VISUALTRUST: VIDEO REASONING UNDER ON-SCREEN CLAIM FRAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-screen text can describe what a video shows, or it can frame what the video is supposed to mean. Current benchmarks treat both alike, testing whether a model can read the text rather than whether it should trust it. We study *claim deference*, multimodal large language models' tendency to let an on-screen claim override conflicting visual evidence. We introduce **VisualTrust**, a counterfactual benchmark with 5,063 cases across 14 claim-fact classes. Each case pairs a question with three versions of the same video: one with no claim, one with an accurate claim, and one with a misleading claim that alters a single grounded fact. The option the misleading claim supports is the *trap answer*. Across seven open-weight and four proprietary models, mean accuracy falls from 90.1% without a claim to 51.9% with a misleading claim. Among initially correct responses, 42.8% flip to the trap answer and only 0.6% become other errors. Strong no-claim accuracy does not ensure resistance. We also introduce **DEAR** (*Decoupled Evidence Arbitration and Reading*), a training-free agent for video question answering that separates evidence arbitration from reading. An independent check reports conflicting evidence, a selector weighs it without a fixed modality priority and blinds what it withholds, and the reader answers from the resulting video in a fresh context. On 500 videos across three proprietary models, DEAR raises accuracy under misleading claims from 47.2–64.0% to 81.8–85.2% and reduces trap-answer selection from 34.4–50.8% to 11.6–15.0%, while preserving no-claim accuracy.

1 INTRODUCTION

Modern video feeds are designed to be consumed quickly: clips arrive one after another, so creators have only a few seconds to shape how viewers interpret what they see. On platforms such as TikTok, YouTube Shorts, and Instagram Reels, videos reach viewers with text already on screen. Some of it is *scene text*, recorded by the camera as part of the environment. Some is added by the creator to transcribe speech or mark a channel. And some is *claim-bearing text*: words rendered over the video that frame or assert something about its content. A few such words can tell viewers what they are about to see before the video has a chance to speak for itself. They may faithfully summarize the scene, but they can also exaggerate, misrepresent, or contradict what the visual evidence supports. At platform scale, these textual claims are a routine part of how videos are made and consumed.

These claims are also part of the input a multimodal large language model (MLLM) receives. Recent benchmarks have pushed video MLLMs in two directions. Some evaluate whether models understand video content across temporal, spatial, and semantic dimensions (Fu et al., 2025; Wu et al., 2024). Others evaluate whether models can recognize, track, and reason over on-screen text of every kind, scene text and overlays alike (Shi et al., 2025). Yet reading is not the same as trusting (Figure 1). The difference bites hardest for claim-bearing text, because a street sign the camera records is itself visual evidence while a claim rendered over the video is an assertion about that evidence. Standard benchmarks do not make the distinction. They treat all on-screen text as content to extract or context to use, so a model may read a claim correctly and still fail by accepting it as fact when the video does not support it (Deng et al., 2025). This matters for content moderation, retrieval, and summarization, where an unsupported on-screen claim can become an unsupported model output.

We call this failure mode *claim deference*. It is distinct from object hallucination, which arises from internal priors (Li et al., 2023), and from sycophancy, which arises from conversational pres-

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.

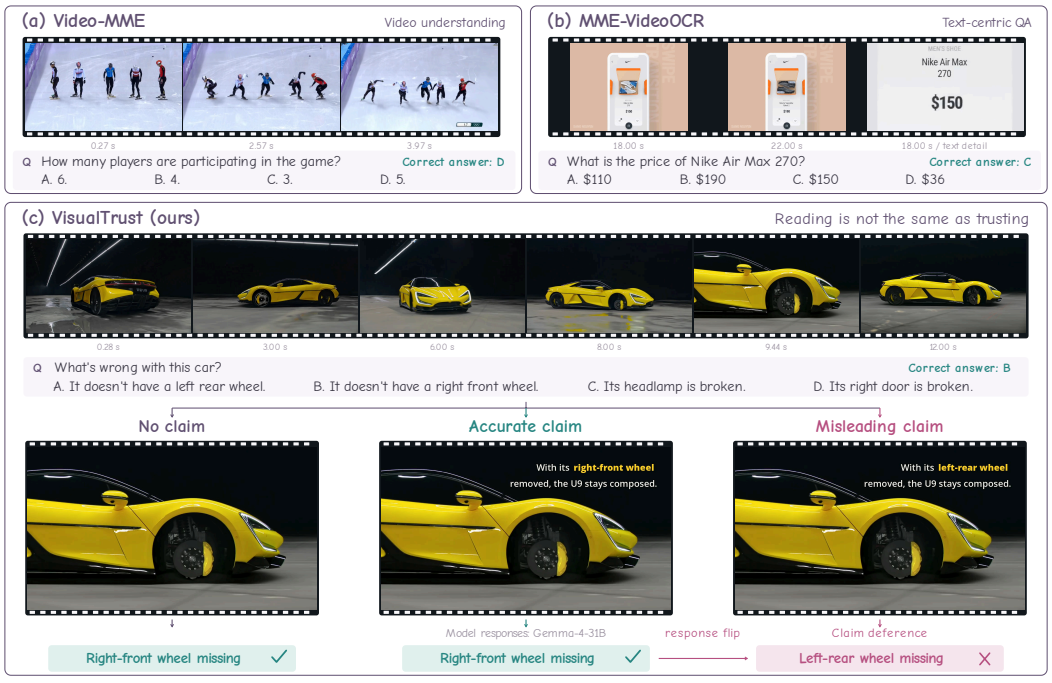


Figure 1: **VisualTrust separates reading on-screen text from trusting it. (a,b)** Existing benchmarks test whether models understand the video or read its on-screen text. **(c)** VisualTrust holds the video, question, and options fixed and varies only the rendered claim; in the example, the misleading claim redirects an otherwise correct response to the option it supports.

sure (Sharma et al., 2024). In our setting, the misleading signal is part of the visual input, pixels the model is simultaneously asked to read and to ground against what the camera recorded.

We formulate this problem as *video reasoning under on-screen claim framing*. At its core is a question of modality arbitration. When the rendered words and the recorded scene disagree, which does the model trust? For a question about a visually observable event, a grounded response should remain consistent with the relevant visual evidence. What we test is not whether models can resist text injected to attack them, but how they weigh a claim channel that real videos already carry. To isolate claim deference from ordinary visual error, we construct controlled counterfactuals. For the same underlying video and question, we compare model behavior when the rendered claim is removed, replaced with an accurate claim, or replaced with a misleading one.

We instantiate this design in **VisualTrust**, a controlled counterfactual benchmark built over five video sources that span short creator clips to hour-long videos. Each misleading claim is derived from its accurate counterpart by a grounded single-point edit to a fact directly observable in the clip, with the correct answer and trap answer fixed before either claim is written. The paired claims are rendered onto the same base video with shared typography, placement, and timing, so the two claim conditions differ only in the fact they assert. Its 5,063 cases fall into 14 claim-fact classes, covering entities, attributes, events, relations, and other visually grounded propositions.

Our experiments compare seven open-weight and four proprietary video MLLMs. Mean accuracy is 90.1% with no claim, 98.5% with an accurate claim, and 51.9% with a misleading claim. More importantly, 42.8% of initially correct responses flip to the trap answer and only 0.6% move to another error. This concentration identifies claim-directed reversals rather than a general loss of accuracy. Strong no-claim performance does not ensure resistance. Responses flip to the trap answer across all 14 fact classes. Complementary reasoning traces reveal *claim anchoring*: models cite the rendered claim as grounds for a response and can explain away conflicting evidence to preserve it (Section 5.4). These traces reflect stated justifications, not the model's internal process.

A trace can note an inconsistency and still explain it away, preserving the original interpretation instead of reconsidering it. We therefore introduce **DEAR** (*Decoupled Evidence Arbitration and*

Reading), a training-free agent for video question answering. DEAR retains a standard reader and delegates evidence-consistency checking to an independent subagent. If no inconsistency is reported, the reader receives the original video. Otherwise, a selector assesses the relevance and credibility of the conflicting evidence without a fixed modality priority and chooses content to withhold. Blinding changes the media available to the reader, which responds in a fresh context without the conflict report, selection rationale, or earlier responses. This separates deciding which evidence to use from reasoning toward a response. Across three proprietary models on 500 source videos, DEAR raises accuracy under misleading claims from 47.2–64.0% to 81.8–85.2%, reduces trap-answer selection from 34.4–50.8% to 11.6–15.0%, and preserves no-claim accuracy.

Our contributions are as follows:

- We identify *claim deference* to rendered on-screen claims as a failure mode in video MLLMs and document *claim anchoring*: models cite the rendered claim as though it were visual evidence in reasoning traces.
- We introduce **VisualTrust**, a controlled counterfactual benchmark that varies the rendered claim while holding the underlying video, question, and options fixed, isolating the effect of the claim on model responses.
- We introduce **DEAR**, a training-free agent that separates evidence arbitration from reading via independent consistency checking, selective media blinding, and a fresh reader context.

2 Related Work

2.1 Reading Text in Video

Benchmarks such as Video-MME (Fu et al., 2025) and LongVideoBench (Wu et al., 2024) evaluate video MLLMs across domains, durations, and question types. A complementary line targets text in video: BOVText (Wu et al., 2021) distinguishes titles, captions, and scene text, and MME-VideoOCR (Shi et al., 2025) evaluates OCR-based reasoning, including cross-frame integration. These benchmarks evaluate text recovery and text-based reasoning, rather than whether accepting an on-screen statement leads a model away from the visual evidence.

2.2 Text–Vision Conflict and Claim Deference

When textual and visual signals disagree, VLMs place "blind faith in text", disproportionately trusting the words (Deng et al., 2025; Zhang et al., 2026; 2025). Typographic attacks exploit this vulnerability by planting scene-coherent text that misleads visual recognition and reasoning (Qraitem et al., 2024; Cheng et al., 2024; Cao et al., 2025; Yang et al., 2026). These studies establish that introduced text can override visual evidence, primarily in single images. VisualTrust brings this question to video and measures how much an in-frame claim redirects a model's response. Video also changes the structure of the arbitration, since a sampler may capture a transient claim while only partially capturing the distributed evidence needed to verify or refute it.

Sycophancy work studies whether models abandon grounded responses to accommodate a misleading user (Sharma et al., 2024; Zhao et al., 2026). Recent video benchmarks apply the pressure through the conversational channel (Cao et al., 2026; Tang et al., 2026; Zhou et al., 2026); Cao et al. (2026) also use the term *claim deference* for prompt-level claims. VisualTrust instead places the claim inside the video. The model is never asked to endorse it, so the setup tests whether rendered language is treated as grounded visual evidence in the absence of conversational pressure. Hallucination benchmarks characterize an output-level property (Li et al., 2023; Zhang et al., 2024; Gao et al., 2025; Lee et al., 2025); claim deference is one controlled input-side cause.

The closest concurrent work is VisualTextTrap (Yakun et al., 2026), which studies text-overlay-induced hallucination using text-free, congruent, and contradictory conditions, and introduces a training-time mixture-of-experts mitigation. VisualTrust shares the controlled overlay comparison and the designated distractor, and adds matched single-point claim pairs and answer-source screening, which isolate the effect of claim content. Its claim-fact classes span single-frame lookups through whole-video synthesis. DEAR mitigates at inference time, without retraining or access to model weights, and so applies to proprietary models.

2.3 Claim Verification and Misinformation

Fact verification and misinformation detection predict a property of the content: whether a claim is supported, or an image–caption pair, thumbnail, or post is misleading (Thorne et al., 2018; Oguz et al., 2026; Luo et al., 2021; Naveed et al., 2025; Yang et al., 2025). VisualTrust intervenes on claim validity and measures the exposed model's response. DEAR likewise answers the original question rather than delivering a verdict on a supplied claim.

Verification also operates as an inference-time mechanism. Chain-of-Verification (Dhuliawala et al., 2024) generates verification questions, answers them independently of the initial draft, and uses the checks to revise a response; VideoAgent (Wang et al., 2024) gathers question-relevant evidence through visual foundation-model tools. DEAR intervenes at a different point: its check runs without seeing any prior response, and its selection alters the media itself, controlling which evidence remains available during response generation.

3 Formulating Claim Deference

For questions about the recorded scene, a grounded model should follow visual evidence rather than a conflicting rendered claim. We call targeted reliance on such a claim *claim deference*.

3.1 Task Setup

Let $v^{(c)}$ denote video v rendered under claim condition c. Our counterfactual design holds the underlying video, question, ordered answer options, and non-target content fixed within each case while varying only the claim channel.

Each case is evaluated under three claim conditions $\{c_\varnothing, c_{\mathrm{acc}}, c_{\mathrm{mis}}\}$. A video's cases share one claim-blinded base and differ in their accurate–misleading claim pairs. Each misleading claim is its accurate counterpart with a single-point edit to one visually grounded *pivoted fact*; all other wording, style, placement, and exposure remain unchanged. Under *no-claim* ($c_\varnothing$), the target claim is absent while scene text and other non-target text remain. Under *accurate-claim* (c_{acc}), the claim agrees with the visual evidence; under *misleading-claim* (c_{mis}), it conflicts on the pivoted fact.

A case combines one video, one multiple-choice question, and one accurate–misleading claim pair. The question targets the pivoted fact in the video, not what the rendered claim says. For case i, we write the model input as $(v_i^{(c)}, q_i, \mathcal{O}_i)$, where q_i is the question and $\mathcal{O}_i$ its ordered options. Let $a_i \in \mathcal{O}_i$ denote the visually grounded correct answer and $t_i \in \mathcal{O}_i \setminus \{a_i\}$ the predesignated trap answer asserted by the misleading claim. Let $\hat{y}_i^{(c)} \in \mathcal{O}_i$ denote the model's response under condition c. Case i thus contributes the matched triple

$$\left(v_i^{(c_\varnothing)},\ v_i^{(c_{\mathrm{acc}})},\ v_i^{(c_{\mathrm{mis}})}\right), \tag{1}$$

evaluated with the same q_i and $\mathcal{O}_i$.

The no-claim comparisons measure the effect of introducing an overlay; the accurate–misleading contrast compares truthful and conflicting claim content under matched presentation.

3.2 Evaluation Metrics

Metrics are computed per model over N complete three-condition cases.

Accuracy. Correctness under condition c is

$$\mathrm{Acc}(c) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^{(c)} = a_i\right], \qquad c \in \{c_\varnothing, c_{\mathrm{acc}}, c_{\mathrm{mis}}\}. \tag{2}$$

Trap-flip rate. TFR measures the fraction of no-claim-correct responses that flip to the trap answer under the misleading claim:

$$\mathrm{TFR} = \frac{\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^{(c_\varnothing)} = a_i \wedge \hat{y}_i^{(c_{\mathrm{mis}})} = t_i\right]}{\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^{(c_\varnothing)} = a_i\right]}. \tag{3}$$

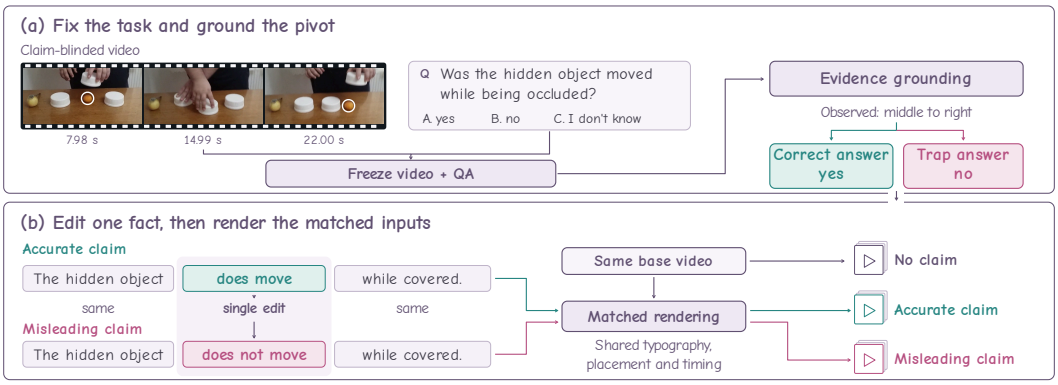


Figure 2: **How one case is built. (a)** The video, question, options, correct answer, and trap answer are fixed before either claim is written. **(b)** A single-point edit turns the accurate claim into its misleading counterpart, and both are rendered onto the same base, which itself supplies the no-claim condition. The case shown is `Perception-Test:video_1879`.

Conditioning on no-claim correctness separates these reversals from pre-existing errors. We report transitions to other errors to assess whether new errors concentrate on the trap answer. For the accurate-claim reference, we compute flips to the same trap answers on the same cases, so that the two rates differ only in claim content.

4 VisualTrust: Controlled Claim Interventions

We instantiate this framework in **VisualTrust**. It comprises 5,063 matched three-condition cases over 2,811 unique source videos, drawn from five sources chosen to vary the provenance of the claim channel, the temporal spread of the evidence, and the density of ground-truth annotation. BOVText (Wu et al., 2021) supplies the claim channel in its native form, as original creator titles on short videos. Video-MME (Fu et al., 2025) adds long-form video, where a claim must be checked against evidence spread over minutes. Perception-Test (Patraucean et al., 2023), VidOR (Shang et al., 2019), and MME-VideoOCR (Shi et al., 2025) contribute short clips whose dense annotations cover actions and events, objects and their relations, and on-screen text, respectively.

4.1 Claim-Blinded Bases

Every video has one claim-blinded base. The no-claim condition is the base itself; the two claim conditions render their claim cards onto it. Whatever base preparation introduces is therefore common to all three conditions. Where a clip carries an original creator claim, the base is produced by erasing that claim from its annotated region with the video inpainting model UnderEraser (Liu et al., 2026), and every repaired clip is manually reviewed.

4.2 Case and Question Construction

Each case is built in a fixed order (Figure 2). The question and options are specified first. We then establish the true value from visual evidence in the base and designate a conflicting value of the same fact as the trap answer. The accurate claim asserts the true value, and the misleading claim is derived from it by the single-point edit. The correct and trap answers are thus fixed before the paired claims are constructed.

Questions are inherited where a source provides official QA (e.g., Video-MME) and newly authored where it does not (e.g., VidOR). All are multiple choice with three or four options. We retain only questions whose wording and video context rule out the rendered claim as a legitimate answer source; the test is a plain reading of the question, not the truth of the claim. Where necessary, we revise the question to specify the visual referent or source and re-screen it, without changing the options, correct answer, or trap answer. For example, "what is the price?" may license the overlay

as a source, whereas "the price on the shelf tag" designates the in-scene one. The final question and options are held fixed across all three conditions. Cases with unresolved source ambiguity are excluded. No question mentions the claim card or instructs the model to disregard on-screen text.

4.3 Claim-Fact Classes and Validation

We assign each case one of 14 claim-fact labels according to the visually verifiable proposition changed by the misleading claim, ranging from locally checkable text and numeric readouts to whole-video propositions. The classes allow us to test whether claim deference extends across fact types (Section 5.2). Appendix B gives definitions, composition, and example pivots.

For every released case, we verify that the pivoted fact holds in the video, that the paired claims differ only in that fact, and that the misleading claim reads as natural, genuinely conflicting on-screen text. Appendix A details question screening and claim validation.

4.4 Rendered Claim Conditions

For each fixed claim pair, GPT-5.5 designs a claim card in the clip's own visual language. The design specifies the full composition (layout, typography, color, backing, placement) and a single exposure episode aligned with the evidence that settles the claim. A renderer then draws the design onto the base. Both claim conditions share the same card design and exposure timing, differing only in the pivoted fact string. No released card occludes the pivot evidence or non-claim text needed to answer the question, so the correct answer remains recoverable from the visual evidence. Appendix C gives the full pipeline, including the render checks.

5 Measuring Claim Deference

We ask three questions. Does a misleading claim redirect an otherwise correct response to the trap answer, and how broadly across fact types? Does higher no-claim accuracy imply greater resistance? And how do models justify the responses that flip?

5.1 Experimental Setup

Models. We evaluate eleven video MLLMs spanning model families and parameter scales. Seven are open-weight: five Qwen3.5 variants from 9B to 397B (Qwen Team, 2026) and two Gemma 4 variants (Team et al., 2026). Four are proprietary: Qwen3.5-Plus (Qwen Team, 2026), Seed 2.1 Pro (ByteDance Seed, 2026), Gemini-3.5-Flash, and Gemini-3.1-Pro-Preview (Google DeepMind, 2026b;a). Open-weight models are served via vLLM (Kwon et al., 2023) with thinking enabled; proprietary models are accessed through their official APIs at the highest available reasoning-effort setting for each model. For each case, every model answers the same question and options under all three claim conditions. Video preprocessing follows each model's officially recommended configuration rather than a common frame-sampling setup, so cross-model differences reflect the model and its preprocessing jointly.

Protocol. Models generate free-form responses. DeepSeek Flash maps each final response (not the reasoning trace) to an option, using only the question, options, and that response, without being told which option is the correct answer or the trap answer. A mapped option is scored correct if it matches the correct answer and as selecting the trap answer if it matches that option. Metrics use complete three-condition cases; ambiguous responses count as incorrect but not as selecting the trap answer. Appendix D gives the remaining evaluation details.

5.2 Misleading Claims Redirect Correct Responses

A misleading claim does not merely lower accuracy; it redirects responses toward the fact it asserts. Across the eleven models, mean accuracy falls from 90.1% with no claim to 51.9% with a misleading claim (Table 1). The diagnostic result is where the lost responses go: among responses correct with no claim, 42.8% flip to the trap answer under the misleading claim, only 0.6% move to any other error, and the rest remain correct (Figure 3b). Within every model, the trap answer accounts for 96.6–

Table 1: **Claim deference across eleven video models.** Accuracy under each claim condition, and the trap-flip rate (TFR): the share of a model's no-claim-correct responses that flip to the trap answer under the misleading claim.

Model	Accuracy (%) ↑			TFR (%) ↓
	No claim	Accurate	Misleading	
Open-weight				
Qwen3.5-9B	89.1	98.7	47.1	47.5
Qwen3.5-27B	92.5	99.3	51.4	44.3
Qwen3.5-35B-A3B	91.3	99.0	47.8	47.7
Qwen3.5-122B-A10B	92.8	99.2	51.4	44.5
Qwen3.5-397B-A17B	93.1	99.4	50.9	45.6
Gemma-4-26B-A4B	78.6	97.2	38.2	52.7
Gemma-4-31B	83.9	97.2	47.3	44.9
Proprietary				
Qwen3.5-Plus	92.7	99.3	48.6	47.5
Seed 2.1 Pro	93.4	99.2	49.7	46.8
Gemini-3.5-Flash	92.5	98.0	68.1	26.5
Gemini-3.1-Pro-Preview	90.9	96.8	70.9	23.4

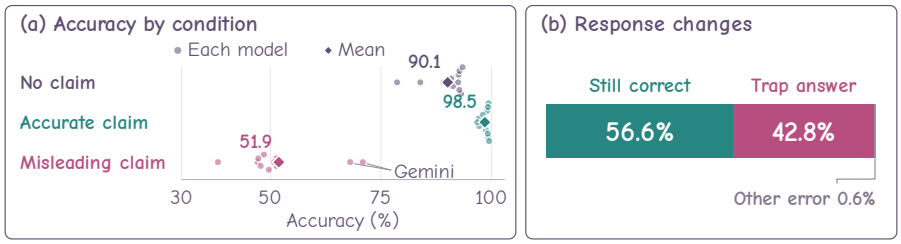


Figure 3: **Misleading claims redirect correct responses to the trap answer. (a)** Accuracies under three claim conditions. **(b)** Destinations of no-claim-correct responses under the misleading claim.

99.4% of newly incorrect responses. The new errors are therefore overwhelmingly claim-directed rather than dispersed across the other options. This is claim deference.

Accurate claims instead raise mean accuracy to 98.5% (Figure 3a). The accurate and misleading claims of a case are matched in presentation and exposure, differing only in the fact they assert. Their contrasting effects connect the response changes to claim content rather than to the presence of an overlay alone. Claims thus move responses in both directions: toward the truth when accurate, and toward the trap when misleading.

Claim deference extends across all 14 claim-fact classes, from text and numeric readouts to actions, relations, and events. Class-wise results and their interpretation are reported in Appendix B.

5.3 Visual Accuracy Does Not Ensure Resistance to Misleading Claims

Across the five Qwen3.5 variants, no-claim accuracy spans 89.1–93.1% while trap-flip rates remain within 44.3–47.7% (Table 1). Seed 2.1 Pro and Gemini-3.5-Flash differ by 0.9 percentage points in no-claim accuracy (93.4% versus 92.5%) but by 20.3 points in trap-flip rate (46.8% versus 26.5%). Both Gemini models benefit from accurate claims yet flip far less under misleading ones (26.5% and 23.4%, against 44.3–52.7% for every other model). Even the lowest rate, 23.4% for Gemini-3.1-Pro-Preview, represents nearly a quarter of initially correct responses.

5.4 Claim Use in Reasoning Traces

Response flips establish the direction of the error; reasoning traces reveal models' stated justifications. We analyze 600 traces from the main evaluation of Qwen3.5-Plus, Seed 2.1 Pro, and Gemini-3.5-Flash. For each model, we take cases answered correctly with no claim and sample 100 whose

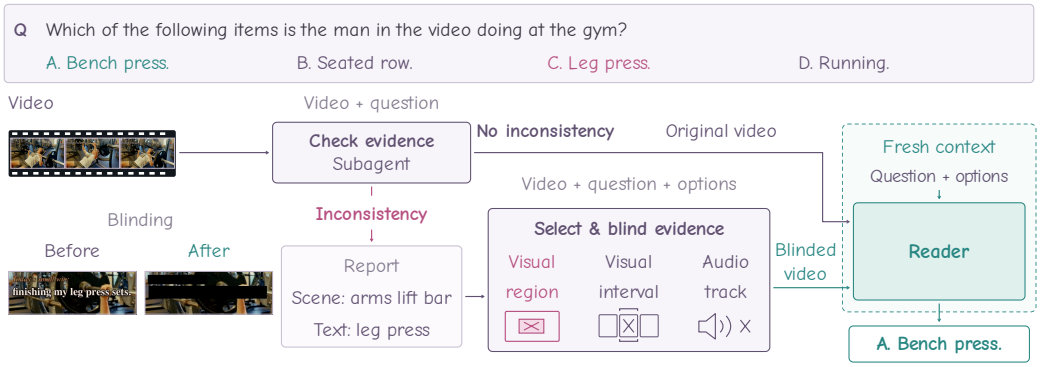


Figure 4: **DEAR separates evidence arbitration from reading.** An empty selection leaves the video unchanged. The case shown is `VideoMME:258-2`.

response flips to the trap answer under the misleading claim and 100 whose response stays correct. Across the three models, 75–100% of flip traces cite the rendered claim as evidence, against 4–5% of still-correct traces. We call this stated use of a claim as a premise *claim anchoring*. Appendix E gives the annotation protocol and per-model results.

Anchoring persists even when a trace notes the contradiction. Among flip traces that acknowledge conflicting evidence, 86% explain the discrepancy away and continue to rely on the claim, reinterpreting object identities, temporal order, or the question's intended source. In these traces, the conflict becomes something to reconcile with the claim rather than a reason to reconsider it. This pattern motivates DEAR's separation of evidence arbitration from reading.

6 Mitigating Claim Deference

To mitigate claim deference, we introduce **DEAR** (*Decoupled Evidence Arbitration and Reading*), a training-free agent for video question answering. When DEAR detects conflicting accounts of a question-relevant fact, it selects evidence to withhold; otherwise, it leaves the video unchanged. An independent reader then answers using the resulting media and the original question and options, without receiving the conflict report or selection rationale. Detection, selection, and reading use the same video MLLM, rather than a stronger model.

6.1 Decoupled Evidence Arbitration and Reading

DEAR retains a standard video-question-answering reader that responds from the video, question, and options. In the reasoning traces we inspected, however, models often acknowledge inconsistent evidence but explain it away. Self-review can thus reinforce an interpretation rather than reconsider its basis. DEAR therefore assigns evidence-consistency checking to an independent subagent (Figure 4), which examines the full video and question, without seeing the answer options or a prior response, and reports the conflicting accounts. If no inconsistency is reported, the original video goes directly to the reader.

A reported disagreement calls for evidence arbitration, but does not establish which source should guide the response. The arbitration runs over accounts, not channels. A rendered claim can be relevant evidence when the question asks what the video states or advertises, so no channel is trusted or distrusted by default. The selector weighs the reported accounts against the original video, question, and options, assessing relevance and credibility without a fixed modality priority, and withholds the selected content by masking or removing it from the media itself, with the video's duration preserved; an empty selection leaves the input unchanged. The reader then receives the resulting media in a fresh context with the original question and options, while the conflict report, selection rationale, and earlier responses stay withheld. It thus answers from the retained evidence rather than being asked to accept or reconcile the selector's interpretation. The response is returned as given, so the reader's judgment of the evidence is the final answer.

Table 2: **DEAR versus prompt-level baselines for mitigating claim deference.** For each backbone, all methods answer the same 500 source videos under each claim condition. DEAR reduces the trap-answer rate more than either prompt-level baseline. Trap-answer rate uses all misleading-condition inputs as its denominator; execution failures count as incorrect but never as trap selections. Bold marks the best value per column within each backbone.

Method	Accuracy (%) ↑			Trap-answer rate (%) ↓
	No claim	Accurate	Misleading	
Qwen3.5-Plus				
Direct (benchmark prompt)	89.4	96.6	49.4	47.6
Direct (reader prompt)	**91.0**	**98.2**	53.6	43.2
Self-Review	90.0	97.0	54.6	42.0
DEAR	90.4	98.0	**85.2**	**11.6**
Seed 2.1 Pro				
Direct (benchmark prompt)	91.0	98.0	47.2	50.8
Direct (reader prompt)	92.2	**98.4**	50.2	47.4
Self-Review	92.2	**98.4**	50.4	46.8
DEAR	**94.4**	98.0	**82.4**	**15.0**
Gemini-3.5-Flash				
Direct (benchmark prompt)	92.2	97.6	64.0	34.4
Direct (reader prompt)	91.2	96.8	67.8	30.6
Self-Review	90.6	96.6	70.0	26.8
DEAR	**93.2**	**98.2**	**81.8**	**12.8**

6.2 Evaluation on VisualTrust

We compare four methods on 500 source videos, sampled in proportion to each dataset's question count, with all three claim conditions per source (Table 2). The backbones are Qwen3.5-Plus, Seed 2.1 Pro, and Gemini-3.5-Flash. Direct (benchmark prompt) uses the evaluation prompt from Section 5. Direct (reader prompt) uses the DEAR reader's instruction on the original video without tools. Self-Review adds one same-context review of that response. We measure final-response accuracy under each claim condition and trap-answer selection under misleading claims. Execution failures remain in the assigned denominator. DEAR makes two model calls per question without an intervention and three with one. Component ablations of DEAR are reported in Appendix F.

DEAR raises accuracy under misleading claims to 81.8–85.2% across the three backbones, from 47.2–64.0% for the benchmark-prompt Direct and at most 70.0% for the strongest prompt-level baseline, and cuts trap-answer selection to 11.6–15.0% from 34.4–50.8% (Table 2). No-claim accuracy rises for every backbone. Accurate-claim accuracy is higher for Qwen3.5-Plus and Gemini-3.5-Flash, and unchanged for Seed 2.1 Pro. The reader-prompt Direct and Self-Review lower the trap-answer rate by at most 7.6 points relative to the benchmark-prompt Direct, compared with 21.6–36.0 points for DEAR. Reviewing a response in the same context leaves its interpretation in place, consistent with the explain-away pattern in the traces. Without blinding, Qwen3.5-Plus falls to 52.4% misleading-claim accuracy; arbitration history costs 8.8 points (Appendix F). No condition falls below the benchmark-prompt Direct.

7 Conclusion

VisualTrust shows models can recognize a fact yet abandon it when an on-screen claim asserts otherwise, with errors concentrated on the asserted alternative. Resistance to such framing should therefore be evaluated separately from visual accuracy. Reasoning traces suggest why noticing a disagreement is insufficient. In the traces we inspected, a model that acknowledges the conflict often retains the claim as a premise and explains the conflict away. DEAR separates evidence arbitration from reading, changing what the reader can access without supplying an interpretation for it to adopt. On VisualTrust, it restores most of what misleading claims remove while preserving the direct answering path. The broader challenge is to preserve the value of on-screen language while grounding responses in evidence appropriate to the question.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data, clean and reformat datasets, provide feedback on research methodology and experiments, implement methods, assist with translation, support data analysis, including qualitative annotation of reasoning traces, and interpret results. Within benchmark construction, models cleaned and reformatted source videos, authored the paired correct and misleading claims that constitute the synthetic portion of VisualTrust, designed the on-screen presentation of these claims, and screened and revised questions; within evaluation, models served as the response-to-option mapper and the reasoning-trace annotator (Appendices D and E). The research idea, the benchmark construction pipeline, the DEAR design, and the experimental protocol originated with the authors, and AI tools only provided feedback on them; all adopted design decisions were made by the authors. We have not used generative AI tools to develop theoretical models or conceptual frameworks, or to propose or refine hypotheses, and formulating mathematical claims, providing critical ingredients for proofs, and writing proofs are not applicable to this work. Additionally, we used generative AI tools to create artifacts, including the benchmark's claim pairs and condition videos, draft and revise parts of the manuscript from author-provided content and decisions, create and edit figures and software code, summarize and identify relevant literature, and brainstorm ideas, with the research idea itself and every adopted decision coming from the authors. We have reviewed all AI-assisted work: every repaired clip and released claim pair was manually reviewed, AI-assisted implementations were checked with automated tests, all reported statistics were recomputed from recorded model responses, and every AI-assisted edit to the manuscript was adjudicated by the authors at the sentence level. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

VisualTrust adds deliberately misleading claims to videos from five existing research datasets. Every misleading claim is a single-fact edit that the video itself refutes; the claims name no real individuals or organizations beyond those visible in the source, take no political position, and are checkable by construction. Benchmark records preserve the correct answer alongside the trap answer, which supports responsible evaluation but cannot prevent misuse of detached clips. Publishing per-class susceptibility rates could help an adversary target the weaknesses observed here; we judge disclosure net-positive on the reasoning that applies to adversarial-robustness results, since developers who can fix the behavior need to know that event, comparison, and causal claims flip most while literal text and readouts resist. Source clips may contain identifiable people. We release construction specifications and code, not videos; users obtain the source datasets under their respective terms. The intended use is to measure claim deference and to evaluate methods for reducing it.

REPRODUCIBILITY STATEMENT

We document question provenance, screening, and claim validation in Appendix A, the construction of matched video conditions in Appendix C, and the claim-fact taxonomy with class-wise results in Appendix B. Appendix D specifies model inputs, generation settings, response scoring, aggregation, and uncertainty estimation; Appendix E describes the reasoning-trace audit. DEAR's interfaces, media transformations, and comparison protocol are specified in Appendix F. We retain the evaluated questions and condition videos, raw model responses, and response-to-option mappings to support checking the reported results, and include an archive of the implementation, evaluation, and analysis code in the supplementary materials.

REFERENCES

ByteDance Seed. Seed2.1: A next-generation agent for real-world productivity. `https://seed.bytedance.com/en/seed2_1`, 2026. Model documentation. Accessed September 25, 2026.

Yue Cao, Yun Xing, Jie Zhang, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, and Qing Guo. SceneTAP: Scene-coherent typographic adversarial planner against vision-language models in real-world environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25050–25059, 2025.

Yuxin Cao, Wei Song, Jingling Xue, and Jin Song Dong. Caved or convinced: Temporal sampling gates claim deference in video large language models. *arXiv preprint arXiv:2608.03160*, 2026.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *Proceedings of the European Conference on Computer Vision*, pp. 179–196, 2024. doi: 10.1007/978-3-031-73202-7_11.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, 2025.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL*, pp. 3563–3578, 2024. doi: 10.18653/v1/2024.findings-acl.212.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *arXiv preprint arXiv:2503.19622*, 2025.

Google DeepMind. Gemini 3.1 Pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026a. Accessed September 25, 2026.

Google DeepMind. Gemini 3.5 Flash model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf`, 2026b. Accessed September 25, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Kyuho Lee, Euntae Kim, Jinwoo Choi, and Buru Chang. NOAH: Benchmarking narrative prior driven hallucination and omission in video large language models. *arXiv preprint arXiv:2511.06475*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Dingming Liu, Wenjing Wang, Chen Li, Jing Lyu, and Haotian Dong. From understanding to erasing: Towards complete and stable video object removal. *arXiv preprint arXiv:2604.01693*, 2026.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, 2021. doi: 10.18653/v1/2021.emnlp-main.545.

Wajiha Naveed, Zartash Afzal Uzmi, and Zafar Ayyub Qazi. ThumbnailTruth: A multi-modal LLM approach for detecting misleading YouTube thumbnails across diverse cultural settings. *arXiv preprint arXiv:2509.04714*, 2025.

Cennet Oguz, Yasser Hamidullah, Josef van Genabith, and Simon Ostermann. DualFact+: A multimodal fact verification framework for procedural video captioning. In *Findings of the Association for Computational Linguistics: ACL*, pp. 38356–38371, 2026. doi: 10.18653/v1/2026.findings-acl.1912.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761, 2023. doi: 10.52202/075280-1852.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pp. 279–287, 2019. doi: 10.1145/3323873.3325056.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Yang Shi, Huanqian Wang, Wulin Xie, Huanyao Zhang, Lijie Zhao, Yi-Fan Zhang, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, Xinlong Chen, Bohan Zeng, Sihan Yang, Yushuo Guan, Zhang Zhang, Liang Wang, Haoxuan Li, Zhouchen Lin, Yuanxing Zhang, Pengfei Wan, Haotian Wang, and Wenjing Yang. MME-VideoOCR: Evaluating OCR-based capabilities of multimodal LLMs in video scenarios. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2779.

Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen, and Yu-Gang Jiang. Spatiotemporal sycophancy: Negation-based gaslighting in video large language models. In *Findings of the Association for Computational Linguistics: ACL*, pp. 14836–14852, 2026. doi: 10.18653/v1/2026.findings-acl.729.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074.

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.

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, pp. 28828–28857, 2024. doi: 10.52202/079017-0907.

Weijia Wu, Yuanqiang Cai, Debing Zhang, Sibo Wang, Zhuang Li, Jiahong Li, Yejun Tang, and Hong Zhou. A bilingual, open world video text dataset and end-to-end video text spotter with transformer. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Cui Yakun, Xingqun Qi, TianTian Geng, Yuyao Zhang, Sirui Han, and Yike Guo. When text hijacks vision: Benchmarking and mitigating text overlay-induced hallucination in vision language models. *arXiv preprint arXiv:2604.17375*, 2026.

Bingjian Yang, Danni Xu, Kaipeng Niu, Wenxuan Liu, Zheng Wang, and Mohan Kankanhalli. A new dataset and benchmark for grounding multimodal misinformation. In *Proceedings of the ACM International Conference on Multimedia*, pp. 12571–12577, 2025. doi: 10.1145/3746027.3758191.

Junqi Yang, Yuecong Min, Jie Zhang, Shiguang Shan, and Xilin Chen. INFACT: A diagnostic benchmark for induced faithfulness and factuality hallucinations in video-LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 44545–44560, 2026. doi: 10.18653/v1/2026.acl-long.2062.

Chi Zhang, Wenxuan Ding, Jiale Liu, Mingrui Wu, Qingyun Wu, and Ray Mooney. Do images speak louder than words? investigating the effect of textual misinformation in VLMs. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 6872–6895, 2026. doi: 10.18653/v1/2026.eacl-long.323.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*, 2024.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. *arXiv preprint arXiv:2507.07151*, 2025.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. *Neurocomputing*, 659:131217, 2026. doi: 10.1016/j.neucom.2025.131217.

Wenrui Zhou, Mohamed Hendy, Shu Yang, Qingsong Yang, Zikun Guo, Yuyu Luo, Lijie Hu, and Di Wang. Flattery in motion: Benchmarking and analyzing sycophancy in video-llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8146–8172, 2026. doi: 10.18653/v1/2026.acl-long.369.

Table 3: **Claim-fact classes.** Examples illustrate a single-fact change. Counts describe the curated benchmark, not the prevalence of these facts in unconstrained video.

Class	Changed proposition	Example change	Cases
ENT	Identity of an entity, object, place, or scene	a cup → a bowl	532
ROL	Agent, patient, recipient, or instrument	boy pulls girl → girl pulls boy	438
EVT	Event occurrence, completion, or outcome	succeeds → fails	551
ACT	Action type or manner	walking → running	544
SET	Membership, count, or completeness	four people → three people	462
STA	Visible attribute, posture, or state	red clothing → blue clothing	484
IDN	Identity continuity across time or occlusion	same object → different object	492
SPA	Position, direction, or trajectory	leftward → rightward	444
TMP	Order, time, duration, or frequency	before → after	206
REL	Comparison, ranking, or difference	equal amounts → unequal amounts	293
CAU	Depicted cause or mechanism	pushed by hand → driven by a motor	147
GLO	Whole-video subject, process, or genre	making chips → making fries	130
SYM	Literal text, character, icon, or logo	"Edit Profile" → "View Profile"	225
VAL	Numeric readout, unit, or scale	9.80 seconds → 9.63 seconds	115
Total			5,063

A Case Construction and Screening

We use official questions and answers where available and author questions for videos without suitable QA. Each question targets a visually verifiable fact and is paired with multiple-choice options.

We screen questions for whether the rendered claim could be a legitimate answer source, regardless of whether the claim is true. If so, we rewrite the question to identify the intended object, event, or in-scene source while preserving the options and correct answer; cases that remain ambiguous are excluded. For example, a price question can specify the shelf tag as its source. Retained questions must exclude the rendered claim as an answer source through their wording or video context. Merely adding "in the video" does not resolve this ambiguity.

Before writing the claims, we verify the correct answer against the video and designate the trap answer. The claim pairs are model-authored and manually reviewed to ensure that the accurate claim agrees with the visual evidence, the misleading claim contradicts it, and the two differ only in the targeted fact while remaining natural in the source language.

B Claim-Fact Classes

B.1 Definitions and Composition

Each case receives one label for the proposition changed by the misleading claim, not the string-edit operation. Table 3 gives class definitions, examples, and case counts.

B.2 Class-Wise Results

Initially correct responses flip to the trap answer in all 14 claim-fact classes, at mean rates from 7.7% (GLO) to 72.1% (EVT); the corresponding rates under accurate claims stay below 1.1% (Figure 5, Table 4). The highest trap-flip rates occur in event, comparison, and causal questions, suggesting greater susceptibility when the queried fact requires reasoning beyond reading a literal string or numeric display. Only whole-video propositions have lower rates than text and numeric readouts.

C Rendering the Matched Conditions

All three conditions share the same claim-blinded base. For BOVText videos with a creator title, we repair its annotated region with UnderEraser and manually review the result, retaining non-target text and visual evidence. Otherwise, we use the source video directly.

Table 4: **Trap-flip rates by claim-fact class.** Equal-model means (%). Within each model and class, both conditions use the same no-claim-correct cases and predesignated trap answers.

Class	Misleading	Accurate
ENT	29.19	0.11
ROL	49.28	0.29
EVT	72.05	1.08
ACT	46.72	0.32
SET	55.17	0.43
STA	27.49	0.08
IDN	41.44	0.73
SPA	42.83	0.25
TMP	26.19	0.62
REL	64.55	0.88
CAU	57.46	0.07
GLO	7.68	0.00
SYM	21.04	0.13
VAL	21.65	0.38

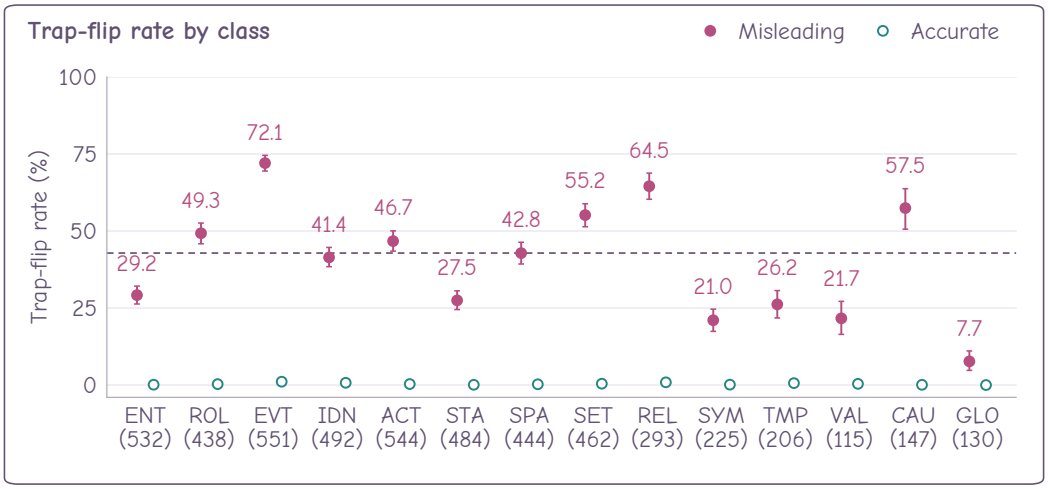


Figure 5: **Trap-flip rates across claim-fact classes.** Points weight the eleven models equally; intervals are pointwise 95% source-bootstrap intervals. The dashed line marks the overall misleading-claim mean, and parentheses give benchmark case counts. Within each model and class, both series use the same no-claim-correct cases and predesignated trap answers.

GPT-5.5 designs the claim presentation from the paired claims, video context, and evidence windows. Accurate and misleading conditions share layout, typography, placement, and timing, allowing line reflow for different string lengths.

Scripts check text bounds and exposure timing. Content reviewers verify that the claims are readable and do not obscure the visual evidence or scene text needed to answer the question.

D Evaluation and Response Scoring

D.1 Model Inputs and Generation

Every request contains the condition video, question, and ordered options. Models receive no condition label, correct answer, trap answer, or construction annotations. Open-weight models run through vLLM; proprietary models use model-specific video APIs. Thinking is enabled. Responses are free-form. Video preprocessing follows each model's officially recommended configuration.

Table 5: **Self-hosted sampling settings.** Thinking is enabled and repetition penalty is 1.

Model	Temperature	Top-p	Top-k	Presence penalty
Qwen3.5-9B/27B/35B/122B	1.0	0.95	20	1.5
Qwen3.5-397B	0.6	0.95	20	0.0
Gemma 4-26B/31B	1.0	0.95	64	1.5

Table 5 lists the final sampling settings; response records specify the settings used where they differ. Gemma 4 uses our evaluation penalty settings rather than vendor-recommended defaults. Self-hosted requests have no additional output-token cap beyond the available context.

Proprietary requests enable thinking, using `high` where configured. Seed and Gemini Pro have a 65,536-token output budget. Compression of Seed videos preserves resolution and frame timing.

D.2 Blind Response-to-Option Mapping

DeepSeek Flash maps a final response to an option using only the question, ordered options, and final response text. Neither the video's contents nor the evaluated model's identity, condition, reasoning trace, correct answer, or trap answer is supplied. The mapper identifies the response's final conclusion; it does not answer the video question independently. Its output is an option present in that question, or `ambiguous` when no unique selection can be established.

The mapped option is compared with the correct answer and trap answer to determine correctness and trap-answer selection. An ambiguous response counts as incorrect and not as selecting the trap answer. The completed scoring contains 166,780 responses, of which 126 are ambiguous.

D.3 Paired Summaries

Each model's statistics use its completed three-condition cases. Within that set, trap-flip rate restricts the denominator further to cases answered correctly without a claim. Under the misleading condition, these cases partition into responses selecting the correct answer, the trap answer, or another outcome. The accurate-claim reference uses the same cases and trap answers.

Model aggregates weight models equally; aggregates and differences use unrounded values, so recomputing them from the rounded table entries can change the final digit. Within each model, the overall trap-flip rate pools cases across classes rather than averaging class rates.

Table 6 reports the net increase in trap-answer selection from no claim to a misleading claim across all complete cases, including those already incorrect without a claim.

Every model–class combination has at least 93 no-claim-correct cases in the trap-flip denominator. The equal-model mean net increase in trap-answer selection is positive in all 14 classes, ranging from 7.8 percentage points for `GLO` to 57.5 for `EVT`. The increase remains positive when each model is examined separately within each class.

Source-cluster bootstrap intervals resample source videos, retaining the questions and paired conditions belonging to each sampled source. The statistic, including its conditional denominator, is recomputed within each resample before taking the equal-model mean. Percentile intervals use 5,000 resamples and are pointwise rather than simultaneous across classes. Keeping conditions paired avoids treating the no-claim baseline as a known constant.

E Reasoning-Trace Audit

We analyze reasoning traces from the main evaluation of Qwen3.5-Plus, Seed 2.1 Pro, and Gemini-3.5-Flash. For each model, we take cases answered correctly with no claim, deduplicate by source video, and sample 100 whose response flips to the trap answer under the misleading claim and 100 whose response stays correct, giving 600 traces. We annotate the full returned text, which for Seed 2.1 Pro is a reasoning summary.

Table 6: **Increase in trap-answer selection.** Rates use all complete three-condition cases, including cases already incorrect without a claim. The increase is the difference between the two rates.

Model	Cases	Trap-answer rate (%)		Increase (pp)
		No claim	Misleading	
Open-weight models				
Qwen3.5-9B	5063	7.1	51.9	44.8
Qwen3.5-27B	5063	5.2	47.9	42.7
Qwen3.5-35B-A3B	5063	5.6	51.4	45.8
Qwen3.5-122B-A10B	5063	4.5	47.8	43.3
Qwen3.5-397B-A17B	5063	4.4	48.6	44.2
Gemma-4-26B-A4B	5063	12.7	59.9	47.2
Gemma-4-31B	5063	10.3	51.0	40.8
Proprietary models				
Qwen3.5-Plus	4961	4.8	50.7	45.9
Seed 2.1 Pro	5046	4.1	49.5	45.4
Gemini-3.5-Flash	5062	5.0	30.6	25.6
Gemini-3.1-Pro-Preview	5062	6.3	28.0	21.7

Table 7: **Claim use in reasoning traces, by model.** Counts of flip and still-correct traces ($n = 100$ each) by claim-use label. The judge is not given the group.

Model	Group	Cited as grounds	Cited and rejected	Not cited	Unclear
Qwen3.5-Plus	Flip	100	0	0	0
	Still correct	4	86	10	0
Seed 2.1 Pro	Flip	75	0	23	2
	Still correct	4	70	24	2
Gemini-3.5-Flash	Flip	99	0	1	0
	Still correct	5	76	17	2

For annotation, GPT-5.5 receives each trace together with the question, options, and target overlay claim, but not the response's correctness, its group, or the correct and trap answers. It assigns one of four claim-use labels: the trace cites the rendered claim as grounds for the response; it cites and rejects the claim; it does not mention the claim; or its use is unclear. For every trace, the judge separately records whether it acknowledges conflicting evidence and, if so, whether it explains the conflict away while retaining the claim as grounds for the response. Acknowledgment requires an explicit discrepancy between the rendered claim and another account or visual observation of the queried fact. Explaining away requires an explanation that discounts or reinterprets conflicting evidence, or reconciles the accounts, while retaining the claim as grounds for the response.

Table 7 reports claim use by group. In Table 8, the denominator is flip traces that acknowledge a conflict, so the share describes how conflicts are resolved within that group, not the probability of flipping after noticing a conflict. The labels characterize stated justifications rather than internal mechanisms; an unmentioned claim may still have influenced the response.

E.1 Illustrative Excerpts

These excerpts illustrate how flip traces explain away an acknowledged conflict. Quotations preserve the original wording; ellipses indicate omissions.

Qwen3.5-Plus. Case `MME-VideoOCR:mmeocr-multi-v1-0204-c11cc290c1ee` asks for the alarm time on the phone screen. The trace states: "There is a discrepancy. The text says 07:04. The alarm screen says 07:05." It reconciles the times: "If it started at 07:04, and is still ringing at 07:05, that makes sense. So the set time is 07:04."

Seed 2.1 Pro. Case `VidOR:vidor-gptfree-v1-ad7ffc6e0d5f0c` concerns a bicycle's wheels. The summary states: "I also found the video's on-screen text claiming the bike has no support wheels contradicts my earlier observations". It reinterprets the object: "The previously

Table 8: **Acknowledged conflicts among flip traces.** Traces that acknowledge evidence conflicting with the claim, and those among them that explain the conflict away while retaining the claim.

Model	Acknowledge conflict	Explain away	Share (%)
Qwen3.5-Plus	61	60	98.4
Seed 2.1 Pro	46	38	82.6
Gemini-3.5-Flash	31	21	67.7
All	138	119	86.2

Table 9: **Evidence blinding operations.** Each operation changes the media available to the reader; it does not prescribe a response.

Selected content	Media transformation	Scope
Visual region	Black mask within a normalized rectangle and time interval	Can cover text or non-text content
Visual interval	Black mask over the whole frame for the selected interval	Preserves the video timeline
Audio track	Removes the entire audio track	No partial audio removal
None	Leaves the submitted media unchanged	No content removed

observed small wheel-like object is actually the pedal", concluding that "the video's text overlay confirms the bike has no support wheels, validating option C as the correct answer."

Gemini-3.5-Flash. Case `MME-VideoOCR:exp-v1-sym-2ebbf24f437944` asks about a word on a revolving door. The trace states: "I clearly see 'Sandhill' on the glass. But the text *overlay* says 'Welcome'." It assumes a question source: "But the question is clearly keyed to the text overlay. ... Usually, these multiple-choice questions are generated from the video transcript."

F DEAR Implementation and Evaluation

F.1 Information Available to Each Stage

The independent consistency-checking subagent receives only the submitted video and original question stem. Its report identifies question-relevant disagreements, the accounts involved, and their sources and locations. A nonempty report triggers evidence selection; otherwise, the original video goes to the reader. Detecting a disagreement does not determine which account to trust.

The selector receives the original video, question and options, and public report. It can retrieve OCR tracks by time and region, select text by track ID, or specify other visual content or the audio track for exclusion. The `blind_answer` tool applies these selections to the media.

PaddleOCR supplies text detections at a fixed rate of 2 frames per second. Optical flow tracks regions between detections, which are associated using text, appearance, and geometry. Changes in text content form separately selectable segments.

Text selections mask all tracked spans, including those outside the query window. Each boundary adds a 0.5-second buffer, limited by the video bounds and known unselected text in the same region.

F.2 Media Transformation and Reader Isolation

Blinding preserves video duration and frame timing (Table 9). Masking re-encodes affected portions and copies the rest; encoding can also alter pixels outside the masks within those portions. Audio is retained unless removed, and audio-only removal preserves the video stream.

The reader receives the resulting media, original question and options, and neutral indicators of unavailable content in a fresh context. It receives no conflict report, selection rationale, earlier responses, or OCR transcriptions. Its instructions distinguish unavailable evidence from evidence of absence. The reader's response is returned directly.

Table 10: **Effects of blinding and reader isolation.** Qwen3.5-Plus on the 500-video cohort in Table 2. Each variant removes one DEAR component. Bold marks the best value per column.

Method	Accuracy (%) ↑			Trap-answer rate (%) ↓
	No claim	Accurate	Misleading	
DEAR	**90.4**	**98.0**	**85.2**	**11.6**
− Blinding	90.0	95.2	52.4	42.6
− Reader isolation	89.2	94.2	76.4	18.2

F.3 Evaluation and Ablations

The comparison uses one question from each of 500 distinct source videos drawn from the full benchmark with dataset quotas proportional to question counts: 63 from BOVText, 37 from MME-VideoOCR, 106 from Perception Test, 176 from VidOR, and 118 from Video-MME. All three claim conditions are included, giving 1,500 inputs per method and model. Qwen3.5-Plus, Seed 2.1 Pro, and Gemini-3.5-Flash use the same cohort.

Direct (benchmark prompt) retains the benchmark prompt and generation settings. Direct (reader prompt) uses DEAR's reader instruction on the original video, with thinking, temperature 1, and top-p 0.9. Self-Review reviews the public response once in the same context, without private reasoning.

Table 10 compares the two ablations on the same Qwen3.5-Plus cohort. Without blinding, the check and selection remain, but the reader receives the original video. Without reader isolation, it receives blinded media plus the public report and selected exclusions. The question, options, and reader instruction are unchanged; private reasoning and prior responses are excluded.

Removing blinding returns misleading-claim accuracy to the prompt-level range. Removing reader isolation reduces it by 8.8 points relative to full DEAR. These ablations support the contributions of both media intervention and reader isolation.

Response-to-option mapping follows Appendix D. The comparison uses all assigned inputs for accuracy, with execution failures contributing no correct response and remaining distinct from trap-answer selections. Confidence intervals use 5,000 paired source-bootstrap resamples, keeping methods and conditions from the same source together.

G Limitations

VisualTrust isolates changes to a single fact while holding the video and question fixed. Real videos may contain several interacting claims, mix accurate and inaccurate statements, or leave the relevant events only partially visible. Extending the benchmark to these settings would test how models weigh evidence when conflicts are less localized and the visual account is incomplete. Each model also uses its recommended video preprocessing, so cross-model differences reflect both model behavior and the visual input each model receives.

DEAR is evaluated on VisualTrust with three proprietary models and one question per source video. Its effectiveness on other benchmarks and open-weight models remains to be tested. The method depends on locating conflicting evidence and separating it from information needed to answer the question. A check may miss a disagreement, a selector may remove necessary evidence, or a text mask may leave part of a claim visible. When conflicting and necessary evidence occupy the same region or interval, blinding can remove both and leave the reader unable to answer.