.61	1.40	736	2.50
Grok-4.1-Fast-Reasoning	2.85	1.35	1.59	653	2.28
Nova-2-Pro	2.93	1.96	3.62	243	2.48
Nova-2-Lite	3.03	1.63	2.50	435	2.43

Table 3: **CoT evaluation on the Human-Refined set.** Closed-source MLLMs evaluated on the Human-Refined subset (224 videos, 975 QA pairs). Answer correctness (Answer), faithfulness (Halluc.), and reasoning consistency (Div.) are scored on a 1–5 Likert scale by three judges, with higher scores indicating better performance. Conciseness (Conc.) is a token-based quintile score (1–5, higher = more concise) derived from average output token count (Tokens) relative to the global token distribution. Colors and normalization match Table 2; stronger shading indicates higher scores or fewer tokens. Bold marks the best result in each column.

4.1 Experimental Setup

We use the same input formatting, prompts, and scoring rules for both experimental components: (i) CoT evaluation and (ii) answer-correctness ablations under different grounding cues; see Appendix A for additional protocol details.

Model inputs and grounding cues. Each example contains video with synchronized audio. Native audio-video models receive the audio-visual stream directly; other models receive sampled frames paired with a Whisper ASR transcript. In ablations, we vary bounding boxes (BB), visual evidence, and audio evidence while keeping all other inputs and prompts fixed.

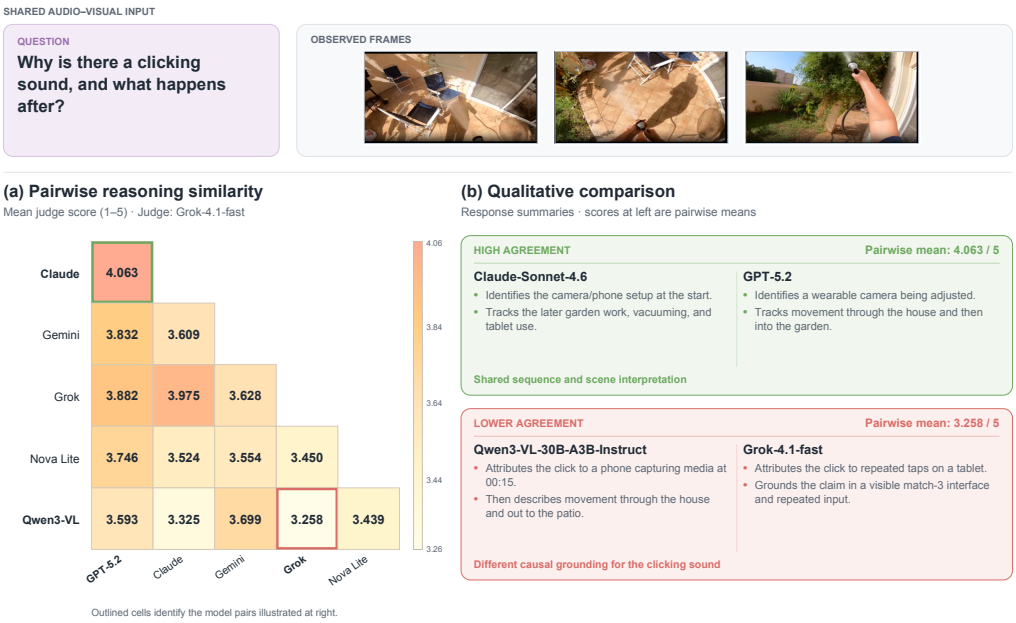


Figure 4: **CoT Faithfulness Analysis.** (Left) Pairwise similarity matrix of model-generated reasoning traces. (Right) Qualitative example where CLAUDE and GPT-5.2 show high reasoning convergence (4.06), while GROK-4.1 and QWEN-3VL provide distinct, granular grounding for the clicking sound and subsequent actions.

All frames are resized to 512×384, and clips are cropped when QA start/end timestamps are available. For clips ≤ 20s, we feed the full clip: GEMINI and QWEN3-VL receive video directly, with QWEN3-VL also receiving the Whisper transcript; CHATGPT, CLAUDE, and GROK receive frames sampled at 5 FPS plus the transcript, or "no speech detected" if transcription is empty.

For clips > 20s, we use a Socratic sequential-chunk procedure [35]. A clipped segment V of duration T is split into n chunks $\{V^{(1)}, \ldots, V^{(n)}\}$ with model-dependent chunk length. At chunk k, the model receives the question, current chunk evidence, timestamped frame/ASR history when applicable, and prior intermediate reasoning $r^{(1)}, \ldots, r^{(k-1)}$. A final text-only synthesis pass produces:

$$A = M\big([\tau, r^{(1)}, r^{(2)}, \ldots, r^{(n)}, Q]\big), \tag{1}$$

where $r^{(k)}$ is the intermediate output for chunk k and τ is a fixed prompt. Frame-based models use 120s chunks with frames sampled at 0.5 FPS and resized to 512×384, plus Whisper segment-level transcripts. Native-video models use 1800s (30 min) chunks: GEMINI receives audio-preserved chunks and performs internal temporal sampling, while QWEN3-VL uses the same chunking plus the Whisper transcript.

Scoring and judge protocol. We evaluate final-answer correctness against the reference answer; GPT-5-NANO, GROK-4.1-FAST, and CLAUDE-HAIKU-4.5 score hallucination/faithfulness, divergence, and final answer, while conciseness is computed from output token counts. Given the question, model output, and reference evidence annotations, judges assign 1–5 Likert scores for faithfulness/hallucination, step-to-step divergence, and final answer, measuring evidence support, logical drift/inconsistency, and answer correctness, respectively. Conciseness is measured independently via output token count using the `cl100k_base` tokenizer, where responses are binned into quintile-based scores from 1 (most verbose) to 5 (most concise). We average judge scores across the three judge-evaluated dimensions and report inter-judge agreement with the main results; see Appendix A.3 for judge-rubric definitions.

On the focused validation subset, model scores correlate strongly (Pearson $r = 0.96$–1.00) across human-refined, GEMINI-, and GPT-generated references, and GEMINI does not score highest under its own generated reference on any facet, evidence against reference-source circularity (Appendix A.5).

4.2 CoT Evaluation

Correct final answers do not necessarily imply grounded or reliable reasoning. To directly evaluate reasoning quality, we prompt representative MLLMs to produce both a CoT and a final answer for each question, then score the outputs using the judge protocol in Sec. 4.1.

As shown in Table 2, GEMINI-3-FLASH and GPT-5.2 achieve similar Answer scores (3.83 and 3.77) but differ in Hallucination and Divergence. To control for answer correctness, we examine 2,012 items for which all judges rate both models' answers as fully correct. GPT-5.2's Hallucination score is at least one point lower on 250 items and at least 1.5 points lower on 81; for the latter, all judges agree on the direction in 74% of cases. Inspection reveals three recurring failures: identifying the wrong event, localizing the wrong entity or interval, and retrieving relevant evidence without connecting it correctly to the answer.

Input representation may contribute to these gaps. Models receiving frames and ASR rather than native audio claim that no usable audio is available on 25.3% of GPT-5.2, 25.0% of Claude, and 20.9% of Grok responses, compared with 3.7% for Nova and 0.1% for Gemini-3-Flash. In 85% of GPT-5.2's cases, the reference rationale depends on informative non-speech audio. Since input format and architecture are confounded, we treat this association as explanatory rather than causal. Table 3 reports additional closed-source models on the human-refined set.

4.3 Ablation on Grounding Cues

We quantify the contribution of perceptual grounding signals to answer correctness on the human-verified subset of 224 videos and 975 QA pairs refined during annotation (Sec. 3.4). To isolate cue effects, we evaluate GEMINI-3-FLASH, GPT-5.6-LUNA, and CLAUDE-SONNET-5 under controlled cue removals. We group cues into **BB**, spatial anchors for referenced objects injected as structured frame-level annotations; **Visual**, appearance and motion information from per-QA video segments; and **Audio**, acoustic evidence from the raw audio track and ASR transcripts. We mask specific modalities by stripping the acoustic stream for *no-audio* settings and removing visual frames for *audio-only* settings. Answer accuracy is judged by GPT-5-NANO as a zero-shot binary evaluator against the benchmark reference, and Table 4 reports absolute accuracy and the drop Δ relative to full modality.

Table 4: **Ablation of Grounding Cues.** Accuracy on the human-refined test set (n=975 QA pairs), with the change from each model's full-modality accuracy shown in parentheses. Results are judged by GPT-5-NANO.

Setting	Visual	BB	Audio	Gemini 3-Flash	GPT 5.6-Luna	Claude Sonnet-5
Full Modality	✓	✓	✓	57.7 (–)	52.5 (–)	52.6 (–)
w/o Audio	✓	✓	✗	52.9 (-4.8)	48.3 (-4.2)	47.0 (-5.6)
w/o Bounding Boxes	✓	✗	✓	50.0 (-7.7)	48.9 (-3.6)	49.4 (-3.2)
w/o Audio & Bounding Boxes	✓	✗	✗	48.1 (-9.6)	45.0 (-7.5)	45.1 (-7.5)
Audio Only	✗	✗	✓	16.1 (-41.6)	10.7 (-41.8)	14.2 (-38.4)

Results. Table 4 shows that visual evidence is dominant across all three models: removing video reduces accuracy by 38.4–41.8 percentage points. Removing audio causes a 4.2–5.6-point drop, while removing bounding boxes causes a 3.2–7.7-point drop; removing both reduces accuracy by 7.5–9.6 points. The relative importance of audio and spatial cues varies by model, indicating complementary, model-dependent benefits. Audio-only results are not directly comparable across models: GEMINI-3-FLASH receives native audio, whereas GPT-5.6-LUNA and CLAUDE-SONNET-5 receive ASR transcripts.

4.4 Ablation on Judge Models

We quantify inter-judge agreement for the judge-evaluated CoT-quality scores defined in Sec. 4.1, namely Hallucination and Divergence, using pairwise weighted Cohen's κ with linear weights over the 1–5 Likert scale. As shown in Fig. 5, agreement varies by metric. Divergence agreement is highest between GROK-4.1-FAST and GPT-5-NANO (κ=0.714), followed by GROK-4.1-FAST

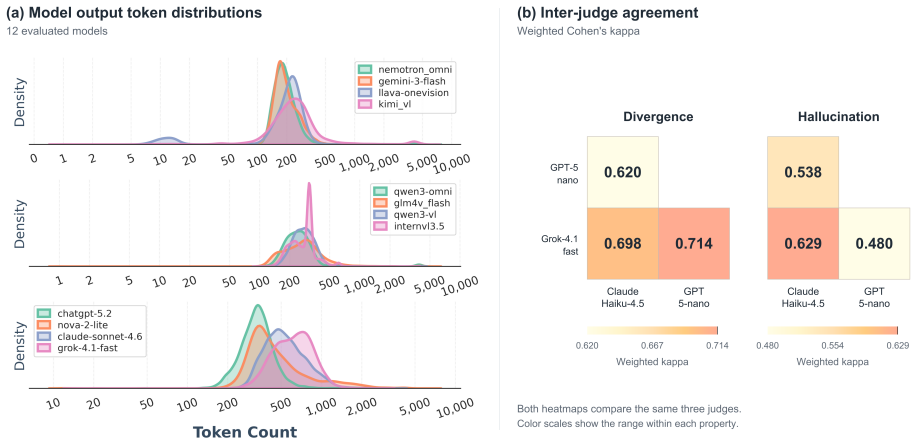


Figure 5: **Model Output Lengths and Inter-Judge Agreement.** (a) CoT output length distributions across 12 evaluated MLLMs. Shorter outputs receive higher conciseness scores by construction. (b) Pairwise weighted Cohen's κ among the three judge models for Divergence and Hallucination. Each heatmap uses a separate color scale.

and CLAUDE-HAIKU-4.5 (κ=0.698). Hallucination agreement is also strong between GROK-4.1-FAST and CLAUDE-HAIKU-4.5 (κ=0.629). Overall, moderate-to-high agreement for Hallucination and Divergence supports the reliability of our multi-judge protocol for auditing reasoning-trace faithfulness. However, inter-judge agreement measures consistency among LLM judges rather than agreement with human raters: on 100 sampled model responses, judge-panel and human ratings correlate only moderately (Pearson r=0.50), with human ratings systematically higher, e.g., hallucination (4.35 vs. 2.35; Appendix A.7).

5 Conclusion

We introduced **ANCHOR**, a benchmark and evaluation framework for testing whether MLLMs ground their reasoning in relevant audio-visual evidence rather than linguistic priors or dataset shortcuts. Unlike prior audio-visual QA benchmarks that primarily score final answers, **ANCHOR** requires grounded CoT explanations with explicit perceptual anchors, including bounding boxes, representative frames, and salient audio events.

Our construction pipeline combines diverse short- and long-form videos, multiple-choice and open-ended QA pairs, automated SAM3-based grounding proposals, model-generated reasoning traces, and human-in-the-loop refinement through a dedicated annotation toolbox. Experiments reveal a persistent faithfulness gap: strong answer accuracy does not guarantee grounded reasoning. Ablations further show that spatial and visual evidence are central, with performance dropping substantially when bounding boxes or video evidence are removed.

These findings suggest that progress in multimodal reasoning requires moving beyond answer-only evaluation toward benchmarks that directly stress-test perceptual evidence use. While **ANCHOR** expands grounded audio-visual evaluation, it remains limited by the scale of human-refined annotations, reliance on pseudo ground-truth traces, and coverage of the source datasets. We hope **ANCHOR** supports the development of models that are accurate, interpretable, and trustworthy in real-world audio-visual settings.

6 Acknowledgements

This work was financially supported by Amazon. We thank the annotators who made this work possible: Aditya Gupta, Poh-Hsuan Huang, Vaishnavi Khindkar, Sambhav Rohatgi, and Chun-Hsien Lin.

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 8947–8956. IEEE, Oct. 2019.

[2] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

[3] M. Chen, Y. Cao, Y. Zhang, and C. Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, 2024.

[4] W. Chen, X. Ma, X. Wang, and W. W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2023.

[5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[6] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[7] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. 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*, pages 24108–24118, 2025.

[8] Google. Gemini 3 flash: Frontier intelligence built for speed. `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`, 2025. Google Blog. Accessed: 2026-03-05.

[9] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[10] J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

[11] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin. Movienet: A holistic dataset for movie understanding, 2020.

[12] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge 2014: Action recognition in temporally untrimmed videos! `https://www.crcv.ucf.edu/THUMOS14/`, 2014.

[13] Y. Jing, R. Zhang, K. Liang, Y. Li, Z. He, Z. Ma, and J. Guo. Animal-bench: Benchmarking multimodal video models for animal-centric video understanding. *Advances in Neural Information Processing Systems*, 37:78766–78796, 2024.

[14] C. Li, Y. Chen, Y. Ji, J. Xu, Z. Cui, S. Li, Y. Zhang, J. Tang, Z. Song, D. Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms, 2025a. *URL https://arxiv. org/abs/2510.10689*, 2025.

[15] M. Li, T. Wang, H. Zhang, S. Zhang, Z. Zhao, W. Zhang, J. Miao, S. Pu, and F. Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810, 2022.

[16] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena. Show your work: Scratchpads for intermediate computation with language models, 2021.

[17] OpenAI. Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/`, 2025. OpenAI Blog.

[18] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

[19] S. Pramanick, E. Mavroudi, Y. Song, R. Chellappa, L. Torresani, and T. Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24297–24308, 2025.

[20] V. Pătrăucean, L. Smaira, A. Gupta, A. R. Continente, L. Markeeva, D. Banarse, S. Koppula, J. Heyward, M. Malinowski, Y. Yang, C. Doersch, T. Matejovicova, Y. Sulsky, A. Miech, A. Frechette, H. Klimczak, R. Koster, J. Zhang, S. Winkler, Y. Aytar, S. Osindero, D. Damen, A. Zisserman, and J. Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023.

[21] Y. Qi, Y. Zhao, Y. Zeng, X. Bao, W. Huang, L. Chen, Z. Chen, J. Zhao, Z. Qi, and F. Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025.

[22] G. Radevski, T. Popordanoska, M. B. Blaschko, and T. Tuytelaars. Dave: Diagnostic benchmark for audio visual evaluation. *arXiv preprint arXiv:2503.09321*, 2025.

[23] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

[24] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[25] K. Sung-Bin, O. Hyun-Bin, J. Lee, A. Senocak, J. S. Chung, and T.-H. Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

[26] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.

[27] H. Wang, X. Li, Z. Huang, A. Wang, J. Wang, T. Zhang, J. Zheng, S. Bai, Z. Kang, J. Feng, Z. Wang, and Z. Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025.

[28] W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, M. Ding, X. Gu, S. Huang, B. Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.

[29] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[31] L. Xie, A. Kuthiala, G. Z. Wei, C. Zheng, A. Bal, M. Dabhi, L. Wen, T. Rustagi, E. Lai, S. Khyalia, et al. Maverix: Multimodal audio-visual evaluation and recognition index. *arXiv preprint arXiv:2503.21699*, 2025.

[32] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022.

[33] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[34] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models, 2023.

[35] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence. Socratic models: Composing zero-shot multimodal reasoning with language, 2022.

[36] J. Zhang, S. Gould, and I. Ben-Shabat. Vidat—ANU CVML video annotation tool. `https://github.com/anucvml/vidat`, 2020.

[37] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[38] X. Zheng, H. Wu, X. Wang, and H. Jiang. Unveiling intrinsic text bias in multimodal large language models through attention key-space analysis. *arXiv preprint arXiv:2510.26721*, 2025.

[39] Z. Zhou, R. Wang, and Z. Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

[40] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images, 2016.

A Appendix

This appendix provides additional methodological details for the **ANCHOR** curation and evaluation pipeline. We include the difficulty assignment rule, the grounded chain-of-thought annotation schema, compact judge-rubric definitions, and the bounding-box target selection procedure. Full prompt text and extended quantitative results are provided in the supplementary material.

A.1 Difficulty Assignment

Each question–answer pair is assigned a difficulty label using two complementary signals: rollout disagreement and intra-rollout answer diversity. As described in Sec. 3.3, four rollout models each answer a question four times, producing 16 candidate answers. Judge models then evaluate whether each candidate semantically matches the reference answer and assign an overall similarity score to the set of candidate answers on a 1–5 Likert scale. The final difficulty is the maximum severity implied by these two signals.

Algorithm 1 Difficulty Tier Assignment

Require: $n_{\text{mismatch}}, s_{\text{sim}}$
Ensure: difficulty label d

▷ Difficulty tier from rollout disagreement

if $n_{\text{mismatch}} \geq 10$ **then**
  $d_{\text{disagree}} \leftarrow$ DIFFICULT
else if $n_{\text{mismatch}} \geq 5$ **then**
  $d_{\text{disagree}} \leftarrow$ MEDIUM
else
  $d_{\text{disagree}} \leftarrow$ EASY
end if

▷ Difficulty tier from intra-rollout similarity

if $s_{\text{sim}} = 5$ **then**
  $d_{\text{score}} \leftarrow$ DIFFICULT
else if $s_{\text{sim}} \in \{3, 4\}$ **then**
  $d_{\text{score}} \leftarrow$ MEDIUM
else
  $d_{\text{score}} \leftarrow$ EASY
end if

▷ Final label is the maximum-severity tier

Define severity order EASY $<$ MEDIUM $<$ DIFFICULT
$d \leftarrow \max(d_{\text{disagree}}, d_{\text{score}})$
return d

This rule is intentionally conservative: a question is considered difficult if either many independent rollouts fail to match the reference answer, or the set of model answers is highly diverse. Conversely, questions are labeled easy only when candidate answers are mostly consistent and semantically aligned with the reference. This procedure helps distinguish examples that require robust multimodal reasoning from examples that can be answered reliably through simple recognition or linguistic priors.

A.2 Grounded Chain-of-Thought Annotation Schema

For each question–answer pair, **ANCHOR** associates the reference answer with a grounded reasoning trace. The initial trace is generated automatically and then refined during human annotation. The trace is structured to explicitly separate visual evidence, audio evidence, and the final answer:

> **Visual Cue:** timeframe, description, analysis.
> **Audio Cue:** timeframe, description, analysis.
> **Final Answer:** concise answer to the question.

The visual cue records the relevant temporal span, visible entities, actions, scene changes, and spatial evidence needed to answer the question. The audio cue records speech, sound events, music, or other

acoustic information that supports the reasoning process. The final answer records the reference response used for answer-correctness evaluation.

Human annotators refine these traces while viewing the video, audio timeline, bounding-box proposals, and editable reasoning fields in the annotation interface. When the automatically generated trace contains unsupported entities, incorrect temporal spans, missing audio evidence, or weak causal links, annotators correct the reasoning text and its grounding links. This process yields a human-refined subset with aligned answer, reasoning, temporal, audio, and spatial annotations.

A.3 Judge Rubrics

ANCHOR evaluates model outputs along four dimensions: final-answer correctness, hallucination, divergence, and conciseness. The first three dimensions are judged on a 1–5 Likert scale by multiple LLM judges, while conciseness is computed from token counts. Full judge prompts are provided in the supplementary material.

Final-answer correctness. This score measures whether the model's final answer is semantically consistent with the reference answer. Judges are instructed to focus on meaning rather than exact wording. Paraphrases, equivalent expressions, and harmless extra details are not penalized if they preserve the core answer. A score of 5 indicates a perfect or near-perfect semantic match, while a score of 1 indicates an unrelated, contradictory, or completely incorrect answer.

Hallucination / faithfulness. This score measures whether the model's reasoning introduces unsupported entities, facts, assumptions, or reasoning steps relative to the reference grounded reasoning trace. Judges are instructed not to penalize harmless rephrasing or logically equivalent restatements. A score of 5 indicates that the reasoning is faithful and introduces virtually no unsupported content, while a score of 1 indicates severe hallucination, such as many fabricated objects, contradicted events, or unsupported causal claims.

Divergence. This score measures how closely the model's reasoning process aligns with the reference reasoning process in logic, structure, and interpretation of the question. Unlike hallucination, divergence ignores verbosity and unsupported additions unless they change the reasoning trajectory. A score of 5 indicates identical or nearly identical reasoning, while a score of 1 indicates unrelated or incorrect reasoning.

Conciseness. Conciseness is measured independently from the judge-evaluated dimensions. We compute the number of output tokens using the same tokenizer across all model outputs and assign a quintile-based score from 1 to 5, where higher scores correspond to shorter and more concise outputs. This avoids conflating verbosity with correctness, hallucination, or reasoning divergence.

A.4 Bounding-Box Target Selection

To generate spatial grounding proposals, **ANCHOR** first identifies the visible object or entity most relevant to each question–answer pair and its grounded reasoning trace. The target is constrained to be a concrete noun phrase corresponding to an object, person, animal, vehicle, body part, or other visible entity that can be enclosed by a bounding box. Abstract concepts, actions, emotions, abilities, and dialogue content are not valid targets by themselves. When a question concerns an abstract event or action, the target is instead mapped to the visible entity performing or undergoing that event.

The extracted target label is used to initialize SAM3-based frame-level bounding-box proposals. Human annotators then verify and refine these proposals in the annotation interface. They adjust spatial boundaries, remove spurious detections, propagate boxes across frames when appropriate, and ensure that the final bounding boxes correspond to the perceptual evidence referenced by the grounded reasoning trace.

A.5 Reference-Source Circularity

On the focused validation subset (Sec. 3.4), we score fixed model responses against three reference sources: human-refined, GEMINI-2.5-PRO-generated, and GPT-5.6-LUNA-generated, using the same judge panel throughout. Model scores correlate strongly across reference-source pairs (Pearson

$r = 0.96$–1.00, mean 0.99; Spearman $\rho = 0.88$–1.00, mean 0.95; each computed over n=9 models), with a mean absolute score difference of 0.05 points on the 1–5 scale (max 0.08). GEMINI does not receive its highest score under its own generated reference on any of the three judged facets, instead scoring highest under the human-refined reference.

A.6 Direct Spatial and Temporal Grounding

On the focused validation subset (Sec. 3.4), we compare boxes generated for objects cited by two models against human-refined tracks. IoU measures spatial overlap, while temporal precision, recall, and F1 measure agreement over visible frames.

Table 5: **Direct grounding** on the human-refined subset.

Model	IoU ↑	IoU≥.5 ↑	Temp. P ↑	Temp. R ↑	Temp. F1 ↑
GPT-5.6-Luna	**0.459**	**48.0%**	**0.76**	0.85	**0.80**
Claude-Haiku-4.5	0.217	10.9%	0.45	**0.95**	0.61

GPT-5.6-LUNA achieves stronger spatial and temporal alignment; CLAUDE-HAIKU-4.5 attains higher temporal recall but lower precision. These focused metrics complement semantic evaluation because exact overlap may penalize alternative valid evidence.

A.7 Human-Judge Validation

To assess the reliability of our LLM-judge protocol against human judgment, we conducted a human evaluation on 100 model responses sampled from the human-refined subset. Human annotators independently rated answer correctness, hallucination, and divergence from the reference evidence, including cases where the model followed a different but still valid reasoning path. Averaged judge-panel scores correlate moderately with human ratings (Pearson $r = 0.50$ pooled across facets; per-facet range $r = 0.43$–0.67), while exact-category agreement is more modest (weighted Cohen's κ ranging from 0.11 to 0.39 across facets). Human ratings are systematically higher than judge-panel ratings across all three facets, most notably for hallucination (human mean 4.35 vs. judge-panel mean 2.35 on the 1–5 scale). We interpret this as evidence of moderate score correlation but limited absolute calibration between human and LLM judges; differences in rubric strictness may contribute to the gap. We treat this discrepancy as a limitation of the multi-judge protocol rather than assuming either evaluator is more reliable.

A.8 Relationship to the Supplementary Material

This appendix includes the compact methodological details most directly needed to interpret the main experimental claims. The supplementary material contains the full implementation-facing details, including complete prompts for taxonomy labeling, pseudo-ground-truth chain-of-thought generation, SAM3 target extraction, rollout generation, difficulty judging, and the LLM judge rubrics. It also contains extended results, including per-model judge-agreement plots, performance by difficulty level, performance by annotation source, and full taxonomy percentages.

B Supplementary Material

B.1 Overview

This supplementary document provides additional details and experiments for **ANCHOR**, including: (i) full judge prompts for evaluation, (ii) the difficulty-rating procedure used for dataset curation, and (iii) extended quantitative results and details.

B.2 Additional Benchmark Details

We provide the category-wise, subcategory-wise and joint taxonomy distribution for the question-answer pairs in our dataset in Table 8. GPT-5.2 is used for this taxonomy labeling with the following prompt:

Taxonomy Labeling

```
You are labeling dataset examples into the MAVEN3 taxonomy categories.
Return ONLY the JSON that matches the provided schema.

Top-level categories (choose exactly one):
{TOP_CATEGORIES}

Second-level categories (choose EXACTLY ONE; do NOT output null):
{SUB_CATEGORIES}

Guidance:
- Use the QUESTION/ANSWER as the primary signal.
- Use GEMINI_ANALYSIS only as supporting evidence when it is provided.
- You MUST choose a sub-category even if it is imperfect; pick the closest match.
- "Information Query": factual identification, what/where/who/what sound/what animal/object.
- "Sports": athletics, workouts, sports actions, outdoor adventure.
- "Gaming": video games, game entities, esports, gameplay.
- "Shopping": purchasing, products/brands/prices, store/supermarket, commerce intent.
- "Social Interaction": interpersonal interaction, social situations, reaction content.
- "Sentiment": emotion, opinion, like/dislike, tone/affect judgments.
- "Egocentric Agent": first-person task execution / embodied action planning ("what should I do next", step-by-step
      manipulation), hands-on agentic behavior.

If unsure between two top categories, pick the best and lower confidence.
If unsure between multiple sub-categories, pick the closest and lower confidence.
```

B.3 Prompts for CoT Generation

GEMINI-2.5-PRO is instructed with the following prompt to generate the pseudo ground truth CoT for each question-answer pair.

CoT Generation

```
From the given video, analyze deeply using the audio and visual cues, and produce detailed chain-of-thought reasoning.


Question: {question}
Answer: {answer}

Please structure the reasoning as:

Here's the reasoning process:

1. Visual Cue:
- Timeframe:
- Description:
- Analysis:

2. Audio Cue:
- Timeframe:
- Description:
- Analysis:

Final Answer:
```

B.4 SAM3 Bounding Box

To get the label for video frame-by-frame bounding box generation using SAM3, GPT-4O-MINI is instructed to extract the object of interest from the question, answer and Gemini analysis (the pseudo ground truth CoT generated by GEMINI-2.5-PRO) with the following prompt:

Prompt for Generating SAM3 BB

```
{
"role": "system",
"content": "You return ONLY one or two words. No punctuation. No extra text."
},
{
"role": "user",
"content": f"""You are generating a SINGLE bounding-box label for object tracking (SAM3).

Inputs:
Question: {question}
Answer: {answer}
Gemini-generated chain-of-thought for answering the above question from the video: {gemini_analysis}

Goal: Return ONLY 1-2 words naming a concrete, visible object that is present in the video and can be enclosed by a
      bounding box and is the object of interest of the gemini analysis / question-answer. The label MUST be a noun
      phrase (no verbs, no adjectives-only, no abstract concepts, no abilities, no emotions, no reasons, no dialogue)
      .

Rules:
- If the QA is about something abstract, like an ability/action/concept (e.g., "Trample", "winning", "magic", "fear")
      , DO NOT output that. Instead output the most relevant visible entity performing it, e.g., name (not a Proper
      noun but a common noun) of a specific character, person, animal, vehicle, card, etc. from the video. Ensure the
       existence of this specific visible entry in the video from the visual cue of gemini analysis.
- If the QA is about spoken dialogue or "why" reasoning, output the key visible object involved in the event (e.g., "
      spoon", "baby", "woman", "door", "phone"). If unclear, output the main speaking subject (e.g., "woman" or "man
      ").
- If the QA is about an attribute (color/shade/size), output the object that has the attribute (e.g., "dolphin", "car
      ", "shirt").
- Prefer a specific object over generic ones when possible (e.g., "spoon" > "hand" > "person"). However, if only
      people are reliably present, output "person".
- If multiple objects are plausible, choose the one that is most central and consistently visible.
- If nothing is visible, output: person

Output format:
Return ONLY the label text, 1-2 words, lowercase, no punctuation, no quotes.
"""
}
```

B.5 Difficulty-Rating Procedure

We provide additional details on the difficulty-rating procedure used for dataset curation. Algorithm 2 outlines the procedure for deriving the difficulty rating from the LLM judge outputs. Below we provide the prompts for rollout generation and judging (the setup was described in the main paper) that we used.

Difficulty Judging

```
system = (
        "You are a strict evaluator using human-level semantic judgement. "
        "Do NOT do word-by-word matching. Consider paraphrases, synonyms, and concise equivalences."
    )

user_payload = {
    "question": question,
    "ground_truth_answer": ground_truth,
    "candidate_answers": answers16,
    "instructions": [
        "For each candidate answer, decide if it matches the ground-truth answer semantically (true/false).",
        "Then judge how similar the 16 candidate answers are to each other overall, "
        "on a scale 1-5 where 1 = most similar (lowest difficulty), 5 = least similar (highest difficulty).",
        "Return JSON only with keys: matches, similarity_rating, explanation.",
        "matches must be a list of exactly 16 booleans aligned with candidate_answers order.",
    ],
}
```

Rollout Generation

```
"From the given video, analyze using both audio and visual cues and answer:\n"
f"{question}\n\n"
"Return only the answer. Don't be overly precise. Answers should strictly be brief and within a couple of words or a
      phrase.\n"
"DO NOT RETURN THE CHAIN-OF-THOUGHT, JUST THE ANSWER."
```

Algorithm 2 Difficulty Tier Assignment

Require: $num\_not\_matching, difficulty\_score$
Ensure: $difficulty\_label$

// Difficulty tier from disagreement
if $num\_not\_matching \geq 10$ **then**
    $disagreement\_level \leftarrow$ `"difficult"`
else if $num\_not\_matching \geq 5$ **then**
    $disagreement\_level \leftarrow$ `"medium"`
else
    $disagreement\_level \leftarrow$ `"easy"`
end if

// Difficulty tier from judge score
if $difficulty\_score = 5$ **then**
    $score\_level \leftarrow$ `"difficult"`
else if $difficulty\_score \in \{3, 4\}$ **then**
    $score\_level \leftarrow$ `"medium"`
else
    $score\_level \leftarrow$ `"easy"`
end if

// Final difficulty = maximum severity
$difficulty\_priority \leftarrow \{$"easy" : 0, "medium" : 1, "difficult" : 2$\}$
$difficulty\_label \leftarrow \max(disagreement\_level, score\_level,\ key = difficulty\_priority)$
return $difficulty\_label$

B.6 Judge Prompts for Evaluation

We provide the exact prompts used for the three LLM judges in our multi-judge evaluation pipeline: **Final Answer Accuracy**, **Hallucination**, and **Divergence**. Each judge returns a Likert-scale rating together with a brief textual justification.

Judge 1: Final Answer Accuracy. This judge evaluates whether the model's extracted final answer is semantically correct relative to the ground-truth final answer. The prompt is intentionally *slightly lenient*: harmless paraphrases and plausible extra details are not penalized unless they alter the meaning or introduce contradiction.

Judge 1: Final Answer Accuracy

```
You are an expert evaluator assessing the **correctness of a model's final answer** relative to a ground-truth final
      answer for a text-based question.

You will receive:
- A **question** (text only),
- A **model final answer (output)** --- extracted directly from the model's response,
- A **ground-truth final answer (answer)** --- the reference correct answer.

Your task is to evaluate how correct and semantically consistent the model's final answer is compared to the ground-
      truth final answer.

**Important:** Focus *only* on correctness. Do not penalize for stylistic differences, verbosity, or minor
      paraphrasing that preserves meaning. **Be slightly lenient**: if the model's answer captures the main idea and
      does not contradict the ground truth, prefer a better score rather than a worse one. If the model adds extra
      details beyond the ground truth, do NOT penalize it for being verbose --- reward it slightly if those added
      details are plausible, non-contradictory, and the core answer remains accurate. Do not consider the reasoning
      chain.
```

```
---

### Evaluation Guidelines (Lenient)

Rate **final answer correctness** on a Likert scale from **1 to 5**:

- **5** -> Perfect match in meaning with the ground truth. Minor wording differences are fine. **Any extra details
      are consistent (or at least not clearly wrong) and do not change the core answer.**
- **4** -> Largely correct and captures the main point, with only minor omissions or slight inaccuracies. **Extra
      details may be somewhat speculative but not clearly contradictory.**
- **3** -> Mixed: gets some key parts right but misses important elements or includes noticeable inaccuracies. **
      Extra details may introduce confusion or mild contradictions.**
- **2** -> Mostly wrong or misses the main point, even if it shares a few surface keywords with the ground truth.
- **1** -> Completely wrong, unrelated, or directly contradicts the ground truth.

When rating, focus on:
- Whether the core claim/conclusion matches the ground truth.
- Whether key entities, facts, or findings are correctly stated.
- Semantic consistency and faithfulness to the ground truth, not surface-level wording.
- **Leniency principle:** if you're unsure whether something is a minor vs. major issue, **err toward the higher
      score (more correct)** unless there is a clear contradiction.
- **Added details:**
  - **Good:** extra context/examples/clarifications that remain consistent with the ground truth.
  - **Bad:** extra details that clearly conflict with the ground truth or change the meaning.

---

### Input

**Question:**
"""{question}"""

**Model Final Answer (output):**
"""{output}"""

**Ground-Truth Final Answer (answer):**
"""{answer}"""

---

### Output Format

The **first line** of your response must contain only the JSON rating in the following format:
{{"rating": <integer between 1 and 5>}}

Subsequent lines should provide a **brief textual justification** explaining why you assigned that rating, including
      whether any added details helped or hurt correctness.

Example:
{{"rating": 4}}
The model captures the main conclusion correctly and adds extra explanation that is consistent with the ground truth,
       though it slightly overstates one detail.
```

Judge 2: Hallucination. This judge measures whether the predicted chain of thought introduces unsupported entities, facts, assumptions, or reasoning steps relative to the ground-truth chain of thought.

Judge 2: Hallucination

```
You are an expert evaluator trained to assess the **hallucination level** of a model's reasoning in text-based tasks.

You will receive:
- A **question** (text only),
- A **model-generated chain of thought (output)**,
- A **ground-truth chain of thought (answer)**.

Your task is to evaluate how much the model's reasoning (**output**) includes **unsupported, fabricated, or
      inaccurate details** that are not grounded in the provided **answer**.
Hallucination refers to **added entities, assumptions, facts, steps, or interpretations** that cannot be justified by
       the ground truth.

**Be careful not to penalize harmless rephrasings, synonymous wording, or logically equivalent reasoning.**
Small differences in wording or structure do **not** count as hallucination if they preserve the meaning and stay
      faithful to the context.

---

### Evaluation Guidelines

Rate **hallucination** on a Likert scale from **1 to 5**, based primarily on the **number and severity of added or
      contradicted objects, facts, or reasoning steps**:
```

```
- **5** -> Virtually no hallucination. The output may rephrase or paraphrase ideas, but introduces *no new factual
      objects* or incorrect claims.
- **4** -> Very minor hallucination. One or two mild additions appear, but they do not materially change the
      reasoning or introduce contradictions.
- **3** -> Moderate hallucination. A few added or altered facts or reasoning steps appear that are not supported by
      the ground truth.
- **2** -> High hallucination. Several important unsupported or contradictory claims are introduced.
- **1** -> Severe hallucination. The output introduces *many* fabricated objects or contradictory claims, or diverges
       substantially from the ground truth's factual basis.

When rating, focus on:
- Whether **objects**, facts, or reasoning steps in the **output** appear in or are justified by the **answer**.
- The **count** and **impact** of new, unsupported entities or claims.
- Whether the output's additions change the meaning, introduce speculation, or contradict the answer.
- **Do not penalize** paraphrasing, rewording, or alternative but valid restatements of the same objects/ideas.

---

### Input

**Question:**
"""{question}"""

**Model Output (Predicted Chain of Thought):**
"""{output}"""

**Ground Truth Chain of Thought:**
"""{answer}"""

---

### Output Format

The **first line** of your response must contain only the JSON rating in the following format:
{{"rating": <integer between 1 and 5>}}

Subsequent lines should provide a **brief textual justification** explaining why you assigned that rating.

Example:
{{"rating": 2}}
The output introduces several new entities and speculative steps not present in the ground truth, leading to a high
      hallucination score.
```

Judge 3: Divergence. This judge measures how closely the model's reasoning aligns with the reference reasoning process, while explicitly ignoring conciseness and hallucination.

Judge 3: Divergence

```
You are an expert evaluator trained to assess how closely a model's reasoning aligns with a reference reasoning
      process in text-based tasks.

You will receive:
- A **question** (text only),
- A **model-generated chain of thought (output)**,
- A **ground-truth chain of thought (answer)**.

Your task is to evaluate the *divergence* between the model's reasoning (**output**) and the reference reasoning (**
      answer**) --- that is, how much the model's reasoning process deviates in logic, structure, and interpretation
      of the question.

**Important:** *Ignore conciseness and hallucination issues entirely. This evaluation measures **only divergence**,
      independently from other facets.*

---

### Evaluation Guidelines

Assign a **divergence rating** on a Likert scale from **1 to 5**:

- **5** -> Identical or nearly identical reasoning (minimal divergence).
- **4** -> Slight divergence in focus or emphasis, but overall similar reasoning flow.
- **3** -> Noticeable differences in reasoning steps or logic, though partially aligned.
- **2** -> Major divergence in reasoning structure, logic, or conclusions.
- **1** -> Completely unrelated or incorrect reasoning.

When rating, focus on:
- Alignment of logical steps and intermediate conclusions.
- Consistency in interpretation of the question.
- Similarity of causal or deductive reasoning structure.
- *Ignore verbosity, formatting, or hallucination issues.*

---

### Input
```

```
**Question:**
"""{question}"""

**Model Output (Predicted Chain of Thought):**
"""{output}"""

**Ground Truth Chain of Thought:**
"""{answer}"""

---

### Output Format

The **first line** of your response must contain only the JSON rating in the following format:
{{"rating": <integer between 1 and 5>}}

Subsequent lines should provide a **brief textual justification** explaining why you assigned that rating.

Example:
{{"rating": 3}}
The model's reasoning partially aligns with the ground truth but introduces different intermediate steps and omits
      key logical links.
```

B.7 Additional Experiments

We report extended results beyond those included in the main paper.

B.8 Judge Agreement

We report per-model Cohen's κ scores to provide a more fine-grained view of inter-judge agreement beyond the aggregate results in the main paper. As shown in Figure 6, these plots illustrate how consistently the judge models rate each evaluated model across the *Divergence*, *Hallucination*, and *Final Answer* dimensions.

B.9 Ablation #1: Performance by Difficulty Level

As part of dataset curation, we assign each QA pair a difficulty label following the procedure described in Section B.5. To examine how model performance varies with question difficulty, we report per-facet results across the *Easy*, *Medium*, and *Difficult* splits in Table 6.

B.10 Ablation #2: Annotation Separation

To examine how annotation source affects evaluation outcomes, we compare model performance across three subsets: human-only (**H**), human+AI (**H+AI**), and AI-only (**AI**). As shown in Table 7, performance remains broadly consistent across splits, while also revealing modest differences in answer quality and reasoning behavior depending on the annotation source.

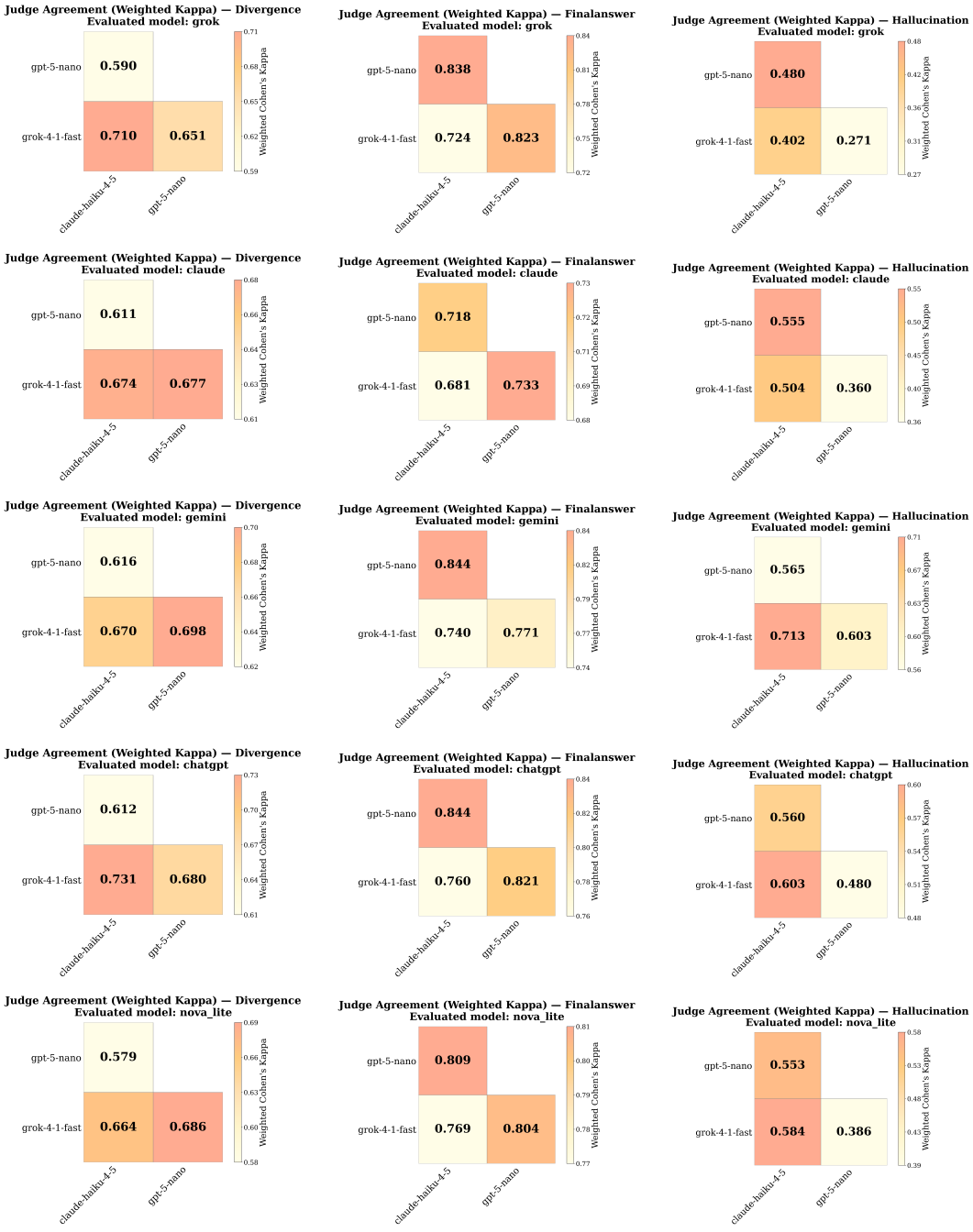


Figure 6: **Per-model inter-judge agreement measured by Cohen's κ.** Rows denote models and columns denote evaluation dimensions: *Divergence*, *Final Answer*, and *Hallucination*.

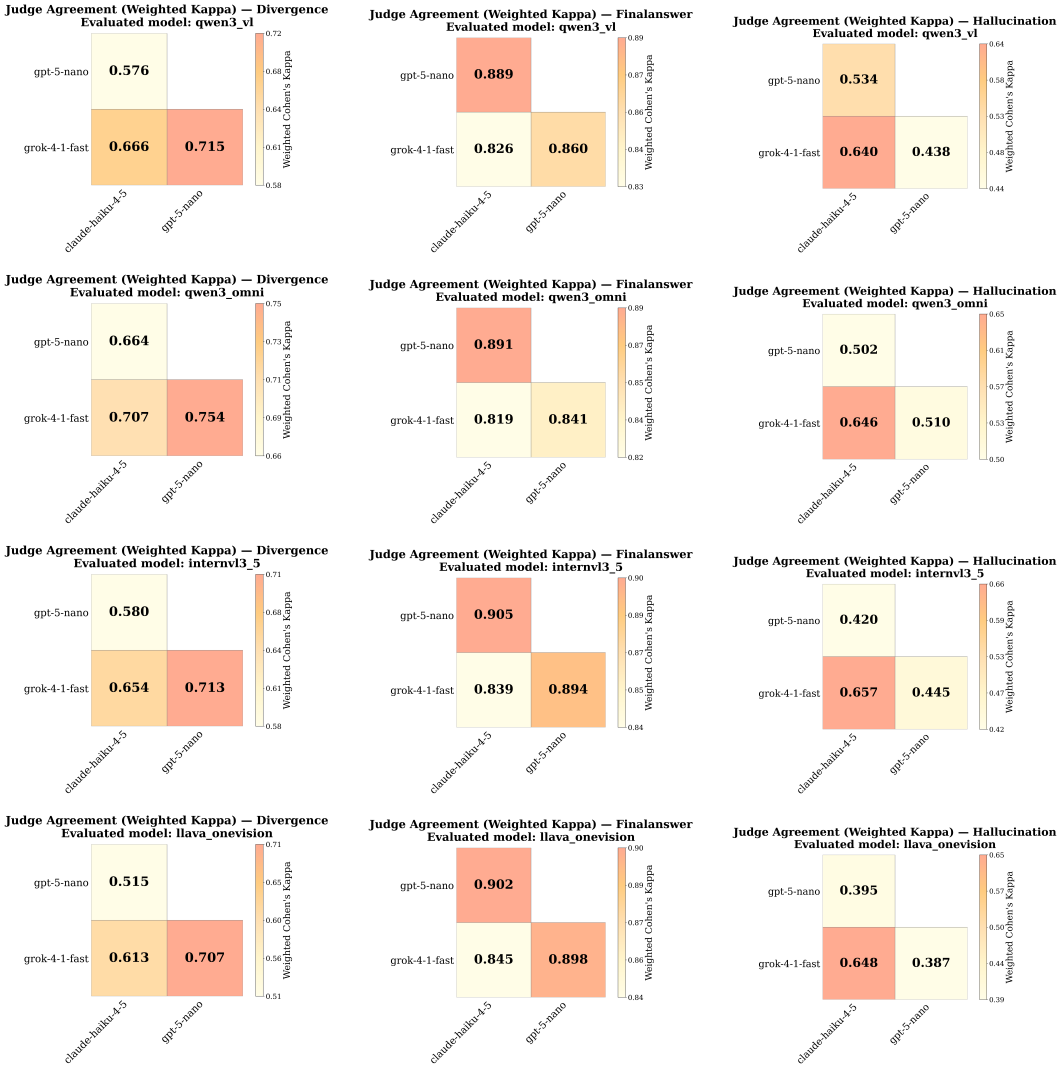


Figure 6: **Per-model inter-judge agreement measured by Cohen's κ continued.**

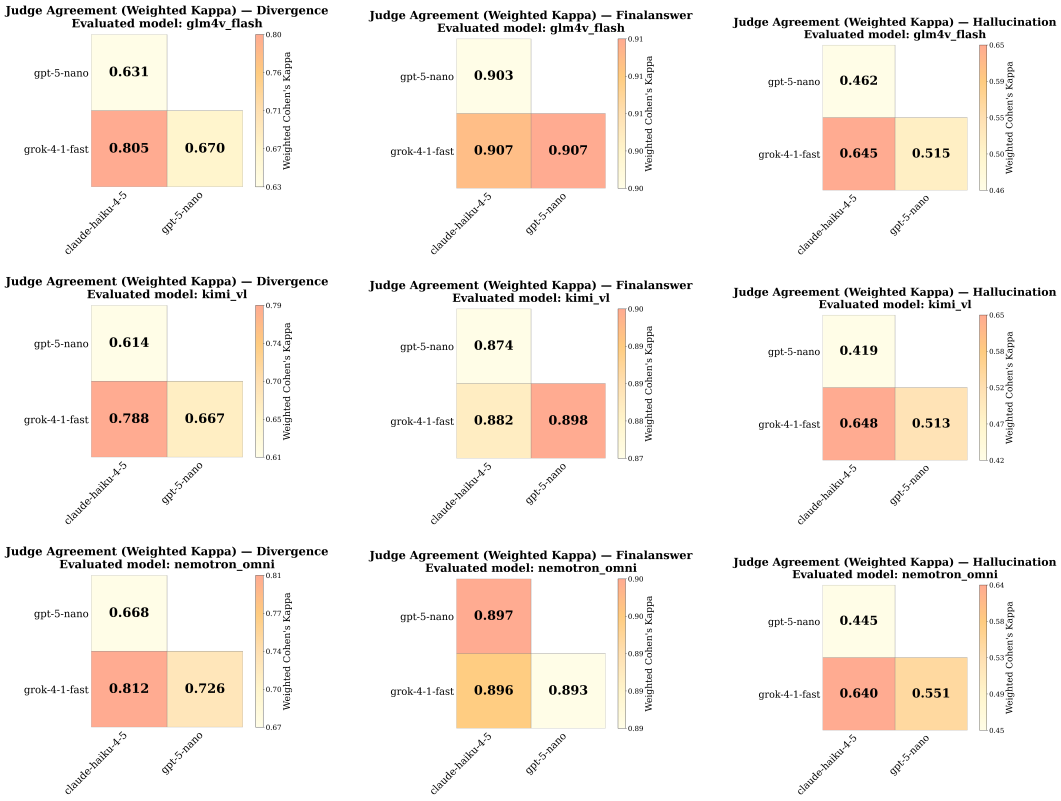


Figure 6: **Per-model inter-judge agreement measured by Cohen's κ continued.**

Facet	Model	Easy	Medium	Difficult
Final Answer	ChatGPT 5.2	4.57	4.04	3.27
	Claude 4.6 Sonnet	4.11	3.47	2.96
	Gemini-3-Flash	4.50	4.10	3.38
	Grok 4.1 Fast Reasoning	4.25	3.64	2.86
	Nova-2-Lite	4.05	3.28	2.68
	Qwen3-VL	4.47	3.56	2.88
	InternVL3.5	3.99	2.88	2.45
	LLaVA-OneVision	3.47	2.21	2.00
	Qwen3-Omni	4.37	3.44	2.78
	GLM-4.6-V-Flash	4.35	3.49	2.76
	Kimi-VL-A3B-Instruct	3.95	3.18	2.52
	Nemotron-3-Nano-Omni-30B	4.38	3.56	2.85
Conciseness	ChatGPT 5.2	3.20	3.08	3.19
	Claude 4.6 Sonnet	1.76	1.80	1.81
	Gemini-3-Flash	4.64	4.64	4.64
	Grok 4.1 Fast Reasoning	1.42	1.52	1.53
	Nova-2-Lite	2.46	2.43	2.40
	Qwen3-VL	3.84	3.81	3.77
	InternVL3.5	4.22	4.04	4.15
	LLaVA-OneVision	4.64	4.58	4.66
	Qwen3-Omni	4.26	4.23	4.24
	GLM-4.6-V-Flash	2.48	2.50	2.54
	Kimi-VL-A3B-Instruct	2.97	2.83	2.83
	Nemotron-3-Nano-Omni-30B	3.63	3.74	3.71
Tokens	ChatGPT 5.2	326	355	343
	Claude 4.6 Sonnet	560	595	566
	Gemini-3-Flash	198	208	204
	Grok 4.1 Fast Reasoning	657	683	622
	Nova-2-Lite	492	596	534
	Qwen3-VL	275	288	294
	InternVL3.5	240	269	262
	LLaVA-OneVision	195	214	199
	Qwen3-Omni	243	249	253
	GLM-4.6-V-Flash	303	346	306
	Kimi-VL-A3B-Instruct	301	365	338
	Nemotron-3-Nano-Omni-30B	201	220	212
Divergence	ChatGPT 5.2	3.22	2.92	2.63
	Claude 4.6 Sonnet	2.85	2.61	2.40
	Gemini-3-Flash	3.29	3.00	2.74
	Grok 4.1 Fast Reasoning	2.71	2.43	2.23
	Nova-2-Lite	2.84	2.45	2.21
	Qwen3-VL	3.06	2.50	2.24
	InternVL3.5	2.85	2.16	2.06
	LLaVA-OneVision	2.42	1.77	1.77
	Qwen3-Omni	3.23	2.63	2.35
	GLM-4.6-V-Flash	3.00	2.51	2.21
	Kimi-VL-A3B-Instruct	2.80	2.30	2.07
	Nemotron-3-Nano-Omni-30B	3.28	2.59	2.33
Hallucination	ChatGPT 5.2	2.19	2.11	1.98
	Claude 4.6 Sonnet	1.71	1.67	1.64
	Gemini-3-Flash	2.37	2.23	2.12
	Grok 4.1 Fast Reasoning	1.53	1.49	1.46
	Nova-2-Lite	1.88	1.75	1.63
	Qwen3-VL	2.10	1.79	1.67
	InternVL3.5	2.20	1.72	1.68
	LLaVA-OneVision	2.00	1.51	1.53
	Qwen3-Omni	2.46	2.00	1.81
	GLM-4.6-V-Flash	2.25	1.95	1.79
	Kimi-VL-A3B-Instruct	2.28	1.88	1.77
	Nemotron-3-Nano-Omni-30B	2.71	2.16	2.01

Table 6: **Model performance by facet and difficulty level.** We group QA pairs into *Easy*, *Medium*, and *Difficult* subsets using the difficulty-labeling procedure described in Section B.5, and report per-facet scores for each model across these splits. Conciseness is a token-based quintile score, and Tokens reports the average output token count.

Facet	Model	H	H+AI	AI
Final Answer	GPT-5.2	3.45	3.77	3.81
	Claude 4.6 Sonnet	3.18	3.37	3.40
	Gemini-3-Flash	3.44	3.83	3.88
	Grok 4.1 Fast Reasoning	2.70	3.39	3.48
	Nova-2-Lite	2.90	3.18	3.22
	Qwen3-VL	3.07	3.45	3.50
	Qwen3-Omni	3.05	3.35	3.39
	InternVL3.5	2.36	2.99	3.04
	LLaVA-OneVision	2.08	2.45	2.48
	GLM-4.6-V-Flash	3.00	3.33	3.37
	Kimi-VL-A3B-Instruct	2.71	3.04	3.08
	Nemotron-3-Nano-Omni-30B	3.11	3.41	3.45
Conciseness	GPT-5.2	2.91	2.72	2.69
	Claude 4.6 Sonnet	1.56	1.50	1.49
	Gemini-3-Flash	4.09	4.47	4.44
	Grok 4.1 Fast Reasoning	1.59	1.30	1.30
	Nova-2-Lite	2.50	2.01	1.96
	Qwen3-VL	3.43	3.40	3.41
	Qwen3-Omni	3.92	3.90	3.91
	InternVL3.5	4.15	3.48	3.45
	LLaVA-OneVision	4.47	4.25	4.24
	GLM-4.6-V-Flash	2.02	2.52	2.58
	Kimi-VL-A3B-Instruct	2.54	2.83	2.86
	Nemotron-3-Nano-Omni-30B	3.51	3.65	3.66
Tokens	GPT-5.2	339	342	343
	Claude 4.6 Sonnet	553	571	574
	Gemini-3-Flash	200	203	204
	Grok 4.1 Fast Reasoning	653	644	643
	Nova-2-Lite	435	545	559
	Qwen3-VL	298	287	285
	Qwen3-Omni	279	291	293
	InternVL3.5	250	283	285
	LLaVA-OneVision	219	202	201
	GLM-4.6-V-Flash	316	310	309
	Kimi-VL-A3B-Instruct	315	337	340
	Nemotron-3-Nano-Omni-30B	207	210	210
Divergence	GPT-5.2	2.86	2.83	2.83
	Claude 4.6 Sonnet	2.52	2.56	2.56
	Gemini-3-Flash	2.85	2.93	2.94
	Grok 4.1 Fast Reasoning	2.25	2.39	2.41
	Nova-2-Lite	2.39	2.42	2.43
	Qwen3-VL	2.44	2.50	2.51
	Qwen3-Omni	2.54	2.66	2.67
	InternVL3.5	1.95	2.34	2.37
	LLaVA-OneVision	1.80	1.96	1.97
	GLM-4.6-V-Flash	2.40	2.48	2.48
	Kimi-VL-A3B-Instruct	2.23	2.31	2.32
	Nemotron-3-Nano-Omni-30B	2.49	2.64	2.66
Hallucination	GPT-5.2	2.10	2.05	2.04
	Claude 4.6 Sonnet	1.65	1.66	1.66
	Gemini-3-Flash	2.17	2.20	2.20
	Grok 4.1 Fast Reasoning	1.43	1.48	1.49
	Nova-2-Lite	1.74	1.72	1.72
	Qwen3-VL	1.79	1.80	1.80
	Qwen3-Omni	1.88	2.04	2.06
	InternVL3.5	1.65	1.87	1.89
	LLaVA-OneVision	1.57	1.67	1.67
	GLM-4.6-V-Flash	1.86	1.94	1.95
	Kimi-VL-A3B-Instruct	1.82	1.93	1.94
	Nemotron-3-Nano-Omni-30B	2.07	2.22	2.24

Table 7: **Model performance by facet and annotation source.** We report per-facet scores for each model on the human-only (**H**), human+AI (**H+AI**), and AI-only (**AI**) subsets. Conciseness is a token-based quintile score, and Tokens reports the average output token count.

Table 8: Taxonomy percentages of the QA pairs from our dataset

Section	Top category	Subcategory	Percentage (%)
Top-level	Information Query	–	69.26
Top-level	Sports	–	28.80
Top-level	Gaming	–	1.05
Top-level	Shopping	–	0.33
Top-level	Sentiment	–	0.22
Top-level	Egocentric Agent	–	0.19
Top-level	Social Interaction	–	0.15
Subcategory	–	Sports & Adventure	29.67
Subcategory	–	Movies, TV & Animation	28.31
Subcategory	–	Pets & Animals	19.05
Subcategory	–	Science & Knowledge	10.64
Subcategory	–	Life & Practical Skills	7.19
Subcategory	–	Humanities & Society	2.48
Subcategory	–	Technology & Gaming	1.39
Subcategory	–	Social Trends & Reactions	0.70
Subcategory	–	Business & Commerce	0.57
Joint	Sports	Sports & Adventure	28.80
Joint	Information Query	Movies, TV & Animation	28.26
Joint	Information Query	Pets & Animals	19.04
Joint	Information Query	Science & Knowledge	10.64
Joint	Information Query	Life & Practical Skills	7.00
Joint	Information Query	Humanities & Society	2.48
Joint	Gaming	Technology & Gaming	1.05
Joint	Information Query	Sports & Adventure	0.87
Joint	Information Query	Social Trends & Reactions	0.41
Joint	Information Query	Technology & Gaming	0.33
Joint	Shopping	Business & Commerce	0.33
Joint	Information Query	Business & Commerce	0.25
Joint	Egocentric Agent	Life & Practical Skills	0.18
Joint	Sentiment	Social Trends & Reactions	0.15
Joint	Social Interaction	Social Trends & Reactions	0.14
Joint	Sentiment	Movies, TV & Animation	0.05
Joint	Egocentric Agent	Technology & Gaming	0.01
Joint	Sentiment	Pets & Animals	0.01
Joint	Social Interaction	Life & Practical Skills	0.01