BETTER SCORES ARE NOT ENOUGH: EVIDENCE-DEPENDENCE EVALUATION FOR VIDEO HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video hallucination is commonly associated with weak visual and temporal grounding, over-dependence on language priors, and response bias. Yet benchmark scores alone do not establish whether correct responses depend on video evidence. When a mitigation method improves a video hallucination benchmark score, does that gain actually depend on video evidence? We propose EVIDENCE-DEPENDENCE EVALUATION (EDE), an evaluation framework that uses controlled interventions on visual and temporal evidence to measure how benchmark performance and mitigation gains depend on that evidence, while preserving each benchmark's original evaluation protocol. EDE combines controlled evidence interventions with complementary diagnostics that assess other evaluation factors relevant to interpreting benchmark scores and mitigation gains, requiring neither additional training nor new benchmark-specific annotations. EDE reveals markedly different evidence-dependence profiles: some benchmarks exhibit substantial performance declines when targeted video evidence is disrupted, whereas others remain stable or even increase after such evidence is removed or disrupted. We evaluate four video LLMs on six video hallucination benchmarks and analyze gains from six mitigation methods on selected benchmarks. Complementary diagnostics show that mitigation gains can coincide with error trade-offs, be partly reproduced by response shifts, and vary substantially across evaluation formats, while predictions can change under meaning-preserving question paraphrases. Together, EDE and these diagnostics reveal dependence on video evidence and sensitivity to other factors that benchmark scores and mitigation gains alone do not capture. We therefore argue that video hallucination evaluation should report benchmark scores and mitigation gains alongside their evidence-dependence profiles.

1 INTRODUCTION

Large Language Models have demonstrated strong capabilities in generation and reasoning, yet hallucination remains a fundamental limitation. As LLMs extend to multimodal settings, this problem becomes more complex because generated outputs must be grounded in richer input evidence across modalities. This is especially so in video understanding, where the input evidence is inherently spatiotemporal, involving not only what appears but also when and how visual content evolves over time.

Video hallucination benchmarks are designed around different forms of visual and temporal evidence (Luo et al., 2026; Zhang et al., 2026; Kong et al., 2025; Li et al., 2025; Wang et al., 2024), ranging from action occurrence and object relations to temporal information and action order (Kong et al., 2025; Wang et al., 2024; Li et al., 2025). We refer to the visual or temporal evidence that a benchmark is designed to probe as its target evidence. However, a benchmark score indicates whether model responses are correct, but not which evidence those responses depend on. Correct responses can be produced without dependence on the target evidence, for example due to language priors (Zhang et al., 2026) or response bias (Kong et al., 2025) (Figure 1). This leaves a gap between a benchmark's target evidence and what its score alone can establish.

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.

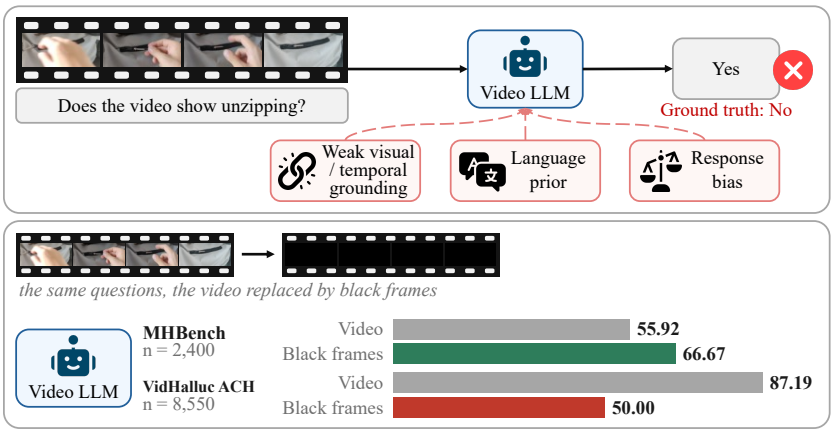

Figure 1: **The same intervention changes benchmark scores in opposite directions. Top:** Factors commonly associated with video hallucination include weak visual and temporal grounding, over-dependence on language priors, and response bias. **Bottom:** For Qwen3-VL-8B, replacing the video with black frames, while keeping the questions and evaluation metric unchanged, increases question-level accuracy on MHBench but decreases it on VidHalluc-ACH.

Mitigation gains likewise do not by themselves establish improved video grounding or reveal their dependence on the target evidence. Moreover, observed gains may be sensitive to other aspects of evaluation, including response tendency, evaluation protocol, and question formulation. Evaluating hallucination mitigation therefore requires examining the evidence dependence of mitigation gains, together with complementary diagnostics of these additional sensitivities.

To address this gap, we propose EVIDENCE-DEPENDENCE EVALUATION (EDE), an evaluation framework that uses controlled interventions on visual and temporal evidence to measure how benchmark performance and condition-matched mitigation gains depend on that evidence, while preserving each benchmark's original evaluation protocol. The resulting changes in benchmark scores and mitigation gains form evidence-dependence profiles. We further interpret these profiles using complementary diagnostics of response tendency, evaluation-protocol sensitivity, and question-formulation sensitivity under meaning-preserving paraphrases. EDE requires neither additional training nor new benchmark-specific annotations, so it can be applied directly to existing benchmarks.

Across six video hallucination benchmarks and four video LLMs, EDE reveals both alignment and mismatches between target evidence and observed dependence. Notably, an analysis of six mitigation methods on a subset of these benchmarks shows that condition-matched gains do not consistently decrease when visual or temporal evidence is removed or disrupted. Complementary diagnostics show that mitigation gains can coincide with error trade-offs, be partly reproduced by response-rate-matched controls, and vary across evaluation protocols. We further find that individual predictions can change under meaning-preserving question paraphrases even when aggregate accuracy remains stable or improves. Together, these findings show that benchmark scores and mitigation gains should be interpreted alongside their evidence dependence and complementary diagnostics.

2 RELATED WORK

2.1 VIDEO HALLUCINATION EVALUATION AND BENCHMARKS

Video hallucination benchmarks are designed to probe different aspects of visual and temporal grounding, while accounting for potential confounds such as language priors and response biases. Examples include visual evidence of entities and their relations in VideoHallucer and Dr.V-Bench (Wang et al., 2024; Luo et al., 2026); visual evidence of events and actions in EventHallusion, MHBench, and VidHalluc's Action Hallucination (VidHalluc-ACH) task (Zhang et al., 2026; Kong et al., 2025; Li et al., 2025); and temporal evidence for action ordering in VidHalluc's Temporal Sequence Hallucination (VidHalluc-TSH) task (Li et al., 2025).

Complementary measures and analyses have also been proposed to interpret question-level accuracy. For example, MHBench reports PairAcc and VideoHallucer uses pair-level overall accuracy over paired examples with opposite expected answers, counting a pair as correct only when both responses are correct (Kong et al., 2025; Wang et al., 2024). Related evaluation concerns have also been studied in image hallucination. In image hallucination, CHAIR and POPE (Rohrbach et al., 2018; Li et al., 2023) introduced widely used evaluations for object hallucination and response bias.

Most recently, concurrent work in the image domain re-examines whether benchmark scores reflect dependence on visual evidence (Lan et al., 2026) and whether mitigation methods reduce hallucination while also describing fewer objects and details present in the image (Fazli et al., 2026). Prior work has shown that video-language models can achieve strong performance using frame-level information without explicit temporal modeling (Buch et al., 2022; Lei et al., 2023). Concurrent work introduces a dedicated diagnostic benchmark for video hallucination under diverse input conditions (Yang et al., 2026). EDE extends this evaluation perspective beyond static images to video while applying it directly to existing video hallucination benchmarks, examining the visual and temporal evidence dependence of both benchmark scores and mitigation gains under their original evaluation protocols.

2.2 Video Hallucination Mitigation

Existing video hallucination mitigation methods span both training-free interventions at inference time and training-based approaches. Training-free methods include contrastive decoding, which contrasts output distributions conditioned on original and altered visual inputs: VCD adds noise to images (Leng et al., 2024), while MotionCD and TCD use reversed and temporally downsampled videos, respectively (Kong et al., 2025; Zhang et al., 2026). Another family of training-free methods intervenes directly on attention or internal representations: DTR reduces excessive decoder attention to dominant frames (Liu et al., 2026), while VTI adjusts visual and textual hidden representations during inference in image-based VLMs (Liu et al., 2025). Among training-based methods, LLaVA-Hound-DPO (hereafter Hound-DPO) uses a language model to score responses against video captions, constructs preference pairs from these scores, and performs direct preference optimization (DPO) to reduce video hallucination (Zhang et al., 2025). These methods are typically evaluated by improvements under the original benchmark condition, leaving unclear whether the reported gains depend on the benchmark's target evidence. We therefore compare each method with its corresponding base model under the matched evidence conditions to examine how mitigation gains change when visual or temporal evidence is removed or disrupted.

3 Method

3.1 Evidence-Dependence Evaluation (EDE) Framework

EDE evaluates how benchmark scores and condition-matched mitigation gains depend on visual and temporal evidence, as illustrated in Figure 2. For each benchmark, we apply controlled interventions to the video input while keeping the questions and evaluation protocol fixed (Figure 2a; Section 3.2). We evaluate benchmark scores under each evidence condition and compare each mitigation method with its corresponding base model under the same condition. The resulting score changes and condition-matched gains form **evidence-dependence profiles** that characterize responses to different interventions (Figure 2b; Section 3.3). Complementary diagnostics examine response tendency, evaluation-protocol sensitivity, and question-formulation sensitivity under meaning-preserving paraphrases (Figure 2c; Section 3.4).

3.2 Controlled Evidence Interventions

We define four controlled interventions T_c on the video input V, yielding $V_c = T_c(V)$. These interventions target **visual** and **temporal** evidence.

Visual evidence. For **Black Frames**, every sampled frame in $V = [f_1, \ldots, f_T]$ is replaced with a black frame, $V_{\text{black}} = [b, \ldots, b]$, removing informative visual content while preserving the presence and structure of the visual input. For **Text-only**, we remove visual input entirely, $V_{\text{text}} = \varnothing$.

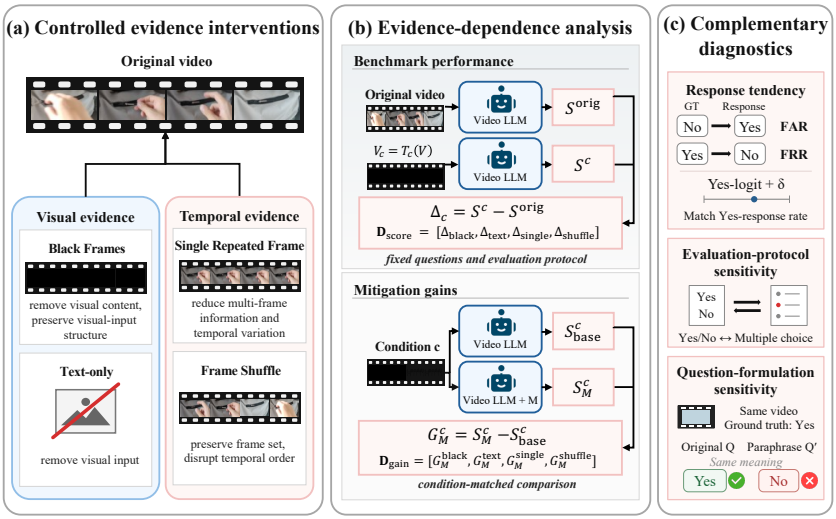

Figure 2: **Overview of the Evidence-Dependence Evaluation (EDE) framework. (a)** Four controlled interventions on visual and temporal evidence, with questions and evaluation protocol held fixed. **(b)** Evidence-dependence analysis of benchmark scores ($\mathbf{D}_{\text{score}}$) and condition-matched mitigation gains ($\mathbf{D}_{\text{gain}}$), interpreted relative to each benchmark's target evidence. **(c)** Complementary diagnostics of response tendency, evaluation-protocol sensitivity, and question-formulation sensitivity.

Temporal evidence. For **Single Repeated Frame**, we select the middle frame from the sampled sequence, f_k with $k = \lfloor T/2 \rfloor + 1$, and repeat it to form $V_{\text{single}} = [f_k, \ldots, f_k]$, preserving the sequence length. For **Frame Shuffle**, we apply a reproducible non-identity permutation π to the sampled frames to form $V_{\text{shuffle}} = [f_{\pi(1)}, \ldots, f_{\pi(T)}]$. Single Repeated Frame reduces multi-frame information and temporal variation, whereas Frame Shuffle preserves the frame content while disrupting temporal order. Implementation details and analyses of alternative frame positions and shuffle seeds are provided in Appendix B.5.

3.3 Evidence-Dependence Analysis

We use the controlled interventions to analyze evidence dependence at two levels: **benchmark performance** and **mitigation gains**.

Evidence dependence of benchmark performance. For a given model and benchmark evaluation set, let S^c denote the score under evidence condition c, with $c = \text{orig}$ denoting the original videos. We define the score change as $\Delta_c = S^c - S^{\text{orig}}$ and the evidence-dependence profile as $\mathbf{D}_{\text{score}} = [\Delta_{\text{black}}, \Delta_{\text{text}}, \Delta_{\text{single}}, \Delta_{\text{shuffle}}]$. We interpret the direction and magnitude of these changes in relation to the benchmark's target evidence.

Evidence dependence of mitigation gains. Let M denote a hallucination mitigation method. For a given benchmark evaluation set and evidence condition c, S^c_M denotes the score with method M, and S^c_{base} denotes the score of its corresponding base model. We define the condition-matched gain as $G^c_M = S^c_M - S^c_{\text{base}}$. We compare G^{orig}_M and G^c_M to assess how the mitigation gain changes under evidence condition c, where visual or temporal evidence is removed or disrupted. The gains across interventions form the evidence-dependence profile $\mathbf{D}_{\text{gain}} = [G^{\text{black}}_M, G^{\text{text}}_M, G^{\text{single}}_M, G^{\text{shuffle}}_M]$.

3.4 Complementary Diagnostics

We use three complementary diagnostics to examine the sensitivity of benchmark scores and mitigation gains to response tendency, evaluation protocol, and question formulation.

Response Tendency. For binary Yes/No benchmarks, we examine **error trade-offs** under the original evidence condition ($c = \text{orig}$) using the false acceptance rate, $\text{FAR} = P(\hat{y} = \text{Yes} \mid y = \text{No})$, and false rejection rate, $\text{FRR} = P(\hat{y} = \text{No} \mid y = \text{Yes})$. We report their changes relative to the corresponding base model to assess whether mitigation gains coincide with reductions in both error types or a trade-off between them. We further construct a **response-rate-matched control** by applying a constant offset to the corresponding base model's Yes-logit so that its overall Yes-response rate matches that of method M. Let ρ_M and $\rho(\delta)$ denote the Yes-response rates of method M and the offset control with offset δ, respectively, on the same evaluation set. We select $\delta_M^* = \arg\min_{\delta \in \mathcal{G}} |\rho(\delta) - \rho_M|$ over a fixed offset grid $\mathcal{G}$ (Table 17), without optimizing benchmark accuracy.

Evaluation-Protocol Sensitivity. We compare mitigation gains under Yes/No and multiple-choice evaluation protocols on matched semantic units, scoring each unit once under each protocol. For each protocol p, we compute $G_M^{(p)} = S_M^{(p)} - S_{\text{base}}^{(p)}$ and report $\Delta G_{\text{protocol}} = G_M^{(\text{MCQ})} - G_M^{(\text{Y/N})}$ to quantify the change in mitigation gains across the protocols.

Question-Formulation Sensitivity. We evaluate three meaning-preserving paraphrases of each question while keeping the video and ground-truth answer fixed (Appendix F.1). We report original-question accuracy and the mean accuracy change across paraphrases, together with correct-to-incorrect and incorrect-to-correct transition rates. We also measure correct-answer retention, C_{para}, defined as the fraction of originally correct questions that remain correct under all three paraphrases. This analysis examines whether changes in question wording alter model predictions despite unchanged video evidence and question meaning.

4 Experiments

4.1 Experimental Settings

Benchmarks and Evidence Requirements. We evaluate EDE on six video hallucination benchmarks comprising seven evaluation tasks: MHBench, EventHallusion, VideoHallucer, Dr.V-Bench, VidHalluc-ACH (Action hallucination) and VidHalluc-TSH (Temporal sequence hallucination), and VidPair-Halluc (Huang et al., 2026a). We identify each task's target evidence from the original benchmark definitions and task construction and map it to the relevant EDE interventions (Table 6) to assess its alignment with observed evidence dependence.

Implementation Details. We evaluate four video LLMs: Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025), and VideoChat2 (Li et al., 2024); additional within-family scale variants are reported in Appendix B.1. Main experiments use eight uniformly sampled frames (sixteen for VideoChat2) and deterministic decoding (Appendix A.2). We use question-level accuracy for consistent comparison across the main evidence-dependence analyses and validate the findings using additional benchmark-defined metrics in Appendix B.2. We evaluate five training-free hallucination mitigation methods: MotionCD (Kong et al., 2025), TCD (Zhang et al., 2026), VCD (Leng et al., 2024), DTR (Liu et al., 2026), and VTI (Liu et al., 2025), along with the training-based method Hound-DPO (Zhang et al., 2025). Training-free methods are applied to Qwen3-VL-8B-Instruct, while Hound-DPO's gains are measured relative to its corresponding Hound-SFT checkpoint. For each method, we use a single fixed configuration across benchmarks and applicable evidence conditions, without benchmark-specific tuning. Method-specific implementation details and applicability across evidence conditions are provided in Appendix A.3.

4.2 Evidence Dependence of Benchmark Performance

RQ1: How sensitive are video hallucination benchmark scores to controlled interventions on visual and temporal evidence?

Evidence-dependence profiles differ markedly across benchmarks. Table 1 shows that removing informative visual evidence substantially reduces scores on several benchmarks. In contrast, MH-Bench scores increase under Black Frames across all four video LLMs, while the direction of score

Table 1: **Evidence-dependence profiles across benchmarks and video LLMs.** *Original* reports question-level accuracy (%) on unmodified videos. The intervention columns report score changes ($\Delta_c = S^c - S^{\text{orig}}$) in percentage points for the same model and benchmark. Positive changes are shown in bold.

			Visual evidence		Temporal evidence	
Benchmark	Video LLM	Original	Black Frames	Text-only	Single Repeated Frame	Frame Shuffle
MHBench	Qwen3-VL-8B	55.92	+**10.75**	+**8.58**	+**2.79**	−7.00
	InternVL3-8B	49.38	+**15.58**	+**1.71**	+**2.42**	−4.33
	MiniCPM-V-4.5	45.58	+**13.17**	−2.67	+**2.58**	−4.21
	VideoChat2	61.17	+**5.50**	−6.63	+**3.17**	−1.79
EventHallusion	Qwen3-VL-8B	71.64	−10.27	−28.36	−7.58	−8.31
	InternVL3-8B	73.84	−29.83	−63.08	−17.11	−3.67
	MiniCPM-V-4.5	79.22	−17.85	−22.98	−13.45	−6.60
	VideoChat2	54.03	+**8.56**	−47.68	−3.67	−6.11
VideoHallucer	Qwen3-VL-8B	76.23	−19.77	−16.91	−8.09	−1.54
	InternVL3-8B	76.13	−19.98	−16.65	−6.35	−1.64
	MiniCPM-V-4.5	76.74	−20.80	−17.37	−9.99	−1.79
	VideoChat2	63.17	−12.30	−8.86	−1.74	+**0.10**
Dr.V-Bench	Qwen3-VL-8B	68.80	−15.50	−12.58	−9.33	−1.25
	InternVL3-8B	73.27	−16.32	−19.21	−11.92	−1.22
	MiniCPM-V-4.5	68.57	−16.58	−13.30	−9.46	−0.95
	VideoChat2	65.35	−14.65	−10.05	−7.00	−0.30
VidHalluc-ACH	Qwen3-VL-8B	87.19	−37.19	−37.19	−13.79	−1.35
	InternVL3-8B	88.32	−38.32	−38.32	−11.23	−0.28
	MiniCPM-V-4.5	89.58	−39.58	−39.58	−13.78	−0.42
	VideoChat2	82.33	−32.35	−32.33	−9.86	−0.62
VidHalluc-TSH	Qwen3-VL-8B	80.00	−22.67	−21.67	−43.33	−24.50
	InternVL3-8B	89.50	−36.50	−29.50	−65.50	−30.33
	MiniCPM-V-4.5	90.33	−28.83	−29.83	−64.92	−32.67
	VideoChat2	72.67	−21.17	−17.00	−21.17	−19.50
VidPair-Halluc	Qwen3-VL-8B	53.43	−7.92	−10.05	−4.60	−1.25
	InternVL3-8B	59.88	−16.20	−15.33	−6.02	−2.25
	MiniCPM-V-4.5	59.05	−14.35	−12.25	−5.30	−0.82
	VideoChat2	58.08	−12.73	−3.50	−5.13	−0.88

changes under Text-only varies across models. Temporal interventions further distinguish these profiles: VidHalluc-ACH shows a substantial score decrease under Single Repeated Frame but only a small decrease under Frame Shuffle; MHBench scores decrease under Frame Shuffle but increase under Single Repeated Frame; and VidHalluc-TSH shows substantial decreases under both temporal interventions. These contrasting responses to temporal interventions are observed across all four video LLMs. These profiles reveal differences in video-evidence dependence that original-condition scores alone do not show. Appendix analyses further examine these profiles across model scales, evaluation metrics, and temporal-intervention settings (Appendices B.1, B.2, and B.5).

RQ2: How do the observed evidence-dependence profiles relate to each benchmark's target evidence?

EDE reveals both alignment and mismatches between target evidence and observed dependence. As summarized in Table 6, VidHalluc-ACH requires identifying the prominent action but not temporal order, consistent with its substantial score decrease under Single Repeated Frame but small decrease under Frame Shuffle. In contrast, VidHalluc-TSH requires the identities and order of two consecutive actions and shows substantial score decreases under both temporal interventions. Within VideoHallucer, despite small aggregate score changes under Frame Shuffle, the temporal subset shows the largest decrease in three of the four video LLMs and all three additional model-size variants (Table 13). MHBench requires distinguishing an action from its opposite and determining whether it is completed or only partially performed, yet its score increases under Black Frames

Table 2: **Condition-matched mitigation gains** ($G^c_M = S^c_M - S^c_{\text{base}}$), reported in percentage points. The base model is Hound-SFT for Hound-DPO and Qwen3-VL-8B-Instruct for all other methods. † and — denote a structural zero and an inapplicable condition, respectively.

			Visual evidence		Temporal evidence	
Benchmark	Method	G^{orig}_M	G^{black}_M	G^{text}_M	G^{single}_M	G^{shuffle}_M
MHBench	MotionCD	+5.92	0.00†	—	0.00†	+0.54
	TCD	+2.96	+0.00	—	+1.67	+0.29
	VCD	+0.46	+0.00	—	+0.50	−1.04
	DTR	−7.88	−6.46	—	−3.42	−6.04
	VTI	+3.46	−0.42	−14.50	+3.50	+10.29
	Hound-DPO	+0.04	+0.00	+0.38	−0.75	+0.29
EventHallusion	MotionCD	+1.96	0.00†	—	0.00†	−0.49
	TCD	+1.96	+0.24	—	−1.22	+1.47
	VCD	+1.47	−5.87	—	+0.98	+3.91
	DTR	+5.38	+4.40	—	+0.00	+4.65
	VTI	−59.90	−55.99	−9.29	−48.66	−51.59
	Hound-DPO	−6.60	+0.00	−1.47	−4.65	−3.18
VidHalluc-ACH	MotionCD	−2.23	0.00†	—	0.00†	−3.57
	TCD	+0.27	+0.00	—	−0.26	−0.44
	VCD	+0.54	+0.00	—	+1.43	+0.47
	DTR	−1.79	+0.00	—	+0.98	−2.02
	VTI	−18.33	+0.00	+0.00	−7.54	−18.39
	Hound-DPO	+0.70	+0.00	+0.00	+1.36	+0.36
VidHalluc-TSH	MotionCD	+7.33	0.00†	—	0.00†	−1.83
	TCD	+2.50	−0.67	—	+0.00	+1.17
	VCD	+1.33	+0.17	—	−2.17	+0.33
	DTR	−28.83	−54.00	—	−24.10	−15.67
	VTI	−21.50	+0.00	+0.00	+2.83	+0.33
	Hound-DPO	+12.67	+5.83	+1.50	+14.17	+10.00

across all four video LLMs. In this benchmark, label imbalance across videos sharing the same question allows high accuracy without distinguishing between videos (Appendix B.3). High benchmark scores alone therefore do not establish dependence on the target evidence and should be interpreted alongside evidence-dependence profiles.

4.3 Evidence Dependence of Mitigation Gains

RQ3: How do hallucination mitigation gains change across visual and temporal evidence conditions?

Mitigation gains do not consistently decrease when visual or temporal evidence is removed or disrupted. On VidHalluc-TSH, MotionCD's gain decreases from $+7.33$ percentage points on original videos to -1.83 under Frame Shuffle (Table 2). On the same task, Hound-DPO retains substantial gains under Single Repeated Frame ($+14.17$) and Frame Shuffle ($+10.00$), compared with $+12.67$ on original videos.

These results indicate that large gains on original videos alone do not establish dependence on target evidence. Structural zeros and inapplicable conditions reflect constraints of the mitigation method rather than demonstrating that its gains depend on video evidence.

We provide additional analyses of mitigation-strength sensitivity in Appendix C.1 and extend the original-condition gain evaluation to additional benchmarks in Appendix C.2.

4.4 Complementary Diagnostics

RQ4: How sensitive are measured benchmark scores and mitigation gains to factors beyond video evidence?

Table 3: **Response-tendency diagnostics on MHBench.** Values are percentage-point changes except italicized base values (%). δ_M^* matches the Yes-response rate; the last two columns report differences from the response-rate-matched control. $*$: 95% CI excludes zero. See Appendices D.1 and D.2.

	Observed gain	Error trade-offs		Response-rate-matched control		
Method	$G_M^{\text{orig}} \uparrow$	ΔFAR$\downarrow$	ΔFRR$\downarrow$	δ_M^*	Acc. vs. control$\uparrow$	Pair acc. vs. control$\uparrow$
Qwen3-VL-8B (absolute)	*55.92*	*57.42*	*30.75*	—	—	—
Existing mitigation methods						
MotionCD	+5.92	−9.50	−2.33	−0.5	+4.75*	+9.92*
TCD	+2.96	−4.17	−1.75	0	+2.96*	+5.50*
VCD	+0.46	−1.50	+0.58	0	+0.46	+2.17*
DTR	−7.88	−1.95	+8.63	−1	−10.13*	−14.67*
VTI	+3.46	+10.58	−17.50	+2	+5.96*	+2.08
Hound-DPO	+0.04	+1.00	−0.88	0	+0.04	+0.75
Mechanistic controls						
Additive steering	+12.38	−42.83	+18.08	−5	+5.17*	+6.08*
Directional ablation	+8.92	−52.83	+35.00	−8	+2.21	+1.92
Random direction	−1.04	+2.50	−0.42	+0.25	−0.63	−1.00
ShuffleCD	+4.46	−7.67	−1.25	−0.5	+3.29*	+7.08*

Table 4: **Evaluation-protocol sensitivity of mitigation gains.** Yes/No and multiple-choice protocols are compared on matched semantic units (1,200 for MHBench; 4,260 for VidHalluc-ACH); Yes/No pair accuracy requires both answers to be correct. Italics report base scores (%); method rows report gains (percentage points), with $\Delta G_{\text{protocol}} = G_M^{(\text{MCQ})} - G_M^{(\text{Y/N})}$.

	MHBench			VidHalluc-ACH		
Method	Yes/No	Multiple-choice	$\Delta G_{\text{protocol}}$	Yes/No	Multiple-choice	$\Delta G_{\text{protocol}}$
Qwen3-VL-8B (absolute)	*28.25*	*61.50*	—	*75.07*	*90.07*	—
MotionCD	+11.42	+1.17	−10.25	−4.01	−1.24	+2.77
TCD	+5.50	+0.08	−5.42	+0.70	−0.28	−0.99
VCD	+2.17	−1.75	−3.92	+1.13	−0.26	−1.38
DTR	−11.17	−7.00	+4.17	−3.08	−2.98	+0.09
VTI	−3.42	−19.00	−15.58	−35.00	−15.02	+19.98
Hound-DPO	+0.75	−0.42	−1.17	+1.34	+0.52	−0.82

Mitigation gains can coincide with error trade-offs and be partly reproduced by response-rate adjustments. On MHBench, VTI improves accuracy by 3.46 percentage points, while FAR increases by 10.58 and FRR decreases by 17.50 (Table 3). Thus, fewer false rejections are accompanied by more false acceptances. MotionCD yields a gain of 5.92 percentage points, of which 1.17 is reproduced by the response-rate-matched control. These results show that raw gains alone do not distinguish improvements in evidence dependence from changes in response tendency.

Mitigation gains vary substantially across evaluation protocols. Table 4 shows differences in gains between paired Yes/No and multiple-choice protocols across matched semantic units. On MHBench, MotionCD's gain decreases from +11.42 percentage points under paired Yes/No to +1.17 under multiple choice, while VCD's gain changes from +2.17 to −1.75. These results show that mitigation gains observed under one protocol need not persist under another and should therefore be interpreted with respect to the evaluation protocol.

Model predictions can change across meaning-preserving question formulations. Table 5 evaluates meaning-preserving question paraphrases while keeping the video and ground-truth answer fixed. On MHBench, mean accuracy increases by 2.61–5.67 percentage points across all four video LLMs, yet only 63.7–68.0% of originally correct questions remain correct under all three paraphrases. Correct-to-incorrect transitions can be offset by improvements on other questions, so aggregate accuracy changes can obscure instability in individual predictions.

Table 5: **Question-formulation sensitivity under meaning-preserving paraphrases.** Each question is evaluated with three paraphrases with the video and ground-truth answer fixed. Δ is mean paraphrased minus original-question accuracy (percentage points). Flip rates are percentages of all retained items averaged over paraphrases; C_{para} is the percentage of originally correct items correct under all three.

				Correctness flips (%)		
Benchmark	Video LLM	Original question	Δ	Correct $\rightarrow$ Incorrect	Incorrect $\rightarrow$ Correct	$C_{\text{para}} \uparrow$
MHBench	Qwen3-VL-8B	55.81	+2.61	9.1	11.7	68.0
	InternVL3-8B	49.22	+3.62	8.1	11.7	67.2
	MiniCPM-V-4.5	45.04	+5.67	7.6	13.2	66.6
	VideoChat2	61.33	+4.06	10.1	14.2	63.7
EventHallusion	Qwen3-VL-8B	71.64	−0.16	5.9	5.8	81.6
	InternVL3-8B	73.84	−7.74	10.7	2.9	69.9
	MiniCPM-V-4.5	79.22	−2.77	7.5	4.7	80.9
	VideoChat2	54.03	−5.38	10.8	5.5	59.3
VideoHallucer	Qwen3-VL-8B	76.19	−2.55	10.5	7.9	74.1
	InternVL3-8B	75.92	−3.86	11.4	7.6	71.8
	MiniCPM-V-4.5	76.92	−3.02	11.0	8.0	71.8
	VideoChat2	63.86	−0.67	8.2	7.6	72.9
Dr.V-Bench	Qwen3-VL-8B	68.81	−1.03	8.5	7.5	75.8
	InternVL3-8B	73.20	−3.15	10.8	7.6	70.8
	MiniCPM-V-4.5	68.51	−1.09	10.4	9.4	69.8
	VideoChat2	65.41	−2.94	12.2	9.3	63.0
VidHalluc-ACH	Qwen3-VL-8B	87.22	−3.14	6.5	3.3	85.2
	InternVL3-8B	88.33	−1.37	5.6	4.2	86.5
	MiniCPM-V-4.5	89.61	−2.10	5.8	3.7	86.0
	VideoChat2	82.37	−7.27	12.8	5.6	66.3
VidHalluc-TSH	Qwen3-VL-8B	80.23	−6.52	8.5	1.9	82.4
	InternVL3-8B	88.95	−2.39	5.6	3.2	86.9
	MiniCPM-V-4.5	90.70	−2.65	5.8	3.2	87.4
	VideoChat2	72.29	−3.17	6.8	3.6	82.3
VidPair	Qwen3-VL-8B	53.27	+1.24	5.8	7.1	78.8
	InternVL3-8B	60.08	−0.31	9.7	9.4	68.6
	MiniCPM-V-4.5	59.06	−1.06	10.5	9.4	66.2
	VideoChat2	58.45	−5.41	13.4	8.0	62.3

5 Conclusion

EDE reveals markedly different evidence-dependence profiles across video hallucination benchmarks. Some tasks show substantial performance declines when target evidence is disrupted, whereas MHBench improves under Black Frames across all four video LLMs. Mitigation gains likewise exhibit different profiles: some disappear or reverse under evidence disruption, while others remain substantial. These findings show that benchmark scores and mitigation gains alone do not establish dependence on the target evidence.

Complementary diagnostics reveal additional sensitivities that aggregate performance can obscure. Mitigation gains can vary with response tendency and evaluation protocol, while on MHBench, mean accuracy increases under paraphrases even though only 63.7–68.0% of originally correct questions remain correct under all three. Video hallucination evaluation should therefore consider not only benchmark scores and mitigation gains, but also their evidence dependence and sensitivity to other evaluation factors. EDE complements standard benchmark evaluation by making this missing evidence-dependence information explicit.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

Generative AI was used for four purposes, one of which contributes to a research artifact.

Synthetic dataset generation. The meaning-preserving paraphrase sets evaluated in §4.4 were generated with `gpt-5.6-luna` through the OpenAI API: for each unique question we generated 20 candidates, validated them in a separate call to the same model, and selected the three passing candidates with the largest surface divergence (Appendix F.1, Table 21). Neither the generator nor the validator received the video, the gold answer, or any evaluated model's prediction, so the construction cannot leak information about the models under evaluation. We did not perform a human review of the selected paraphrases; meaning preservation is therefore supported only by the automated validation criteria, a limitation we state in Appendix F.1.

Research execution. We used Claude to implement parts of the evaluation and aggregation scripts and the paraphrase-generation pipeline, to give feedback on the reporting protocol and the presentation of results, and to assist in interpreting the computed results. We did not use generative AI to propose or refine the research questions, the EDE framework, or the design of the evidence interventions, and the remaining tasks with required disclosure are not applicable to this work.

Retrieval and discovery. We used ChatGPT and Claude to search for and shortlist related work. Every AI-suggested reference was checked by the authors against the original source; items we could not verify were not cited.

Writing assistance. We used ChatGPT and Claude to polish author-written text and improve its readability.

All AI-assisted work was reviewed by the authors. Reported numbers and table entries were not taken from AI output: each was recomputed from the stored per-item predictions through a single scoring path and checked against the tables. The authors take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work trains no new models; it evaluates publicly available checkpoints on existing benchmarks under their original evaluation protocols. Appendix A.2 gives the checkpoints for the four video LLMs and three scale variants, the frame budget (eight frames; sixteen for VideoChat2), each model's native preprocessing, deterministic decoding held fixed across all benchmarks and conditions, and the scoring convention: question-level accuracy throughout, with VideoHallucer computed over its five core subsets (1,952 questions, excluding `fact_detect`). Appendix A.1 maps benchmarks to target evidence, and Appendix B.2 recomputes the profiles under benchmark-specific metrics on the same evaluation sets.

The evidence interventions are deterministic (Appendix B.5): the permutation seed is derived from each item's prompt string, so it depends on neither the global random state nor the evaluation order, and a rerun assigns the same permutation to the same item; sensitivity to the frame position and to the seed is reported in the same section. Appendix A.3 gives the mitigation implementations, including every departure from the original releases. For the diagnostics, see Appendices D.1 and D.2 for the offset grid and matching quality, §3.4 and Appendix E for the matched semantic units, and Appendix F.1 for paraphrase generation. Confidence intervals use a paired bootstrap with 10,000 resamples, over items and over pairs for pair accuracy.

We store per-item predictions, but not the per-item frame indices selected or permutations applied; reproducing that step relies on the deterministic rules in Appendix B.5. Code and per-item outputs will be released upon publication.

REFERENCES

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

arXiv:2511.21631, 2025.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022.

Mehrdad Fazli, Sina Mansouri, Mohit Marvania, and Ziwei Zhu. Does playing it safe count as faithfulness? Reassessing LVLM hallucination mitigation methods. *arXiv preprint arXiv:2609.01888*, 2026.

Haojian Huang, Harold Haodong Chen, Meng Luo, Junjia Du, Shanqing Xu, Ziheng Chen, Yanxiang Huang, Yinchuan Li, and Ying-Cong Chen. No place to hide: Benchmarking video hallucination with background-controlled pairs. *arXiv preprint arXiv:2606.31933*, 2026a.

Yiyang Huang, Yitian Zhang, Yizhou Wang, Mingyuan Zhang, Liang Shi, Huimin Zeng, and Yun Fu. Distorted or fabricated? a survey on hallucination in video LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26599–26611. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.1325. URL https://aclanthology.org/2026.findings-acl.1325/.

Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. MHBench: Demystifying motion hallucination in VideoLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4401–4409, 2025.

Zixuan Lan, Luzhe Sun, Matthew R. Walter, and Jiawei Zhou. Seeing without looking: Do vision-language benchmarks really test vision? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026. arXiv:2605.22903.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 487–507, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206. IEEE, 2024.

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

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025.

Zijian Liu, Sihan Cao, Pengcheng Zheng, Kuien Liu, Caiyan Qin, Xiaolin Qin, Jiwei Wei, and Chaoning Zhang. Relaxing anchor-frame dominance for mitigating hallucinations in video large language models. *arXiv preprint arXiv:2604.12582*, 2026. URL https://arxiv.org/abs/2604.12582.

Meng Luo, Shengqiong Wu, Liqiang Jing, Tianjie Ju, Li Zheng, Jinxiang Lai, Tianlong Wu, Xinya Du, Jian Li, Siyuan Yan, et al. Dr. V: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. *International Journal of Computer Vision*, 134(6):278, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

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 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44545–44560. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2062. URL https://aclanthology.org/2026.acl-long.2062/.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2409.16597.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G. Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.30. URL https://aclanthology.org/2025.naacl-long.30/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Experimental Details

A.1 Benchmarks and Target Evidence

The evaluated benchmarks differ in both their original task definitions and the visual and temporal evidence implicated by those task designs. Table 6 summarizes the original task categories, their benchmark-level placement in the video-hallucination taxonomy of Huang et al. (2026b), the target evidence identified from the original task designs, and the EDE interventions most relevant to probing that evidence.

A.2 Video LLMs and Inference Settings

Our main evaluation includes four video LLMs from different model families: Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025), and VideoChat2 (Li et al., 2024). We additionally evaluate Qwen3-VL-4B, Qwen3-VL-32B, and InternVL3-14B as within-family scale variants. For VideoChat2, we use the HD stage-4 Mistral-7B release. The main experiments use eight uniformly sampled video frames. VideoChat2 is the sole exception, for which we follow its standard sixteen-frame inference configuration. Each video LLM retains its native preprocessing pipeline rather than being forced into a shared preprocessing implementation; for Qwen3-VL, we use a 448×448 frame resolution. Qwen3-VL, InternVL3, and VideoChat2 use greedy decoding, whereas MiniCPM-V-4.5 uses deterministic beam search with `num_beams=3`. Generation settings are held fixed within each video LLM across all benchmarks and evidence conditions.

For multiple-choice evaluation, the benchmark questions and answer choices are left unchanged; we only constrain the output to the option format required by the corresponding evaluation parser.

Table 6: Benchmark task structure and target evidence. Original task categories are reported separately from the taxonomy of Huang et al. (2026b). Target evidence is identified from the original task designs before examining EDE results, and the listed EDE interventions indicate relevant diagnostic probes rather than expected performance changes.

Benchmark	Original Task / Categories	Survey Taxonomy	Target Evidence	Relevant EDE Intervention(s)
MHBench	Motion Hallucination	Event Misordering	Action-identity evidence that distinguishes an action from its antonym and completed from partially executed actions.	Black; Text-only; Single; Shuffle
Event Hallusion	Event hallucination (Entire / Mix / Misleading)	Scene–Event	Category-dependent event evidence: evidence for uncommon events in Entire/Mix, and evidence sufficient to reject a linguistically primed but absent event in Misleading.	Black; Text-only; Single
Video Hallucer	Intrinsic: Object-Relation, Temporal, Semantic Detail; Extrinsic: Factual, Non-factual	Duration Distortion	Category-dependent video evidence spanning object/relation and fine-grained visual content and temporal position/order/duration; extrinsic subsets additionally involve external knowledge.	Black; Text-only; Single; Shuffle
Dr.V-Bench	Perceptive / Temporal / Cognitive Hallucination	Not covered	Category-dependent evidence spanning static perception, actions and temporal relations, and spatiotemporal evidence combined with commonsense or causal reasoning.	Black; Text-only; Single; Shuffle
VidHalluc-ACH	Action Hallucination (ACH)	Event Misordering†	Evidence identifying the prominent action when surface appearance is an unreliable cue by construction.	Black; Text-only; Single
VidHalluc-TSH	Temporal Sequence Hallucination (TSH)	Event Misordering†	The temporal order of two consecutive actions, together with evidence for their identities.	Shuffle; Single
VidPair-Halluc	Spatial and temporal reasoning	Not covered	Fine-grained foreground semantics across spatial and temporal aspects; temporal pairs additionally require evidence about event order.	Black; Text-only; Single; Shuffle

Note. The survey taxonomy follows benchmark-level placements in Huang et al. (2026b); ACH and TSH are not separately classified, and Dr.V-Bench and VidPair-Halluc are not covered. Target evidence and corresponding EDE interventions are identified from the original benchmark definitions and task construction prior to examining the EDE results.

We report question-level accuracy throughout the main analysis for consistent comparison across all evaluation tasks. For VideoHallucer, accuracy is computed over its five core hallucination subsets (1,952 questions), excluding `fact_detect`. For benchmarks that additionally define pair-based or aggregate evaluation metrics, we recompute the EDE profiles under those benchmark-specific metrics in Table 8. These comparisons use exactly the same underlying evaluation questions as the main analysis; in particular, VideoHallucer uses the same five core subsets (976 pairs / 1,952 questions).

A.3 Hallucination Mitigation Implementations

We evaluate six hallucination-mitigation methods spanning contrastive decoding, attention intervention, latent-space steering, and preference-based training. For the main evaluation, each method uses one fixed configuration across benchmarks and evidence conditions without benchmark-specific tuning. Unless otherwise noted, inference-time methods use Qwen3-VL-8B-Instruct.

MotionCD. MotionCD (Kong et al., 2025) contrasts predictions from the original video and its temporally reversed frames. We use the released implementation with contrastive strength $\lambda = 10$

and plausibility threshold $\beta = 0.1$; the implementation parameter `mcd_alpha=20` is halved when forming the contrastive logits.

TCD. TCD (Zhang et al., 2026) contrasts predictions against a temporally downsampled video; without an official implementation, we reimplement it from the paper and adapt it to Qwen3-VL-8B-Instruct. We use the reported best contrastive strength $\alpha = 0.5$ and $\beta = 0.1$, a threshold evaluated in the paper's ablation, since no single final threshold is specified. Our $8 \to 4$ downsampling retains the four-frame target of the original VideoChat2 setting ($16 \to 4$), adapting the ratio because the paper provides no general downsampling rule.

VCD. We extend image-based VCD (Leng et al., 2024) to video by independently applying its released diffusion-noise operator to preprocessed frame tensors and contrasting predictions from the original and corrupted inputs. We use $\alpha = 1.0$, $\beta = 0.1$, and diffusion step $T = 500$. This diffusion step follows the original paper's generative evaluation setting rather than its discriminative setting of $T = 999$.

DTR. DTR (Liu et al., 2026) modifies frame-level attention logits before softmax to reduce excessive attention to dominant frames; we use the published parameters $\alpha = 0.5$, $\beta = 0.4$, and $\epsilon = 10^{-6}$. Although the paper evaluates Qwen3-VL-8B, the released intervention code supports only Video-LLaVA, so we transfer the intervention range by relative decoder depth: layers 18–31 of its 32-layer decoder map to layers 20–35 of the 36-layer Qwen3-VL decoder. We use eager attention to materialize the attention matrices required by the intervention.

VTI. VTI (Liu et al., 2025) steers latent representations using separately estimated visual and textual directions. Our Qwen3-VL adaptation uses $\alpha = \beta = 0.9$, a masking ratio of 0.99, 50 masking trials, and directions estimated from 50 demonstration examples with eight frames per video. The textual direction uses paired non-hallucinated and hallucinated answers from our video benchmark pool in place of the original image-caption pairs.

Hound-DPO. For Hound-DPO (Zhang et al., 2025), the only training-based method, we compare the released preference-optimized checkpoint with its corresponding Hound-SFT checkpoint, which serves as the baseline for computing gains. Both checkpoints use eight frames and greedy decoding.

Condition applicability. MotionCD, TCD, VCD, and DTR require visual inputs and are undefined under Text-only evaluation. VTI retains only its textual steering component under Text-only, while Hound-DPO remains applicable through its trained weights. Under Black Frames and Single Repeated Frame, temporal reversal leaves the input unchanged, so MotionCD's two branches coincide and its contrastive term vanishes. We mark such structurally inactive cases with †; unmarked numerical zeros denote measured outcomes.

B Evidence-Dependence Validation and Robustness

This appendix supports the evidence-dependence profiles in Table 1 (§4.2) across six aspects: model scale, additional benchmark-defined metrics, the structural ceiling of the no-video conditions, item-level prediction agreement between Black Frames and Text-only, temporal-intervention construction, and responses across benchmark subsets. Unless otherwise noted, results use Qwen3-VL-8B with eight sampled frames and greedy decoding. VideoHallucer is evaluated on the five core hallucination subsets described in Appendix A.2.

B.1 Scale Variants within Video LLM Families

Table 1 reports one model size per video LLM family. To examine whether the main qualitative patterns in these profiles persist across model sizes, we apply the full intervention set to Qwen3-VL-4B, Qwen3-VL-32B, and InternVL3-14B (Table 7).

The main qualitative patterns in Table 1 persist across these three additional video LLMs. MHBench improves under Black Frames for all three variants ($+2.79$ to $+4.92$ percentage points) while declining under Frame Shuffle (-12.13 to -7.46). VidHalluc-ACH declines by more than 37 percentage

Table 7: **Evidence-dependence profiles across model sizes within video LLM families.** The same evidence interventions as in Table 1 are applied to three additional video LLMs: Qwen3-VL-4B, Qwen3-VL-32B, and InternVL3-14B. *Original* reports question-level accuracy (%) on unmodified videos; the remaining columns report the change (Δ) from *Original* under each intervention in percentage points.

			Visual evidence		Temporal evidence	
Benchmark	Video LLM	Original	Black Frames	Text-only	Single Repeated Frame	Frame Shuffle
MHBench	Qwen3-VL-4B	63.25	+3.42	+2.71	+0.79	−8.46
	Qwen3-VL-32B	63.88	+2.79	+2.79	−0.17	−12.13
	InternVL3-14B	61.75	+4.92	−8.75	+4.96	−7.46
EventHallusion	Qwen3-VL-4B	69.93	−12.96	−17.11	−3.42	−4.89
	Qwen3-VL-32B	67.97	−24.94	−5.62	−7.09	−6.36
	InternVL3-14B	70.42	−17.85	−23.72	−15.40	−4.40
VideoHallucer	Qwen3-VL-4B	74.59	−18.65	−16.91	−6.51	−2.56
	Qwen3-VL-32B	76.90	−19.16	−21.36	−6.56	−2.15
	InternVL3-14B	76.13	−17.06	−14.14	−6.71	−2.05
Dr.V-Bench	Qwen3-VL-4B	69.59	−17.04	−15.86	−8.77	−1.41
	Qwen3-VL-32B	73.04	−15.30	−22.46	−9.82	−0.89
	InternVL3-14B	75.40	−21.18	−22.10	−14.09	−1.84
VidHalluc-ACH	Qwen3-VL-4B	87.05	−37.05	−37.05	−13.63	−1.20
	Qwen3-VL-32B	89.89	−39.89	−39.89	−11.45	−0.70
	InternVL3-14B	89.27	−39.27	−39.27	−12.43	−0.62
VidHalluc-TSH	Qwen3-VL-4B	87.83	−26.50	−26.83	−47.00	−34.00
	Qwen3-VL-32B	95.33	−28.17	−26.67	−55.83	−36.50
	InternVL3-14B	92.50	−34.33	−28.17	−69.17	−37.67
VidPair	Qwen3-VL-4B	53.85	−8.38	−9.18	−5.83	−2.05
	Qwen3-VL-32B	56.65	−11.15	−11.48	−7.23	−1.55
	InternVL3-14B	60.13	−15.73	−14.08	−6.77	−0.40

points under Black Frames but by at most 1.20 under Frame Shuffle. VidHalluc-TSH declines under both temporal interventions, with larger reductions under Single Repeated Frame (-69.17 to -47.00). Although Qwen3-VL-32B achieves higher original-condition accuracy than Qwen3-VL-4B on six of the seven evaluation tasks, these qualitative contrasts remain consistent across the tested model sizes.

B.2 Validation with Additional Benchmark Metrics

We report question-level accuracy throughout the main analysis for consistent comparison across evaluation tasks. For benchmarks that additionally define pair-based or aggregate metrics, we also evaluate the EDE profiles under those metrics using the same evaluation items (Table 8). Specifically, we report pair accuracy for MHBench, VideoHallucer, and VidHalluc-ACH, and wAcc for VidPair, alongside question-level accuracy. This comparison examines whether the observed intervention effects persist under these additional scoring rules.

Three of the four evaluation tasks preserve the sign of every intervention effect under these additional metrics. MHBench improves under Black Frames ($+10.75$ for question-level accuracy, $+5.08$ for pair accuracy) and Text-only ($+8.58$, $+7.50$), and declines under Frame Shuffle (-7.00, -9.92). VideoHallucer and VidHalluc-ACH also preserve the sign across all four interventions, with larger absolute declines under pair accuracy. In particular, VidHalluc-ACH reaches 0.00% pair accuracy under both Black Frames and Text-only: a pair is correct only when both items are answered correctly, whereas the model gives identical answers to the two items with opposite gold labels under these conditions.

VidPair shows metric-dependent sign changes. Under wAcc, the Black Frames and Single Repeated Frame effects are positive ($+0.38$ and $+0.19$), whereas their question-level accuracy changes are

Table 8: **EDE profiles under additional benchmark-defined metrics.** For benchmarks that additionally define pair-based or aggregate metrics, we report EDE results under those metrics alongside question-level accuracy. All results use Qwen3-VL-8B and the same evaluation items across metrics. Δ reports the change from the original condition in percentage points under the corresponding metric. VidPair's wAcc is computed using our implementation of Eqs. (2)–(4) of the original paper rather than its released evaluation code.

		Per-question acc.		Additional metric	
Benchmark	Condition	Score	Δ	Score	Δ
MHBench (pair acc.)					
	Original	55.92	+0.00	28.25	+0.00
	Black	66.67	+10.75	33.33	+5.08
	Text-only	64.50	+8.58	35.75	+7.50
	Single	58.71	+2.79	31.00	+2.75
	Shuffle	48.92	−7.00	18.33	−9.92
VideoHallucer (pair acc.)					
	Original	76.23	+0.00	57.38	+0.00
	Black	56.45	−19.77	17.73	−39.65
	Text-only	59.32	−16.91	24.59	−32.79
	Single	68.14	−8.09	39.86	−17.52
	Shuffle	74.69	−1.54	54.00	−3.38
VidHalluc-ACH (pair acc.)					
	Original	87.19	+0.00	75.09	+0.00
	Black	50.00	−37.19	0.00	−75.09
	Text-only	50.00	−37.19	0.00	−75.09
	Single	73.40	−13.79	48.65	−26.43
	Shuffle	85.85	−1.35	72.37	−2.71
VidPair (wAcc)					
	Original	53.43	+0.00	36.90	+0.00
	Black	45.50	−7.92	37.28	+0.38
	Text-only	43.38	−10.05	34.00	−2.90
	Single	48.83	−4.60	37.09	+0.19
	Shuffle	52.18	−1.25	35.66	−1.24

negative (-7.92 and -4.60). Its evidence-dependence profile is therefore sensitive to metric choice and should be interpreted using both question-level accuracy and wAcc.

B.3 Structural Ceiling of the No-Video Conditions

When the same question appears across different videos, a system that gives one fixed answer to that question without distinguishing the videos may not answer every item correctly. For example, if ten items share a question and seven have the gold answer Yes while three have No, the highest accuracy attainable with one fixed answer is 70%. Table 9 reports the structural ceiling obtained by aggregating these maximum correct counts over all evaluation items. This ceiling assumes deterministic decoding and identical remaining input conditions, and is computed from question-level label distributions alone, without model outputs.

The structural ceiling differs across benchmarks. VidHalluc-ACH has a ceiling of 50.00% because its pairs present the same question against two videos with opposite gold answers. In contrast, EventHallusion has a ceiling of 100.00% because gold labels are consistent within each group of identical questions. Scores and score changes after removing video evidence should therefore be interpreted in light of these differences in question and label structure.

MHBench illustrates how high scores can be obtained without video evidence. Its $2{,}400$ evaluation items contain only 200 unique questions, and imbalanced labels within question groups yield a structural ceiling of 68.21%, above its original accuracy of 55.92%. Black Frames reaches 66.67%, only 1.54 percentage points below this ceiling. VidHalluc-ACH reaches its ceiling of 50.00%; the benchmarks other than MHBench and VidHalluc-ACH remain at least 24.92 percentage points below their ceilings under Black Frames. These results support the interpretation that MHBench score

Table 9: **Structural ceiling of the no-video conditions.** Under deterministic decoding, identical question text and remaining input conditions yield identical answers. *Ceiling* is the maximum question-level accuracy attainable by assigning one fixed answer to each unique question, computed from gold-label distributions without model outputs. *Original*, *Black Frames*, and *Text-only* report Qwen3-VL-8B accuracy under the corresponding conditions. Accuracies and ceilings are reported in %; *Gap to ceiling* is the ceiling minus Black Frames accuracy, in percentage points. *Unique questions* counts distinct question strings, and *Items* counts evaluation items.

Benchmark	Ceiling	Original	Black Frames	Text-only	Gap to ceiling	Unique questions	Items
MHBench	68.21	55.92	66.67	64.50	1.54	200	2,400
EventHallusion	100.00	71.64	61.37	43.28	38.63	159	409
VideoHallucer	89.04	76.23	56.45	59.32	32.58	1,440	1,952
Dr.V-Bench	96.72	68.80	53.30	56.22	43.42	2,503	3,045
VidHalluc-ACH	50.00	87.19	50.00	50.00	0.00	3,819	8,550
VidHalluc-TSH	99.33	80.00	57.33	58.33	42.00	590	600
VidPair	70.43	53.43	45.50	43.38	24.92	1,005	4,000

Table 10: **Item-level prediction differences between Black Frames and Text-only.** Results use Qwen3-VL-8B. Black Frames removes informative visual content while preserving the visual-input structure; Text-only removes visual input entirely. Accuracy is reported in %, and *Diff.* is Black Frames accuracy minus Text-only accuracy in percentage points. *Disagree* is the percentage of all evaluation items whose extracted answers differ between the conditions. The two columns under *Correct only under* report the percentage of all items answered correctly under the indicated condition and incorrectly under the other. *Items* is the number of evaluation items.

	Accuracy and difference				Correct only under		
Benchmark	Black Frames	Text-only	Diff.	Disagree	Black Frames	Text-only	Items
MHBench	66.67	64.50	+2.17	10.0%	6.1%	3.9%	2,400
EventHallusion	61.37	43.28	+18.09	22.0%	20.0%	2.0%	409
VideoHallucer	56.45	59.32	−2.87	34.5%	15.8%	18.7%	1,952
Dr.V-Bench	53.30	56.22	−2.92	36.7%	16.9%	19.8%	3,045
VidHalluc-ACH	50.00	50.00	+0.00	48.7%	24.3%	24.3%	8,550
VidHalluc-TSH	57.33	58.33	−1.00	7.3%	3.2%	4.2%	600
VidPair	45.50	43.38	+2.13	11.8%	7.0%	4.9%	4,000

gains need not indicate stronger use of video evidence, but do not identify which non-visual cues the model actually uses.

B.4 Black Frames and Text-only Are Distinct Conditions

Black Frames replaces sampled frames with black frames, removing informative visual content while preserving the visual-input structure. Text-only removes the visual input entirely. Table 10 examines whether similar aggregate accuracy under these conditions is accompanied by agreement in item-level predictions.

On VidHalluc-ACH, both conditions yield 50.00% accuracy, yet their predictions differ on 48.7% of evaluation items. Items answered correctly only under Black Frames and only under Text-only each account for 24.3% of all items. VideoHallucer and Dr.V-Bench exhibit accuracy differences of -2.87 and -2.92 percentage points, respectively, alongside prediction disagreement rates of 34.5% and 36.7%. For EventHallusion, 20.0% of items are correct only under Black Frames, compared with 2.0% only under Text-only, consistent with the accuracy difference of $+18.09$ percentage points. These results show that aggregate accuracy alone does not characterize prediction differences between the two conditions and support reporting Black Frames and Text-only as separate interventions.

B.5 Construction of the Temporal Interventions

Single Repeated Frame requires selecting a frame, whereas Frame Shuffle requires selecting a permutation. We evaluate repeated frames selected using relative positions 0.25, 0.50, and 0.75 in the sampled frame sequence and three deterministic permutation settings (Table 11).

Across the tested seeds, the maximum within-row variation in the Frame Shuffle effect is 0.79 percentage points on MHBench, 0.05 on VidHalluc-ACH, and 4.17 on VidHalluc-TSH. Single Repeated Frame is more sensitive to frame selection: on VidHalluc-TSH with Qwen3-VL-8B, its effect ranges from -54.00 to -18.00 percentage points, while on MHBench it changes sign, from -0.04 at Point 1 to $+8.46$ at Point 3.

Repeating a single frame removes information across frames while retaining content specific to the selected frame. Frame Shuffle preserves the frame set and disrupts its order. We therefore interpret Single Repeated Frame as a probe of performance with reduced multi-frame information, whose outcome depends on the selected frame, and use Frame Shuffle to probe sensitivity to temporal order.

Table 11: **Sensitivity to temporal-intervention construction.** For Single Repeated Frame, Points 1–3 are the frames at relative positions 0.25, 0.50, and 0.75 of the *sampled* frame sequence (zero-based index $\min(n-1, \lfloor np \rfloor)$ among n sampled frames; with $n = 8$, these correspond approximately to 28.6%, 57.1%, and 85.7% of the clip); the selected frame is then repeated n times, leaving the frame count unchanged. For Frame Shuffle, Points 1–3 use offsets 0, 1, and 2 added to a deterministic seed derived from each prompt; reruns reproduce the same permutations. *Original* reports question-level accuracy (%); Points 1–3 report changes from Original in percentage points. * marks the setting used in the main tables, which is Point 2 for Single Repeated Frame and Point 1 for Frame Shuffle.

Video LLM	Benchmark	Intervention	Original	Point 1	Point 2	Point 3
Qwen3-VL-8B	MHBench	Single	55.92	−0.04	+2.79*	+8.46
Qwen3-VL-8B	MHBench	Shuffle	55.92	−7.00*	−6.21	−6.33
Qwen3-VL-8B	VidHalluc-ACH	Single	87.19	−14.35	−13.79*	−14.96
Qwen3-VL-8B	VidHalluc-ACH	Shuffle	87.19	−1.35*	−1.38	−1.36
Qwen3-VL-8B	VidHalluc-TSH	Single	80.00	−18.00	−43.33*	−54.00
Qwen3-VL-8B	VidHalluc-TSH	Shuffle	80.00	−24.50*	−21.00	−22.33
InternVL3-8B	MHBench	Single	49.38	−1.17	+2.42*	+7.25
InternVL3-8B	MHBench	Shuffle	49.38	−4.33*	−4.58	−4.67
InternVL3-8B	VidHalluc-ACH	Single	88.32	−10.98	−11.23*	−12.28
InternVL3-8B	VidHalluc-ACH	Shuffle	88.32	−0.28*	−0.30	−0.33
InternVL3-8B	VidHalluc-TSH	Single	89.50	−58.00	−65.50*	−69.50
InternVL3-8B	VidHalluc-TSH	Shuffle	89.50	−30.33*	−27.83	−32.00

B.6 Within-Benchmark Decomposition

To examine how intervention effects vary with task-specific evidence requirements, Table 12 decomposes Δ using each benchmark's subset annotations. Results use Qwen3-VL-8B, with frame count, answer format, and decoding held fixed within each benchmark.

In VideoHallucer, the `temporal` subset, which includes event position, relative order, and duration, shows the largest decline under Frame Shuffle (-2.84 percentage points); the other four subsets decline by at most 1.50 percentage points. The temporal subset is also least affected by Black Frames (-5.68), whereas `semantic_detail` and `object_relation` decline by -35.00 and -31.75, respectively. These contrasting responses are consistent with sensitivity to different evidence requirements.

Table 13 extends the VideoHallucer Frame Shuffle analysis to all seven evaluated video LLMs, including the three additional model-size variants. The temporal subset shows the largest score decrease in three of the four main video LLMs and all three additional variants. For InternVL3-8B and MiniCPM-V-4.5, temporal-subset decreases are 7.95 and 7.10 percentage points, respectively, compared with aggregate decreases of 1.64 and 1.79. VideoChat2 is the exception: its largest decrease

occurs in the external factual subset (2.00 percentage points), while the temporal subset decreases by 0.28. For Qwen3-VL-4B, the temporal decrease exceeds that of the external non-factual subset by only 0.23 percentage points. These comparisons describe numerical rankings rather than statistical significance.

In Dr.V-Bench, `factual_prediction` (-9.05), `dynamic_attribute` (-3.33), and `dynamic_relation` (-2.67) show the largest Frame Shuffle declines. In contrast, `static_relation` ($+1.27$), `object` ($+2.00$), and `counterfactual_prediction` ($+0.00$) do not decline. The counterfactual subset changes by 0.00 to $+3.33$ percentage points across all four interventions. These results indicate limited sensitivity under the tested interventions, but do not establish that its questions are answerable from text alone.

EventHallusion exhibits substantial variation across subsets. Its `mix` subset declines under every intervention (-40.41, -46.63, -23.32, -17.10), whereas `entire` improves under Black Frames ($+27.19$) and Single Repeated Frame ($+12.28$). The `misleading` subset is unchanged under both temporal interventions. The aggregate profile therefore combines distinct subset-level responses.

Table 12: **Evidence-dependence profiles by benchmark subset.** Results use Qwen3-VL-8B and each benchmark's subset annotations, with rows sorted by Frame Shuffle within each benchmark. Subsets with fewer than 80 items are omitted. VideoHallucer includes the five core hallucination subsets. n is the number of evaluation items, *Original* is question-level accuracy (%), and the intervention columns report changes from Original in percentage points.

Benchmark	Subset	n	Original	Black	Text-only	Single	Shuffle
VideoHallucer	temporal	352	75.85	−5.68	−9.66	−10.80	−2.84
VideoHallucer	external_nonfactual	400	75.75	−15.25	−11.25	−10.75	−1.50
VideoHallucer	external_factual	400	63.25	−9.50	−7.75	−5.00	−1.50
VideoHallucer	object_relation	400	81.25	−31.75	−20.00	−4.50	−1.00
VideoHallucer	semantic_detail	400	85.00	−35.00	−35.00	−9.75	−1.00
Dr.V-Bench	factual_prediction	210	70.48	−13.33	−4.29	−15.71	−9.05
Dr.V-Bench	dynamic_attribute	90	76.67	−26.67	−7.78	−22.22	−3.33
Dr.V-Bench	dynamic_relation	450	51.11	−5.56	−5.11	−8.67	−2.67
Dr.V-Bench	color	149	75.17	−18.79	−20.81	−9.40	−1.34
Dr.V-Bench	number	150	83.33	−12.00	−27.33	+0.00	−1.33
Dr.V-Bench	location	150	90.67	−37.33	−29.33	−1.33	−1.33
Dr.V-Bench	action	360	64.17	−15.83	−2.78	−11.39	−1.11
Dr.V-Bench	context_based_explanation	300	69.67	−15.67	−16.67	−13.33	−1.00
Dr.V-Bench	counterfactual_prediction	150	60.00	+0.67	+3.33	+2.67	+0.00
Dr.V-Bench	knowledge_based_explanation	180	70.56	−15.00	−9.44	−10.56	+0.00
Dr.V-Bench	ocr	150	73.33	−18.67	−26.00	−5.33	+0.67
Dr.V-Bench	static_relation	236	78.81	−27.97	−30.93	−6.78	+1.27
Dr.V-Bench	object	350	69.14	−16.29	−10.00	−12.86	+2.00
EventHallusion	mix	193	63.21	−40.41	−46.63	−23.32	−17.10
EventHallusion	entire	114	64.91	+27.19	−8.77	+12.28	−0.88
EventHallusion	misleading	102	95.10	+4.90	−15.69	+0.00	+0.00

C Mitigation Robustness and Coverage

This appendix supports §4.3 by examining sensitivity to intervention strength and extending original-condition results beyond the four evaluation tasks in Table 2. Results use Qwen3-VL-8B with eight sampled frames and greedy decoding, except Hound-DPO, whose gains use the matched Hound-SFT checkpoint.

Table 13: **Frame Shuffle sensitivity across VideoHallucer subsets and video LLMs.** Values report question-level accuracy changes from the original condition in percentage points. All models are evaluated on the same five core subsets: 352 temporal items and 400 items in each remaining subset. *Overall* reports the change across all 1,952 items. Bold values indicate the largest subset-level decrease within each model, without implying statistical significance.

Video LLM	Temporal	Object relation	Semantic detail	External factual	External non-factual	Overall
Qwen3-VL-8B	**−2.84**	−1.00	−1.00	−1.50	−1.50	−1.54
InternVL3-8B	**−7.95**	+1.50	−0.25	−1.25	−1.00	−1.64
MiniCPM-V-4.5	**−7.10**	+1.25	−1.25	−1.75	−0.75	−1.79
VideoChat2	−0.28	+2.00	0.00	**−2.00**	+0.75	+0.10
Qwen3-VL-4B	**−3.98**	+0.75	−3.00	−3.00	−3.75	−2.56
Qwen3-VL-32B	**−4.83**	−2.25	−1.25	−0.25	−2.50	−2.15
InternVL3-14B	**−7.95**	+0.25	−1.00	−2.75	+0.50	−2.05

Table 14: **Sensitivity of original-condition gains to intervention strength.** Results cover MotionCD and selected control interventions on MHBench using Qwen3-VL-8B. For each intervention, the upper row lists strength values and the lower row gives G^{orig} in percentage points. Direction-based interventions use directions estimated with eight frames. Empty trailing cells reflect different numbers of evaluated strength settings.

Method	Strength and G^{orig}						
Additive steering	0.05	0.2	0.5	1.0			
	+0.92	+2.38	+5.33	+12.38			
Directional ablation	0.05	0.2	0.5	1.0			
	−0.96	−4.33	−9.42	+8.92			
Random direction	0.05	0.2	0.5	1.0			
	−0.50	−0.33	−0.96	−1.04			
TextVTI	0.05	0.2	0.5	0.8			
	+0.38	+1.50	+3.50	+5.54			
MotionCD	0	0.2	0.5	1	2	5	10
	−0.04	+3.08	+4.67	+4.92	+5.46	+5.67	+5.92
NoiseCD	0.2	0.5	1	2	5	10	
	−8.33	−9.29	−9.29	−9.29	−9.29	−9.29	
ShuffleCD	0.2	0.5	1	2	5	10	
	+1.54	+2.92	+3.50	+3.83	+4.38	+4.46	

C.1 Sensitivity to Intervention Strength

Table 14 evaluates original-condition gains on MHBench across intervention strengths for MotionCD and selected control interventions, holding the video LLM and decoding settings fixed.

MotionCD improves from -0.04 at $\lambda = 0$ to $+4.92$ at $\lambda = 1$, with a further increase of 1.00 percentage point to $+5.92$ at $\lambda = 10$. ShuffleCD increases from $+1.54$ to $+4.46$ over the tested range, whereas NoiseCD remains at -9.29 for $\lambda \geq 0.5$. These sweeps characterize sensitivity within the evaluated range; they do not establish invariance to settings outside that range.

Additive steering increases monotonically from $+0.92$ to $+12.38$ over the tested strengths. Directional ablation is non-monotonic, declining to -9.42 at $\alpha = 0.5$ before increasing to $+8.92$ at $\alpha = 1.0$. Random direction produces small negative gains throughout (-1.04 to -0.33), with an absolute magnitude of at most 1.04 percentage points. TextVTI increases from $+0.38$ to $+5.54$. These results show that original-condition gains can depend substantially on intervention strength.

Table 15: **Original-condition gains on the remaining evaluation tasks.** G^{orig} is reported in percentage points for evaluated method–task combinations on the three tasks not shown in Table 2. Hound-DPO uses its matched Hound-SFT baseline; other methods use Qwen3-VL-8B. Combinations that were not run are omitted.

Benchmark	Method	Video LLM	G^{orig}
Dr.V-Bench	MotionCD	Qwen3-VL-8B	-1.64
	Hound-DPO	Hound	-0.23
VideoHallucer	MotionCD	Qwen3-VL-8B	-4.61
	TCD	Qwen3-VL-8B	-0.15
	VCD	Qwen3-VL-8B	-1.18
	DTR	Qwen3-VL-8B	-5.79
	Hound-DPO	Hound	$+3.23$
VidPair	MotionCD	Qwen3-VL-8B	$+0.50$
	TCD	Qwen3-VL-8B	$+1.23$
	VCD	Qwen3-VL-8B	$+1.25$
	DTR	Qwen3-VL-8B	-6.90
	Hound-DPO	Hound	$+2.15$

C.2 Coverage beyond the Four Evaluation Tasks

Table 2 reports condition-matched gains on four evaluation tasks. Table 15 extends coverage to original-condition gains on the three remaining tasks, reporting the method–task combinations evaluated among the six mitigation methods. TCD, VCD, DTR, and VTI were not run on Dr.V-Bench, and VTI was not run on VideoHallucer or VidPair; these combinations are omitted.

Seven of the twelve combinations have negative gains, and the largest positive gain is $+3.23$ percentage points. On VideoHallucer, the inference-time methods in this table yield negative gains: MotionCD -4.61, TCD -0.15, VCD -1.18, and DTR -5.79. Hound-DPO yields $+3.23$ relative to its corresponding Hound-SFT checkpoint. On VidPair, MotionCD, TCD, and VCD yield $+0.50$, $+1.23$, and $+1.25$, respectively, whereas DTR yields -6.90 and Hound-DPO yields $+2.15$. On Dr.V-Bench, the two evaluated methods, MotionCD and Hound-DPO, yield -1.64 and -0.23.

Together with Table 2, these results show that the evaluated methods do not provide uniformly positive gains across tasks, even under the original video condition.

D Response-Tendency Diagnostics

This appendix supports §4.4 with matching-quality estimates, residual confidence intervals, the Yes-logit offset grid, and additional control comparisons. Results use Qwen3-VL-8B on MHBench with eight frames and greedy decoding. Hound-DPO uses its corresponding Hound-SFT baseline and a separate offset grid built from that checkpoint.

D.1 Matching Quality and Uncertainty of the Residual

We write $E^{\text{acc}} = S_M - S_{\delta^*_M}$ and $E^{\text{pair}} = S^{\text{pair}}_M - S^{\text{pair}}_{\delta^*_M}$ for the method's accuracy and pair accuracy minus those of the response-rate-matched control, where a pair is correct only if both answers are correct. These residuals compare each method with a baseline whose predicted-Yes rate is matched as closely as possible on the offset grid.Table 16 reports both the matching error and bootstrap confidence intervals for these residuals.

Nine of the ten methods match within 1.5 percentage points, and five within 0.7. VTI has the largest error (2.63 percentage points): its predicted-Yes rate of 77.38% exceeds the maximum grid rate of 74.75% at $\delta = +2$. Directional ablation lies below the other end of the grid (19.42% versus 20.88% at $\delta = -8$; error 1.46 percentage points). Residuals for methods near the grid boundaries should therefore be interpreted with the remaining rate mismatch in mind.

Table 16: **Response-rate matching quality and residual uncertainty.** Results are on MHBench. ρ_M is the method's predicted-Yes rate (%), and the matching error is $|\rho_M - \rho(\delta^*)|$ in percentage points. The residual columns report E^{acc} and E^{pair} with 95% confidence intervals from 10,000 paired-bootstrap resamples. Residuals and interval endpoints are in percentage points; * indicates an interval excluding zero. Hound-DPO uses an offset grid built from its own Hound-SFT baseline.

	Matching quality			Residual with 95% CI	
Method	ρ_M	δ^*	Match err (pp)	E^{acc}	E^{pair}
MotionCD	59.75	-0.5	0.67	$+4.75^*$ [$+3.33$, $+6.21$]	$+9.92^*$ [$+7.75$, $+12.25$]
TCD	62.13	$+0$	1.21	$+2.96^*$ [$+2.00$, $+3.92$]	$+5.50^*$ [$+3.92$, $+7.08$]
VCD	62.29	$+0$	1.04	$+0.46$ [-0.67, $+1.62$]	$+2.17^*$ [$+0.25$, $+4.08$]
DTR	58.06	-1	0.31	-10.13^* [-12.08, -8.17]	-14.67^* [-17.33, -11.92]
Hound-DPO	56.17	$+0$	0.93	$+0.04$ [-1.46, $+1.54$]	$+0.75$ [-1.08, $+2.58$]
VTI	77.38	$+2$	2.63	$+5.96^*$ [$+3.75$, $+8.08$]	$+2.08$ [-0.83, $+4.83$]
Additive steering	32.88	-5	0.08	$+5.17^*$ [$+2.58$, $+7.75$]	$+6.08^*$ [$+1.92$, $+10.08$]
Directional ablation	19.42	-8	1.46	$+2.21$ [-0.25, $+4.67$]	$+1.92$ [-2.42, $+6.25$]
Random direction	64.79	$+0.25$	0.13	-0.63 [-1.54, $+0.29$]	-1.00 [-2.42, $+0.42$]
ShuffleCD	60.12	-0.5	0.29	$+3.29^*$ [$+1.88$, $+4.75$]	$+7.08^*$ [$+4.83$, $+9.33$]

Six of the ten E^{acc} confidence intervals exclude zero. The intervals for Random direction (-0.63, $[-1.54, +0.29]$) and Hound-DPO ($+0.04$, $[-1.46, +1.54]$) include zero. Although Directional ablation has a raw gain of $+8.92$ (Table 3), its residual interval also includes zero ($+2.21$, $[-0.25, +4.67]$). Additive steering retains a positive residual interval ($+5.17$, $[+2.58, +7.75]$). The same distinction holds for their pair-accuracy residuals.

D.2 The Yes-Logit Offset Grid

Table 17 reports the full offset grid on MHBench. Each setting adds a constant to the Yes logit without changing video processing; accuracy, FAR, FRR, and the predicted-Yes rate are recorded.

Accuracy ranges from 53.42% to 63.54%, a spread of 10.12 percentage points. Relative to the baseline accuracy of 55.92%, the best grid score represents a gain of 7.62 percentage points. Thus, substantial score gains can arise from an output-side response shift without modifying video processing. The accuracy-maximizing offset is $\delta = -4.5$, but no method is matched to this point. Instead, δ^* is selected by predicted-Yes rate, with selected offsets ranging from -8 to $+2$.

Positive offsets extend the grid to higher predicted-Yes rates and improve matching for methods such as Random direction (64.79%) and VTI (77.38%). Without these offsets, residuals could still be computed, but against baselines with larger response-rate mismatches.

D.3 Additional Response-Tendency Diagnostics

Table 18 reports additional contrastive-decoding variants and controls beyond those in Table 3.

With contrastive strength $\lambda = 0$, MotionCD yields a gain and residual of -0.04 with $\delta^* = 0$. DTR at $\alpha = \beta = 0$ yields a gain and residual of $+0.21$. These reported effects are small but nonzero.

NoiseCD and BlankCD have identical reported values for gain (-9.29), ΔFAR ($+18.33$), selected offset ($+1.5$), and E^{acc} (-7.38). These values associate the score decline with increased false acceptance but do not establish its internal cause. At $\lambda = 1$, ShuffleCD yields a smaller gain than MotionCD ($+3.50$ versus $+4.92$), with reduced FAR and little change in FRR.

The reported E^{acc} values using 12-frame versus 8-frame directions are $+5.13$ versus $+5.17$ for Additive steering, $+2.88$ versus $+2.21$ for Directional ablation, and -1.63 versus -0.63 for Random direction. The residual signs are unchanged for these comparisons.

Table 17: **Yes-logit offset grid on MHBench.** Each row adds a constant δ to the Qwen3-VL-8B baseline's Yes logit without modifying video processing. Accuracy, FAR, FRR, and predicted-Yes rate are percentages. δ^* is selected by the closest predicted-Yes rate, not by maximum accuracy. The last column identifies matched methods; empty cells indicate that no listed method selects that offset. Hound-DPO uses a separate Hound-SFT grid and selects $\delta^* = 0$.

Offset δ	Acc.	FAR	FRR	Pred. Yes (%)	Selected for
-8	62.62	8.25	66.50	20.88	Directional ablation (α=1.0, 8f dir)
-7	62.83	11.83	62.50	24.67	
-6	63.25	15.25	58.25	28.50	
-5	63.12	19.67	54.08	32.79	Additive steering (α=1.0, 8f dir)
-4.5	63.54	21.42	51.50	34.96	
-4	63.25	24.58	48.92	37.83	
-3	62.54	31.67	43.25	44.21	
-2.5	61.62	35.92	40.83	47.54	
-2	60.00	41.25	38.75	51.25	
-1.5	59.12	45.50	36.25	54.62	
-1	58.17	49.58	34.08	57.75	DTR (α=0.5, β=0.4, published)
-0.5	57.08	53.33	32.50	60.42	MotionCD (λ=10, published); ShuffleCD (λ=10)
+0	55.92	57.42	30.75	63.33	TCD (α=0.5, published); VCD (α=1.0, published)
+0.25	55.50	59.17	29.83	64.67	Random direction (α=1.0, 8f dir)
$+0.5$	55.54	60.83	28.08	66.38	
$+1$	54.96	64.67	25.42	69.62	
$+1.5$	54.00	68.25	23.75	72.25	
+2	53.42	71.33	21.83	74.75	VTI (adapted, α=β=0.9)

E Evaluation-Protocol Sensitivity

This appendix supports §4.4 by examining protocol sensitivity on full evaluation splits and under option permutations. Results use Qwen3-VL-8B with eight frames and greedy decoding, except Hound-DPO, which uses its matched Hound-SFT baseline. Direction-based interventions use directions estimated with eight frames.

E.1 Evaluation before Semantic Matching

Table 4 compares paired Yes/No and multiple-choice evaluation on matched semantic units. Table 19 reports each protocol's full split, using question-level accuracy for Yes/No rather than pair accuracy. The evaluation sets and scoring rules differ, so we examine the sign of ΔG rather than directly comparing its magnitude.

On MHBench, the sign of ΔG is preserved for all six published methods. MotionCD, TCD, VCD, and VTI have negative values (-4.75, -2.88, -2.21, and -22.46), as in the matched comparison. DTR is positive ($+0.88$ versus $+4.17$ matched), and Hound-DPO is negative (-0.46 versus -1.17). On VidHalluc-ACH, the sign is preserved for five of six methods, with DTR as the exception. Thus, protocol sensitivity is also observed without semantic matching, although its direction is not invariant in every case.

The controls show larger differences on MHBench. Additive steering yields $+12.38$ under Yes/No and -19.50 under multiple choice ($\Delta G = -31.88$), while Directional ablation yields $+8.92$ and -28.92, respectively. Random direction yields smaller negative gains of -1.04 and -1.75.

E.2 Option-Order Control

Table 20 tests sensitivity to option position by holding answer content fixed and permuting the options. Each gain is computed against the base model under the same permutation.

On VidHalluc-ACH, base accuracy varies by at most 0.51 percentage points across the three orders (90.04%, 90.55%, and 90.53%). Additive steering yields gains from -18.71 to -17.19, and Direc-

Table 18: **Response-tendency diagnosis of the remaining interventions.** The decomposition of Table 3 applied to the contrastive-decoding variants, steering configurations, and null controls not shown there. Columns are as in that table: ΔFAR and ΔFRR are changes in false acceptance and false rejection rates, ρ_M is the method's own predicted-Yes rate, δ^* is the Yes-logit offset selected to make the base model's predicted-Yes rate as close as possible to ρ_M on the evaluated grid, without optimizing accuracy; E^{acc}, E^{pair} are the method's accuracy and pair accuracy minus those of that response-matched baseline. $*$: the E value's 95% CI excludes zero (10,000 paired-bootstrap resamples with seed 42; question-level resampling for E^{acc} and pair-level resampling for E^{pair}). *12f* and *8f* denote the frame budget used to estimate the steering direction, not the evaluation input, which is eight frames throughout. All rows are on MHBench, 2,400 questions (1,200 pairs) per row with no unparsed outputs; the full offset grid is in Table 17.

	Observed gain	Error attribution			Response-matched analysis		
Method	$G^{\mathrm{orig}}\uparrow$	ΔFAR↓	ΔFRR↓	ρ_M	δ^*	$E^{\mathrm{acc}}\uparrow$	$E^{\mathrm{pair}}\uparrow$
Base model (absolute)	*55.92*	*57.42*	*30.75*	*63.33*	—	—	—
MotionCD (λ=1)	+4.92	−9.67	−0.17	58.58	−1	+2.67*	+6.00*
NoiseCD (λ=1)	−9.29	+18.33	+0.25	72.38	+1.5	−7.38*	−9.92*
NoiseCD (λ=10)	−9.29	+18.33	+0.25	72.38	+1.5	−7.38*	−9.92*
BlankCD (λ=1)	−9.29	+18.33	+0.25	72.38	+1.5	−7.38*	−9.92*
BlankCD (λ=10)	−9.29	+18.33	+0.25	72.38	+1.5	−7.38*	−9.92*
ShuffleCD (λ=1)	+3.50	−6.75	−0.25	60.08	−0.5	+2.33*	+5.25*
TextVTI (α=0.8, 8f)	+5.54	−7.92	−3.17	60.96	−0.5	+4.38*	+7.50*
Additive steering (α=1.0, 12f)	+12.33	−44.58	+19.92	31.08	−5	+5.13*	+5.17*
Directional ablation (α=1.0, 12f)	+9.58	−54.58	+35.42	18.33	−8	+2.88*	+1.92
Random direction (α=1.0, 12f)	−2.04	+3.67	+0.42	64.96	+0.25	−1.63*	−2.00*
Null controls							
MotionCD (λ=0)	−0.04	+0.08	+0.00	63.38	+0	−0.04	+0.00
DTR (α=β=0)	+0.21	−0.50	+0.08	63.04	+0	+0.21	+0.17

Table 19: **Evaluation-protocol sensitivity before semantic matching.** Results use each protocol's full split. Yes/No uses question-level accuracy, whereas the matched comparison in Table 4 uses pair accuracy. The evaluation sets and scoring rules differ across comparisons. The italicized base row reports accuracy (%); method rows report gains in percentage points, with $\Delta G = G^{\mathrm{MCQ}} - G^{\mathrm{Y/N}}$. Hound-DPO uses its matched Hound-SFT baseline.

	MHBench			VidHalluc-ACH		
Method	Y/N acc.	Multiple-choice	ΔG	Y/N acc.	Multiple-choice	ΔG
Base model (absolute)	*55.92*	*61.50*	—	*87.19*	*90.04*	—
Existing mitigation methods						
MotionCD	+5.92	+1.17	−4.75	−2.23	−1.24	+0.99
TCD	+2.96	+0.08	−2.88	+0.27	−0.30	−0.57
VCD	+0.46	−1.75	−2.21	+0.54	−0.26	−0.80
DTR	−7.88	−7.00	+0.88	−1.79	−2.99	−1.20
Hound-DPO	+0.04	−0.42	−0.46	+0.70	+0.49	−0.21
VTI	+3.46	−19.00	−22.46	−18.33	−14.94	+3.38
Mechanistic controls						
Additive steering	+12.38	−19.50	−31.88	−23.74	−17.19	+6.55
Directional ablation	+8.92	−28.92	−37.83	−37.08	−51.80	−14.72
Random direction	−1.04	−1.75	−0.71	−0.14	−0.07	+0.07

tional ablation from -52.46 to -51.80. On MHBench, only the A↔B swap is admissible because the third option refers to the first two. Base accuracy declines by 1.83 percentage points, while gains change from -19.50 to -17.33 for Additive steering and from -28.92 to -24.67 for Directional ablation. Both remain negative.

Table 20: **Option-order control for multiple-choice evaluation.** Answer content is fixed while option positions are permuted. *Base* reports Qwen3-VL-8B accuracy (%); method columns report gains in percentage points relative to the base under the same permutation. Additive, Dir-ablation, and Random denote Additive steering, Directional ablation, and Random direction. MHBench permits the A↔B swap because its third option refers to the first two; VidHalluc-ACH also permits cyclic and reversed orders.

Benchmark	Option order	Base	MotionCD	Additive	Dir-ablation	Random
MHBench	original	61.50	+1.17	−19.50	−28.92	−1.75
	A↔B swapped	59.67	+1.75	−17.33	−24.67	−2.08
VidHalluc-ACH	original	90.04	−1.24	−17.19	−51.80	−0.07
	cyclic shift	90.55	−1.54	−18.03	−52.17	−0.23
	reversed	90.53	−1.89	−18.71	−52.46	−0.44

The gain reversals persist under the tested option permutations, indicating that they cannot be explained solely by the position of the correct option.

F Question-Formulation Sensitivity

This appendix supports §4.4 by describing paraphrase generation and filtering and extending the stability analysis to additional video LLM sizes.

F.1 Paraphrase Generation and Filtering

To probe sensitivity to question wording, the paraphrase procedure is designed to preserve the question's meaning and required evidence while keeping the video and gold answer fixed during evaluation. Table 21 summarizes generation and filtering.

Both generation and validation use `gpt-5.6-luna` through the OpenAI API with a fixed decoding seed. For each unique question we request 20 candidates and validate them in a separate call, so that the generator does not assess its own output. The validator judges each candidate on four criteria: whether the correct answer could differ from the original for some video, whether the answer format and question structure are preserved, whether the candidate is well-formed, and whether it adds or removes information; candidates marked uncertain on any criterion are discarded. Among the candidates that pass, we select the three with the largest surface divergence from the original question, where surface divergence is the Jaccard distance between token sets (lowercased alphanumeric tokens, with CJK characters treated as individual tokens). Neither the video, the gold answer, nor any evaluated model's prediction is provided during generation, validation, or selection.

Candidate rejection rates range from 25.1% for VideoHallucer to 59.7% for VidHalluc-TSH. TSH has the lowest complete-set success rate (85.9%) and candidate pass rate (40.3%). Complete-set success rates for the other tasks are at least 94.0%. Mean surface divergence of the selected paraphrases ranges from 0.697 to 0.800. These statistics describe filtering and coverage; they do not by themselves establish that every retained paraphrase preserves meaning.

F.2 Paraphrase Stability across Video LLM Sizes

Table 22 extends the paraphrase evaluation to the three additional video LLMs using the definitions in Table 5.

Across the 21 rows, C_{para} ranges from 60.5% to 92.7%. Thus, even for the most stable combination, InternVL3-14B on VidHalluc-TSH, 7.3% of originally correct items become incorrect under at least one paraphrase. The corresponding fraction is 39.5% for Qwen3-VL-4B on MHBench. Aggregate accuracy changes by at most 3 percentage points in 16 rows, while C_{para} is below 80% in 13 rows.

Within the Qwen family, C_{para} improves from 4B to 32B on six of seven evaluation tasks, including MHBench (60.5% to 71.9%) and VidHalluc-TSH (75.8% to 89.5%). VidPair is the exception (81.7% to 78.1%). Larger model size therefore does not uniformly improve paraphrase stability.

Table 21: **Paraphrase generation and filtering statistics.** We request 20 candidates per unique question, validate candidates in a separate call, and select the three passing candidates with the largest surface divergence. *Questions* counts unique source questions; *Built* counts complete sets of three selected paraphrases; *Success* is Built/Questions. *Cand.* and *Passed* count recorded and accepted candidates; *Pass* is Passed/Cand. *Diverg.* is the mean surface divergence of selected paraphrases, measured as the Jaccard distance between token sets (0 = identical tokens, 1 = disjoint). Neither videos, gold answers, nor evaluated-model predictions are used during generation or selection. Counts cover the full question pool; VideoHallucer evaluation uses its five core subsets.

Benchmark	Questions	Built	Success	Cand.	Passed	Pass	Diverg.
MHBench	200	188	94.0%	3,999	2,342	58.6%	0.800
EventHallusion	159	159	100.0%	3,180	2,370	74.5%	0.779
VideoHallucer	1,540	1,492	96.9%	30,800	23,078	74.9%	0.725
Dr.V-Bench	2,503	2,490	99.5%	50,052	34,786	69.5%	0.697
VidHalluc-ACH	3,819	3,770	98.7%	76,380	51,546	67.5%	0.790
VidHalluc-TSH	590	507	85.9%	11,771	4,739	40.3%	0.727
VidPair	1,005	986	98.1%	20,100	13,518	67.3%	0.773

Table 22: **Paraphrase stability across additional video LLM sizes.** The evaluation follows Table 5. *Original* is accuracy (%) on the original wording of retained items, and Δ is mean paraphrased accuracy minus Original in percentage points. Flip rates are percentages of all retained items, averaged over three paraphrases. C_{para} is the percentage of originally correct items that remain correct under all three paraphrases. n counts retained evaluation items.

				Prediction flips (%)			
Benchmark	Video LLM	Original	Δ	Correct $\rightarrow$ wrong $\downarrow$	Wrong $\rightarrow$ correct $\downarrow$	$C_{\text{para}} \uparrow$	n
MHBench	Qwen3-VL-4B	63.28	−3.14	13.9	10.7	60.5	2,247
	Qwen3-VL-32B	63.82	−2.79	10.5	7.7	71.9	2,247
	InternVL3-14B	61.73	+0.46	8.7	9.2	72.8	2,247
EventHallusion	Qwen3-VL-4B	69.93	−5.87	10.4	4.5	66.8	409
	Qwen3-VL-32B	67.97	+1.14	6.0	7.2	82.0	409
	InternVL3-14B	70.42	−3.02	7.5	4.5	78.1	409
VideoHallucer	Qwen3-VL-4B	74.71	−2.79	11.0	8.3	71.5	1,898
	Qwen3-VL-32B	76.92	−2.55	9.6	7.1	77.0	1,898
	InternVL3-14B	76.13	−2.39	8.1	5.7	80.1	1,898
Dr.V-Bench	Qwen3-VL-4B	69.54	−1.85	9.3	7.4	73.4	3,030
	Qwen3-VL-32B	72.94	−1.02	7.6	6.6	78.5	3,030
	InternVL3-14B	75.31	−2.00	8.7	6.7	77.7	3,030
VidHalluc-ACH	Qwen3-VL-4B	87.04	−2.96	6.6	3.6	84.6	8,448
	Qwen3-VL-32B	89.89	−1.91	4.6	2.7	89.2	8,448
	InternVL3-14B	89.28	−0.86	4.3	3.5	89.8	8,448
VidHalluc-TSH	Qwen3-VL-4B	87.40	−9.30	13.0	3.7	75.8	516
	Qwen3-VL-32B	95.74	−6.46	7.3	0.8	89.5	516
	InternVL3-14B	92.44	−1.61	3.4	1.8	92.7	516
VidPair	Qwen3-VL-4B	53.75	−0.16	5.2	5.0	81.7	3,935
	Qwen3-VL-32B	56.52	−0.64	6.5	5.8	78.1	3,935
	InternVL3-14B	60.23	+0.00	11.0	11.0	66.9	3,935

InternVL3-14B on VidPair has $\Delta = +0.00$ with correct-to-wrong and wrong-to-correct flip rates of 11.0% each. On MHBench, the same model has $\Delta = +0.46$ with flip rates of 8.7% and 9.2%. These cases show how stable or nearly stable aggregate accuracy can coexist with substantial changes in individual predictions.