# RiskCueBench: Benchmarking Anticipatory Reasoning from Early Risk Cues in Video-Language Models

Anonymous ACL submission


## Abstract

With the rapid growth of video centered social media, the ability to anticipate risky events from visual data is a promising direction for ensuring public safety and preventing real world accidents. Prior work has extensively studied supervised video risk assessment across domains such as driving, protests, and natural disasters. However, many existing datasets provide models with access to the full video sequence, including the accident itself, which substantially reduces the difficulty of the task. To better reflect real world conditions, we introduce a new video understanding benchmark RiskCueBench in which videos are carefully annotated to identify a risk signal clip, defined as the earliest moment that indicates a potential safety concern. Experimental results reveal a significant gap in current systems' ability to interpret evolving situations and anticipate future risky events from early visual signals, highlighting important challenges for deploying video risk prediction models in practice.


## 1 Introduction

Advanced vision language models (VLMs) have surfaced with remarkable abilities to comprehend complex spatial relationships, temporal sequences, and visual narratives. However, the specific needs of safety critical applications, particularly situational risk assessment and prediction, still differ significantly from those of general video understanding benchmarks. As shown in Figure 1, most existing benchmarks emphasize post hoc understanding, where models analyze or describe events after they have fully unfolded, such as answering questions or generating captions given complete visual context (Caba Heilbron et al., 2015; Zhou et al., 2025; Hong et al., 2025a).

In contrast, predictive reasoning requires models to anticipate future events from partial and often ambiguous early signals. While current VLMs per- form well on descriptive tasks, their capacity for such anticipatory reasoning remains largely unex- plored. Recent benchmarks targeting next event prediction typically frame evaluation as multiple choice question answering, where models select from predefined outcomes rather than reason freely about future risks (Wu et al., 2024; Wang et al., 2025; Xun et al., 2025; Cheng et al., 2025), which lack the open-ended nature of real-world risk. Other work focuses narrowly on traffic accidents via synthetic data or non-reasoning metrics (Kung et al., 2024; Fatima et al., 2021; Bao et al., 2020; Hussain et al., 2024), leaving domains like crowd dynamics unaddressed.

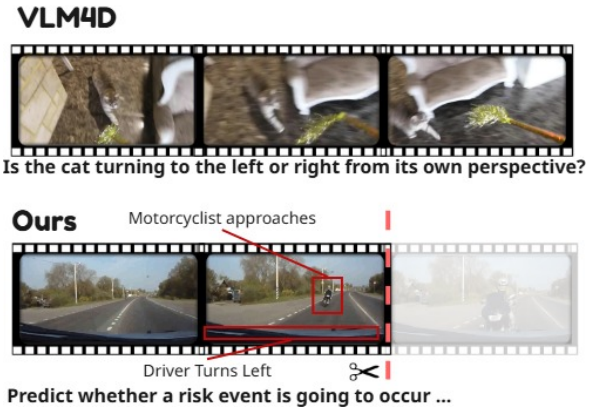


Figure 1: Many existing benchmarks ask about events which occur during the video. Our RiskCueBench focuses on predicting future risky events (details in §3).

To bridge this gap, we propose RiskCueBench, a benchmark evaluating whether VLMs can anticipate emerging safety concerns from early video signals. Using a Question-Reasoning-Answer (QRA) framework, we explicitly capture model decision-making. Our analysis reveals significant limitations in state-of-the-art VLMs' predictive abilities. Our contributions include:

- We introduce a challenging video risk prediction benchmark RiskCueBench that requires VLMs

| Dataset | Temporal Reasoning Sec 4.3.3 | Object Grounding Sec 4.3.1 | Reasoning Metrics Sec 4.3.2 | Event Forecasting Sec 4.3.4 | Risk Events Sec 3.3 |
|---|---|---|---|---|---|
| TimeLogic | ✓ | ✓ | ✗ | ✗ | ✗ |
| ReXTime | ✓ | ✓ | ✗ | ✗ | ✗ |
| MotionBench | ✓ | ✗ | ✗ | ✗ | ✗ |
| MVBench | ✓ | ✓ | ✗ | ✗ | ✗ |
| VLM4D | ✓ | ✗ | ✗ | ✗ | ✗ |
| RTV-Bench | ✓ | ✓ | ✗ | ✓ | ✗ |
| FutureBench | ✓ | ✗ | ✗ | ✓ | ✗ |
| RiskBench | ✓ | ✓ | ✗ | ✓ | ✓ |
| **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Overview of benchmarks for spatio-temporal reasoning in vision–language models. Unlike prior work, our analysis centers on risk events, evaluating models' ability to forecast under uncertainty, while also examining reasoning traces and object grounding.

to infer potential danger from subtle visual cues.

- We develop an efficient workflow that leverages model disagreement and LLM filtering to automatically identify difficult cases and construct the benchmark (Section 3).
- We conduct an extensive analysis of model reasoning using custom metrics on our curated videos, highlighting strengths and limitations of current state-of-the-art VLMs (Section 4).

## 2 Related Work

**Video Language Models and Reasoning Systems** Advanced capabilities across temporal visual understanding tasks have been demonstrated by recent developments in video language models. The state-of-the-art in combining language modeling with video encoding mechanisms for intricate temporal reasoning is represented by VideoLLaMA 3 (Zhang et al., 2025), Video LLaVA (Lin et al., 2023), Qwen-VL (Bai et al., 2025), and Apollo (Zohar et al., 2025). With the introduction of reinforcement learning techniques for video comprehension by MiMo-RL (Team et al., 2025), the resolution of long-form video analysis problems by LongVILA-R1 (Chen et al., 2025), and the incorporation of explicit reasoning mechanisms that mimic human cognitive processes by GLM-4.1V-Thinking (Hong et al., 2025b), specialized reasoning systems have further advanced this field.

**Spatio-Temporal Reasoning** Recent work has introduced numerous benchmarks aimed at evaluating the spatio temporal reasoning capabilities of vision language models in video understanding (Table 1). These efforts probe diverse aspects of temporal and spatial cognition, including temporal logic and event ordering, fine grained motion understanding, and dynamic scene interpretation in domains such as egocentric video and autonomous driving. Benchmarks such as TimeLogic (Swetha et al., 2025) and ReXTime (Chen et al., 2024) focus on logical and causal reasoning over event sequences, while MotionBench (Hong et al., 2025a) and related diagnostic tasks evaluate sensitivity to motion dynamics and temporal direction. MVBench (Li et al., 2024), VLM4D (Zhou et al., 2025), and STSBench (Fruhwirth-Reisinger et al., 2025), assess joint spatial and temporal reasoning across diverse video scenarios, yet many tasks can still be solved using static cues or retrospective access to the full video.

Crucially, existing benchmarks primarily evaluate model performance using multiple-choice accuracy, providing limited insight into how models arrive at their predictions. As a result, they do not explicitly assess the quality, faithfulness, or grounding of model reasoning. In contrast, we include a dedicated evaluation of model reasoning with metrics designed to assess whether risk predictions are supported by temporally coherent and visually grounded explanations. This enables a more fine-grained understanding of spatio-temporal reasoning capabilities and exposes limitations that are not captured by answer accuracy.

## 3 Benchmark Construction

We propose a reproducible, domain-agnostic framework for constructing risk-centric video benchmarks, emphasized by temporally grounded signals. The four-stage process includes: (1) large-scale collection via structured queries (§3.1); (2) multi-stage

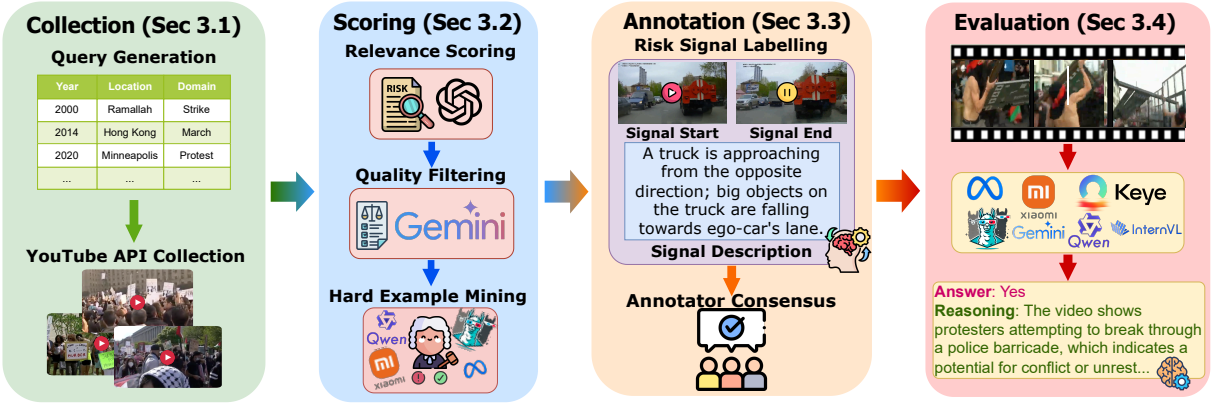


Figure 2: Overview of our pipeline to curate risk signal clips from real-world incidents and evaluate VLM reasoning. **Collection:** We collect a large candidate set of YouTube videos using domain-relevant keywords. **Scoring:** The collected videos are filtered to only retain those with potential risk and high difficulty. **Annotation:** Human annotators label the filtered videos to identify the risk signal clip. **Evaluation:** Popular VLMs are presented with only the risk signal clip, and their output reasoning traces are evaluated.

filtering for risk relevance (§3.2); (3) fine-grained human annotation using a temporal protocol (§3.3); and (4) a set of evaluation metrics for video reasoning, as shown in Figure 2. While instantiated for protests and traffic incidents, the framework is extensible to scenarios like natural disasters or workplace hazards, ensuring both diverse risk context and detailed temporal depth.

### 3.1 Data Collection

We first curate diverse risk-relevant text queries for the YouTube Data API and collect an initial pool of candidate videos and their metadata.

**Query Curation.** To systematically evaluate models, we construct a structured list of real-world events across diverse regions and time. First, we identify safety-critical domains relevant to public safety and social risk(e.g., protests, traffic incidents). For each, we curate an event list by prompting GPT-4V for year–location pairs, followed by manual verification via Google Search to ensure actual news coverage. The validated events are compiled into a structured list of year–location combinations.Each event is paired with a predefined vocabulary of domain-relevant terms (e.g., "protest," "march," ). We convert each structured event into a text query using a template; the full template and examples are provided in Appendix A.1. This process maximizes coverage while reducing bias toward specific regions or languages, enabling the dataset to reflect a variety of sociopolitical and environmental conditions under which risk unfolds.

**Video Collection.** Using these queries, we retrieve videos and their metadata (e.g., title, description, upload date) through the YouTube Data API. This ensured replicable sampling of publicly available video, rather than relying on proprietary or manually sourced data. Metadata enables traceability and allows filtering or stratification by contextual variables such as time and location.

### 3.2 Dataset Filtering

We apply a sequence of automated filtering steps to retain videos that depict active risk events and meet minimum visual quality standards.

**Text-based Relevance Filtering.** Not all retrieved videos are useful for analyzing real-time risk dynamics; many contain news commentary or post-event summaries. To identify videos that plausibly contain *in-action* risk events, we use an LLM (GPT-4o) to evaluate each video's title and description pair. The model is prompted to assess the likelihood that the video depicts an ongoing risk event rather than commentary, lectures, or summaries:

> **Prompt:** On a scale from 1 to 10, how likely is this video to contain [risk event] video in action but not lecture/tutorial/slides, etc.? Answer with a single integer from 1 (very unlikely) to 10 (very likely).

Videos scoring below a predefined threshold of 9 out of 10 are filtered out.

**Fine-Grained Visual Quality Filtering.** Even among relevant videos, production artifacts such as subtitles, overlays, or heavy editing can introduce spurious shortcuts for visual models. To ensure

models perform reasoning on risk-relevant visual content rather than these shortcuts, we use Gemini 2.5 Flash Lite to evaluate videos along 12 visual quality dimensions as in Appendix A.2 (i.e., logo, location, time/date, reporter presence, SNS overlay, image quality, temporal continuity, consequence text, title/banner, subtitle, camera perspective). Videos that fail to obtain Score 2 in at least 10 out of the 12 quality criteria are removed, ensuring that the benchmark prioritizes authentic, unedited videos and emphasizes visual perception and reasoning under realistic conditions.

**Hard Example Mining.** To explicitly include challenging and ambiguous cases, we perform hard example mining using five VLMs (Gemini 2.5, Qwen-7B, MiMo-RL, MiMo-SFT, InternVL). Each model is prompted to predict whether a given video contains a clear risk event solely based on the visual content, using a standardized instruction prompt (provided in the Appendix A.5). We define a video as a *hard example* if more than three out of five models disagree with each other. Videos with the highest disagreement represent the most ambiguous or complex scenarios, meaning cases that often challenge both human and AI perception. Including these "hard examples" strengthens the dataset's ability to test model generalization and resilience in uncertain or borderline conditions.

### 3.3 Annotation

The final dataset is annotated by trained human annotators using a structured protocol designed to capture the temporal and visual progression of risk.

**Annotation Protocol.** Each video is annotated along eight dimensions that jointly consider temporal boundaries, visual cues, and semantic content (annotation example see Appendix A.3)

1. **Risk Signal Start.** Timestamp of the first observable visual signal indicating potential risk (or no risk).

2. **Risk Signal End.** Timestamp when clear risk (or no-risk) indicators cease.

3. **Risk Visual Indicator.** Initial visual cues used to judge risk/no risk.

4. **Risk Signal Description.** Full narrative description of the signals using temporal markers (*first*, *then*, *afterwards*).

5. **Accident Start Frame.** Timestamp of the first observable moment of incident.

6. **Accident End Frame.** Timestamp when the incident or scene fully concludes.

7. **Accident Description.** One or more sentences summarizing the incident.

8. **Risk Label.** Binary assignment of "yes" (clear risk observed) or "no" (peaceful protest, normal driving).

This manual annotation process ensures the final dataset retains high-granularity temporal and semantic information about risk onset, escalation, and resolution. By combining timestamped cues, narratives, and categorical labels, the annotation schema supports both fine-grained visual reasoning analysis and binary classification evaluation.

**Instantiation.** Using our pipeline, we collect video data from two risk domains: protest and traffic incidents. We report statistics in Table 2. Depending on the domain and data source, certain steps in the pipeline may be omitted. For example, for the car crash dataset, which was already preprocessed and cleaned, we skipped several cleaning and filtering steps (e.g., YouTube API collection, visual quality filtering).

| Video Type | Average Signal Length | # Videos |
|---|---|---|
| Car Crash (Normal) | 3.49 ± 0.58 | 250 |
| Car Crash (High Risk) | 1.38 ± 0.77 | 252 |
| Protest (Normal) | 17.77 ± 14.81 | 267 |
| Protest (High Risk) | 5.99 ± 7.07 | 217 |
| Total | | 986 |

Table 2: Risk Signal Length for Our Risk Prediction Dataset.

The length of risk signal vary by domain. The Protest dataset exhibits a broad distribution with an average high-risk signal length of 12.5 seconds, reflecting the gradual escalation typical of crowd dynamics. In contrast, the Car Crash dataset features much more abrupt transitions, with high-risk signals averaging only 2.4 seconds (see Appendix A.6). These statistics shows the differing "decision windows" available for VLMs to perform successful risk prediction in each scenario.

### 3.4 Evaluation

**Risk Prediction (F1).** We first evaluate the binary risk classification of each model using the standard F1 score.

**Reasoning Grounding Accuracy (RGA).** To quantify how well model reasoning is anchored in relevant visual evidence, we compute semantic alignment between judge-extracted decision items and human-annotated risk visual indicators using SentenceTransformer embeddings (all-MiniLM-L6-v2). For each decision item $\hat{o}_{ij} \in \hat{\mathcal{O}}_i$, we compute its maximum cosine similarity against all ground-truth visual indicators $\mathcal{O}_i = \{o_{i1}, \ldots, o_{im}\}$:

$$s_{ij} = \max_{k \in \{1,\ldots,m\}} \cos(\mathbf{e}_{\hat{o}_{ij}}, \mathbf{e}_{o_{ik}}) \quad (1)$$

where $\mathbf{e}_{\hat{o}_{ij}}$ and $\mathbf{e}_{o_{ik}}$ denote the embedding vectors for decision item $\hat{o}_{ij}$ and visual indicator $o_{ik}$, respectively. Using an F1-optimized threshold $\tau$ derived from ROC analysis, we determine whether each decision item is semantically grounded:

$$g_{ij} = \mathbb{I}\left[s_{ij} \geq \tau\right] \quad (2)$$

The overall reasoning grounding accuracy (RGA) metric is computed as the average percentage of grounded decision items across all samples:

$$\text{RGA} = \frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} \frac{1}{|\hat{\mathcal{O}}_i|} \sum_{j=1}^{|\hat{\mathcal{O}}_i|} g_{ij} \times 100\% \quad (3)$$

**Temporal Reasoning Difference (TRD).** To analyze the temporal reasoning capabilities, we augmented each video in 3 different ways and observed the effect on performance:

- **Frames Shuffled:** All the frames in the video are shuffled into random order.
- **Half-Swaped:** The first half of the video is swapped with the second half.
- **Frames Reversed:** The frames of the video are presented to the model in reverse order.

For each augmentation type $a \in \mathcal{A} = \{$"shuffled", "swapped", "reversed"$\}$, we compute the absolute F1 score difference between the original video and its augmented counterpart. The TRD metric is then calculated as the average absolute F1 difference across all augmentation types:

$$\text{TRD} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left|\text{F1}_{\text{original}} - \text{F1}_a\right| \quad (4)$$

A higher TRD value indicates sensitivity to temporal ordering, suggesting the model relies on temporal information for risk prediction. Conversely, a TRD near zero implies the model is invariant to temporal structure, potentially relying on static visual features rather than dynamic reasoning.

**Self-Correction Degradation (SCD).** To quantify the impact of model hesitation on prediction quality, we partition the dataset into two subsets based on whether the judge-extracted confusion count is non-zero: $\mathcal{V}^+ = \{v_i \mid c_i > 0\}$ (samples with self-correction markers) and $\mathcal{V}^- = \{v_i \mid c_i = 0\}$ (samples without). We then compute a weighted F1 gap that accounts for subset size:

$$\text{SCD} = P(\mathcal{V}^+) \cdot \text{F1}(\mathcal{V}^+) - P(\mathcal{V}^-) \cdot \text{F1}(\mathcal{V}^-) \quad (5)$$

where $P(\mathcal{V}^k) = \frac{|\mathcal{V}^k|}{|\mathcal{V}|}$ is the proportion of samples in subset $k$, and $\text{F1}(\cdot)$ denotes the F1 score over the respective subset. This formulation automatically normalizes for subset size, enabling fair comparison across models. A negative SCD value indicates that self-correction behavior degrades model performance, with larger negative values suggesting greater F1 loss due to overthinking.

## 4 Experiments and Analysis

### 4.1 Experimental Setup

Videos are sampled at 1 frame per second (fps) to balance computational efficiency with temporal coverage of the risk signal clips. All models are evaluated using the same standardized prompt template to ensure fair comparison, with one exception: for reasoning-enhanced models, we omit explicit reasoning instructions as these capabilities are integrated directly into their chat templates.

**Baseline Models** We evaluate 16 state-of-the-art models across two distinct categories.
*(1) Standard VLMs* perform direct video-to-text generation; include Video LLaVA, VideoLLaMA 3, InternVL 3.5, Qwen-VL, ChatUniVi, Gemini 2.5 Flash Lite, Apollo, and MiMo-SFT.
*(2) Reasoning-Enhanced VLMs* leverage explicit reasoning chains or reinforcement learning-based training to bolster predictive accuracy. These are represented by Qwen-VL Thinking, MiMo-RL, LongVILA-R1, and GLM-4.1V-Thinking.

### 4.2 Overall Performance

Performance varies significantly across domains. On Car Crashes, models achieve F1 scores of

| Model | Car Crash | | | | Protest | | | |
|---|---|---|---|---|---|---|---|---|
| | F1 | RGA | TRD | SCD | F1 | RGA | TRD | SCD |
| ***Non-Reasoning Models*** | | | | | | | | |
| Video LLaVA | 0.53 ± 0.04 | 48.2 ± 5.62 | 1.84 ± 0.63 | - | 0.44 ± 0.02 | 33.5 ± 10.54 | 1.87 ± 0.83 | - |
| VideoLLaMA 3 | 0.59 ± 0.05 | 54.1 ± 6.21 | 1.65 ± 0.58 | - | 0.46 ± 0.05 | 39.2 ± 8.73 | 2.42 ± 0.71 | - |
| Qwen3-VL-8B | 0.64 ± 0.03 | 57.3 ± 5.14 | 1.48 ± 0.52 | - | 0.48 ± 0.04 | 42.8 ± 7.92 | 2.23 ± 0.69 | - |
| MiMo-SFT | 0.58 ± 0.06 | 51.4 ± 7.38 | 2.12 ± 0.74 | - | 0.47 ± 0.04 | 38.1 ± 9.45 | 1.76 ± 0.81 | - |
| Apollo-7B | 0.57 ± 0.05 | 53.6 ± 6.84 | 1.83 ± 0.61 | - | 0.45 ± 0.03 | 37.4 ± 8.91 | 2.59 ± 0.76 | - |
| ***Reasoning Models*** | | | | | | | | |
| InternVL-3.5 | 0.61 ± 0.04 | 64.2 ± 4.87 | 2.87 ± 0.68 | -0.28 ± 0.05 | 0.47 ± 0.05 | 47.8 ± 6.32 | 2.54 ± 0.73 | -0.42 ± 0.06 |
| Qwen3-VL-8B-T | 0.66 ± 0.03 | 68.7 ± 4.23 | 2.21 ± 0.57 | -0.29 ± 0.04 | 0.57 ± 0.04 | 56.5 ± 5.41 | 2.08 ± 0.64 | -0.36 ± 0.05 |
| MiMo-RL | 0.63 ± 0.05 | 61.3 ± 6.95 | 3.67 ± 0.82 | -0.35 ± 0.07 | 0.51 ± 0.06 | 49.9 ± 7.84 | 3.42 ± 0.89 | -0.45 ± 0.08 |
| Keye-VL 1.5 | 0.57 ± 0.04 | 63.8 ± 5.56 | 3.15 ± 0.71 | -0.21 ± 0.04 | 0.59 ± 0.05 | 52.3 ± 6.18 | 2.91 ± 0.77 | -0.29 ± 0.05 |
| GLM 4.1 V-T | 0.58 ± 0.05 | 65.1 ± 5.28 | 3.39 ± 0.76 | -0.26 ± 0.05 | 0.56 ± 0.05 | 49.1 ± 6.72 | 3.18 ± 0.81 | -0.34 ± 0.06 |
| ***Closed-Source Models*** | | | | | | | | |
| Gemini-Flash-3 | 0.69 ± 0.02 | - | - | - | 0.67 ± 0.03 | - | - | - |
| ***Baselines*** | | | | | | | | |
| Random Guess | 0.50 | - | - | - | 0.44 | - | - | - |
| Human | 0.98 ± 0.01 | - | - | - | 0.97 ± 0.01 | - | - | - |

Table 3: Performance statistics of current SOTA models on the binary risk prediction task for both Car Crash and Protest scenarios. Certain metrics remain blank for Gemini-Flash-3 as its internal reasoning traces are not accessible for evaluation.

0.27–0.71, with Qwen3-VL-8B leading (0.7151), followed by InternVL-3.5 (0.6585) and VideoLLaMA 3 (0.6047). Protest scenarios show markedly lower performance (F1: 0.26–0.46), suggesting models struggle with nuanced social and behavioral cues compared to structured vehicular patterns. MiMo-RL performs best on protests (0.4560) but only achieves 0.4256 on car crashes.

Different models exhibit distinct failure modes. Video LLaVA and ChatUniVi show low precision on protests (high false positives), while MiMo-SFT demonstrates low recall on car crashes (conservative detection). Reasoning-enhanced models do not consistently outperform standard VLMs, suggesting that explicit reasoning mechanisms are not effectively calibrated for early risk signal detection.

Overall, results reveal a critical performance ceiling, particularly for socially-embedded scenarios. F1 scores of 45–55% on protest footage indicate substantial gaps from human-level performance and fundamental limitations in interpreting early warning signals and predicting safety outcomes from partial information.

**Takeaway: Model performance varies sharply by domain.** While leading vision–language models achieve moderate success on car-crash prediction, they struggle substantially on protest scenarios that require interpreting social and behavioral cues.

### 4.3 Reasoning Chain Analysis

To analyze the failure modes of VLMs in risk prediction, we study two aspects of model behavior: self-correction patterns and reasoning grounding. We employ Gemini-Pro-3 as a judge model to systematically evaluate model reasoning chains.

For each video $v_i \in \mathcal{V}$, we obtain human ground-truth risk annotations $a_i = (d_i, \mathcal{O}_i, l_i)$, where $d_i$ is the human-written risk description, $\mathcal{O}_i = \{o_{i1}, o_{i2}, \ldots, o_{im}\}$ is the set of annotated risk visual indicators, and $l_i$ is the risk label. The judge model parses each model-generated reasoning chain $r_i$ to extract: (1) a confusion count $c_i$, indicating the presence of self-correction markers (e.g., "wait...", "actually..."), and (2) a structured set of predicted decision items $\hat{\mathcal{O}}_i = \{\hat{o}_{i1}, \hat{o}_{i2}, \ldots, \hat{o}_{in}\}$, representing the objects and entities the model references when justifying its risk prediction.

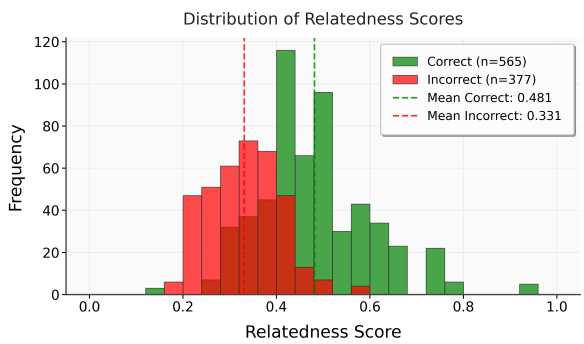


Figure 3: Distribution of Relatedness Score. Correct predictions (green) exhibit significantly higher mean relatedness scores than incorrect ones (red), indicating that accurate risk anticipation is strongly tied to better visual grounding.

#### 4.3.1 Reasoning Grounding

Correct predictions exhibit stronger grounding than incorrect ones. In the protest dataset, correct predictions achieve a mean relatedness score of 0.48 compared to 0.33 for incorrect ones, while car crash scenarios show a similar gap (0.40 vs. 0.32). Additionally, incorrect predictions contain approximately 50% more ungrounded decision items. These results indicate grounding failures are prevalent in VLM risk prediction errors.

**Takeaway: VLM reasoning is poorly grounded in relevant risk objects.** Prediction accuracy strongly correlates with whether the model explicitly references relevant visual risk indicator objects.

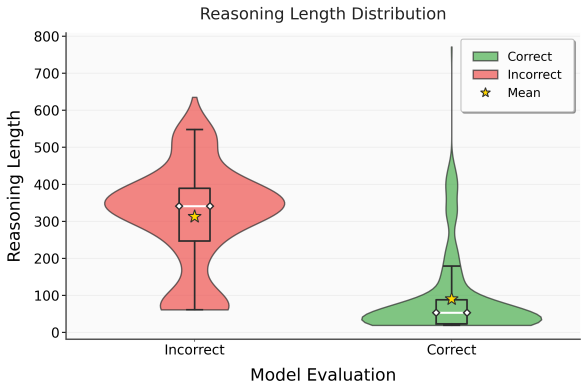


Figure 4: Reasoning Length Distribution. Incorrect predictions (red) are associated with significantly longer and more complex reasoning traces compared to correct ones (green), suggesting that model's overthinking or circular deliberation often leads to performance degradation.

#### 4.3.2 Self-Correcting Reasoning Analysis.

Models frequently engage in self-correction during risk prediction, as reflected by elevated confusion counts. However, self-correction consistently reduces accuracy by 15–26 percentage points across all experimental conditions. Qualitative inspection reveals two dominant patterns: (1) rethinking fails to correct an initially wrong perception, and (2) rethinking overrides an initially correct judgment with speculative reasoning. This contrasts with prior work where deliberation improves performance, suggesting that uncertainty in early risk prediction stems from insufficient visual evidence rather than inadequate reasoning.

| Video Type | Car Crash | Protest |
|---|---|---|
| Basic | 0.67 ± 0.2 | 0.59 ± 0.2 |
| Shuffled | 0.68 ± 0.1 | 0.57 ± 0.3 |
| Swap Halves | 0.65 ± 0.3 | 0.58 ± 0.3 |
| Reversed | 0.68 ± 0.3 | 0.58 ± 0.3 |

Table 4: Model performance across temporal perturbations. Model performance is nearly identical across original and modified temporal sequences, suggesting that VLMs rely more on static frame content than genuine temporal reasoning for risk assessment.

**Takeaway: Overthinking degrades performance in risk prediction.** Unlike traditional reasoning tasks, additional deliberation consistently lowers accuracy.

#### 4.3.3 Temporal Reasoning Analysis

Table 4 illustrates the performance for each augmentation. Across both protest and car crash datasets, reversing or shuffling video frames results in only minor performance degradation (1–3%). For example, protest scenarios show a ∼2% difference between the basic and shuffled settings, while car crash performance remains largely unchanged. This suggests that VLM predictions are driven by the presence of salient frames rather than by temporal progression or causal ordering, revealing a lack of robust temporal reasoning.

**Takeaway: Current VLMs lack genuine temporal reasoning.** Performance differences between original, shuffled, and reversed video inputs are small, indicating that models rely primarily on static frame content rather than temporal order.

#### 4.3.4 Effect of Risk Signal Length

As the temporal gap between the risk signal and incident increases from 1 to 20 seconds, model accuracy systematically degrades, typically dropping from 45–50% to 36–44%. This demonstrates VLMs struggle to anticipate future risky events when predictive cues are subtle and temporally remote. Among evaluated models, MiMo-RL maintains stable performance across temporal distances, indicating improved robustness to early signals.

**Takeaway: Longer temporal distance weakens risk prediction.** Model accuracy declines as the risk signal becomes temporally distant from the incident.

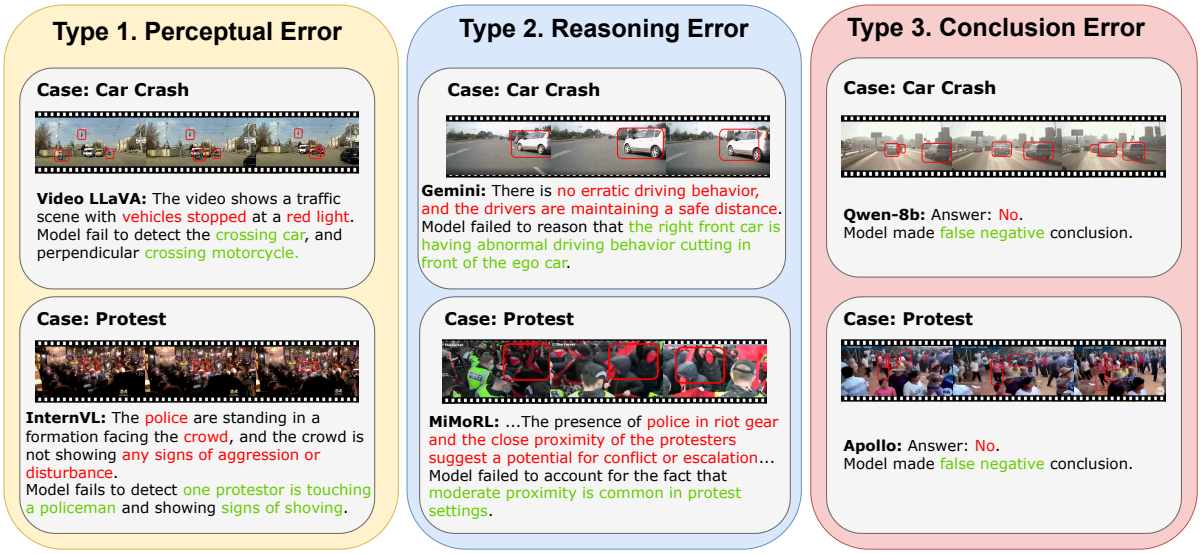


Figure 5: Taxonomy of VLM failure modes in situational risk assessment, categorized by perceptual, reasoning, and conclusion errors. Examples include perceptual errors (Type 1) miss critical cues like aggressive shoving, reasoning errors (Type 2) like failing to interpret normal vs. abnormal behaviors, and incorrect answers (Type 3).

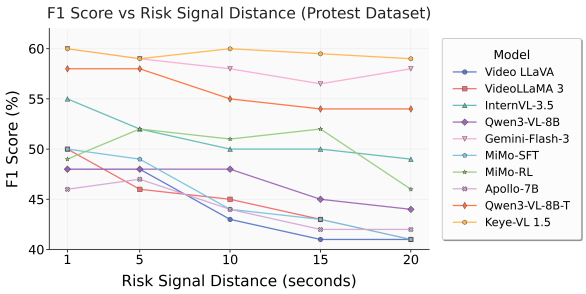


Figure 6: Model Accuracy vs. Risk Signal Distance for Protest Dataset. F1 scores generally decline as the temporal distance between the risk cue and the incident increases from 1 to 20 seconds, suggesting that VLMs struggle to maintain predictive accuracy when cues are temporally far.

### 4.4 Qualitative Case Study: Error Analysis of Gemini on Car-Crash Videos

To analyze VLM failures, we conducted a qualitative study of 120 incorrect predictions of best-performing model Gemini. By manually inspecting reasoning traces, we derived an error taxonomy from recurring patterns (Figure 5):

- *Perceptual Errors* (38.33%): Missing or misinterpreting critical visual cues, such as aggressive gestures or unexpected vehicle entries.
- *Reasoning Errors* (92.50%): Failing to integrate detected cues, leading to causal misattribution or incomplete inference.
- *Conclusion Errors*: Premature decisions, overconfidence, or internal contradictions.

These non-mutually exclusive categories indicate that while perceptual misalignment is common, reasoning failures where models apply inappropriate causal explanations to detected elements, remain the primary bottleneck for state-of-the-art VLMs.

## 5 Conclusion

We introduce a video benchmark RiskCueBench designed to explicitly evaluate models' ability to predict real-life risk events, supported by a carefully curated, high quality dataset constructed through model disagreement and LLM based filtering. Beyond the benchmark itself, we develop a principled framework for identifying challenging risk scenarios and define interpretable evaluation metrics that capture temporal sensitivity, visual grounding, and reasoning behavior. Our evaluation shows that existing VLMs substantially lag behind human performance, relying on static visual cues for temporal reasoning and performing similarly even under significant temporal perturbations. Moreover, models struggle to justify predicted risks through accurate visual grounding, and unlike in domains such as mathematics or coding, more elaborate reasoning traces frequently lead to degraded performance rather than improvements. We hope that our benchmark, data curation pipeline, and accompanying evaluation code provide a strong foundation for evaluating VLMs in situational risk assessment.

## Limitations

Our study focuses on two categories of risk videos, Protest and Car Crash. While these domains capture important safety scenarios, they do not comprehensively cover all realistic risk situations encountered in video understanding. In particular, RiskCueBench does not include other non physical risk types such as misinformation or AI generated content, which remain important directions for future study. In addition, parts of our evaluation pipeline rely on results extracted from an LLM based judge. Although this enables scalable and interpretable analysis, the judge may occasionally produce incorrect assessments, which could impact some of the reported metrics and findings. Finally, the current annotation framework requires manual labeling of risk signal clips, including identifying when risk becomes relevant. This process is difficult to scale to larger datasets, and future work may explore more automated approaches for localizing risk signals.

## Ethical Considerations

This research adheres to the ACL Code of Ethics. The data collection and annotation process was conducted under the following ethical considerations:

**Data Collection and Privacy** The video data used in this benchmark were sourced from public platforms (YouTube) following the platform's Terms of Service. We have ensured that no private or personally identifiable information is explicitly highlighted or utilized beyond the scope of situational risk assessment.

**Annotation Process** The human annotations for the risk signal clips and reasoning traces were performed entirely by the authors of this paper. As the researchers themselves conducted the labeling, there are no concerns regarding the recruitment of vulnerable populations or the adequacy of participant compensation. This internal annotation process ensured high-quality control and a deep alignment with the specialized domain expertise required for situational risk reasoning.

**Funding and Support** This work was supported by a funded project.

**Potential Misuse** While RiskCueBench aims to improve public safety through early risk anticipation, we acknowledge that risk prediction models could potentially be used for unauthorized surveillance. We advocate for the use of this technology strictly within the bounds of legal and ethical safety-critical frameworks.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2682–2690.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems*, volume 37, pages 28662–28673. Curran Associates, Inc.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2025. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*.

Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. 2025. Video-as-answer: Predict and generate next video event with joint-grpo. *arXiv preprint arXiv:2511.16669*.

Mishal Fatima, Muhammad Umar Karim Khan, and Chong-Min Kyung. 2021. Global feature aggregation for accident anticipation. In *2020 25th International conference on pattern recognition (ICPR)*, pages 2809–2816. IEEE.

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. *arXiv preprint arXiv:2506.06218*.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025a. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8450–8460.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025b. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Fizza Hussain, Yasir Ali, Yuefeng Li, and Md Mazharul Haque. 2024. A bi-level framework for real-time crash risk forecasting using artificial intelligence-based video analytics. *Scientific Reports*, 14(1):4121.

Chi-Hsi Kung, Chieh-Chi Yang, Pang-Yuan Pao, Shu-Wei Lu, Pin-Lun Chen, Hsin-Cheng Lu, and Yi-Ting Chen. 2024. Riskbench: A scenario-based benchmark for risk identification. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14800–14807. IEEE.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. 2025. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2025. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, and 1 others. 2025. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *arXiv preprint arXiv:2505.02064*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. 2025. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8600–8612.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, and 1 others. 2025. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901.

# A Appendix

### A.1 Example List for Query Construction (Truncated for Readability)

| Year | Location | Specific Cities | Protest Synonym | Example Query |
|---|---|---|---|---|
| 2000 | Palestine | Ramallah, Nablus, Hebron, Gaza City, . . . | civil unrest, protest, rally, sit-in, strike, march, demonstration | 2000 Palestine civil unrest; 2000 Ramallah civil unrest,... |
| 2001 | Philippines | Manila, Quezon City, Makati, . . . | | 2001 Philippines rally; 2001 Manila civil unrest, ... |
| 2003 | Global | New York City, London, Rome, Tokyo, . . . | | 2003 Global civil unrest; 2003 New York City march; ... |
| 2010 | Tunisia | Sidi Bouzid, Tunis, Kasserine, . . . | | 2010 Sidi Bouzid civil unrest;... |
| 2011 | Egypt | Cairo, Alexandria, Suez, . . . | | 2011 Cairo demonstration;... |
| 2014 | Ukraine | Kyiv, Lviv, Kharkiv, . . . | | 2014 Kyiv march;... |
| 2014 | Hong Kong | Hong Kong | | 2014 Hong Kong strike; ... |
| 2016 | USA | Standing Rock | | 2016 Standing Rock rally;... |
| 2020 | United States | Minneapolis, New York City, Los Angeles, . . . | | 2020 Minneapolis civil unrest;... |
| 2022 | Iran | Tehran, Sanandaj, Mashhad, . . . | | 2022 Tehran strike;... |
| 2023 | France | Paris, Lyon, Marseille, . . . | | 2023 Paris demonstration;... |
| 2024 | India | New Delhi, Amritsar, Chandigarh, . . . | | 2024 New Delhi protest;... |
| 2025 | UK (Essex) | Clacton-on-Sea, Colchester, . . . | | 2025 UK (Essex) civil unrest; ... |

## A.2 Visual Quality and Authenticity Scoring Criteria

| No. | Dimension | Criterion | Score 0 | Score 1 | Score 2 |
|---|---|---|---|---|---|
| 1 | Logo Assessment | Video should have no news organization branding or features | Obvious news logos, watermarks, or media branding visible | Ambiguous logos or possible reposts with unclear branding | No branding visible; appears bystander- or CCTV-style |
| 2 | Location Information | Video should contain no location mentioned in text or audio | Explicit location references (city, state, country) | Ambiguous or hypothetical location references | No location information present |
| 3 | Time/Date Information | Video should contain no time or date mentioned in text or audio | Explicit time/date references (e.g., dates, timestamps) | Ambiguous or unrealistic temporal references | No time/date information present |
| 4 | Reporter Presence | Video should include no reporter, anchor, or journalist | Clearly identifiable reporter or journalist present | Possibly media-affiliated speaker or unclear role | No formal speaker; only participants or bystanders |
| 5 | SNS Engagement Overlays | Video should contain no social media overlays or engagement icons | Likes, shares, emojis, or UI elements visible | Minimal or transient overlays | No overlays or engagement elements |
| 6 | Natural Image Quality | Video should show no signs of professional editing | Heavy editing, stylized transitions, montage effects | Some editing artifacts or unclear transitions | Continuous, shaky, or natural camera movement |
| 7 | Natural Temporal Continuity | Video should be a continuous recording | Obvious jump cuts, stitched scenes, or time gaps | Possible continuity, but breaks unclear | One uninterrupted shot with natural temporal flow |
| 8 | Consequence Text | Video should include no embedded text describing consequences | Text describing severity (arrests, crackdowns, deaths) | Minor or unclear commentary text | No consequence text present |
| 9 | Title / Description / Banner Text | Video should have no inflammatory or descriptive banners | Mentions violence, arrests, or protest names | Vague or unclear text | Only factual metadata (e.g., place, date) |
| 10 | Subtitle Text | Video should contain no speech transcription subtitles | Subtitles clearly present | Partial or unclear subtitle presence | No subtitles visible |
| 11 | Camera Perspective | Video should appear from participant, bystander, or CCTV view | Professional media vantage point | Drone or distant zoomed footage | Ground-level, handheld, immersed, or surveillance view |
| 12 | Post-Production Effects | Video should have no post-processing effects | Filters, slow motion, stabilization effects visible | Minor or uncertain processing | Completely raw, unfiltered footage |

### A.3 Examples of Risk and Non-Risk Annotations Across Protest and Car Crash Scenarios

| **Label** | **Protest (Risk)** | **Protest (Non-Risk)** | **Car Crash (Risk)** | **Car Crash (Non-Risk)** |
|---|---|---|---|---|
| Risk Signal Start | 00:03:00 | 00:04:00 | 00:05:12 | 00:39:38 |
| Risk Signal End | 00:12:58 | 00:08:55. | 00:10:11. | 00:44:00 |
| Risk Visual Indicator | Throwing gestures, aggressive body postures, raised batons, riot shields. | Singing, dancing, sign-holding, and stationary police lines. | Crossing vehicle, red light, arrow on the ground indicating traffic flow. | Lane mark, steady speed, stop sign |
| Risk Signal Description | First police vehicles appear and protestors intend to throw objects; then officers raise baton toward protesters. | Protesters hold signs still, some sing and dance while police observe from a distance without engagement. | First a car enters the intersection against traffic flow; then ego continues when traffic light is red. | Vehicles slow appropriately at stop sign and proceed steadily according to traffic signals within lane mark correctly. |
| Accident Start Frame | 00:12:59 | / | 00:10:12 | / |
| Accident End Frame | 00:15:00 | / | 00:13:00 | / |
| Accident Description | A smoke device is deployed into the crowd, causing protesters to disperse. | / | Following a collision, the vehicle spins and stops near the roadside. | / |
| Risk Label | Yes | No | Yes | No |

## A.4 Prompt Used for Quality Filtering of 12 Selection Criteria.

**Selection Criteria**

You are a video analysis expert tasked with evaluating YouTube videos based on 12 specific criteria. Analyze the provided video and assign scores for each category based on the detailed scoring rubric below.

**SCORING SYSTEM**

For each category, assign exactly one score:
Score 0 = Fails to meet criterion
Score 1 = Partially meets or unclear
Score 2 = Fully meets criterion

**EVALUATION CRITERIA**

**1. Logo Assessment**
Criterion: Video should have no news organization branding or features
Score 0: Obvious news marks, logos, or clear media branding visible
Score 1: Ambiguous news marks/logos or possible reposts with unclear branding
Score 2: No branding visible, appears to be bystander-style or CCTV-style capture

**2. Location Information**
Criterion: Video should have no location mentioned in text or audio
Score 0: Obvious location references in text or audio (city, state, country)
Score 1: Ambiguous location references (hypothetical/unrealistic places)
Score 2: No location information present in text or audio

**3. Time/Date Information**
Criterion: Video should have no time/date mentioned in text or audio
Score 0: Obvious time/date references (e.g., "May 25th, 2025", "13:00pm")
Score 1: Ambiguous temporal references (e.g., "a thousand years ago/later")
Score 2: No time/date information present in text or audio

**4. Reporter Presence**
Criterion: Video should have no reporter, anchor, or journalist present
Score 0: Reporter, journalist, or anchor clearly present and identifiable
Score 1: Possibly media-affiliated speaker or unclear professional presence
Score 2: No formal speaker present, only participants/bystanders

**5. SNS Engagement Overlays**
Criterion: Video should have no social media overlays or engagement icons
Score 0: Social media overlays/pop-ups clearly visible (likes, shares, emojis)
Score 1: Minimal or transient overlays present
Score 2: No overlays or engagement elements visible

**6. Natural Image Quality**
Criterion: Video should show no signs of professional editing
Score 0: Highly edited with stylized transitions, cuts, or montage effects
Score 1: Some editing artifacts present or unclear transitions
Score 2: Continuous, shaky, or natural camera movement with no professional editing

**7. Natural Temporal Continuity**
Criterion: Video should be continuous recording without interruptions
Score 0: Obvious jump cuts, stitched scenes, or time gaps
Score 1: Possibly continuous but breaks/transitions unclear
Score 2: One uninterrupted shot with natural temporal flow

**8. Consequence Text**
Criterion: Video should have no embedded text about protest consequences
Score 0: Text describing consequences/severity (arrests, crackdowns, deaths)
Score 1: Unclear or minor commentary text present
Score 2: No consequence text; only basic metadata like location/time allowed

**9. Title/Description/Banner Text**
Criterion: Video should have no inflammatory title/description banners
Score 0: Text mentioning violence, specific protest names, arrests
Score 1: Title/description is unclear or vague
Score 2: Only factual metadata like place/date present

**10. Subtitle Text**
Criterion: Video should have no speech transcription subtitles
Score 0: Subtitles clearly present showing speech transcriptions
Score 1: Some subtitle presence but hard to distinguish
Score 2: No subtitles visible at all

**11. Camera Perspective**
Criterion: Video should appear to be from participant, bystander, or CCTV perspective
Score 0: Clearly taken from media zone or professional media position
Score 1: Drone footage or distant zoomed footage
Score 2: Ground-level, hand-held, immersed in crowd, or surveillance camera

**12. Post-Production Effects**
Criterion: Video should have no post-processing effects applied
Score 0: Clear post-processing (filters, slow motion, stabilization effects)
Score 1: Some smoothing effects or unknown processing present
Score 2: Completely raw, unfiltered footage

**OUTPUT FORMAT**

Provide your analysis as a JSON array containing exactly 12 objects, one for each category. Each object must include:
category_number: integer (1-12)
category_name: string (exact name from criteria above)
score: integer (0, 1, or 2)

**ANALYSIS INSTRUCTIONS**

1. Watch the entire video carefully
2. Listen to all audio content
3. Examine all visible text and overlays
4. Assess video quality and editing characteristics
5. Score each category independently
6. If uncertain between two scores, choose the lower score
7. Ensure all 12 categories are evaluated and included in response

## A.5 Prompt Used for Reasoning Chain Evaluation

**Reasoning Trace Evaluation Prompt**

You are analyzing a reasoning chain from Video-Language Models to evaluate its quality and characteristics. Please analyze the following reasoning chain and provide the requested metrics.
GROUND TRUTH: {ground_truth}
MODEL PREDICTION: {model_pred}
REASONING CHAIN: {reasoning}

**Please analyze this reasoning chain and provide:**
**1. Confusion Count:** Count the number of times the reasoning shows confusion, uncertainty, or self-correction. Look for phrases like:

- "wait..."
- "no let me think again"
- "actually..."
- "hold on..."
- "let me reconsider"
- "I'm confused"
- "that doesn't seem right"
- Similar expressions of uncertainty or backtracking

**2. Decision Items:** Extract ALL items, objects, or things that the model specifically mentions from the video as part of its reasoning process for reaching its conclusion. This includes:

- Physical objects mentioned in the reasoning
- Items that influenced the decision (whether risky or safe)
- Specific things the model identified or considered
- Objects that were part of the analysis
- Any concrete items/things mentioned that contributed to the final decision
- Both safe items (for non-risk scenarios) and dangerous items (for risk scenarios)

**OUTPUT FORMAT**
Provide your analysis in the requested JSON format.

## A.6 Risk Signal Temporal Distribution

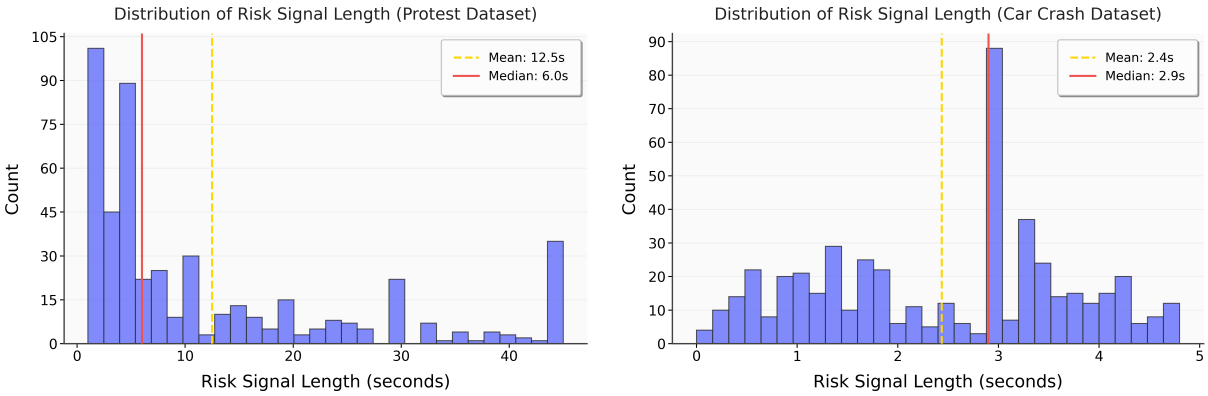


Figure A.6: Distribution of risk signal lengths for protest (left) and car crash (right) scenarios.

The histograms above illustrate the distribution of risk signal lengths across the two primary datasets used in this study. A "risk signal" is defined as the temporal window during which a model can identify potential danger before an actual incident occurs.

**Protest Dataset.** The data shows a significantly broader distribution of signal lengths. The mean duration is 12.5s, while the median is 6.0s. This suggests that risk in protest scenarios often involves a gradual escalation, with some signals extending beyond 40 seconds.

**Car Crash Dataset.** In contrast, the risk signals for vehicular accidents are highly concentrated and brief. The mean length is 2.4s and the median is 2.9s, indicating that the critical decision-making window for safety systems in these scenarios is very narrow.

### A.7 Representative Failure Cases

We present representative failure cases from Gemini to illustrate how perceptual and reasoning errors manifest in practice. These examples are selected from the car-crash domain and are intended to highlight common patterns rather than exhaustively enumerate all error types.

**Perceptual Errors**

**Case 1: Snow-caused loss of control misperceived as a normal intersection.**

*Annotation.* "First, a black car on the opposing lane slips on snow, then it drifts out of control."

*Model output.* Gemini describes a routine urban intersection with functioning traffic lights and a yellow taxi. *Diagnosis.* The model fails to indicate the hazardous road condition of snow and the resulting loss of traction, instead substituting an unrelated scene with different objects and traffic structure. The core visual indicators are entirely absent from the model's perception.

**Case 2: Wrong-way driving replaced by an empty parking garage.**

*Annotation.* "First, the car in front drives in the opposing direction, then it continues into the driver's lane."

*Model output.* The model describes an empty underground parking garage with no moving vehicles.

*Diagnosis.* The active roadway, opposing traffic flow, and dynamic agents are replaced by a static and unrelated environment.

**Reasoning Errors**

**Case 3: "No collision yet" interpreted as no risk.**

*Annotation.* "The white car in front brakes suddenly while the ego vehicle is still approaching."

*Model output.* Gemini concludes that no risk is present because traffic lights are functioning and no crash is visible.

*Diagnosis.* Although the sudden braking event is acknowledged, the model incorrectly equates the absence of an observed collision with safety, ignoring the temporal consequence of risk.

**Case 4: Normal lane change framed as risky via hypothetical hazards.**

*Annotation.* "Ego vehicle changes to the right lane at normal speed."

*Model output.* The model predicts risk by invoking slippery roads or poor weather conditions not observed.

*Diagnosis.* The model introduces speculative weather hazards instead of reasoning based on the described driving behavior, applying a generic safety heuristic without causal grounding of visual indicators in the video.

**Conclusion Errors**

**Case 5: Collision labeled as no risk.**

*Annotation.* "First, a black car in front signals left and turns, at the same time it does not yield to the ego vehicle."

*Model output.* Gemini predicts no risk.

*Diagnosis.* The cars collide after the signal, while the model assigns a negative risk label.

**Case 6: Benign stop at red light labeled as risky.**

*Annotation.* "During the red light, vehicles are waiting, and vehicles going straight from the right intersection are moving at normal speed the whole time."

*Model output.* Gemini predicts a risk.

*Diagnosis.* The model incorrectly flags a compliant and stationary driving scenario as risky.