CCTVBench: Contrastive Consistency Traffic VideoQA Benchmark for Multimodal LLMs

Xingcheng Zhou[1]*, Hao Guo[1], Rui Song[2], Walter Zimmer[2], Mingyu Liu[1], André Schamschurko[1], Hu Cao[1], Alois Knoll[1]

[1]Technical University of Munich [2]University of California, Los Angeles

Abstract

Safety-critical traffic reasoning requires contrastive consistency: models must detect true hazards when an accident occurs, and reliably reject plausible-but-false hypotheses under near-identical counterfactual scenes. We present CCTVBENCH, a **C**ontrastive **C**onsistency **T**raffic **V**ideoQA **Bench**mark built on paired real accident videos and world-model-generated counterfactual counterparts, together with minimally different, mutually exclusive hypothesis questions. CCTVBENCH enforces a single structured decision pattern over each video question quadruple and provides actionable diagnostics that decompose failures into positive omission, positive swap, negative hallucination, and mutual-exclusivity violation, while separating video versus question consistency. Experiments across open-source and proprietary video LLMs reveal a large and persistent gap between standard per-instance QA metrics and quadruple-level contrastive consistency, with unreliable none-of-the-above rejection as a key bottleneck. Finally, we introduce C-TCD, a contrastive decoding approach leveraging a semantically exclusive counterpart video as the contrast input at inference time, improving both instance-level QA and contrastive consistency.

1 Introduction

Vision–language models (VLMs) are increasingly applied to traffic scene understanding, roadway monitoring, and language-conditioned perception and decision making in autonomous driving systems (Zhou et al., 2024; Cui et al., 2026). However, current Video-LLMs are prone to hallucinating non-existent entities or events and to missing subtle but critical visual cues, while recent multi-modal studies therefore begin to systematically evaluate and mitigate such hallucinations (Guan et al., 2024; Leng et al., 2023; Zhang et al., 2025b). In safety-critical traffic settings, this behavior manifests as false alarms and missed hazards that can propagate into unstable or unsafe downstream decisions. Beyond improving per-instance accuracy, it is therefore essential to ensure that models exhibit contrastive consistency: when the decisive outcome changes under tightly controlled scene edits, their predictions must adapt coherently to the underlying evidence. Concretely, models should behave consistently under scene-matched counterfactual edits, reject plausible-but-false alternatives, and support a calibrated none-of-the-above state instead of defaulting to overconfident but unsupported answers.

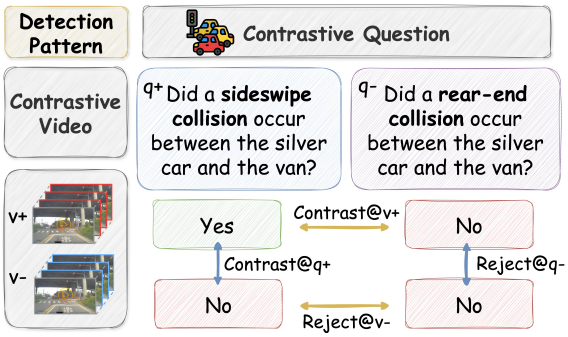


Figure 1: Illustration of the contrastive quadruple evaluation protocol in CCTVBench. Given a contrastive video pair (v^+, v^-) and a mutually exclusive question pair (q^+, q^-), a consistent model should answer Yes only for (v^+, q^+) and reject all other combinations.

At the same time, most existing traffic video QA benchmarks evaluate models on isolated video question pairs and report standard classification metrics (Qian et al., 2023; Zhou et al., 2025; Xu et al., 2021; Sima et al., 2023), which provide limited signal on whether a model maintains a coherent decision rule when the decisive outcome changes while the surrounding scene context is largely preserved. This leaves contrastive failures under scene-matched counterfactual edits underexplored. A model can achieve reasonable per-

*Correspondence: xingcheng.zhou@tum.de

Findings of the Association for Computational Linguistics: ACL 2026, pages 21665–21684

instance accuracy yet still answer *Yes* on counterfactual negatives, endorse a plausible distractor on the same scene, or even accept mutually exclusive hypotheses, all unacceptable for risk-aware traffic decision making. We therefore introduce CCTVBENCH, a **C**ontrastive **C**onsistency **T**raffic **V**ideoQA **Bench**mark.

CCTVBENCH is designed to make failures actionable. Its metric suite measures video and question consistency, enforces a strict quadruple-level decision pattern, reports vision-language sensitivity, and decomposes failed quadruples into positive omission, positive swap, negative hallucination, and mutual-exclusivity violation, as shown in Fig. 1. This allows targeted diagnosis of whether errors stem from missed hazard detection on v^+, false positives on v^-, or confusion between paired hypotheses. The same counterfactual pairing also provides a stronger contrast signal at inference time. We therefore present C-TCD, which introduces the semantically exclusive counterpart video as a contrast input, showing that counterfactual pairs are not only for evaluation but also effective to improve consistency. Our contributions can be summarized as follows:

- We introduce CCTVBENCH, a contrastive traffic VideoQA benchmark built on real anomaly footage, with contrastive video and question pairs across six categories.
- We propose a quadruple evaluation protocol with a diagnostic suite that measures cross-video and cross-question consistency and attributes quadruple failures to major patterns.
- We benchmark a broad set of video LLMs and reveal the gap between classification quality and contrastive consistency, highlighting robust rejection remains the main bottleneck.
- We propose C-TCD for inference-time contrastive decoding with semantically exclusive counterfactual pairs, demonstrating that counterfactual video pairs are not only for evaluation but also effective to improve model consistency.

2 Related Work

Traffic and Driving Vision-Language Benchmarks. VideoQA benchmarks in the traffic domain primarily study traffic event-centric reasoning and multi-agent interactions in real-world monitoring scenarios. SUTD-TrafficQA (Xu et al., 2021) provides large-scale traffic video QA covering diverse traffic accidents, while TUMTraffic-VideoQA (Zhou et al., 2025) studies incident-centric questions in roadside surveillance footage. In the driving domain, multiple works (Qian et al., 2023; Sima et al., 2023; Marcu et al., 2023) extend VQA to multi-view driving perception, targeting instruction-following and reasoning grounded in driving scenes understanding. Despite existing contrastive-pair benchmarks in general VQA (Li et al., 2025; Qiu et al., 2024; Chandu et al., 2024), neither the traffic nor the driving VideoQA lines provide benchmarks explicitly designed for contrastive consistency evaluation, which is critical for VLM-based models.

Consistency and Hallucination Evaluation of Multimodal LLMs. Despite strong generation and instruction following, multimodal LLMs often hallucinate non-existent entities or events. In images, POPE (Li et al., 2023b) proposes a polling-based protocol for stable hallucination measurement. Some benchmarks and mitigation work further motivate evidence-grounded evaluation Guan et al., 2024; Lee et al., 2024. In the video understanding domain, hallucinations are amplified by temporal ambiguity and long-range dependencies. VidHal (Choong et al., 2025) evaluates temporal hallucination via ordering-based probes, while VidHalluc (Li et al., 2025) targets temporally grounded hallucinations. Complementary to QA-only evaluation, contrastive setting formulations naturally support consistency checks (Liu et al., 2020; Lei et al., 2020b,a; Xiao et al., 2024).

Controllable Driving World Models. World models for autonomous driving aim to simulate or forecast future observations under actions and interventions. GAIA-2 (Russell et al., 2025) formulates driving world modeling as autoregressive sequence prediction conditioned on video, text, and actions. DrivingDiffusion (Li et al., 2023a) synthesizes layout-guided multi-view driving videos using latent diffusion with cross-view and cross-frame consistency mechanisms. Drive-WM (Wang et al., 2024a) generates controllable multi-view future videos and explores planning via imagined rollouts. MagicDrive (Gao et al., 2024) provides diverse 3D geometry control for street-view generation and supports multi-view consistency. Recent world foundation model (NVIDIA et al., 2025;

Zheng et al., 2024) further scales controllable video world simulation and controllable video editing.

3 CCTVBench

In real-world driving, traffic accidents and anomalies are rare corner cases, and we empirically find that current world models struggle to synthesize physically plausible accident videos, whereas they can generate high-quality normal sequences.

3.1 Dataset Curation

As illustrated in Fig. 2, CCTVBench is curated to evaluate faithfulness via contrastive video pairs and mutually exclusive question pairs. We start from open-source traffic and driving anomaly datasets and select representative accident scenes as positive videos v_s^+. For each v_s^+, annotators segment the clip into three parts, as shown in Fig. 4. A *base* segment that does not affect the outcome, *key frames* that determine the event trajectory, and a *contrastive* segment targeted for counterfactual synthesis. Besides, we run off-the-shelf detectors and trackers to extract traffic participants and coarse trajectories across the video. Annotators then refine instance identities and provide scene-level meta annotations, including involved instances, major causes, and critical spatio-temporal relations among key entities. To obtain minimally different yet semantically decisive video pairs, we generate the counterfactual variant v_s^- using a driving world model (NVIDIA et al., 2025). We condition the generation on the shared base video prefix and synthesize the contrastive segment such that the accident does not occur, while preserving layout, illumination, and camera viewpoint as much as possible. We then temporally align (v_s^+, v_s^-) and apply quality filtering to remove videos with obvious artifacts, viewpoint drift, and key-entity inconsistencies, ensuring v_s^- remains a valid negative control for the target hypotheses.

Given the aligned video pair and the human meta annotations, we adopt GPT-5 to generate paired contrastive questions across multiple categories, spanning event, key entity, temporal, spatial, spatio-temporal, causal, and counterfactual. We further constrain $q_{s,k}^-$ to be a plausible but verifiably false distractor on v_s^+. By design, $(q_{s,k}^+, q_{s,k}^-)$ are mutually exclusive but not collectively exhaustive: they cannot both be true, and on the negative-control video v_s^- we expect both to be false with respect to the target hypotheses. By construction, each

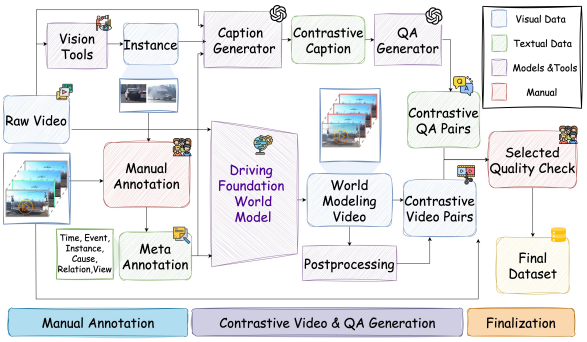


Figure 2: Dataset curation pipeline of CCTVBench.

QA pair follows a fixed four-way decision pattern: on the positive video v_s^+, the positive question $q_{s,k}^+$ should be answered *Yes* while its counterpart $q_{s,k}^-$ should be *No*; on the counterfactual negative-control video v_s^-, *both* questions should be answered *No*. Finally, we conduct human validation and remove samples with generation artifacts, semantic mismatch, viewpoint drift, key-entity inconsistency, or ambiguous question grounding, yielding the finalized dataset. This review ensures that each (v^+, v^-) pair remains a valid contrastive control and that each question pair is visually grounded and mutually exclusive.

3.2 Dataset Overview

CCTVBench is a contrastive traffic video QA benchmark with 305 contrastive scenes, 305 positive–negative video pairs v^+ and v^-, and 1,776 mutually exclusive question pairs, yielding 7,104 binary video question instances across six categories: Event, Key-Entity, Spatial, Spatial-Temporal, Causal, and Counterfactual.

Fig. 3 (left) summarizes the distribution of high-level crash types such as rear-end, T-bone, and sideswipe across various accident video datasets (Fang et al., 2019; Chan et al., 2016; Bao et al., 2020), covering a broad range of recording sources. Fig. 3 (right) shows the question vocabulary, dominated by accident-centric words such as car, vehicle, and collision.

3.3 Counterfactual Quality and Reliability

We quantify generator-induced bias and distribution shift in counterfactual videos using two complementary analyses: artifact leakage and distributional similarity.

Artifact Leakage. Using frozen R3D-18 features and a linear classifier, distinguishing counterfactual from real videos yields near-chance perfor-

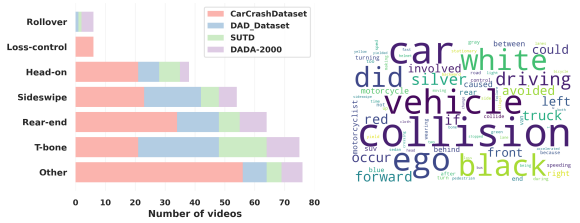


Figure 3: Left: Distribution of event types from various sources. Right: Word cloud of event scene caption.

mance (AUC 56.7%, accuracy 53.6%), indicating that counterfactual videos do not exhibit strong generator-specific artifacts that could be exploited as shortcut signals.

Distributional Similarity. We measure FVD and FID between counterfactual and real videos. Counterfactual videos are substantially closer to real accident videos (FVD 22.44, FID 32.63) than to unrelated normal driving footage (FVD 35.73, FID 105.00), suggesting a moderate distribution shift.

Overall, these results indicate that generator-induced bias is limited and that counterfactual videos remain reasonably aligned with the real data distribution.

3.4 Contrastive Consistency Evaluation

The core of CCTVBENCH is a contrastive quadruple structure that evaluates whether a model can maintain a coherent decision pattern across contrastive paired videos and mutually exclusive questions. Let K_s be the number of question pairs for scene s, $N_{\text{pairs}} = \sum_s K_s$ the total number of (s,k) pairs, and $\sum_{s,k}$ denote summation over all such pairs. For brevity, we denote the binary outputs

$$\begin{aligned} \hat{y}^{++}_{s,k} &= \hat{y}(v^+_s, q^+_{s,k}), \quad \hat{y}^{+-}_{s,k} = \hat{y}(v^+_s, q^-_{s,k}), \\ \hat{y}^{-+}_{s,k} &= \hat{y}(v^-_s, q^+_{s,k}), \quad \hat{y}^{--}_{s,k} = \hat{y}(v^-_s, q^-_{s,k}) \end{aligned} \tag{1}$$

with $\hat{y}(\cdot) \in \{0,1\}$ denoting the final binary prediction.

Quadruple Correctness. The benchmark is constructed such that each contrastive quadruple has a single consistent semantic pattern: $q^+_{s,k}$ is true only on v^+_s, $q^-_{s,k}$ is false on both videos, and v^-_s is a genuine *none-of-the-above* counterfactual. We count a pair (s,k) as fully consistent only if the model recovers the entire pattern:

$$\begin{aligned} \text{QuadCr}(\cdot) &= \mathbf{1}\left[\hat{y}^{++} = 1 \wedge \hat{y}^{+-} = 0 \wedge \hat{y}^{-+} = 0 \wedge \hat{y}^{--} = 0\right], \\ \text{QuadAcc} &= \frac{1}{N_{\text{pairs}}} \sum_{s,k} \text{QuadCr}(s,k) \end{aligned} \tag{2}$$

QuadAcc therefore measures how often a model simultaneously detects the target event in the positive video, rejects the mutually exclusive alternative on the same scene, and assigns negative answers to both hypotheses on the counterfactual video. Unlike single-video accuracy, it only gives credit when the entire contrastive VQA pair is correct, making it the strictest consistency metric in our benchmark.

Video Consistency. Video consistency evaluates whether model predictions genuinely respond to changes in the video when question is kept fixed. We therefore consider two complementary quantities. For positive query q^+, the desired behavior is a clean transition from *Yes* on the positive video v^+_s to *No* on the counterfactual v^-_s:

$$\text{Contr@q}^+ = \tfrac{1}{N_{\text{pairs}}} \textstyle\sum_{s,k} \mathbf{1}\left[\hat{y}^{++}_{s,k} = 1 \wedge \hat{y}^{-+}_{s,k} = 0\right] \tag{3}$$

For the mutually exclusive alternative query q^-, the model should consistently reject the hypothesis on both videos:

$$\text{Reject@q}^- = \tfrac{1}{N_{\text{pairs}}} \textstyle\sum_{s,k} \mathbf{1}\left[\hat{y}^{+-}_{s,k} = 0 \wedge \hat{y}^{--}_{s,k} = 0\right] \tag{4}$$

In the tables, *Video Consistency* is reported as the average of Contr@q$^+$ and Reject@q$^-$, summarizing how reliably a model exploits the contrast between v^+_s and v^-_s under fixed queries, both required by the quadruple pattern.

Question Consistency. Question consistency, in turn, probes whether the model separates the two *linguistic* hypotheses $q^+_{s,k}$ and $q^-_{s,k}$ when the video is fixed. On the positive video v^+_s, the desired pattern is to endorse the target proposition and reject its mutually exclusive alternative:

$$\text{Contr@v}^+ = \tfrac{1}{N_{\text{pairs}}} \textstyle\sum_{s,k} \mathbf{1}\left[\hat{y}^{++}_{s,k} = 1 \wedge \hat{y}^{+-}_{s,k} = 0\right] \tag{5}$$

On the counterfactual video v^-_s, both formulations should be rejected:

$$\text{Reject@v}^- = \tfrac{1}{N_{\text{pairs}}} \textstyle\sum_{s,k} \mathbf{1}\left[\hat{y}^{-+}_{s,k} = 0 \wedge \hat{y}^{--}_{s,k} = 0\right] \tag{6}$$

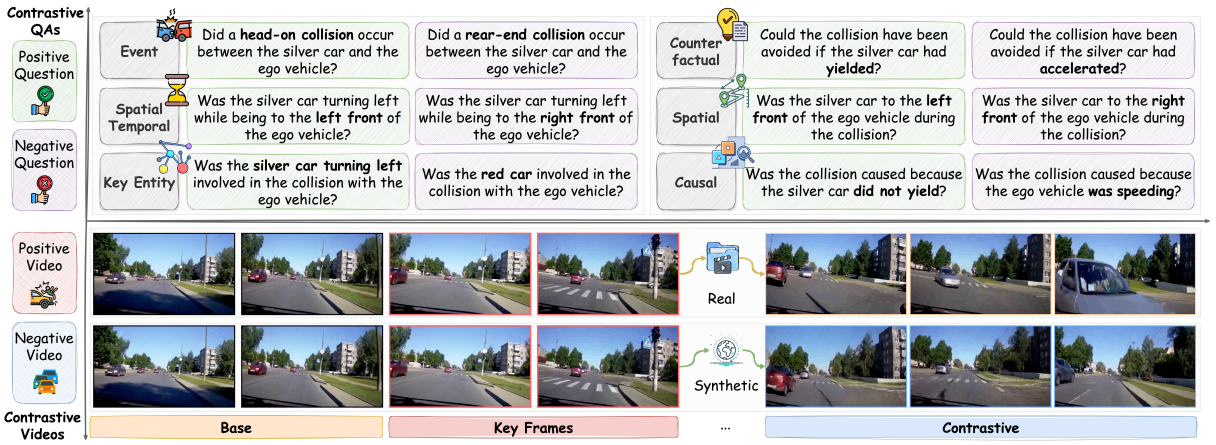


Figure 4: Quadruple Example of CCTVBench with contrastive video pair (v_s^+, v_s^-) and corresponding contrastive question pairs $(q_{s,k}^+, q_{s,k}^-)$ across different categories.

We report *Question Consistency* as the average of Contr@v$^+$ and Reject@v$^-$, balancing mutual-exclusivity handling on v^+ and none-of-the-above rejection on v^-.

Global Classification Metrics. To relate this contrastive protocol to standard QA performance, we additionally report balanced accuracy (BaAcc) and a squashed Matthews correlation coefficient

$$\text{MCCScore} = \left(\frac{\text{MCC}+1}{2}\right)^2 \tag{7}$$

computed over all VQA instances. MCCScore follows the setting of $Score_{cls}$ in (Li et al., 2025). Both summarize plain binary classification quality on our data, while the contrastive metrics above reveal how that quality decomposes when video and question are systematically perturbed.

3.5 Behavior Diagnosis

Contrastive consistency scores indicate whether a model succeeds on a quadruple, but not how it fails. To analyze typical error patterns in traffic reasoning, we introduce two diagnostic families tailored to our setting: failure-mode decomposition over failed quadruples, and vision-language sensitivity within the same quadruple structure.

Failure Mode Decomposition. To expose the typical failure reasons of existing Video LLMs, we focus on the subset of failed question pairs F and decompose them into 4 major failure modes: positive omission, positive swap, negative hallucination, and mutual-exclusivity violation. For each failure mode, we report its failure rate as the proportion of $(s,k) \in F$ that satisfy the corresponding condition.

Positive omission measures how often the model misses the target hypothesis on positive video v_s^+:

$$\text{PosOmiss} = \frac{1}{|F|}\sum \mathbf{1}[\hat{y}^{++} = 0] \tag{8}$$

Positive swap measures how often the model answers *Yes* to $q_{s,k}^-$ on positive video v_s^+:

$$\text{PosSwap} = \frac{1}{|F|}\sum \mathbf{1}[\hat{y}^{+-} = 1] \tag{9}$$

Negative hallucination quantifies violations of the none-of-the-above requirement on v_s^-:

$$\text{NegHall} = \frac{1}{|F|}\sum \mathbf{1}[\hat{y}^{-+} = 1 \vee \hat{y}^{--} = 1] \tag{10}$$

Mutual-exclusivity violation measures how often the model simultaneously breaks mutually exclusive hypotheses on the same video, both on v_s^+ and v_s^-:

$$\text{MEViol} = \frac{1}{|F|}\sum \mathbf{1}\left[(\hat{y}^{++} = 1 \wedge \hat{y}^{+-} = 1) \vee (\hat{y}^{-+} = 1 \wedge \hat{y}^{--} = 1)\right] \tag{11}$$

Note that a failed pair can satisfy multiple reasons, so the 4 failure rates are not mutually exclusive and do not necessarily sum to 1.

Vision-Language Sensitivity. Vision-language sensitivity metrics show how strongly model predictions depend on visual evidence or textual modality. Video sensitivity (VS) measures the effect of swapping v_s^+ and v_s^- under the positive query:

$$\text{VS} = \frac{1}{N_{\text{pairs}}}\sum_{s,k} \left|\hat{y}_{s,k}^{++} - \hat{y}_{s,k}^{-+}\right| \tag{12}$$

while question sensitivity (QS) measures the effect of swapping $q_{s,k}^+$ and $q_{s,k}^-$ on the positive video:

$$\text{QS} = \frac{1}{N_{\text{pairs}}}\sum_{s,k} \left|\hat{y}_{s,k}^{++} - \hat{y}_{s,k}^{+-}\right| \tag{13}$$

We aggregate them into a visual reliance index and a global vision-language balance score:

$$\mathrm{VRI} = \frac{\mathrm{VS}}{\mathrm{VS}+\mathrm{QS}}, \quad \mathrm{GVRS} = 2 \cdot \mathrm{VRI} \cdot \mathrm{QS} \tag{14}$$

VRI quantifies the relative reliance on the video modality, whereas GVRS captures joint sensitivity to both visual changes and question wording. It attains high values only when the model reacts strongly to perturbations in both the vision and language branches.

$$\mathrm{SVE} = \frac{1}{N_{\text{pairs}}} \sum_{s,k} \mathbf{1}\Big[(\hat{y}^{++}_{s,k} \neq \hat{y}^{-+}_{s,k}) \wedge (\hat{y}^{+-}_{s,k} = \hat{y}^{--}_{s,k})\Big] \tag{15}$$

To capture more selective visual grounding, we additionally report the selective video effect (SVE), which counts cases where changing the video flips the answer only for the positive query q^+ while leaving responses to q^- unchanged.

4 Experiments

4.1 Models

We evaluate a diverse set of video LLMs spanning multiple well-known VLMs. Our open-source lineup includes InternVL3.5 (Wang et al., 2025), Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), PLLaVA (Xu et al., 2024), VideoLLaMA3 (Zhang et al., 2025a), VideoChatGPT (Maaz et al., 2024), and phi-4-multimodal (Microsoft et al., 2025). We also include proprietary API-based models from Gemini-2.5 (Comanici et al., 2025) and GPT-5 (OpenAI et al., 2024), i.e., Gemini-2.5 flash/flash-lite; GPT-5 mini/nano.

4.2 Experimental Setup

All open-sourced models are evaluated under a unified binary QA interface. Each query is formatted with the instruction: *You are given a traffic video and a question. Answer based only on the video with exactly one word: Yes or No. Question: {QUESTION}.* We instruct the model to answer with a single word (`Yes` or `No`) and map the output to a binary label. We use the default model decoding strategy with a temperature as 0. For proprietary models, we run the model in a video-level batch to reduce the cost of closed-source API evaluations, while then parsing each QA as an independent binary response.

4.3 Contrastive Decoding Inference

Recent work has shown that training-free contrastive decoding can mitigate hallucinations in large VLMs by contrasting the output distributions of an original input with those of a deliberately degraded counterpart during inference. Therefore, we follow VCD (Leng et al., 2023) and TCD (Zhang et al., 2025b) to examine how these methods behave under our contrastive evaluation protocol. Building on the paired counterfactual videos provided by our benchmark, we additionally present contrastive temporal contrastive decoding (C-TCD). Instead of constructing the contrast input by generic visual corruption or temporal degradation, C-TCD uses the semantically exclusive counterpart video from the same scene as the contrast signal. At each decoding step t for a query (v, q), we obtain logits under the original video z_t^{ori} and under a contrastive video z_t^{con}, and fuse them as

$$z_t = (1+\alpha)\, z_t^{\text{ori}} - \alpha\, z_t^{\text{con}}, \tag{16}$$

where $\alpha \geq 0$ controls the contrast strength and $\alpha = 0$ reduces to vanilla decoding.

The three variants differ in how z_t^{con} is constructed. VCD and TCD rely on heuristic degradations of the same input video. In VCD, z_t^{con} is computed on a visually corrupted video v^-_{vcd} obtained by injecting diffusion noise into frames, which preserves coarse layout but removes fine-grained evidence. In TCD, z_t^{con} is computed on a temporally degraded video v^-_{tcd} produced by strong temporal downsampling, which blurs temporal cues while keeping appearance largely intact.

In contrast, C-TCD uses a semantically decisive, scene-matched counterpart video provided by CCTVBENCH. We compute z_t^{con} on the paired counterfactual video from the same scene, denoted $v^-_{\text{c-tcd}}$: for a positive video v_s^+ we use its paired negative v_s^-, and for v_s^- we use v_s^+. C-TCD therefore

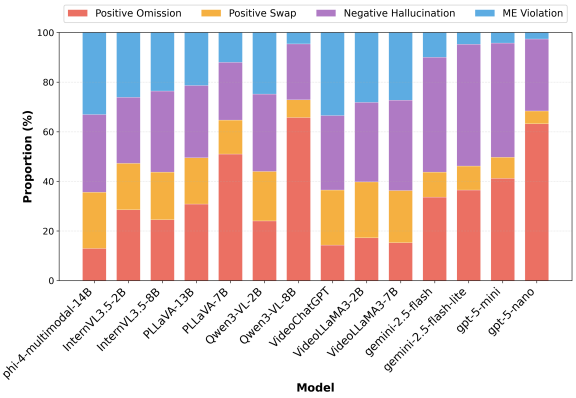


Figure 5: Normalized four failure-mode composition across various models.

Table 1: Quadruple-consistency evaluation on CCTVBENCH. All values are percentages. Each entry is reported as mean with confidence interval, where the interval is the larger absolute deviation, in percentage points, between the mean and the upper bound of the 95% bootstrap confidence interval.

Model	Size	Video Consistency			Question Consistency			BaAcc↑	MCCScore↑	QuadAcc↑
		Contr@q+↑	Reject@q-↑	Overall↑	Contr@v+↑	Reject@v-↑	Overall↑			
Open-source Models										
InternVL3.5	2B	$13.05_{\pm 1.25}$	$63.79_{\pm 1.76}$	$38.42_{\pm 1.06}$	$24.23_{\pm 1.57}$	$53.54_{\pm 1.77}$	$38.88_{\pm 1.01}$	$58.46_{\pm 0.97}$	$33.17_{\pm 1.02}$	$7.05_{\pm 0.98}$
InternVL3.5	8B	$20.88_{\pm 1.45}$	$64.92_{\pm 1.77}$	$42.90_{\pm 1.06}$	$\mathbf{38.58}_{\pm 1.81}$	$48.49_{\pm 1.83}$	$43.53_{\pm 1.09}$	$\mathbf{64.85}_{\pm 1.10}$	$\mathbf{39.89}_{\pm 1.24}$	$\mathbf{11.38}_{\pm 1.16}$
PLLaVA	7B	$8.89_{\pm 1.04}$	$80.75_{\pm 1.45}$	$44.82_{\pm 0.74}$	$22.35_{\pm 1.47}$	$68.95_{\pm 1.62}$	$45.65_{\pm 0.74}$	$57.38_{\pm 0.90}$	$33.35_{\pm 1.04}$	$3.90_{\pm 0.71}$
PLLaVA	13B	$13.87_{\pm 1.26}$	$66.67_{\pm 1.68}$	$40.27_{\pm 0.95}$	$29.06_{\pm 1.58}$	$52.39_{\pm 1.84}$	$40.73_{\pm 0.94}$	$59.81_{\pm 1.09}$	$34.70_{\pm 1.14}$	$6.19_{\pm 0.90}$
Qwen3-VL	2B	$17.02_{\pm 1.37}$	$63.51_{\pm 1.77}$	$40.26_{\pm 1.00}$	$34.21_{\pm 1.70}$	$47.94_{\pm 1.76}$	$41.07_{\pm 0.99}$	$62.96_{\pm 1.01}$	$37.77_{\pm 1.08}$	$7.62_{\pm 0.98}$
Qwen3-VL	8B	$13.01_{\pm 1.26}$	$\mathbf{91.05}_{\pm 1.03}$	$\mathbf{52.03}_{\pm 0.75}$	$27.85_{\pm 1.65}$	$\mathbf{76.38}_{\pm 1.54}$	$\mathbf{52.11}_{\pm 0.74}$	$60.26_{\pm 1.00}$	$38.62_{\pm 1.26}$	$10.65_{\pm 1.13}$
VideoLLaMA3	2B	$\mathbf{21.65}_{\pm 1.40}$	$31.16_{\pm 1.66}$	$26.41_{\pm 1.11}$	$27.44_{\pm 1.62}$	$28.28_{\pm 1.63}$	$27.86_{\pm 1.13}$	$55.38_{\pm 1.17}$	$29.89_{\pm 1.10}$	$6.48_{\pm 0.88}$
VideoLLaMA3	7B	$20.04_{\pm 1.48}$	$36.80_{\pm 1.78}$	$28.42_{\pm 1.12}$	$36.41_{\pm 1.84}$	$21.81_{\pm 1.49}$	$29.11_{\pm 1.13}$	$58.47_{\pm 1.26}$	$32.91_{\pm 1.24}$	$6.58_{\pm 0.92}$
VideoChatGPT	7B	$16.21_{\pm 1.32}$	$21.45_{\pm 1.50}$	$18.83_{\pm 1.01}$	$18.59_{\pm 1.50}$	$19.81_{\pm 1.41}$	$19.20_{\pm 1.01}$	$49.85_{\pm 1.14}$	$24.87_{\pm 1.02}$	$4.13_{\pm 0.71}$
Phi-4-Multimodal	14B	$11.23_{\pm 1.16}$	$40.07_{\pm 1.83}$	$25.65_{\pm 1.01}$	$27.26_{\pm 1.67}$	$25.38_{\pm 1.52}$	$26.32_{\pm 1.02}$	$57.07_{\pm 1.09}$	$31.61_{\pm 1.07}$	$2.92_{\pm 0.60}$
Proprietary Models										
Gemini-2.5	flash	$24.48_{\pm 1.48}$	$83.79_{\pm 1.32}$	$54.13_{\pm 0.99}$	$55.56_{\pm 1.75}$	$48.91_{\pm 1.75}$	$52.24_{\pm 1.10}$	$69.66_{\pm 1.18}$	$46.43_{\pm 1.45}$	$20.87_{\pm 1.39}$
Gemini-2.5	flash-lite	$21.43_{\pm 1.46}$	$86.95_{\pm 1.25}$	$54.19_{\pm 0.94}$	$\mathbf{59.42}_{\pm 1.79}$	$48.86_{\pm 1.86}$	$54.14_{\pm 0.93}$	$70.37_{\pm 1.16}$	$47.71_{\pm 1.39}$	$18.28_{\pm 1.39}$
GPT-5	nano	$13.92_{\pm 1.27}$	$\mathbf{91.19}_{\pm 1.03}$	$52.55_{\pm 0.81}$	$26.24_{\pm 1.58}$	$\mathbf{68.39}_{\pm 1.66}$	$47.31_{\pm 1.00}$	$57.46_{\pm 1.04}$	$34.43_{\pm 1.35}$	$12.60_{\pm 1.18}$
GPT-5	mini	$\mathbf{29.07}_{\pm 1.66}$	$87.13_{\pm 1.26}$	$\mathbf{58.10}_{\pm 1.04}$	$56.64_{\pm 1.88}$	$55.47_{\pm 1.76}$	$\mathbf{56.05}_{\pm 1.08}$	$\mathbf{70.43}_{\pm 1.23}$	$\mathbf{48.43}_{\pm 1.56}$	$\mathbf{25.85}_{\pm 1.57}$

replaces generic corruption or temporal degradation with an aligned counterfactual contrast signal, which directly matches the benchmark design and provides a more targeted semantic regularizer at inference time.

Assumptions and interpretation. C-TCD is a training-free inference-time calibration method. It does not modify model parameters or improve intrinsic visual representations, but instead adjusts prediction confidence using semantically matched counterfactual evidence. It does not assume that the counterfactual video perfectly replicates the original scene with only the target signal removed. Instead, it relies on a weaker condition: the counterfactual reduces causal evidence for the positive hypothesis while preserving most background context. Under this setting, the logit difference between the original and counterfactual inputs acts as a contrastive calibration signal.

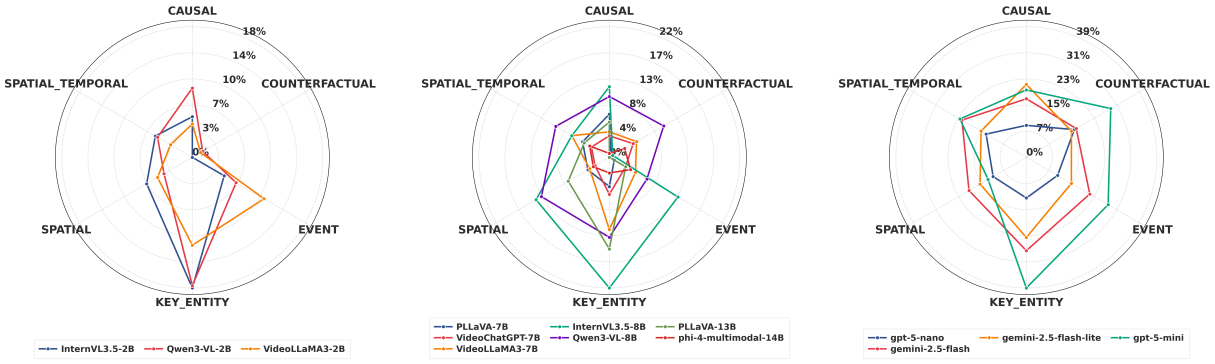
4.4 Results and Analysis

Statistical Reliability. To quantify statistical reliability under the benchmark scale, we perform scene-level bootstrap resampling with 2,000 replicates and report 95% confidence intervals for all metrics. Across models, interval widths are typically around 1–2 percentage points, indicating low variance. For example, GPT-5-mini achieves a QuadAcc of 25.85% with a 95% CI of [24.32, 27.42], while its video consistency and question consistency are 58.10% [57.07, 59.14] and 56.05% [54.98, 57.12], respectively. We further evaluate ranking stability under bootstrap resampling and observe that model rankings remain consistent, with GPT-5-mini ranked first on QuadAcc across resampled datasets.

Large gap between global classification and contrastive consistency. Tab. 1 shows that most models reach moderate binary QA quality, with BaAcc commonly in the 50–70% range and non-trivial MCCScore, yet QuadAcc remains low across the board. For open-source models, QuadAcc is typically below 12%, and even the best open-source entries only approach low double digits despite BaAcc around 60% and above. The gap is also visible among proprietary models, which indicates that many models can produce plausible answers for isolated video question instances, but do not implement a stable decision rule that generalizes across the structured quadruple.

Video and question consistency expose distinct weaknesses. The consistency breakdown in Tab. 1 reveals two common behaviors. First, several models achieve very high Reject@q^- but low Contr@q^+, meaning they reliably reject the mutually exclusive alternative yet frequently miss the true event on v^+. Second, other models show stronger Contr@v^+ but weaker Reject@v^-, meaning they can separate q^+ from q^- on the positive video but still produce excessive *Yes* answers on the counterfactual video, violating the none-of-the-above requirement.

Failure-mode composition across models. Fig. 5 decomposes failed quadruples into four failure patterns. Model size affects the omission-hallucination balance but not in a consistent direction across families. Within GPT-5, the



(a) Small models (b) Medium models (c) Large/commercial models

Figure 6: Category-wise QuadAcc radar plots for (a) small, (b) medium, and (c) large/commercial model groups.

Table 2: Vision-language sensitivity and binding diagnostics across models on CCTVBENCH. All values are percentages. Values are reported as mean percentages, with 95% bootstrap confidence intervals estimated by bootstrap. Shaded cells indicate the best result within each model group, and boldface marks the best overall result in each column.

Model	Size	VL Sensitivity & Binding Diagnostics			
		VS↑	QS↑	GVRS↑	SVE↑
Open-source Models					
InternVL3.5	2B	16.05	29.92	20.79	10.69
InternVL3.5	8B	27.75	36.11	31.27	14.41
PLLaVA	7B	16.33	30.45	21.14	9.35
PLLaVA	13B	18.02	30.42	22.53	10.30
Qwen3-VL	2B	20.34	29.08	23.82	11.01
Qwen3-VL	8B	16.36	29.03	20.84	14.02
VideoLLaMA3	2B	**40.89**	42.54	41.61	**23.27**
VideoLLaMA3	7B	38.06	**48.13**	**42.41**	21.01
VideoChatGPT	7B	31.92	31.61	31.67	16.63
Phi-4-Multimodal	14B	19.08	34.36	24.42	13.41
Proprietary Models					
Gemini-2.5	flash-lite	25.70	**83.71**	39.25	18.32
Gemini-2.5	flash	36.06	63.70	45.98	**31.38**
GPT-5	nano	33.00	40.68	36.36	28.68
GPT-5	mini	**36.67**	68.67	**47.74**	30.97

smaller variant becomes markedly more conservative and fails mainly by Positive Omission, while the larger variant shifts toward Negative Hallucination. In Qwen3-VL, scaling to 8B also increases Positive Omission, suggesting that improved fluency or safety calibration can collapse decisions toward No rather than improving contrastive binding.

High reactivity does not imply correct binding. Regarding sensitivity diagnostics, Tab. 2 shows that strong modality sensitivity alone is insufficient. VideoLLaMA3 attains the highest VS and QS, and also leads on SVE among open-source models, indicating that its predictions change substantially under controlled swaps. Yet its QuadAcc in Tab. 1

Table 3: Effect of contrastive decoding variants on classification and quadruple consistency. Subscripts denote signed change relative to Base.

Model	Decoding	BaAcc↑	$\text{Score}_{\text{cls}}$↑	QuadAcc↑
Qwen3-VL-2B	Base	62.96	37.77	7.62
	VCD	$58.76_{-4.20}$	$33.19_{-4.58}$	$8.46_{+0.85}$
	TCD	$63.56_{+0.60}$	$38.18_{+0.41}$	$7.56_{-0.06}$
	C-TCD	$65.84_{+2.88}$	$40.69_{+2.92}$	$9.85_{+2.23}$
Qwen3-VL-8B	Base	60.26	38.62	10.65
	VCD	$63.20_{+2.95}$	$39.69_{+1.06}$	$14.62_{+3.97}$
	TCD	$61.82_{+1.56}$	$39.62_{+1.00}$	$12.70_{+2.05}$
	C-TCD	$64.82_{+4.56}$	$43.03_{+4.41}$	$18.63_{+7.97}$

remains limited, suggesting that the direction of these changes is often misaligned with the contrastive constraints. Besides, GPT-5-mini combines high GVRS with the highest SVE and also achieves the best QuadAcc. Gemini-2.5-flash-lite shows extremely high QS but weaker QuadAcc, consistent with over-reacting to wording differences without reliably enforcing rejection on v^-.

Category-wise analysis highlights persistent gaps in counterfactual and causal reasoning. Fig. 6 summarizes QuadAcc by question category. Key-Entity tends to be the easiest, likely because it reduces to identifying salient participants and outcomes in accident scenes. Event and Spatial categories are intermediate. Counterfactual and Causal remain consistently hard, because they require rejecting plausible hypotheses on counterfactual videos and resisting language priors that associate accident scenes with stereotypical causes. Proprietary models excel across all categories, but the radar shapes still show imbalance, with most models concentrating successes in a subset of categories rather than achieving uniform consistency.

Contrastive decoding benefits most from semantically decisive counterfactual pairs. Tab. 3

compares training-free contrastive decoding variants on Qwen3-VL. VCD can improve QuadAcc but may reduce BaAcc and $\text{Score}_{\text{cls}}$, indicating that visually corrupted negatives can over-suppress useful evidence in strict Yes/No classification. TCD yields smaller and less consistent gains. C-TCD improves all reported metrics simultaneously and produces the largest QuadAcc gains. These results suggest that using a semantically exclusive counterpart video provides a more targeted contrast signal than generic corruption or temporal degradation, aligning well with the benchmark's design.

4.5 Discussion

Why QuadAcc is hard: enforcing none-of-the-above without collapsing to conservatism. The quadruple protocol is intentionally asymmetric: q^+ is true only on v^+, while both questions must be rejected on v^-. Fig. 5 shows that many models do not handle this none-of-the-above state robustly. A common pattern is that rejecting q^- is much easier than answering q^+ correctly. The other bottleneck is rejecting both questions on v^-. Even strong models with high Contr@v$^+$ still leak *Yes* on the counterfactual, and lead to QuadAcc fail.

Language priors vs. visual grounding: imbalance shows up differently across metrics. Strong language grounding yields reasonable BaAcc, since many queries align with accident patterns and can be answered from text priors alone. However, when decisions are driven mainly by question wording and not by the video, QuadAcc remains low, because predictions fail to flip correctly across v^+ and v^-. In contrast, models with stronger visual grounding react more to video changes and can better separate positive and counterfactual scenes, which helps QuadAcc, but they still lose contrastive consistency if they do not distinguish mutually exclusive formulations. High BaAcc, therefore, does not imply high QuadAcc unless visual evidence and language understanding are jointly calibrated, as reflected by GVRS and SVE in Tab. 2.

Implications: Counterfactual world modeling is useful beyond evaluation. The consistent gains of C-TCD indicate that counterfactual pairs can serve as a strong inference-time anchor. More broadly, they suggest an interesting direction: align models using paired (v^+, v^-) with objectives that explicitly enforce mutual exclusivity on v^+, rather than relying solely on single-video QA supervision. It also shows that driving world models facilitates model consistency ability, leveraging contrastive video pairs.

Benchmark scope. CCTVBench is a controlled contrastive benchmark rather than a large-scale distributional dataset. It focuses on event-level paired scenes with minimal counterfactual edits, preserving causal structure but limiting coverage of long-tail driving scenarios. Compared to existing traffic VideoQA benchmarks that emphasize large-scale QA under natural distributions, CCTVBench provides a complementary setting for evaluating consistency under near-identical counterfactual conditions, trading coverage for diagnostic precision.

5 Conclusion

We introduced CCTVBench, a contrastive traffic VideoQA benchmark built from real accident footage paired with generated counterfactual normal videos. The contrastive quadruple protocol enforces a single coherent decision rule and explicitly probes calibrated rejection under minimal scene changes for risk-aware reasoning. We also provide actionable diagnostics that separate cross-video and cross-question consistency and attribute failures to 4 failure patterns. Experiments across diverse video LLMs reveal a persistent gap between per-instance classification scores and quadruple-level contrastive consistency. We further proposed C-TCD, showing that semantically contrastive video pairs serve not only as an evaluation construct but also as an effective inference-time contrast signal that improves consistency without retraining.

Limitations

CCTVBench has several limitations. First, counterfactual negatives are generated using a single driving world model, as our setting requires temporally consistent video-conditioned generation with stable camera geometry, which is not supported by most publicly accessible alternatives. Although we mitigate artifacts through filtering and human review, some generator-specific bias may remain. Second, the benchmark scale is constrained by the cost of collecting rare accident scenes and constructing paired counterfactuals, limiting statistical coverage and external validity. Third, ensuring strictly mutually exclusive yet not collectively exhaustive video-text pairs is challenging, and some residual ambiguity may persist despite manual validation.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM Multimedia Conference*.

Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. 2016. Anticipating accidents in dashcam videos. In *Asian Conference on Computer Vision*, pages 136–153. Springer.

Khyathi Raghavi Chandu, Linjie Li, Anas Awadalla, Ximing Lu, Jae Sung Park, Jack Hessel, Lijuan Wang, and Yejin Choi. 2024. Certainly uncertain: A benchmark and metric for multimodal epistemic and aleatoric awareness. *Preprint*, arXiv:2407.01942.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. 2025. Vidhal: Benchmarking temporal hallucinations in vision llms. *Preprint*, arXiv:2411.16771.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Can Cui, Yunsheng Ma, Sung-Yeon Park, Zichong Yang, Yupeng Zhou, Peiran Liu, Juanwu Lu, Juntong Peng, Jiaru Zhang, Ruqi Zhang, Lingxi Li, Yaobin Chen, Jitesh H. Panchal, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, and Ziran Wang. 2026. Llm4ad: Large language models for autonomous driving – concept, review, benchmark, experiments, and future trends. *Preprint*, arXiv:2410.15281.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. 2019. Dada-2000: Can driving accident be predicted by driver attention*f* analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, page 4303–4309. IEEE Press.

Christian Fruhwirth-Reisinger, Dušan Malić, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. 2025. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. *Preprint*, arXiv:2506.06218.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. 2024. MagicDrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, Mexico City, Mexico. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020b. What is more likely to happen next? video-and-language future event prediction. *Preprint*, arXiv:2010.07999.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *Preprint*, arXiv:2412.03735.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. 2023a. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *Preprint*, arXiv:2310.07771.

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

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. *Preprint*, arXiv:2003.11618.

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. 2024. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *Preprint*, arXiv:2407.17140.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. 2023. Lingoqa: Visual question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, and 60 others. 2025. Cosmos world foundation model platform for physical ai. *Preprint*, arXiv:2501.03575.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2023. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. *Preprint*, arXiv:2404.13874.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. 2025. Gaia-2: A controllable multi-view generative world model for autonomous driving. *Preprint*, arXiv:2503.20523.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. 2025. EGOILLUSION: Benchmarking hallucinations in egocentric video understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28461–28480, Suzhou, China. Association for Computational Linguistics.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024a. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14749–14759.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arxiv*.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. 2025. Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6585–6597.

Li Xu, He Huang, and Jun Liu. 2021. SUTD-TrafficQA: A Question Answering Benchmark and an Efficient

Network for Video Reasoning Over Traffic Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *Preprint*, arXiv:2404.16994.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *Preprint*, arXiv:2501.13106.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. 2025b. Eventhallusion: Diagnosing event hallucinations in video llms. *Preprint*, arXiv:2409.16597.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *Preprint*, arXiv:2412.20404.

Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois Knoll. 2025. TUMTraf videoQA: Dataset and benchmark for unified spatio-temporal video understanding in traffic scenes. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. 2024. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pages 1–20.

A Appendix

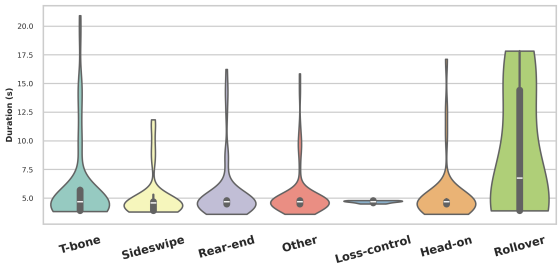
A.1 Additional Related Work

Recent work has moved beyond generic VQA accuracy to explicitly probe faithfulness, meaning answers must be grounded in visual evidence, and to diagnose hallucination. Early image-centric benchmarks (Sun et al., 2023; Chen et al., 2024; Qiu et al., 2024) typically evaluate models via LLM-based judging or curated prompts, focusing on object, attribute, and relation errors. A clear trend is the use of paired or controlled questions for consistency evaluation under minimal linguistic perturbations (Guan et al., 2024; Chandu et al., 2024), which contrastive QA settings that expose strong language priors and poor calibration on unanswerable cases. Video-based domain is less explored but rapidly emerging (Wang et al., 2024b; Zhang et al., 2025b; Seth et al., 2025). However, most faithfulness benchmarks remain single-image or single-video centered. They measure whether a model answers each query correctly, but do not enforce a global decision pattern across mutually exclusive conditions.

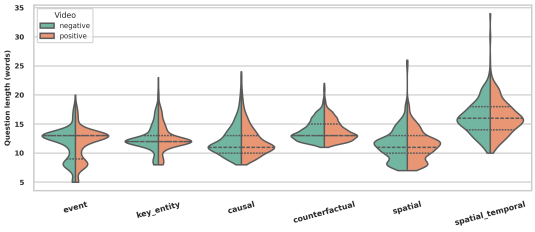
As summarized in Tab. 4, most existing video faithfulness datasets still lack scene-matched counterfactual video pairs that differ only in the critical event, which is essential to disentangle genuine visual grounding from plausible narrative completion. In parallel, traffic and autonomous driving VQA benchmarks have grown in scale and diversity (Xu et al., 2021; Zhou et al., 2025; Qian et al., 2023; Xie et al., 2025; Fruhwirth-Reisinger et al., 2025). These benchmarks primarily emphasize capability in scene understanding, multi-agent reasoning, and planning-related queries via multiple-choice or mixed-format VQA, but they are not designed for contrastive faithfulness evaluation.

A.2 Data Distribution and Statistics

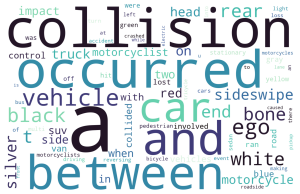
CCTVBENCH contains 7,104 binary video-question instances, comprising 305 positive videos and 305 paired generated negative videos. The benchmark is organized into 1,776 complete contrastive quadruples, where each quadruple includes four instances (v^+, q^+), (v^+, q^-), (v^-, q^+), and (v^-, q^-). All samples fall into six categories: `event`, `key_entity`, `spatial`, `spatial_temporal`, `causal`, and `counterfactual`. The dataset is strictly paired with equal numbers of positive and negative videos and has 100% quadruple coverage.



(a) Clip duration distributions across major types



(b) Question length distributions by category



(c) Word cloud distribution of traffic event caption

Figure 7: CCTVBench dataset statistics.

Fig. 7 summarizes several basic properties of CCTVBENCH. Fig. 7 (a) plot reports the distribution of clip durations across major anomaly types, showing that most clips concentrate around a short time window with a long tail of longer scenarios. Fig. 7 (b) plot compares question length across the six categories, indicating that the paired questions are concise and largely consistent in length, with moderate variation depending on whether the category emphasizes entities, spatial relations, or temporal descriptions. Fig. 7 (c) plot visualizes the overall question vocabulary as a word cloud, which is dominated by accident-centric terms such as *collision*, and *vehicle*.

A.3 Dataset Generation Details

CCTVBENCH is constructed via a hybrid pipeline that couples (i) structured traffic video annotations, (ii) world-model-based counterfactual video synthesis to obtain video pairs, and (iii) LLM-assisted contrastive caption and QA generation.

Manual meta-data annotation. We annotate each positive clip v_s^+ with visual instance tracks

Table 4: **Comparison of faithfulness-oriented and traffic VQA benchmarks.** #Ques = #QA pairs ; #Img/#Video = unique images / video clips. Eval: LLM = LLM-judge; B-QA = binary QA; MCQ = multiple-choice QA; OE = open-ended. CP (contrastive pairing): Video CP = contrastive/paired videos; Text CP = contrastive/paired questions.

Benchmark	Modality	#Ques	#Img	#Video	Eval	Video CP	Text CP
Faithfulness VQA benchmarks							
MMHal-Bench (Sun et al., 2023)	Image	96	96	-	LLM	✗	✗
EasyDetect (Chen et al., 2024)	Image	420	420	-	LLM	✗	✗
VideoHallucer (Wang et al., 2024b)	Video	1,800	-	948	B-QA/OE	✗	✓
EventHallusion (Zhang et al., 2025b)	Video	711	-	400	B-QA/OE	✗	✗
VALOR (Qiu et al., 2024)	Image	211	211	-	LLM	✗	✓
HallusionBench (Guan et al., 2024)	Image, Video	1,129	-	346	LLM	✗	✓
CertainlyUncertain (Chandu et al., 2024)	Image	178,000	-	95,800	VQA	✗	✓
EGOILLUSION (Seth et al., 2025)	Video	8,000	-	1,400	Closed+Open	✗	✗
VidHalluc (Li et al., 2025)	Video	9,295	-	5,002	B-QA/MCQ/OE/Sort	✓	✓
Traffic VQA benchmarks							
SUTD-TrafficQA (Xu et al., 2021)	Video	62,535	-	10,080	MQA	✗	✗
TUMTraffic-VideoQA (Zhou et al., 2025)	Video	85,000	-	1,000	MCQ	✗	✗
NuScenes-QA (Qian et al., 2023)	Multi-view Image, LiDAR	460,000	34,000	-	VQA	✗	✗
DriveBench (Xie et al., 2025)	Image	20,498	19,200	-	VQA	✗	✗
STSBench (Fruhwirth-Reisinger et al., 2025)	Video	971	-	43	VQA	✗	✗
CCTVBench (Ours)	Video	7,104	-	610	B-QA	✓	✓

and structured scene meta data. We first run RT-DETR-L (Lv et al., 2024) at 1 fps and ByteTrack (Zhang et al., 2022) to obtain coarse trajectories, which annotators refine into consistent instance identities and key participants, along with hypothesized causes and critical spatio-temporal relations. Annotators also segment v_s^+ into base and contrastive segment. The resulting track IDs, event segments, and motion cues constrain downstream caption and QA synthesis during LLM-assisted QA generation.

Counterfactual video synthesis. To construct minimally different yet semantically decisive video pairs, we leverage NVIDIA Cosmos Predict-2 Video2World-2B to synthesize a counterfactual video v_s^- in which the accident does not occur while preserving layout, illumination, and camera viewpoint as much as possible. We condition generation on the preserved prefix using a small set of conditioning frames, typically five, and falling back to one when the prefix is short. We execute generation at 720p by default, and use 16 frames per second for videos with sufficiently high frame rates and 10 frames per second otherwise. We show view-specific safe prompts in Fig. 8 and negative prompt in Fig. 9.

For long videos, we perform auto-regressive multi-segment generation with a maximum segment length of 5 seconds and re-condition each segment on the last conditioning frames from the previous segment. Prompts are selected based on the camera view, driving or roadside, to encourage normal continuation without collisions or violations, while a negative prompt list suppresses failure modes such as reckless behavior, anomalies, and temporal artifacts. After synthesis, we remove duplicated conditioning frames at segment boundaries, normalize the two videos to a consistent frame rate and resolution, and match their durations by truncating both videos to the shorter one, yielding temporally aligned pairs for contrastive evaluation.

Safe prompts

```
SAFE_PROMPTS = {
    "driving": [
        "Continue the driving scene with
    smooth, normal traffic flow. "
        "Show vehicles moving safely and
    predictably without any accidents, "
        "collisions, or traffic violations.",
    ],
    "roadside": [
        "Continue the roadside view showing
    normal traffic passing by safely. "
        "Maintain the fixed camera position
    with smooth vehicle movements and "
        "no traffic incidents. The objects
    in the scene remain in the same "
        "relative positions",
    ]
}
```

Figure 8: Safe prompts for counterfactual video synthesis

Negative prompt

```
NEGATIVE_PROMPT = (
    "reckless behavior, traffic jam,
     emergency vehicles, police chase, road
     rage, speeding, illegal maneuvers, "
    "traffic anomaly, traffic anomalies,
     anomaly, anomalies, inconsistency,
     inconsistencies, temporal inconsistency,
     temporal artifacts"
)
```

Figure 9: Negative prompt for counterfactual video synthesis.

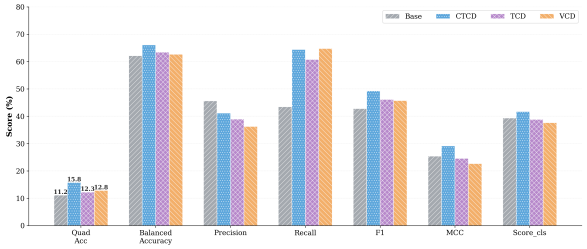


Figure 10: Qwen3-VL contrastive decoding comparison.

Figure 11: Grouped bar charts for Qwen3-VL-7B contrastive decoding methods, Quad Acc and classification metrics.

A.3.1 Category-wise Breakdown of Main Results

Fig. 12 summarizes the category-wise classification score $\text{Score}_{\text{cls}}$. Across model groups, BaAcc is comparatively high and shows a relatively smooth profile across Event, Spatial, Spatial-Temporal, and Key-Entity, indicating that many models can reach moderate instance-level QA quality when evaluated independently. However, the consistency-oriented views in Fig. 13 and Fig. 14 expose much sharper category gaps and model separations. Key-Entity is the most stable category across groups, while Causal and Counterfactual remain the most challenging and also exhibit the largest variance across models, consistent with the benchmark requirement to reject plausible-but-false hypotheses under near-identical counterfactual scenes.

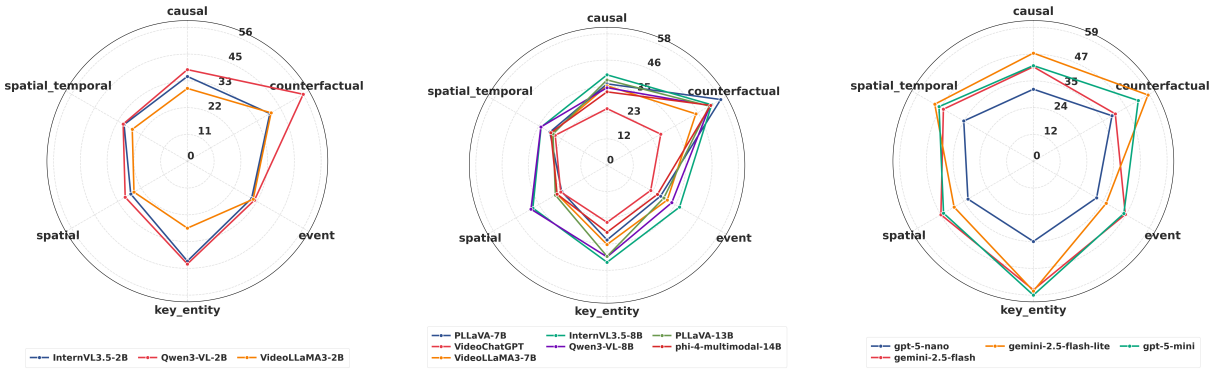
Within open-source models, the medium group shows the clearest stratification. Qwen3-VL-8B presents the most uniformly high profiles in both question and video consistency, whereas VideoChatGPT and Phi-4-Multimodal display consistently low and uneven shapes, suggesting that improvements in standard QA accuracy do not necessarily translate into stable contrastive decision patterns. The small-model group further highlights this mismatch: despite moderate BaAcc. For commercial models, the group shows substantially more balanced radar shapes in consistency metrics, with GPT-5-mini and Gemini-2.5 variants maintaining stronger and more uniform performance across categories, while GPT-5-nano remains weaker and less stable, especially on the harder Counterfactual and Causal axes.

A.3.2 Detailed Contrastive Decoding Results

Fig. 10 reveal that contrastive decoding improves standard QA scores but is not always aligned with quadruple consistency. On Qwen, C-TCD yields the strongest overall gains, increasing QuadAcc while also improving balanced accuracy, F1, MCC, and $\text{Score}_{\text{cls}}$. The main effect is a large boost in recall, which indicates that Qwen is partially omission-limited on positive scenes and benefits from a semantically decisive contrast signal provided by the paired counterfactual video. In comparison, VCD and TCD also raise recall but reduce precision more noticeably, suggesting that generic corruption or temporal degradation can over-suppress useful evidence and shift the decision boundary in a less targeted way than C-TCD.

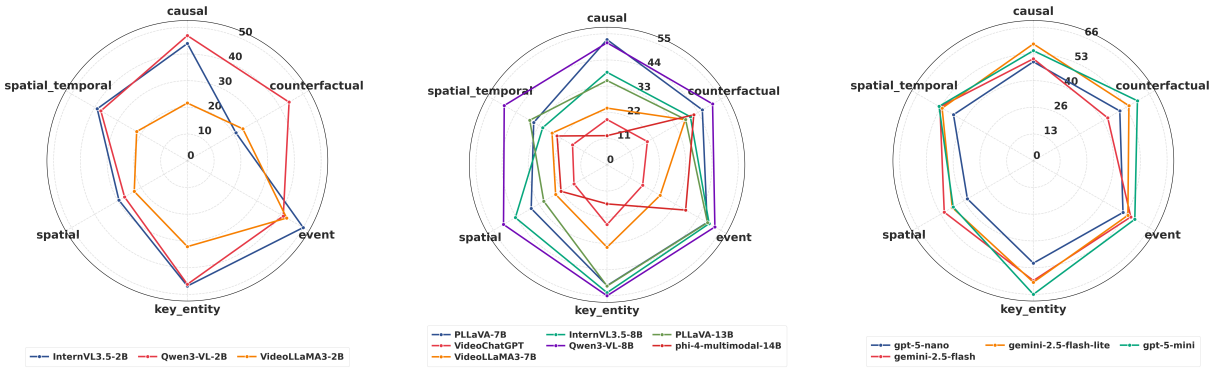
A.3.3 Hyperparameter Analysis of Contrastive Decoding

Fig. 15 analyzes the effect of the contrast strength α for Qwen3-VL-2B under VCD, TCD, and C-TCD. Overall, C-TCD is the most α-sensitive but also the most rewarding: increasing α steadily improves both video consistency and question consistency, and these gains translate to higher BalancedAcc and QuadAcc up to a clear sweet spot around $\alpha \approx 0.75$–1.0. Beyond this range, performance drops across metrics, suggesting that overly strong contrast begins to over-suppress valid evidence and destabilizes the strict quadruple decision pattern. TCD is comparatively stable across α with modest fluctuations, indicating that temporal degradation provides a weaker and less targeted contrast signal, so scaling α yields limited returns. In contrast, VCD degrades as α increases, with both consistency scores and QuadAcc trending downward, which is consistent with visually corrupted negatives becoming increasingly harmful when their influence is amplified. These results support using a moderate contrast strength for reli-



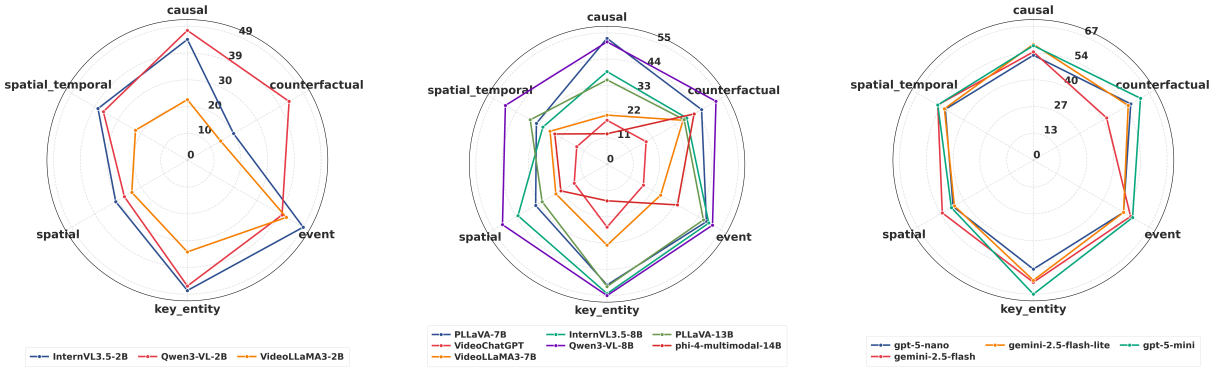
(a) Small models (b) Medium models (c) Large/commercial models

Figure 12: Category-wise classification score ($\text{Score}_{\text{cls}}$) radar plots for small, medium, and large/commercial model groups.



(a) Small models (b) Medium models (c) Large/commercial models

Figure 13: Category-wise question consistency radar plots for small, medium, and large/commercial model groups.



(a) Small models (b) Medium models (c) Large/commercial models

Figure 14: Category-wise video consistency radar plots for small, medium, and large/commercial model groups.

able improvements, and they further highlight that semantically decisive counterfactual pairs make C-TCD substantially more effective and tunable than generic corruption or temporal degradation.

A.4 Qualitative Results

Fig. 16 shows an intersection scenario with a T-bone collision between the black car and the ego vehicle in the positive video, paired with a counterfactual negative where the collision is removed. The quadruple structure exposes two common inconsistencies: some models correctly fire on the target event but still answer *Yes* on the counterfactual negative under the same query, while others confuse the mutually exclusive alternative collision type and trigger a false positive on the rear-end hypothesis on the same scene. The spatial and spatial-temporal pairs further reveal brittle left–right and

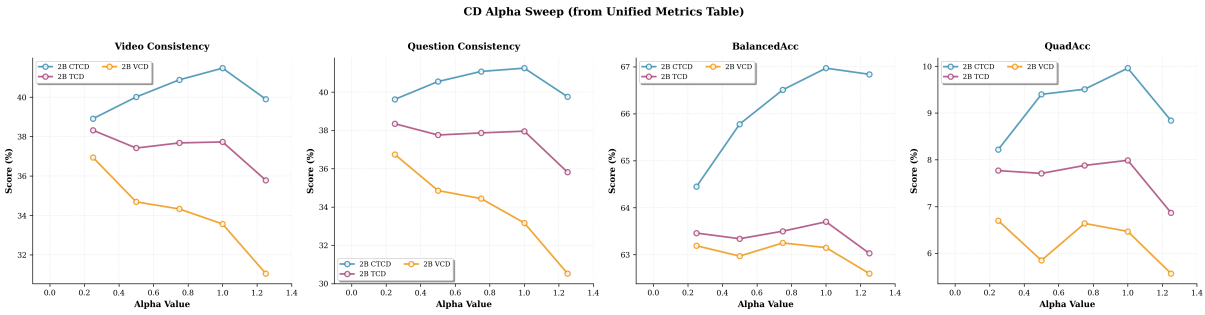


Figure 15: Effect of contrast strength α on Qwen3-VL-2B under VCD, TCD, and C-TCD. We report video consistency, question consistency, balanced accuracy, and QuadAcc.

front–behind grounding under near-identical views, where models may change answers across videos but in the wrong direction.

Fig. 17 presents a snowy road case where the silver car's turning maneuver is central to the accident trajectory. This scene highlights how temporal-localized motion cues and turning intent can dominate the downstream causal and counterfactual questions: even when models reject the alternative event type, they may still fail the none-of-the-above requirement by producing affirmative answers on the counterfactual negative for maneuver-dependent or explanation-style queries. The case illustrates that counterfactual and causal questions are especially sensitive to subtle premise mismatches and to ambiguity in what visual evidence is sufficient to justify a causal attribution.

Fig. 18 shows an urban scene involving a red motorcycle, where the target event is a head-on collision in the positive video and the paired negative removes the collision while preserving the surrounding layout. Here, a frequent failure mode is *omission* on the positive video, where models default to *No* for the target event and propagate that conservatism to related key-entity and counterfactual questions. At the same time, some models still hallucinate on the counterfactual negative for counterfactual-condition queries, suggesting that plausible textual hypotheses can override the intended none-of-the-above state even when the paired video edit removes the critical event.

	Positive Video	Negative Video	Positive Video	Negative Video
Event	Did a T-Bone collision occur between the black car and the ego vehicle?		Did a rear-end collision occur between the black car and the ego vehicle?	
InternVL	Yes	No	Yes	No
Qwen3-VL	No	No	No	No
GPT-5	Yes	Yes	No	No
Gemini	Yes	Yes	No	No
Spatial Temporal	Was the black car to the right of the ego vehicle while it was driving forward?		Was the black car to the left of the ego vehicle while it was driving forward?	
InternVL	No	No	No	No
Qwen3-VL	Yes	No	No	Yes
GPT-5	No	Yes	Yes	No
Gemini	No	No	Yes	Yes
Key Entity	Was the black car involved in the collision with the ego vehicle?		Was the white truck involved in the collision with the ego vehicle?	
InternVL	Yes	No	No	No
Qwen3-VL	No	No	No	No
GPT-5	Yes	Yes	No	No
Gemini	Yes	Yes	No	No
Counter factual	Could the collision have been avoided if the ego vehicle had slowed down at the intersection?		Could the collision have been avoided if the ego vehicle had sped up at the intersection?	
InternVL	Yes	Yes	No	No
Qwen3-VL	Yes	No	No	No
GPT-5	Yes	Yes	No	No
Gemini	Yes	Yes	No	No
Causal	Was the collision caused by the ego vehicle's failure to yield?		Was the collision caused by the black car's failure to yield?	
InternVL	Yes	No	No	Yes
Qwen3-VL	Yes	No	No	No
GPT-5	Yes	No	No	Yes
Gemini	Yes	No	No	Yes
Spatial	Was the black car to the right of the ego vehicle?		Was the black car to the left of the ego vehicle?	
InternVL	No	No	No	No
Qwen3-VL	Yes	No	No	Yes
GPT-5	No	Yes	Yes	No
Gemini	No	No	Yes	Yes

Figure 16: Qualitative example 1: Intersection T-bone collision between the black car and the ego vehicle.

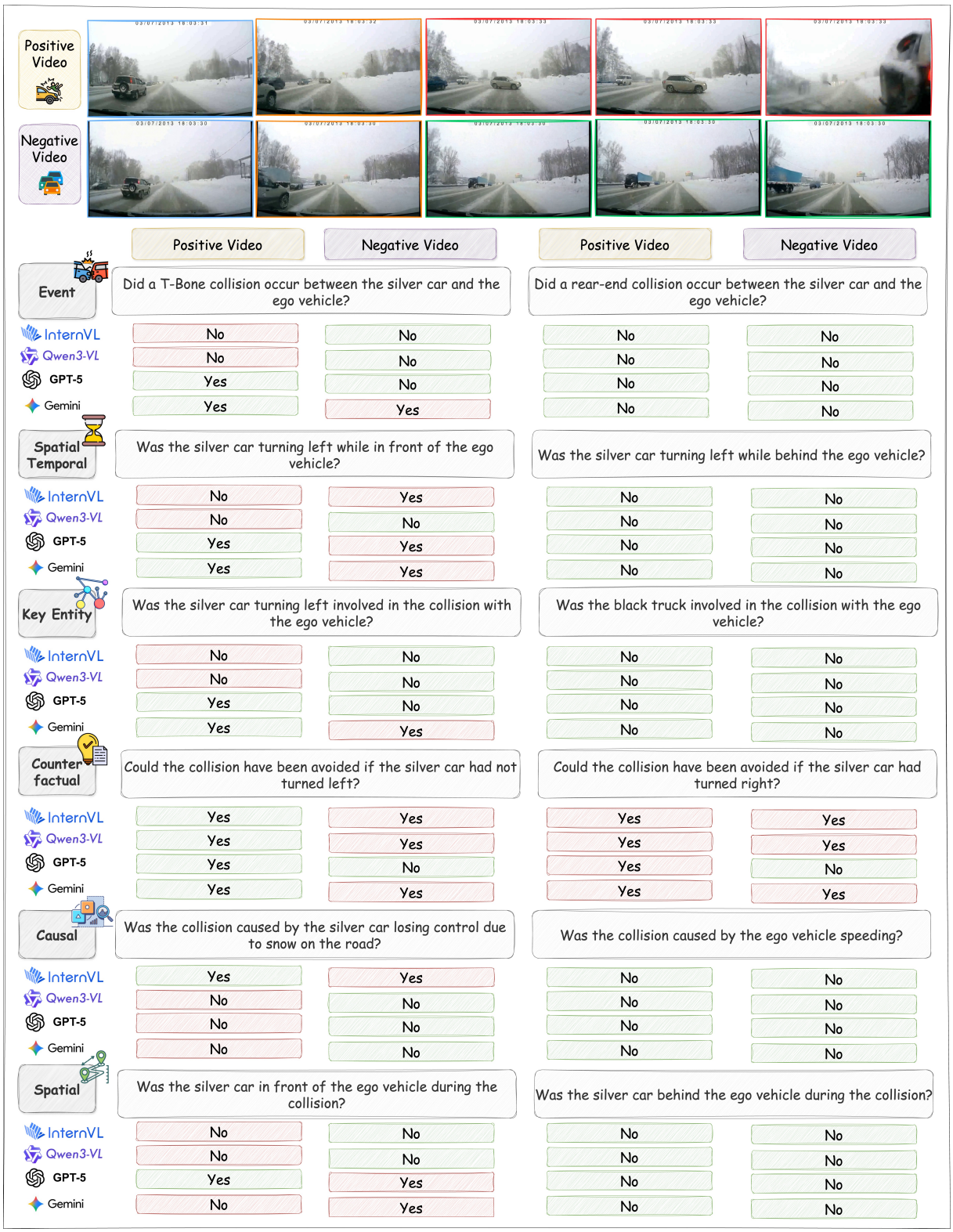


Figure 17: Qualitative example 2: Snowy-road turning scenario with loss of control collision.

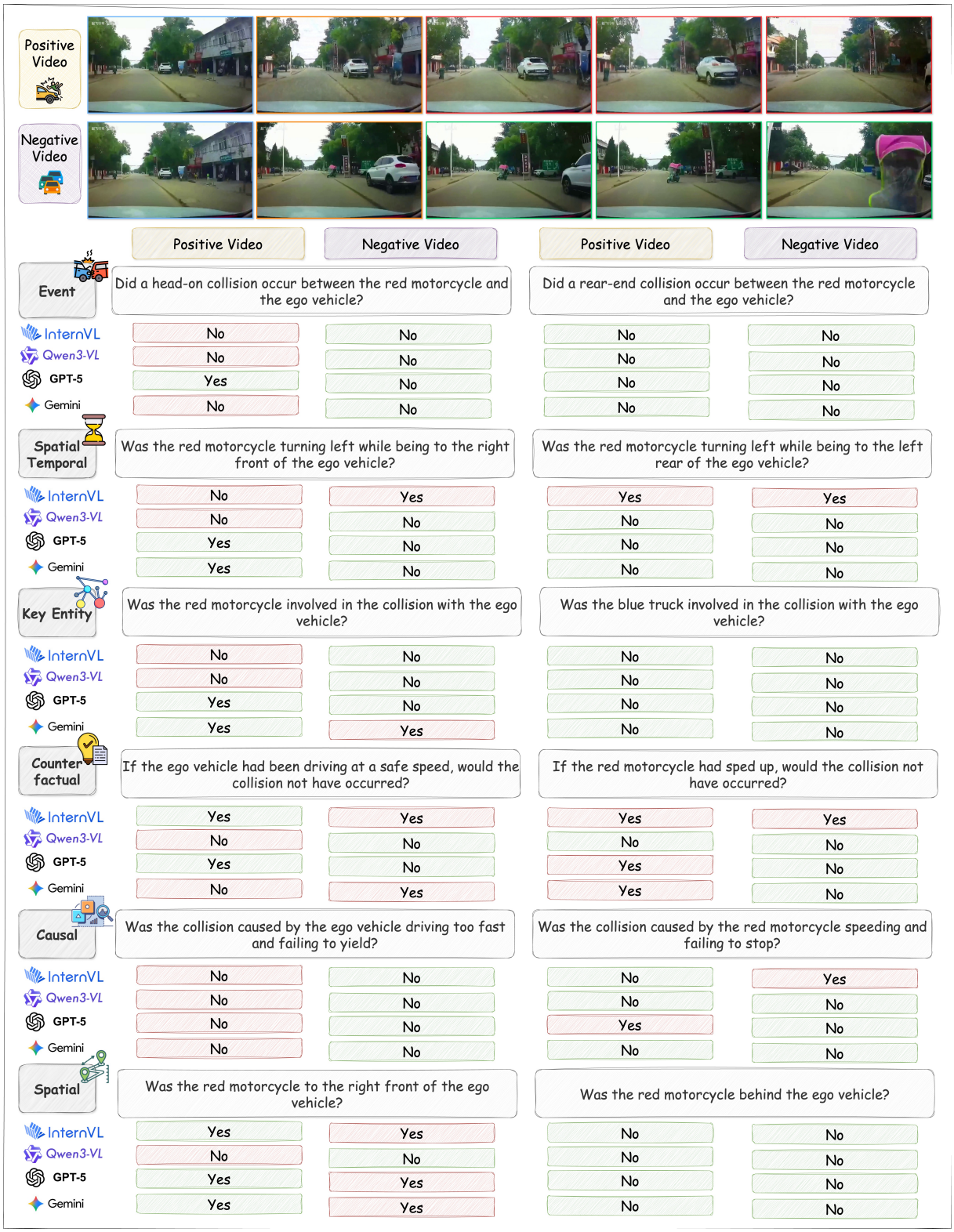


Figure 18: Qualitative example 3: Urban scenario with a head-on collision between a red motorcycle and ego vehicle.