HOMEACE: A COMPREHENSIVE EVALUATION BENCHMARK FOR HOME ANOMALY VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding safety risks in household videos is essential for intelligent systems that support risk prevention and timely response. Achieving this requires a model that can as a *forecaster* of the household: perceiving subtle risk precursors, inferring causal chains of danger escalation, localizing the temporal window for intervention, and formulating actionable mitigation strategies. However, existing evaluations of multimodal large language models (MLLMs) focus predominantly on general video understanding, with no systematic assessment of the broad range of capabilities required in home settings. To address this gap, we introduce **HomeACE**, a large-scale **Home A**nomaly **C**omprehensive **E**valuation dataset comprising more than 14K household videos, including nearly 3K real-world recordings and over 11K generated videos that span normal, risk-only (an emerging risk), and abnormal (a realized anomaly) situations across diverse home scenes. Based on this dataset, we design a systematic evaluation protocol that assesses models along four complementary dimensions: Perception, Cognition, Temporal, and Planning. Extensive experiments reveal substantial limitations in current MLLMs: limited anomaly perception, weak causal reasoning, imprecise temporal localization, and mitigation plans that are often insufficiently grounded in the observed hazards. HomeACE provides a standardized platform for diagnosing these deficiencies and advancing home anomaly understanding toward the vision of prevention-oriented intelligent home systems.

1 INTRODUCTION

Homes are the environments in which people spend the majority of their time, yet they contain a spectrum of safety risks, including elderly falls, child injuries, fires, electrical shocks, and accidental poisoning, leading to numerous incidents reported every year. As home cameras become increasingly prevalent and companion robots are poised to enter households in the near future, the role of in-home vision systems must evolve from passive recording to active understanding and prediction of safety risks. This transition represents a societal need and a crucial step in building intelligent home-safety systems. A long-term vision is to develop systems that integrates scene understanding, reasoning about possible consequences, and planning of interventions. Toward this vision, we focus on the video-grounded risk-understanding capabilities needed for household safety.

Multimodal large language models (MLLMs) (Liu et al., 2023; Wang et al., 2024; Chen et al., 2024), which unify visual understanding with language-based reasoning over long video contexts (Maaz et al., 2024; Li et al., 2024a; Cheng et al., 2024), are promising candidates for supporting household risk understanding. This raises a central question: *do current MLLMs possess the video-grounded perception, reasoning, localization, and planning capabilities required for household safety?* In practice, current MLLMs are predominantly developed and evaluated on general video understanding (Fu et al., 2025; Li et al., 2024b; Mangalam et al., 2023; Wu et al., 2024), visual question answering (Xiao et al., 2021; Yu et al., 2019), and temporal grounding (Gao et al., 2017; Krishna et al., 2017), and whether they possess the four core capabilities required for home-safety understanding, namely *risk perception*, *causal cognition*, *risk temporal localization*, and *mitigation planning*, has not yet been systematically studied. This gap stems from two key deficiencies: (1) **On the data side**, there is no dedicated video dataset for risk behaviors in home environments. Existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

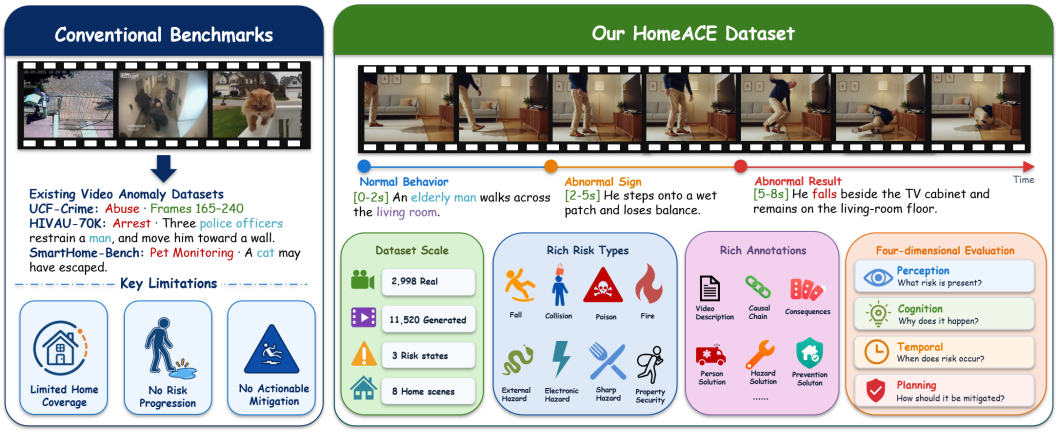

Figure 1: Comparison between conventional video anomaly detection benchmarks and HomeACE. Existing benchmarks provide limited coverage of home environments, cannot model risk progression, and lack annotations for risk-process causal chains, actionable mitigation measures, etc.

anomaly datasets are confined to public surveillance (Sultani et al., 2018b; Luo et al., 2017; Lu et al., 2013), failing to cover the diversity of household scenes and the stochastic nature of domestic risks. (2) **On the evaluation side**, there is no comprehensive evaluation protocol designed around real home-safety needs. Existing benchmarks are built around generic scenes and routine daily activities (Das et al., 2019; Sigurdsson et al., 2016), unable to jointly assess perception, cognition, temporal, and planning. Together, this dual deficiency constitutes a key bottleneck preventing the large-scale deployment of intelligent home-safety systems.

To address the limitations of existing data and evaluation, we construct **HomeACE**, a large-scale benchmark for home anomaly and risk understanding. HomeACE organizes real-world and generated videos under a unified risk definition. In terms of risk progression, the dataset distinguishes normal activities, risk-only precursors that have not yet caused actual consequences, and realized abnormal events. In terms of data diversity, the risks cover diverse home scenes, subjects, and risk categories. Overall, HomeACE contains more than 14K household videos, including nearly 3K real-world recordings and over 11K videos generated through a structured generation pipeline, covering diverse home scenes, subjects, and risk types.

Based on this dataset, we further develop a systematic evaluation protocol for the household safety decision-making chain. Rather than reducing household risk understanding to anomaly classification, the protocol provides fine-grained evaluation along the complete process of "identifying risk evidence, explaining risk progression, localizing intervention windows, and planning mitigation measures," covering four complementary dimensions: Perception, Cognition, Temporal, and Planning. Under this protocol, we conduct a comprehensive evaluation of representative open-source and proprietary multimodal large language models (MLLMs), revealing substantial gaps in their ability to recognize emerging risks, explain causal relations, and localize hazards in household videos. The fixed inference prompts and dimension-specific judging criteria are provided in Appendix A.4.

In summary, the main contributions are threefold:

1. **Large-scale home-scene dataset.** We construct a dataset of more than 14K household videos combining real-world and generated scenes, covering normal, abnormal, and risk situations across diverse home environments, addressing the limitations of existing resources in scale, realism, and scenario coverage.

2. **Comprehensive four-dimensional evaluation protocol.** We design a comprehensive evaluation protocol spanning Perception, Cognition, Temporal, and Planning that enables fine-grained, dimension-specific diagnosis of MLLMs' home anomaly understanding capabilities, moving beyond the limitations of conventional general video understanding benchmarks. Under this protocol, we evaluate a wide range of representative open-source and proprietary MLLMs, providing a systematic comparison of their capabilities and remaining performance gaps in home anomaly understanding.

3. **Diagnostic findings.** Our experiments reveal several structural limitations of current MLLMs. First, they recognize realized incidents far better than emerging risks. Second, they exhibit strong factual recognition but weak causal reasoning. Third, they suffer from imprecise temporal localization that propagates to other capabilities. These findings identify key bottlenecks and offer concrete, actionable guidance for developing next-generation home-safety models.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) combine visual representations with language reasoning for open-ended understanding and dialogue. GPT-4o (Hurst et al., 2024) and Gemini (Team et al., 2024) exemplify this direction. For video, the central challenges are visual recognition, long-range temporal modeling, and language-based reasoning and planning.

Existing methods address these challenges along three complementary dimensions. *Video perception* methods improve entity and behavior recognition through temporal connectors, multimodal positional encoding, video sampling, and token resampling. VideoLLaMA2 (Cheng et al., 2024) and Qwen2-VL (Wang et al., 2024) are representative, while Apollo explores efficient spatiotemporal modeling (Zohar et al., 2025). *Long-range temporal understanding* methods model relationships across extended contexts. LongVA expands the visual context (Zhang et al., 2024b). Video-XL (Shu et al., 2025) and Oryx (Liu et al., 2025) investigate visual-token compression and efficient long-video processing. *Reasoning and planning* methods support instruction following and task-oriented reasoning. Video-ChatGPT (Maaz et al., 2024), LLaVA-Video (Zhang et al., 2024c), and LLaVA-OneVision (Li et al., 2024a) improve open-ended video dialogue. VideoChat-R1 (Li et al., 2025), Video-R1 (Feng et al., 2026), and Video-KTR (Wang et al., 2026) explore instruction and reinforcement fine-tuning. Despite these advances, our benchmark reveals substantial gaps in existing MLLMs' risk perception, causal reasoning, temporal localization, and solution planning.

2.2 Video Understanding Datasets and Benchmarks

Dataset and benchmark research relevant to household risk understanding spans three related areas. (1) *General video understanding.* Video-MME (Fu et al., 2025), VideoVista (Li et al., 2024c), and MVBench (Li et al., 2024b) assess video question answering, perception, and reasoning, while MLVU (Zhou et al., 2025) and LongVideoBench (Wu et al., 2024) focus on longer video contexts. (2) *Video anomaly analysis.* UCF-Crime (Sultani et al., 2018b) and XD-Violence (Wu et al., 2020b) support anomaly or violence detection, while ShanghaiTech (Luo et al., 2017), UBnormal (Acsintoae et al., 2022b), and MSAD (Zhu et al., 2024b) evaluate anomaly detection or localization across different scenes and data sources. HIVAU-70K extends this line toward semantic anomaly understanding with multi-granular descriptions (Zhang et al., 2025b). (3) *Household-scene understanding.* Toyota Smarthome (Das et al., 2019) and YouHome (Pan et al., 2022) cover daily activities; UR Fall Detection (Kwolek & Kepski, 2014) and Multiple Cameras Fall Dataset (Auvinet et al., 2010) target falls. SmartHome-Bench introduces household video descriptions and normality/anomaly reasoning (Zhao et al., 2025a). However, joint evaluation of household risk perception, causal understanding, temporal localization, and mitigation remains limited. HomeACE addresses this gap across risk-only and realized-anomaly scenarios. A detailed comparison with existing datasets is provided in Section 3.5.

3 HomeACE

3.1 Overview

HomeACE combines real-world and generated household videos, with a four-level semantic risk tree guiding generated scenario coverage. Three states—*normal*, *risk-only*, and *abnormal*—distinguish routine activities, visible risk precursors without a realized anomalous outcome, and realized incidents. Structured annotations connect risk evidence, event semantics, temporal intervals, and mitigation responses, supporting the four-dimensional evaluation in Section 4.

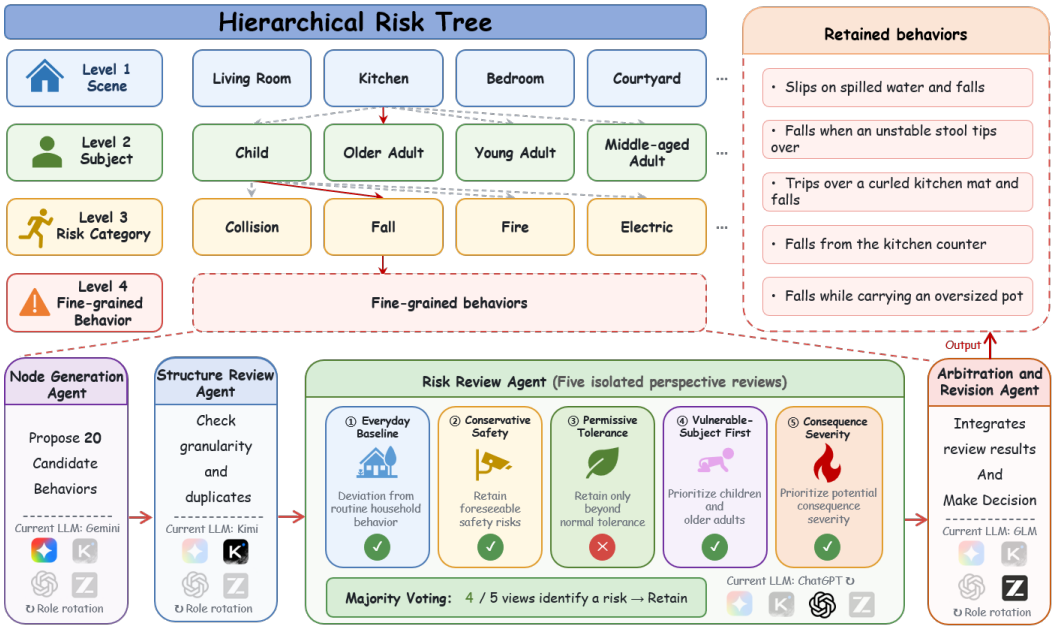

Figure 2: Multi-agent construction of the hierarchical household-risk tree. For each parent configuration, 20 candidate behaviors are generated and reviewed through structural checks, five-perspective risk voting, and final arbitration, resulting in five valid and non-redundant leaf behaviors.

3.2 Data Collection

Real-World Video Collection. Since individual household datasets typically focus on specific activities or incidents, we combine public datasets with online videos to broaden scenario coverage. We select relevant household videos from five datasets: SmartHome-Bench Zhao et al. (2025a), the Fall Detection Dataset Kwolek & Kepski (2014); Auvinet et al. (2010), the YouHome Dataset Pan et al. (2022), and Toyota Smarthome Das et al. (2019). We also collect publicly available videos from Bilibili Bil, Douyin Dou, Kuaishou Kua, Xigua Video Xig, YouTube You, TikTok Tik, and Dailymotion Dai.

Generated Video Collection. To extend the limited coverage of real-world household risk videos, we construct a generated subset. However, model biases and low-variation prompts can limit the diversity and coverage of LLM-generated scenarios (Wang et al., 2025a). To organize generation systematically, we first structure target scenarios with a hierarchical risk tree, then use its root-to-leaf paths to guide video generation. The manually defined schema is *scene* → *subject* → *risk category* → *specific anomalous behavior*; LLMs generate concrete nodes conditioned on their parents.

Risk judgments can vary with safety standards and subject vulnerability. To make these judgments explicit while checking structural validity, a four-role multi-agent framework generates and reviews candidate behaviors for each parent configuration. The node generation agent proposes 20 candidates, which are subsequently processed by structure review, five-perspective risk review, and final arbitration. A candidate passes risk review when at least three of the five isolated perspectives recommend retention, and five valid and non-redundant behaviors are finally retained as leaf nodes. Gemini (Google DeepMind, 2026), Kimi (Moonshot AI, 2024), ChatGPT (OpenAI, 2026a), and GLM (Zhipu AI, 2024) cyclically rotate across the four agent roles to reduce persistent coupling between any single model and a fixed role. Figure 2 illustrates the workflow, with further details provided in Appendix A.1.

For each retained leaf, we combine the leaf description with those of its ancestors to form a complete scenario specification. To distinguish risk precursors from realized incidents, DeepSeek-V4 (Xu et al., 2026) uses this specification to generate three temporally structured prompts: *normal*, *risk-only*, and *abnormal*. These variants represent progressively escalating risk manifestations, corresponding to the stages `[Normal Behavior]`, `[Abnormal Sign]`, and `[Abnormal Result]`. Each state prompt is generated by three models drawn from Sora 2 (OpenAI, 2025),

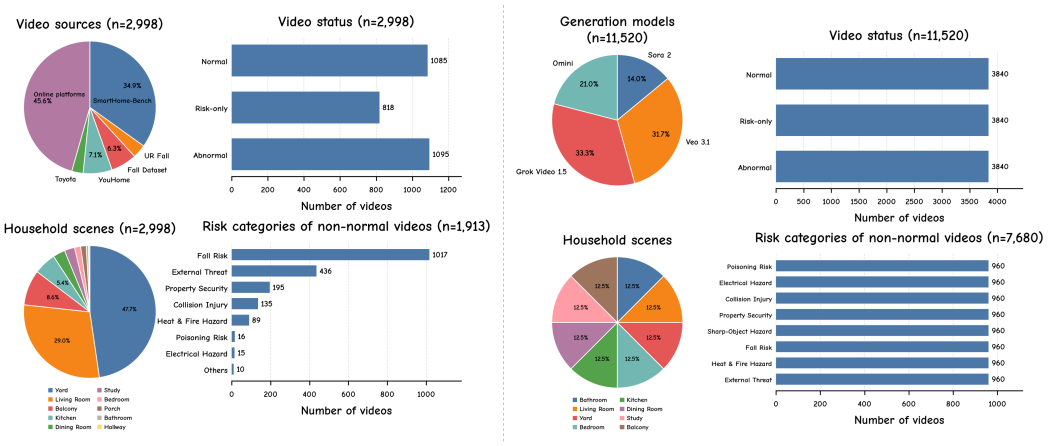

Figure 3: Statistics of the real-world and generated-video subsets. The left panel summarizes the real-world videos, and the right panel summarizes the generated videos.

Veo 3.1 (Google DeepMind, 2025a), Grok Video 1.5 (xAI, 2025), and Omini (Google DeepMind, 2025b), producing nine videos for each retained behavior. Videos rejected during human verification are regenerated until the target number of accepted samples has been reached. A full example of this path-to-prompt expansion is provided in Appendix A.2.

3.3 Data Annotation

The two subsets require different initial annotation procedures: real-world videos need a shared risk schema across heterogeneous sources, whereas generation scripts may differ from the events actually visible in generated videos. We therefore obtain source-specific preliminary annotations and use human inspection to establish the final video-grounded labels. Five open-source video MLLMs (Ye et al., 2025; Yu et al., 2025; Team, 2026; Wang et al., 2025b; Zhang et al., 2025a) assist this process.

Real-World Video Annotation. For real-world videos, each video is first processed by the five models to produce candidate annotations, and human annotators then select, merge, or rewrite the candidates to produce a preliminary structured annotation. The resulting annotation for each video includes the risk category, scene category, event subject, risk event type, temporal localization of the risk interval, and a risk description.

Generated Video Annotation. For generated videos, the risk-tree path supplies the initial scene, subject, and risk category, while the three-stage script provides behavior and temporal annotations. The five models provide prompt–video consistency scores for risk-only and abnormal videos as supporting information; normal videos bypass automatic scoring. Human annotators inspect every generated video, decide whether to retain or discard it, and correct retained annotations and risk intervals against the visible content. Further details are provided in Appendix A.3.

Four computer science Ph.D. students divide the annotation work across both subsets, reviewing and correcting all machine-assisted annotations. Two additional annotators review the results, with disagreements resolved through joint discussion before finalization.

3.4 Dataset Statistics

Figure 3 summarizes the scale, composition, and annotation distribution of HomeACE. The dataset contains 2,998 real-world videos and 11,520 generated videos, for a total of 14,518 household videos. The real-world subset covers diverse household scenes such as yards, living rooms, kitchens, and balconies, with 1,085 normal, 818 risk-only, and 1,095 abnormal videos. The most frequent risk categories among non-normal videos are fall risk, external threats, and stranger theft. The generated subset spans eight household scenes and eight risk categories, with a strictly balanced distribution of 3,840 videos per risk state.

Table 1: Comparison with representative video anomaly datasets. "Home" denotes a household-centered scope; "Temp.", "Progress.", "Reason.", and "Mitig." denote temporal localization, explicit risk progression, textual/causal reasoning, and mitigation annotations, respectively.

Dataset	Venue	#Videos	Source	Home	Taxonomy	Temp.	Progress.	Reason.	Mitig.
UCF-Crime Sultani et al. (2018a)	CVPR'18	1,900	Real	–	13 classes	✓	–	–	–
XD-Violence Wu et al. (2020a)	ECCV'20	4,754	Films + Real	–	6 classes	✓	–	–	–
ShanghaiTech Liu et al. (2018)	ICCV'17	437	Real	–	13 scenes	✓	–	–	–
UBnormal Acsintoae et al. (2022a)	CVPR'22	543	Synthetic	–	22 types	✓	–	–	–
MSAD Zhu et al. (2024a)	NeurIPS'24	720	Real	–	55 types	✓	–	–	–
HIVAU-70K Zhang et al. (2024a)	CVPR'25	5,443	Real	–	3 levels	✓	–	✓	–
SmartHome-Bench Zhao et al. (2025b)	CVPRW'25	1,203	Real	✓	2 levels	–	–	✓	–
HomeACE (Ours)	–	**14,518**	**Real + Syn.**	✓	**4 levels**	✓	✓	✓	✓

3.5 Dataset Comparison

Table 1 compares datasets along household scope, taxonomy, temporal localization, risk progression, reasoning, and mitigation annotations. HomeACE combines a household focus with separation of risk-only and realized anomalies, linking risk sources, actions, and consequences to temporal intervals and mitigation responses. This coverage supports diagnosis of whether incident description is accompanied by precursor recognition, causal understanding, and appropriate responses.

4 Evaluation Protocol

Based on this dataset and its annotations, we further develop a systematic evaluation protocol. Following the complete decision-making chain involved in handling household safety events, the protocol decomposes model capabilities into four interconnected stages: *identifying risk-related evidence*, *explaining risk evolution*, *localizing risk timing*, and *planning risk mitigation measures*. These stages correspond to four evaluation dimensions: Perception, Cognition, Temporal, and Planning. This decomposition covers the complete process from risk detection to action planning while distinguishing capability differences across different stages. To avoid masking specific capability weaknesses with a single overall score, we report the results for each dimension and its corresponding submetrics separately. The following sections introduce these four dimensions and their respective evaluation submetrics in detail.

Perception measures whether the model recognizes risk-related evidence, including risk state (**Sta**; *normal*, *risk-only*, or *abnormal*), risk type (**Typ**; e.g., fall risk or external threat), risk sources (**Src**), abnormal actions (**Act**), and affected objects (**Obj**). Separating these fields distinguishes a correct overall risk judgment from identification of the specific entities and behaviors involved.

Cognition measures whether the model explains the risk process through risk description (**Desc**), consequence understanding (**Csq**), factual-boundary consistency (**Fact**), and causal understanding (**Caus**). Whereas Perception identifies individual evidence fields, Cognition examines how they are connected in an event account. Description and causal understanding are assessed separately because recounting an event does not establish its source–action–consequence relation. Fact distinguishes observed outcomes from potential consequences, which is particularly relevant when a risk-only video contains no realized anomalous outcome.

Temporal measures when a risk or its precursor is visible, using intersection over union (**IoU**), start-time accuracy (**St**), end-time accuracy (**End**), and duration accuracy (**Dur**). A correct semantic account does not establish precise localization; interval overlap and boundary-related measures provide complementary descriptions of temporal performance.

Planning measures the appropriateness of textual mitigation advice for the affected person (**Per**), the hazard source (**Haz**), and overall prevention (**Ovr**). These targets distinguish immediate protection, control of the current hazard, and prevention of recurrence. Unlike consequence understanding, which concerns what may happen or has happened, Planning concerns what should be done in response. The **AVG** column denotes the mean of the submetrics within the corresponding dimension.

The evaluation protocol differs by risk state. For abnormal and risk-only videos, we evaluate all four dimensions of Perception, Cognition, Temporal, and Planning. Scoring follows a hybrid strategy:

(1) risk status and risk type are first normalized to the canonical label space and then scored by exact match against the ground truth; (2) temporal localization uses interval IoU together with normalized start-, end-time, and duration errors; and (3) all other semantic outputs, including risk descriptions, causal chains, mitigation plans, and fine-grained recognition fields, are scored by Qwen3-8B (Yang et al., 2025) acting as a judge model under fixed, dimension-specific criteria (Appendix A.4). The judge assigns each semantic output a score from 0 to 5, which is normalized to a percentage by multiplying by 20. We report both overall performance and fine-grained results across the four evaluation dimensions. For ground-truth fields not covered by the preliminary annotations, GPT-5.5 (OpenAI, 2026a) generates candidate annotations conditioned on the existing fields.

5 EXPERIMENTS

5.1 SETTINGS

To comprehensively evaluate the capabilities of current multimodal large language models (MLLMs) in home anomaly understanding, we conduct extensive experiments on the real-world and generated-video subsets of HomeACE with ten representative models. The evaluated models include three proprietary MLLMs (Claude-Opus-5 (Anthropic, 2026), Gemini-3.1-Pro (Google DeepMind, 2026), and GPT-5.6-Terra (OpenAI, 2026b)) and seven open-source MLLMs (Qwen3.5-9B (Team, 2026), InternVL3.5-8B (Wang et al., 2025b), Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Tarsier2-7B (Yuan et al., 2025), VideoLLaMA2.1-7B (Cheng et al., 2024), VideoLLaMA3-7B (Zhang et al., 2025a), and Qwen3.8-27B (Qwen Team, 2026)). For each video, we uniformly sample 32 frames as the visual input, except for VideoLLaMA2.1-7B, for which we sample 16 frames, and apply the evaluation protocol described above. We report the results averaged over the real-world and generated-video subsets, together with fine-grained analyses of the four evaluation dimensions.

5.2 MAIN RESULTS

Table 2 reports the performance of all ten models on abnormal and risk-only videos, averaged over the real-world and generated subsets; subset-specific results are reported in Appendix A.6, and normal-video results are deferred to the appendix. Proprietary models lead, while the strongest open-source models remain competitive. Performance differences across dimensions appear to reflect visual grounding more than nominal task difficulty: Planning is relatively strong and varies little across model tiers, whereas Temporal localization is the main bottleneck. Cognition combines near-ceiling factual-boundary accuracy with weak causal understanding. Overall, models describe incidents and propose generic responses better than they proactively determine *where*, *when*, and *why* risks emerge.

The following four points provide a detailed analysis of the four evaluation dimensions.

Perception. *Models recognize realized incidents better than emerging risks.* The gap is driven primarily by the risk-status sub-metric (Sta), which drops sharply from abnormal to risk-only videos for both open-source and proprietary models, as illustrated by Qwen3.5-9B (84.0 to 13.0) and Gemini-3.1-Pro (82.1 to 25.5) (Table 2). Risk-only videos contain abnormal signs but lack the salient consequence that models rely on as a classification anchor, so without a visible outcome they tend to fall back to the "normal" prior. This suggests that current MLLMs can recognize what has already happened but struggle to anticipate what is about to go wrong, which is precisely the capability most needed for home safety.

Cognition. *Factual recognition does not imply causal understanding, especially when the risk has not yet produced a visible outcome* (Figure 4). A pronounced dissociation emerges within the Cognition dimension: models reach near-ceiling factual-boundary consistency (Fact) yet remain weak on causal understanding (Caus). Even GPT-5.6-Terra shows a substantial gap between factual and causal understanding on abnormal videos (Fact 97.4 versus Caus 50.6) (Table 2). This pattern shows that a high Fact score, which reflects telling an observed consequence from a merely possible one, does not imply a usable risk model: the low Caus score reveals that models cannot chain the visible cues into an ordered source–action–consequence relation. The deficit is worst exactly where anticipation matters most, since causal understanding remains weaker on risk-only videos. This is consistent with the intuitive difficulty of inferring an emerging consequence from precursor cues before the outcome has materialized.

Table 2: Results on the HomeACE benchmark, averaged over the real-world and generated subsets. All values are percentages; "AVG" denotes the within-dimension mean.

Model	Score	Perception						Cognition					Temporal					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	IoU	St	End	Dur	AVG	Per	Haz	Ovr
Abnormal																					
GPT-5.6-Terra	**60.8**	**69.0**	69.2	**76.6**	**58.9**	**70.2**	**70.1**	**74.0**	**81.1**	**66.9**	**97.4**	**50.6**	25.1	20.0	39.8	20.9	19.7	**74.9**	**74.3**	**77.1**	**73.5**
Claude-Opus-5	58.3	61.5	64.2	68.6	51.0	58.4	65.4	68.5	75.8	58.8	95.4	43.6	33.4	27.6	**49.3**	29.4	28.2	69.9	68.6	70.7	71.2
Gemini-3.1-Pro	57.0	65.5	**82.1**	72.2	50.8	58.1	64.0	64.2	71.9	51.2	94.2	39.9	**35.6**	**28.7**	48.2	**32.4**	**33.2**	62.8	60.0	64.4	64.0
Qwen3.8-27B	**57.2**	64.0	79.9	68.4	**50.8**	**59.7**	**61.3**	**66.9**	**75.6**	**53.9**	**95.8**	**42.3**	30.0	23.9	45.2	22.6	28.2	**67.9**	**65.9**	**69.2**	**68.5**
Qwen3.5-9B	55.8	**64.2**	**84.0**	70.1	48.2	58.4	60.4	64.5	73.3	51.4	94.0	39.4	30.0	23.0	46.4	21.9	28.7	64.2	62.8	66.7	63.3
InternVL3.5-8B	53.0	51.5	36.7	**72.0**	41.9	50.1	57.1	59.0	62.8	43.9	93.3	36.2	**40.5**	**29.0**	**53.9**	**30.7**	**48.7**	60.8	57.7	63.8	60.8
Qwen2.5-VL-7B-Instruct	36.1	27.5	22.4	33.0	22.8	25.1	34.1	45.4	43.4	25.9	92.0	20.3	27.4	24.5	52.1	16.9	16.2	44.2	42.5	45.0	45.1
Tarsier2-7B	34.5	28.2	46.0	23.1	18.0	26.5	27.7	43.5	45.4	22.7	85.4	20.2	28.3	23.2	47.1	17.3	25.5	37.9	37.7	37.9	38.0
VideoLLaMA3-7B	29.4	18.4	30.3	23.5	8.4	15.4	14.1	37.8	34.5	15.2	88.8	12.6	24.3	21.4	31.7	20.9	23.1	37.0	36.4	37.3	37.2
VideoLLaMA2.1-7B	25.6	22.2	28.9	46.0	5.1	14.1	17.1	32.5	31.0	11.1	78.8	9.1	15.3	13.0	27.1	9.8	11.3	32.5	31.8	32.7	33.2
Risk-only																					
GPT-5.6-Terra	**62.0**	**63.7**	**54.7**	**72.0**	**59.2**	**65.2**	**67.4**	**69.8**	**77.8**	**67.0**	**86.2**	**48.4**	40.2	34.5	49.5	38.1	38.9	**74.2**	**73.9**	**74.3**	**74.3**
Claude-Opus-5	53.2	53.1	39.9	62.3	50.2	55.8	57.8	62.9	71.9	57.4	81.8	40.2	28.5	24.8	32.6	29.8	25.6	68.1	66.7	67.9	69.9
Gemini-3.1-Pro	52.5	50.4	25.5	65.1	50.3	51.1	59.7	55.8	65.0	50.4	70.7	37.0	**41.6**	**35.3**	**49.7**	**38.5**	**42.9**	62.3	60.5	62.9	63.7
Qwen3.8-27B	**52.5**	47.2	25.9	59.2	**47.6**	**48.8**	**54.8**	**57.4**	**67.5**	**49.7**	74.2	**37.8**	39.4	33.8	47.2	36.0	40.3	**66.0**	**65.0**	**65.8**	**67.1**
InternVL3.5-8B	52.1	**51.4**	**71.1**	**59.8**	37.6	41.1	47.4	55.2	58.5	40.4	90.8	30.8	**45.0**	**34.6**	**55.9**	**42.5**	**46.8**	56.9	55.9	58.0	56.9
Qwen3.5-9B	45.9	39.6	13.0	53.0	40.3	44.4	47.2	50.9	59.6	42.2	71.2	30.6	34.7	29.0	42.7	32.7	34.4	58.3	57.2	59.5	58.3
Tarsier2-7B	32.1	15.6	18.2	10.2	11.8	20.0	17.9	38.8	40.9	15.6	84.9	13.6	37.9	31.4	55.1	31.4	33.7	36.3	36.2	35.6	37.1
Qwen2.5-VL-7B-Instruct	30.9	19.8	23.9	21.3	14.8	15.1	24.0	40.9	38.6	17.9	94.5	12.7	22.6	21.8	33.2	18.1	17.4	40.2	39.2	40.2	41.2
VideoLLaMA2.1-7B	25.6	11.7	0.0	39.9	2.0	7.9	8.4	32.9	28.0	5.4	93.5	4.4	26.8	25.3	37.9	21.6	22.0	31.1	30.2	30.8	32.2
VideoLLaMA3-7B	21.2	6.0	0.4	10.8	3.2	8.0	7.8	34.4	28.7	7.6	**95.4**	5.7	11.1	9.0	15.3	9.8	10.2	33.6	32.7	33.5	34.6

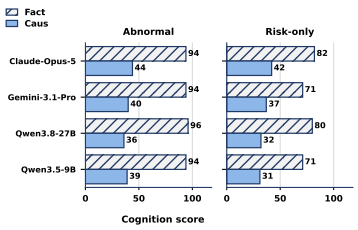

Figure 4: Fact vs. causal understanding on abnormal and risk-only videos.

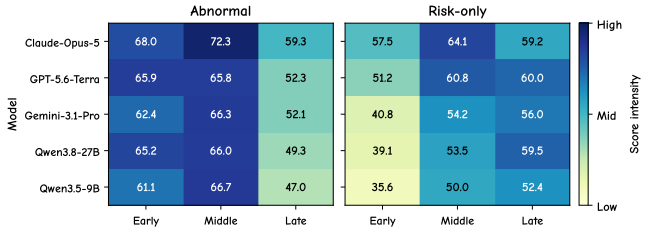

Figure 5: Overall score by risk-event position (Early/Middle/Late thirds of the video) for the top proprietary and open-source models.

Temporal. Temporal localization is affected by both search difficulty and the position of the risk event within the clip. *(1) Temporal windows mainly help weak models.* To measure how much temporal localization contributes to overall understanding, we provide models with the ground-truth risk window as an auxiliary hint. Annotated risk interval is mapped to the nearest uniformly sampled frames and appended to the perception, cognition, and planning prompts together with the frame-to-time correspondence. Since the temporal window is given, we exclude the Temporal dimension and report the average over Perception, Cognition, and Planning (Table 3). The window mainly helps weak models, and the gains concentrate in coarse perception: on abnormal videos, Perception AVG rises substantially for Qwen2.5-VL (+27.6) but only marginally for GPT-5.6-Terra (+3.2). On risk-only videos the gains are smaller because an emerging risk still lacks a visible outcome even once its interval is known. This reflects relieving a *search* burden rather than adding knowledge: a window frees weak models from scanning a long, redundant clip, while strong models already localize the informative region unaided. The primary deficit therefore lies in *where to seek information* rather than *how to reason*. This suggests that risk-region proposal and query-conditioned frame sampling are more valuable research directions than directly improving models' reasoning capabilities.

(2) Effect of risk event position. Event position further affects performance (Figure 5). We divide videos into Early, Middle, and Late groups by the normalized midpoint of the ground-truth interval, corresponding to the first, middle, and last thirds of the video. Scores generally peak when the event sits in the Middle third for abnormal videos, because it is flanked by enough preceding footage to establish a baseline and enough following footage to observe how the event develops. For risk-only videos, scores tend to be lowest in the Early group, since an early precursor is diluted by the long normal remainder that follows.

Planning. *Planning appears relatively strong, but its scores are less diagnostic of genuinely video-grounded understanding.* For example, Qwen3.8-27B achieves Planning scores of 67.9 and 66.0 on

Table 3: Results when ground-truth risk windows are provided as frame-level hints, averaged over the real-world and generated subsets. The Temporal dimension is excluded because the ground-truth window is given.

Model	Score	Perception						Cognition					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	Per	Haz	Ovr
Abnormal																
GPT-5.6-Terra	**75.8**	72.2	70.3	**81.3**	**61.0**	**74.3**	**74.1**	**77.4**	**86.7**	**70.2**	**99.0**	**53.8**	**77.7**	**77.0**	**79.8**	**76.2**
Gemini-3.1-Pro	71.6	**73.3**	**91.3**	79.7	57.4	66.6	71.4	71.6	83.0	59.4	96.1	47.9	69.8	65.3	71.0	72.9
Claude-Opus-5	69.4	68.1	85.4	74.0	50.0	67.1	63.9	68.7	78.5	60.8	90.1	45.2	71.5	69.5	71.8	73.0
Qwen3.5-9B	**71.9**	**73.7**	95.5	**78.3**	53.7	**66.6**	**74.2**	71.8	**84.5**	60.6	95.2	46.8	70.3	68.7	73.1	69.3
Qwen3.8-27B	71.7	70.3	83.4	76.7	**56.9**	66.0	68.6	**72.3**	83.0	**61.8**	97.0	**47.6**	**72.5**	**71.1**	**74.8**	**71.6**
InternVL3.5-8B	62.8	60.0	54.4	77.5	48.7	55.6	63.5	65.0	70.8	51.0	97.6	40.6	63.5	63.4	65.5	61.7
Qwen2.5-VL-7B-Instruct	55.6	55.1	68.2	66.8	39.2	49.3	51.7	57.0	61.5	37.6	96.1	32.4	54.8	53.9	56.9	53.7
VideoLLaMA2.1-7B	48.7	59.4	**99.6**	73.4	31.4	53.0	39.8	42.4	51.1	15.9	85.7	16.7	44.3	45.9	43.8	43.1
Tarsier2-7B	45.5	48.1	69.6	57.8	27.4	42.3	43.3	45.2	52.5	25.3	78.6	24.2	43.5	44.6	42.2	43.5
VideoLLaMA3-7B	37.5	30.4	45.9	33.4	17.2	26.8	28.7	44.7	43.0	20.8	**97.8**	17.1	37.5	37.3	38.3	36.8
Risk-only																
GPT-5.6-Terra	**74.7**	**70.7**	**65.8**	**78.3**	**64.5**	**71.5**	**73.7**	**75.2**	**85.6**	**71.8**	**89.9**	**53.3**	**77.9**	**78.1**	**78.0**	**77.6**
Gemini-3.1-Pro	65.5	60.8	32.9	76.5	61.2	61.9	71.3	63.8	77.0	61.5	70.5	46.2	71.7	68.9	72.0	74.2
Claude-Opus-5	63.5	57.2	34.3	71.0	51.8	61.6	67.0	61.6	74.0	61.9	66.1	44.7	71.6	69.2	71.5	74.0
Qwen3.8-27B	**62.7**	55.3	29.7	69.3	**56.3**	**56.4**	**64.6**	**63.8**	**74.0**	**61.0**	75.1	**44.8**	**69.2**	**68.1**	**69.8**	**69.5**
InternVL3.5-8B	61.3	**59.6**	**71.8**	**71.3**	48.9	50.7	55.5	63.0	68.6	51.7	93.0	38.9	61.2	61.9	61.7	60.1
Qwen3.5-9B	58.2	51.0	11.7	69.3	54.0	55.5	64.3	58.5	70.8	57.5	63.9	42.0	65.2	64.3	66.4	65.2
Qwen2.5-VL-7B-Instruct	47.1	41.0	33.2	54.5	35.3	37.9	44.3	51.2	54.4	36.0	87.0	27.8	49.2	49.4	49.9	48.2
Tarsier2-7B	38.1	38.2	42.7	51.2	21.9	33.8	41.4	35.8	44.4	21.2	58.6	18.9	40.2	41.6	38.7	40.2
VideoLLaMA2.1-7B	35.8	35.1	0.7	63.7	27.6	40.9	42.6	31.5	44.5	14.2	53.2	14.0	40.8	42.2	39.9	40.2
VideoLLaMA3-7B	25.4	10.9	0.3	16.1	9.1	14.6	14.6	36.3	30.7	12.1	**94.0**	8.5	28.8	28.9	29.1	28.4

abnormal and risk-only videos, respectively, compared with 64.2 and 58.3 for Qwen3.5-9B (Table 2). This overlap suggests that Planning is less discriminative of capability than the visually grounded dimensions. A plausible explanation is that models can produce fluent mitigation language from general safety priors without fully grounding the advice in the observed hazard. Accordingly, a fluent mitigation plan should not be interpreted as evidence of reliable visual risk understanding.

6 Discussion and Future Work

The findings of HomeACE suggest four directions for future research. *First*, several high-performing models perform worse on videos containing only risk precursors, suggesting an over-reliance on visible event outcomes. Contrastive learning with paired normal, precursor, and anomalous samples may improve risk-state discrimination. *Second*, a gap remains between factual boundary consistency and causal understanding. Models can distinguish observed events from potential consequences, but struggle to connect visible risk cues into a coherent causal chain. Supervision on structured risk trajectories could link risk sources, human actions, and potential consequences, particularly for precursor-only videos. *Third*, temporal localization remains challenging, while temporal cues improve performance. A coarse-to-fine strategy could first identify candidate risk intervals and then increase the sampling frame rate within them for finer temporal analysis. *Finally*, planning performance remains relatively strong despite limited perception and temporal localization, suggesting that plausible interventions may lack sufficient visual evidence. Intervention generation should therefore be constrained by identified people, risk sources, and temporal evidence, with their consistency explicitly verified. Overall, future research should move from post-hoc event description toward risk evolution prediction and evidence-grounded intervention generation.

7 Conclusion

We introduce HomeACE, a household-video benchmark combining real-world data with human-verified generated videos produced through structured risk discovery. HomeACE evaluates Perception, Cognition, Temporal, and Planning. Experiments across ten MLLMs expose structural limitations: models act more as post-hoc incident describers than proactive risk monitors, exhibit strong factual recognition but weak causal understanding, and suffer from temporal localization errors that propagate to other capabilities. Planning remains relatively strong but is less discriminative of model capability and may not be grounded in the observed hazard. These findings identify key limitations in household risk understanding by current MLLMs, and we hope HomeACE will support the development of more dependable video understanding for practical home-safety applications.

AI USE STATEMENT

We used generative AI tools in dataset development, evaluation, and manuscript preparation. As described in Section 3 and Appendices A.1–A.3, LLMs assisted with generating and reviewing risk-tree candidates, expanding scenario specifications into state-specific prompts, producing preliminary annotations, completing annotation fields, and checking prompt–video consistency. Video generation models produced the generated subset. Human annotators inspected every generated video and determined its inclusion, correcting retained annotations and risk intervals against the visible content. Four computer science Ph.D. students divided the annotation work across both subsets, and two additional annotators reviewed the results; disagreements were resolved through joint discussion. For evaluation, Qwen3-8B served as an automated judge for semantic outputs, as described in Section 4; these scores are model-based assessments rather than human ratings.

We also used LLM-based assistants for literature retrieval and synthesis, Chinese–English translation, drafting and revising portions of the manuscript, improving readability, suggesting manuscript structure, and providing feedback on evaluation design and result interpretation. These uses included research feedback as well as writing assistance. We have reviewed all AI-assisted work. The authors retain responsibility for the research decisions and the final content of the paper, including its references, claims, data, and AI-assisted text and artifacts. Mathematical proof assistance is not applicable to this work.

REFERENCES

Bilibili. https://www.bilibili.com/. Online video platform, accessed August 2, 2026.

Dailymotion. https://www.dailymotion.com/. Online video platform, accessed August 2, 2026.

Douyin. https://www.douyin.com/. Online video platform, accessed August 2, 2026.

Kuaishou. https://www.kuaishou.com/new-reco. Online video platform, accessed August 2, 2026.

Tiktok. https://www.tiktok.com/explore. Online video platform, accessed August 2, 2026.

Xigua video. https://www.ixigua.com/. Online video platform, accessed August 2, 2026.

Youtube. https://www.youtube.com/. Online video platform, accessed August 2, 2026.

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022a.

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20143–20153, 2022b.

Anthropic. Claude opus 5. https://www.anthropic.com/news/claude-opus-5, 2026.

Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. *DIRO-Université de Montréal, Tech. Rep*, 1350:24, 2010.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 833–842. IEEE, 2019.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Google DeepMind. Veo 3.1. `https://deepmind.google/models/veo/`, 2025a.

Google DeepMind. Gemini omini. `https://deepmind.google/models/gemini/`, 2025b.

Google DeepMind. Gemini 3.1 pro. `https://deepmind.google/models/gemini/`, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3):489–501, 2014. doi: 10.1016/j.cmpb.2014.09.005.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. VideoVista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations*, volume 2025, pp. 85485–85507, 2025.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2720–2727, 2013.

Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 341–349, 2017.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Moonshot AI. Kimi. `https://kimi.moonshot.cn/`, 2024.

OpenAI. Sora 2. `https://openai.com/index/sora-2/`, 2025.

OpenAI. Gpt-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 2026a.

OpenAI. Gpt-5.6. `https://openai.com/index/gpt-5-6/`, 2026b.

Junhao Pan, Zehua Yuan, Xiaofan Zhang, and Deming Chen. Youhome system and dataset: Making your home know you better. In *2022 IEEE International Symposium on Smart Electronic Systems (iSES)*, pp. 414–420, 2022. doi: 10.1109/ISES54909.2022.00091.

Qwen Team. Qwen3.8-27b. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169. IEEE, 2025.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 2016.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018a.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sheng Wang, Pengan Chen, Jingqi Zhou, Qintong Li, Jingwei Dong, Jiahui Gao, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. TreeSynth: Synthesizing diverse data from scratch via tree-guided subspace partitioning. *arXiv preprint arXiv:2503.17195*, 2025a. URL https://arxiv.org/abs/2503.17195.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Jiang Xudong. Video-ktr: Reinforcing video reasoning via key token attribution. In *International Conference on Learning Representations*, volume 2026, pp. 71553–71570, 2026.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020a.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–339, 2020b.

xAI. Grok video 1.5. https://x.ai/, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

A Xu, B Lin, B Xue, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 98891–98913, 2025.

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

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

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

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024a.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Xinyi Zhao, Congjing Zhang, Pei Guo, Wei Li, Lin Chen, Chaoyue Zhao, and Shuai Huang. Smarthome-bench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. *arXiv preprint arXiv:2506.12992*, 2025a.

Xinyi Zhao, Congjing Zhang, Pei Guo, Wei Li, Lin Chen, Chaoyue Zhao, and Shuai Huang. Smarthome-bench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 3975–3985, June 2025b.

Zhipu AI. Glm. `https://chatglm.cn/`, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18891–18901. IEEE, 2025.

A Appendix

A.1 Multi-Perspective Risk Review for Generated Videos

Motivation. Whether a household behavior should be considered risky is not always unambiguous. For example, a child jumping on a sofa may be regarded as ordinary play under a permissive standard, but as a meaningful fall risk under a conservative standard or when the vulnerability of the child is emphasized. A single unconstrained judgment may therefore reflect only one implicit safety preference. To make the decision process more explicit and robust, we introduce a perspective-conditioned multi-view review procedure within the risk review agent.

Shared input and isolated execution. For every candidate fine-grained behavior, the risk review agent performs five separate evaluations using the same underlying LLM. All five evaluations receive exactly the same candidate and contextual information, including its scene, subject, parent risk category, and behavioral description. The only difference is the definition of the review perspective included in the instruction. Each perspective is executed in an isolated model call: it cannot observe the decisions or rationales produced by the other perspectives. Every call returns (i) a binary decision indicating whether the candidate should be retained and (ii) a textual rationale supporting that decision. We preserve the output of every perspective for subsequent aggregation and auditing.

The five review perspectives are operationalized as follows:

Table 4: Definitions of the five perspectives used for candidate risk review.

Perspective	**Operational definition**
Everyday Baseline	Judges whether the candidate meaningfully deviates from activities that are commonly acceptable in an ordinary household under the given scene and subject. The presence of a potentially hazardous object alone is insufficient if the described behavior remains routine and no meaningful exposure is established.
Conservative Safety	Applies a precautionary standard and retains a candidate when it involves a reasonably foreseeable possibility of injury or escalation, even if the harmful outcome has not yet occurred. This perspective prioritizes early warning and sensitivity to potential risks.
Permissive Tolerance	Allows minor hazards and common behavioral variation in daily life. A candidate is retained only when the danger is sufficiently explicit, exceeds ordinary household risk tolerance, or has a clear pathway to harm. This perspective limits the over-classification of normal activities as anomalies.
Vulnerable-Subject First	Evaluates the behavior in relation to the subject's ability to recognize, avoid, or recover from danger. Behaviors involving children, older adults, people with limited mobility, or pets may be retained even when the same behavior would pose limited risk to a healthy adult.
Consequence Severity	Prioritizes the magnitude of the plausible outcome rather than its likelihood alone. A low-probability event may be retained when it can plausibly lead to severe injury, poisoning, electric shock, fire, drowning, or death.

Majority aggregation. Let $d_i \in \{0, 1\}$ denote the binary decision of perspective i, where $d_i = 1$ indicates "retain" and $d_i = 0$ indicates "reject." The aggregated risk-review decision D is

$$D = \mathbb{I}\left(\sum_{i=1}^{5} d_i \geq 3\right), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Thus, a candidate passes risk review when at least three of the five perspectives recommend retention. Otherwise, it is excluded from the final risk set. The aggregation uses the binary decisions, while the associated rationales are retained as supporting records rather than converted into numerical scores.

Arbitration and decision boundaries. After multi-perspective risk review, the arbitration and revision agent receives the candidate behavior, the structure-review result, and all five perspective-specific decisions and rationales. Its role at this stage is restricted to determining whether the specific candidate behavior is retained as a final leaf node. It applies the $3/5$ risk-review rule together with the structural requirements: a retained behavior must belong to the intended semantic level, match its parent configuration, and not duplicate another retained behavior. The arbitration agent does not introduce a new risk criterion that replaces the majority decision. It instead consolidates the available evidence, removes structurally invalid or redundant candidates, and outputs a normalized final behavior description. Starting from 20 proposed behaviors for each parent configuration, this process retains five valid and mutually non-redundant leaf behaviors.

Interpretation and limitation. The five judgments should not be interpreted as votes from five independent models. They are five isolated, perspective-specific evaluations performed by the same

risk-review model assigned in the current multi-agent rotation. Consequently, they may still share model-specific knowledge, preferences, and failure modes. We use this procedure as a structured form of multi-perspective self-consistency: its purpose is to expose competing yet reasonable safety standards and make the inclusion decision auditable, rather than to claim statistical independence among reviewers. Across construction batches, different LLMs rotate into the risk-review role, while the five perspective definitions and the $3/5$ aggregation rule remain unchanged.

Illustrative example. Consider the candidate behavior "a child jumping on a sofa." Everyday Baseline may reject it as common play, and Permissive Tolerance may also reject it when no additional environmental hazard is described. In contrast, Conservative Safety may retain it because of a foreseeable fall, while Vulnerable-Subject First may emphasize the child's limited balance and self-protection ability. Consequence Severity determines its decision from the plausible injury severity given contextual factors such as sofa height and nearby sharp objects. The candidate is retained only if at least three perspectives ultimately return a positive decision. This example illustrates why all perspectives receive the same contextual description and why their individual rationales are preserved.

A.2 Root-to-Leaf Path Expansion for Video Prompts

Path extraction. Each retained leaf node is paired with all of its ancestors to form a complete semantic path,

$$p = (s, o, c, b), \tag{2}$$

where s denotes the household scene, o the subject, c the risk category, and b the retained fine-grained behavior. The path provides a compact and traceable specification of what should appear in the generated video. Using the full path, rather than the leaf description alone, reduces ambiguity about the intended environment, involved subject, and type of risk.

Prompt expansion. We input the structured path into DeepSeek-V4 (Xu et al., 2026) and instruct it to generate three state-specific records. Each record contains a state label, a temporally segmented video prompt, and three mitigation fields. The video prompt uses the following shared stage vocabulary:

- `[Normal Behavior]`: ordinary behavior before any visible warning sign;
- `[Abnormal Sign]`: a visible hazard, unstable condition, or unsafe tendency indicating that the subject is exposed to risk;
- `[Abnormal Result]`: the realized anomalous event and its observable immediate consequence.

Every included stage is followed by an explicit interval in seconds. The normal record contains only `[Normal Behavior]`; the risk-only record contains `[Normal Behavior]` followed by `[Abnormal Sign]`, but excludes `[Abnormal Result]`; and the abnormal record contains the complete progression through all three stages. The stage intervals are ordered according to the temporal development of the event. The remaining fields describe an immediate solution for the affected person, a solution for the hazard source, and a follow-up prevention solution.

Expansion example. Consider the following structured path before expansion:

> **Scene:** Living room → **Subject:** Child → **Risk category:** Fall and instability → **Fine-grained behavior:** A child slips and falls from the sofa while bouncing on it.

DeepSeek-V4 expands this path into the following three records:

Normal record (`R000001_normal`).

> **Video prompt:** `[Normal Behavior][0--12s]` An East Asian boy, about 5 years old, sits calmly on a beige fabric sofa in a living room. He is watching television, occasionally shifting his position or playing with a soft toy in his hands. The sofa cushions are firm and flat. The boy's feet are on the floor or tucked under him on the sofa.
>
> **Mitigation:** Solution for the person: none. Solution for the hazard source: none. Follow-up prevention solution: none.

Risk-only record (`R000001_risk_only`).

> **Video prompt:** `[Normal Behavior][0--3s]` An East Asian boy stands on a beige fabric sofa in the living room, holding a toy. `[Abnormal Sign][3--12s]` He starts bouncing lightly on the sofa cushions. The cushions are soft and compress under his weight. A small, smooth decorative pillow slides off the sofa onto the floor near his feet. He is not holding onto anything for support as he bounces, and his movements are becoming less controlled.
>
> **Mitigation:** Solution for the person: Immediately instruct the child to stop bouncing and carefully sit down or step off the sofa onto the floor. Solution for the hazard source: Remove the slippery decorative pillow from the floor and fluff or firm up the sofa cushions to reduce instability. Follow-up prevention solution: Use non-slip pads under sofa cushions and keep the sofa area clear of toys or objects that could be tripped over. Supervise play and reinforce rules about safe sofa use.

Abnormal record (`R000001_abnormal`).

> **Video prompt:** `[Normal Behavior][0--2s]` An East Asian boy is energetically jumping up and down on a beige fabric sofa in the living room. `[Abnormal Sign][2--5s]` On one high jump, his foot lands on the edge of a loose, smooth cushion, causing it to slide out from under him. He loses his balance, arms flailing. `[Abnormal Result][5--12s]` He slips sideways off the sofa, falling onto the hardwood floor with a thud. He lands on his side and starts crying, sitting on the floor rubbing his arm. A toy is knocked off the sofa beside him.
>
> **Mitigation:** Solution for the person: Comfort the child and check for injuries. Do not move him if a serious injury is suspected; call for medical help if needed. If he is okay, help him stand up away from the sofa. Solution for the hazard source: Immediately stop the jumping activity. Secure all sofa cushions and remove any objects from the floor around the sofa. Follow-up prevention solution: Consider placing a soft, non-slip rug in front of the sofa to cushion potential falls. Consistently supervise and redirect high-energy play to safer areas such as a carpeted floor or play mat.

The three records are grounded in the same root-to-leaf path, but differ in how far the event progresses. The explicit temporal boundaries provide direct control over the ordering and duration of the normal behavior, warning sign, and anomalous result, while the associated mitigation fields support evaluation of immediate response, hazard handling, and long-term prevention.

A.3 Video Annotation and Human Verification

Annotation Interface and Human Verification. We develop a dedicated annotation platform to support video annotation and human verification for both real-world and generated videos. The platform 6 integrates video browsing, structured video annotation, annotation-progress statistics, and online collaborative annotation by multiple annotators. For real-world videos, annotators compare the candidate annotations generated by five video understanding MLLMs and select the most accurate one. If none of the candidates is sufficiently accurate, they rewrite the annotation or refine its wording. For generated videos, annotators inspect the video together with the automatic prompt–video consistency evaluation results and make a binary retain-or-discard decision.

Real-Video Candidate Annotation. The five MLLMs—mPLUG-Owl3 Ye et al. (2025), MiniCPM-V 4.5 Yu et al. (2025), Qwen3.5-9B Team (2026), InternVL3.5-8B Wang et al. (2025b), and VideoLLaMA3 Zhang et al. (2025a)—independently generate a complete candidate annotation for each real-world video under the same instruction. Each candidate specifies whether a household risk is present; the observed state (normal, risk-only, or abnormal); the Level-1 scene, Level-2 subject, and Level-3 risk category; the temporal localization of the risk-related event; an objective video description; and the three mitigation solutions for the affected subject, hazard source, and recurrence prevention. A human annotator then watches the complete video and reviews all five candidates. The annotator either selects the candidate that best matches the visible evidence or writes a new annotation when none is sufficiently accurate. The selected or newly written result is normalized into the four-dimensional HomeACE schema and treated as the final ground truth.

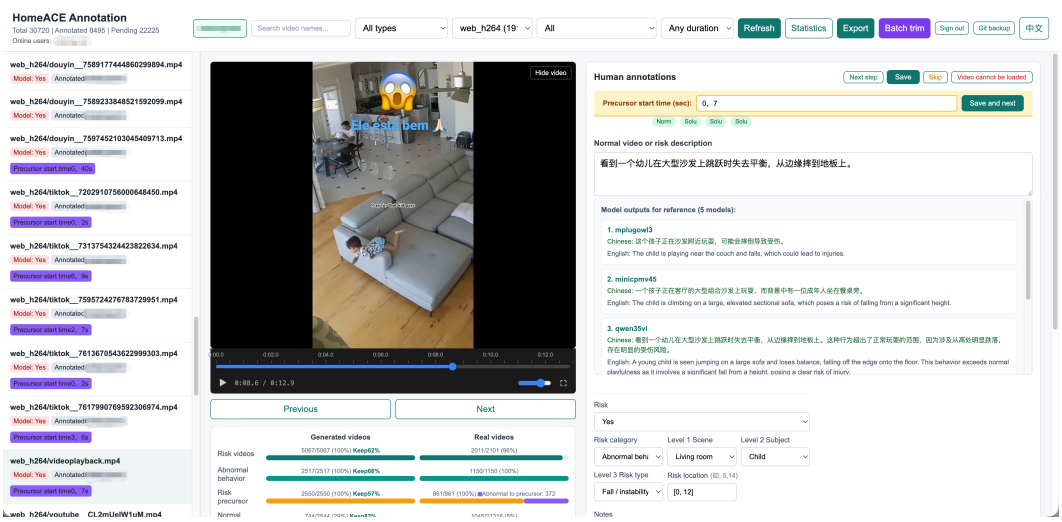

Figure 6: Overview of the HomeACE annotation platform. The platform supports video browsing, structured video annotation, annotation-progress statistics, and online collaborative annotation by multiple annotators. The left panel provides video navigation and annotation status, the center panel supports video playback, and the right panel presents preliminary outputs from five video understanding MLLMs together with the human annotation fields.

Generated-Video Consistency Screening. Before inclusion, generated videos undergo a separate prompt–video consistency check. For every risk-only or abnormal sample, each of the five MLLMs evaluates scene consistency, subject consistency, risk-source consistency, event-type consistency, and overall semantic consistency, and returns an initial score with a brief explanation. The observed video state is retained as auxiliary evidence rather than used as a hard acceptance gate; normal samples bypass this automatic scoring stage under the current protocol. A human annotator subsequently inspects every generated video and makes a binary decision to retain or discard it. For each retained risk-only or abnormal video, the annotator also checks the actual visible risk interval and corrects its start and end times when they deviate from the scripted intervals. Retained videos enter the dataset with their corrected temporal annotations and generation records, whereas discarded videos are treated as unsuccessful generations and returned to the video-generation pipeline. Regeneration continues until the planned number of accepted videos is reached, so the final retained set preserves the target size. The automatic scores only support efficient inspection; the human decision determines final inclusion.

A.4 Details of Model Inference

The model-inference procedure uses a set of structured prompts to decompose household risk understanding into Perception, Cognition, Temporal, and Planning. Each prompt instructs the model to inspect the video through its uniformly sampled input frames and return exactly one JSON object without explanations, prefixes, suffixes, or Markdown. The prompts define the permissible risk states, required fields, state-dependent output conditions, and the format of temporal intervals and mitigation responses. This section summarizes the shared output schema and the instructions used by the evaluation pipeline.

Prompt templates. The following excerpts reproduce the operational content of the prompts used for model inference. The original prompts are written in Chinese; they are translated here for readability while preserving their required fields and decision rules. The shared instruction is: "You are a video safety analysis expert. Inspect the video through its uniformly sampled input frames and output exactly one JSON object, without explanations, prefixes, suffixes, or Markdown." The four dimension-specific prompts then request the following outputs:

> **Perception.** Classify the video as `normal`, `risk_only`, or `abnormal`, and output `risk_status`, `risk_type`, `risk_sources`, `abnormal_actions`, and

`affected_objects`. Normal videos must have an empty risk type and empty lists. Risk-only videos must contain a visible risk source or unsafe behavior but no confirmed harmful consequence. Abnormal videos must contain a realized anomalous event or consequence.

Cognition. The risk state has been determined by a preceding step. Output `risk_description`, `consequence_understanding`, `causal_chain`, `causal_chain_text`, and `event_summary`. Describe the relation "risk source → abnormal action → consequence." Use prospective language for risk-only videos and indicative language for observed abnormal events.

Temporal. Output `risk_status` and `time_spans` as integer-second intervals. For risk-only videos, localize the interval in which the risk precursor is visible; for abnormal videos, localize the anomalous event; for normal videos, return an empty list.

Planning. Output `risk_status` and a three-part `solution` covering the affected person, the hazard source, and overall prevention. Normal videos use an empty solution.

These three mutually exclusive risk states determine the required fields and linguistic framing: *risk-only* outputs use conditional or prospective expressions such as "may" or "poses a risk," whereas *abnormal* outputs may describe observed events in the indicative form. Normal outputs contain no risk-related content.

The prompts therefore require the model to ground its semantic fields in the visual evidence of the input video. In particular, the Cognition prompt does not ask the model to complete a causal sentence from text alone: it asks the video model to inspect the video through its uniformly sampled input frames and describe the risk source, visible action or state, and observed or potential consequence. The resulting causal output is subsequently parsed into structured fields for evaluation.

Perception output. The Perception output uses the predefined risk-type vocabulary and records the visible risk sources, abnormal actions, and affected persons or objects. The affected-person vocabulary is constrained to children, young adults, middle-aged adults, older adults, or all people, while explicitly visible threatened objects or facilities may also be reported. The state-dependent empty-value and evidence requirements follow the rules in the prompt template.

Cognition output. The Cognition output is generated by the video model from the complete input video, rather than by applying a text-only causal-completion rule. Its causal-chain field is required to express the relation

$$\text{risk source} \;\rightarrow\; \text{abnormal action} \;\rightarrow\; \text{consequence}.$$

For risk-only videos, the description and causal relation must preserve the distinction between visible evidence and a possible future consequence. For abnormal videos, they describe the observed event while retaining only consequences supported by the video or annotation. Normal videos use empty values for all cognition fields. The subsequent Qwen3-8B judge does not receive the video pixels again; it evaluates the model-produced fields against the structured ground truth, the ground-truth risk description, and the raw model response. Thus, Caus measures video-grounded causal-chain consistency in the model's inference output, while the judge provides semantic scoring of that output rather than an independent visual causal analysis.

The machine-readable representation used by the released evaluation code is a list of links, each with three string fields:

```
"causal_chain": [
  {"source": "...", "action": "...", "consequence": "..."}
]
```

When a model returns a single causal-chain string, the evaluation parser converts the arrow-separated segments into the same source–action–consequence representation. The free-form `causal_chain_text` and the full raw response are retained as auxiliary textual evidence.

Temporal output. The resulting `time_spans` are represented as integer-second intervals. For risk-only videos, the interval covers the visible risk precursor from its onset to its end; for abnormal videos, it covers the anomalous event and its visible duration. Normal videos return an empty interval list. The corresponding scoring formulas are given in Section A.5.

Table 5: Criteria for the four top-level evaluation dimensions.

Dimension	Evaluation criterion
Perception	Recognition of risk status, risk type, risk sources, abnormal actions, and affected objects. Status and type use objective matching, while the remaining fields allow semantically equivalent descriptions.
Cognition	Understanding of the risk description, consequences, factual boundary, and causal chain. The output should distinguish observed facts from possible consequences and preserve the annotated risk progression.
Temporal	Localization of the risk-related interval. The predicted interval is evaluated using IoU, start-time accuracy, end-time accuracy, and duration accuracy.
Planning	Appropriateness, actionability, and target consistency of the mitigation responses for the affected person, hazard source, and overall prevention.

Planning output. The `solution` field contains three separately labeled lines:

```
Person solution: ...
Hazard-source solution: ...
Overall prevention solution: ...
```

The three lines respectively address immediate protection of the affected person, removal or control of the hazard source, and longer-term environmental or behavioral prevention. Normal videos use an empty solution.

Ground-truth temporal hints. In the GT-hint setting, the complete video and the same dimension prompts are retained. Before Perception, Cognition, and Planning inference, the prompt is appended with the manually annotated risk interval converted to the nearest sampled-frame indices. When enabled, the prompt also includes the mapping from each sampled frame to its timestamp. The hint is phrased as supplementary information and asks the model to focus on the indicated frames; it does not provide the risk type, causal chain, consequence, or mitigation answer. Ground-truth hints are not appended to the Temporal prompt because Temporal evaluation is skipped in this setting. Thus, GT-hint results isolate the effect of reducing temporal search while keeping the semantic outputs subject to the same evaluation.

Judge prompt. For each semantic evaluation, the Qwen3-8B judge receives a system instruction equivalent to the following:

> You are a video-safety evaluation expert. Given the ground truth and the model output, evaluate only the specified dimension. Assign a score from 0 to 5, allowing up to two decimal places. Use fine-grained scoring based on the available evidence rather than defaulting to integer levels. Return exactly one JSON object in the form {`"score": 0.00--5.00, "reason": "..."`} and do not output any additional text or Markdown.

The corresponding user message contains four components: `dimension`, the dimension-specific `rubric`, the relevant `ground_truth` fields, and the model `prediction`. The prediction includes the structured model output and, when available, the raw response. The judge therefore evaluates the semantic consistency and quality of the model-generated answer against the supplied annotations; it does not receive or rewatch the video frames.

In the GT-hint setting, the same judge prompt and dimension-specific rubrics are used. The evaluation script only changes the inference and aggregation configuration: it injects the annotated temporal window into the Perception, Cognition, and Planning prompts, restricts evaluation to abnormal and risk-only videos, and excludes the Temporal dimension. It does not provide the judge with additional video evidence or alter the semantic scoring rubric.

The criteria for the four top-level dimensions are summarized in Table 5.

For the Caus submetric, the judge evaluates whether the predicted risk source, abnormal action, and consequence form a complete causal chain consistent with the annotated relation. A score of 5

Table 6: Results on normal videos. "Score" is the overall normal-video score, while "Status" and "Desc." denote status accuracy and description quality, respectively. All score metrics are percentages.

Real Videos			
Model	**Score**	**Status**	**Description**
Gemini-3.1-Pro	**74.3**	**78.2**	**70.6**
GPT-5.6-Terra	66.3	62.8	69.8
Claude-Opus-5	62.7	60.4	65.0
Qwen2.5-VL-7B-Instruct	**85.7**	99.0	72.4
Qwen3.5-9B	83.2	86.4	**80.0**
VideoLLaMA2.1-7B	80.1	98.6	61.8
VideoLLaMA3-7B	79.9	**100.0**	60.0
Qwen3.8-27B	77.9	80.0	75.8
Tarsier2-7B	71.3	86.4	56.2
InternVL3.5-8B	58.6	59.6	57.6

Generated Videos			
Model	**Score**	**Status**	**Description**
Gemini-3.1-Pro	**86.2**	**89.0**	**83.6**
Claude-Opus-5	85.4	87.8	83.0
GPT-5.6-Terra	79.6	77.2	82.0
Qwen3.5-9B	**96.2**	98.6	**93.8**
Qwen3.8-27B	91.6	92.4	90.8
Qwen2.5-VL-7B-Instruct	88.9	99.6	78.0
VideoLLaMA3-7B	88.1	**100.0**	76.2
VideoLLaMA2.1-7B	87.5	99.8	75.0
Tarsier2-7B	86.8	99.8	73.8
InternVL3.5-8B	82.2	87.0	77.4

indicates a complete and correct chain, whereas 0 indicates that no causal relation is captured; intermediate scores reflect partial correctness. Unsupported common-sense details should not receive full credit merely because they are plausible. For risk-only videos, a potential consequence must be expressed prospectively, while claiming that it has already occurred is a factual-boundary error assessed by the separate Fact submetric. The judge score is converted to a percentage by multiplying by 20. When an LLM judge is unavailable, the evaluation code falls back to component matching over the annotated source, action, and consequence fields.

A.5 Details of Protocol

We report all evaluation scores as percentages. Binary objective fields, including risk status and risk type, receive 5 points for an exact match and 0 points otherwise. Subjective semantic fields are evaluated by Qwen3-8B (Yang et al., 2025), which assigns a score $s \in [0, 5]$ according to the corresponding rubric. All five-point scores are converted to percentages as $20s$.

For abnormal and risk-only videos, each dimension score is the equal-weight arithmetic mean of its submetrics. The Perception score is the mean of its five submetrics:

$$\text{Perception} = \frac{1}{5}(\text{Sta} + \text{Typ} + \text{Src} + \text{Act} + \text{Obj}), \tag{3}$$

where each submetric is reported on the percentage scale. The Cognition score is the equal-weight mean of its four submetrics:

$$\text{Cognition} = \frac{1}{4}(\text{Desc} + \text{Csq} + \text{Fact} + \text{Caus}). \tag{4}$$

Temporal localization is evaluated from the predicted and ground-truth risk intervals. Let $D_{\text{gt}} = \max(|T_{\text{gt}}|, 1)$ denote the ground-truth interval duration, and let e_{st}, e_{end}, and e_{dur} denote the absolute start-time, end-time, and duration errors, respectively. The four temporal submetrics are computed on a five-point scale as

$$\text{IoU} = 5\,\text{IoU}(T_{\text{gt}}, T_{\text{pred}}), \tag{5}$$

$$\text{St} = 5\max\left(0, 1 - \frac{e_{\text{st}}}{D_{\text{gt}}}\right), \tag{6}$$

$$\text{End} = 5\max\left(0, 1 - \frac{e_{\text{end}}}{D_{\text{gt}}}\right), \tag{7}$$

$$\text{Dur} = 5\max\left(0, 1 - \frac{e_{\text{dur}}}{D_{\text{gt}}}\right). \tag{8}$$

They are then converted to percentages and averaged with equal weights:

$$\text{Temporal} = \frac{1}{4}(\text{IoU} + \text{St} + \text{End} + \text{Dur}). \tag{9}$$

When a risk interval exists but no interval is predicted, IoU is set to zero and each temporal error is set to D_{gt}, yielding zero endpoint and duration scores. For normal videos, Temporal is not included in the overall score; an absent predicted interval is treated as the correct outcome.

Table 7: Results on real-world abnormal and risk-only videos. All values are percentages; "AVG" denotes the within-dimension mean. Light blue and orange rows indicate abnormal and risk-only videos, respectively.

Model	Score	Perception						Cognition					Temporal					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	IoU	St	End	Dur	AVG	Per	Haz	Ovr
Claude-Opus-5	**65.1**	**63.0**	**70.4**	71.2	**48.0**	**64.8**	**60.8**	**70.8**	**81.6**	**63.6**	94.8	**43.2**	53.7	43.2	64.6	54.8	52.4	**72.9**	**73.2**	**73.4**	**72.4**
GPT-5.6-Terra	58.8	59.7	54.0	**73.0**	47.4	64.0	60.2	69.9	77.4	63.4	**95.4**	**43.2**	33.7	26.2	38.2	39.4	30.8	71.8	72.0	73.2	70.2
Gemini-3.1-Pro	58.5	57.5	68.6	66.2	42.4	57.0	53.2	60.3	70.4	44.6	92.2	34.2	**58.3**	**47.4**	**67.2**	**62.6**	**55.8**	58.0	56.0	59.4	58.6
Qwen3.8-27B	**57.2**	55.6	67.4	64.8	**42.4**	54.8	48.6	**62.9**	**72.8**	**47.8**	95.8	**35.4**	44.1	**36.4**	57.8	**42.4**	39.6	**66.4**	**64.4**	**67.6**	**67.2**
Qwen3.5-9B	56.0	**57.8**	**77.0**	66.2	40.2	**55.8**	49.6	61.4	71.6	46.0	94.0	34.0	42.0	32.4	58.8	36.4	40.4	62.7	61.0	64.6	62.6
InternVL3.5-8B	54.3	49.5	30.4	73.0	38.8	52.8	**52.8**	59.8	66.0	42.8	**96.4**	33.8	**45.9**	34.0	**66.6**	41.0	**42.2**	61.8	58.6	65.0	61.8
Qwen2.5-VL-7B-Instruct	31.4	18.9	16.8	20.6	14.2	20.2	22.6	42.0	38.8	19.4	96.2	13.8	24.0	21.2	31.0	22.6	21.2	40.8	39.0	41.4	41.8
Tarsier2-7B	28.8	17.5	21.8	11.4	13.6	20.8	20.2	40.1	39.6	16.0	91.2	13.6	23.2	15.2	39.2	16.0	22.4	34.5	34.0	34.2	35.2
VideoLLaMA2.1-7B	25.1	21.7	13.8	**78.0**	1.2	9.0	6.2	30.2	26.4	5.0	85.6	3.8	18.4	15.6	24.2	17.6	16.2	30.1	29.2	30.4	30.8
VideoLLaMA3-7B	18.7	6.0	8.6	9.8	1.8	5.8	3.8	32.6	26.0	4.6	96.2	3.8	6.0	3.8	10.8	4.2	5.0	30.3	29.2	30.2	31.4
Claude-Opus-5	**62.6**	54.0	29.8	68.6	**52.4**	**61.6**	57.6	**65.2**	**77.8**	**58.8**	81.6	**42.4**	**56.4**	**49.6**	**65.2**	**59.6**	**51.2**	**75.0**	**75.4**	**73.8**	**75.8**
GPT-5.6-Terra	59.6	**55.7**	**41.8**	**69.2**	49.0	59.8	**58.8**	65.0	74.0	57.4	**87.8**	40.8	46.7	40.4	46.8	55.4	44.2	70.9	72.2	69.6	71.0
Gemini-3.1-Pro	52.9	45.1	16.8	62.8	44.2	49.2	52.4	54.2	64.2	41.0	80.4	31.2	54.2	48.2	59.2	58.6	51.0	58.0	56.6	57.8	59.8
Qwen3.8-27B	**52.6**	41.4	14.2	59.0	**40.8**	**46.8**	**46.2**	**55.0**	**67.0**	**40.4**	80.4	**31.8**	**49.0**	**44.2**	53.0	**52.8**	**45.6**	**65.0**	**64.8**	**63.8**	**66.4**
InternVL3.5-8B	52.0	**49.8**	**64.2**	62.6	34.8	42.6	44.8	53.1	58.6	33.2	93.2	27.4	48.5	38.4	**65.0**	46.8	43.8	56.8	56.2	57.8	56.6
Qwen3.5-9B	48.6	37.9	4.2	56.2	38.0	45.2	46.0	50.8	61.6	37.2	76.4	28.0	45.6	39.4	54.4	47.0	41.8	60.1	59.8	60.6	60.0
Tarsier2-7B	33.3	18.6	23.8	6.2	14.4	23.6	25.0	40.3	44.8	16.2	86.0	14.2	35.7	26.8	54.0	28.4	33.4	38.8	39.0	38.0	39.4
Qwen2.5-VL-7B-Instruct	28.9	14.0	12.8	16.2	9.8	12.8	18.2	38.6	36.0	12.8	96.6	9.0	22.7	21.4	25.2	22.8	21.4	40.2	39.4	40.2	41.0
VideoLLaMA2.1-7B	28.5	19.1	0.0	**76.4**	1.2	8.6	9.2	33.0	28.0	4.6	95.0	4.4	29.0	27.6	31.2	30.0	27.0	33.0	32.4	33.0	33.6
VideoLLaMA3-7B	21.0	6.1	0.0	12.8	1.4	8.2	8.2	34.6	28.8	6.8	**97.8**	5.0	8.3	5.6	13.4	6.0	8.2	35.2	34.4	35.2	35.8

Planning contains three independently judged mitigation responses for the affected person, the hazard source, and overall prevention. Its score is their equal-weight mean:

$$\text{Planning} = \frac{1}{3}(\text{Per} + \text{Haz} + \text{Ovr}). \tag{10}$$

The overall score for abnormal and risk-only videos is the equal-weight mean of the four dimension scores:

$$\text{Score} = \frac{1}{4}(\text{Perception} + \text{Cognition} + \text{Temporal} + \text{Planning}). \tag{11}$$

For normal videos, only risk-status accuracy and description quality are evaluated. The former is an exact-match score, while the latter is judged by Qwen3-8B on the five-point scale and converted to a percentage. The overall normal-video score is their equal-weight mean, $\text{Score} = (\text{Status} + \text{Desc.})/2$.

A.6 Additional Results

Normal Videos Results. For normal videos, we evaluate risk-status classification and video description. Since normal videos do not contain a realized risk interval or mitigation target, the Perception and Cognition dimensions are represented by status accuracy and description quality, respectively. The overall normal-video score is their equal-weight mean, $\text{Score} = (\text{Status} + \text{Desc.})/2$. We report results separately for real-world and generated videos.

- *Real Videos Results.* Table 6 reports the results on real-world normal videos. The overall score is computed from status accuracy and description quality according to the normal-video evaluation protocol.
- *Generated Videos Results.* The same table provides the corresponding results for generated normal videos.

Abnormal and Risk-only Videos Results.

- *Real Videos Results.* Table 7 reports the results on real-world abnormal and risk-only videos, where light blue and orange rows denote abnormal and risk-only videos, respectively.
- *Generated Videos Results.* Table 8 provides the corresponding results for generated abnormal and risk-only videos.

Table 8: Results on generated abnormal and risk-only videos. All values are percentages; "AVG" denotes the within-dimension mean. Light blue and orange rows indicate abnormal and risk-only videos, respectively.

Model	Score	Perception						Cognition					Temporal					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	IoU	St	End	Dur	AVG	Per	Haz	Ovr
GPT-5.6-Terra	**62.8**	**78.3**	84.4	**80.2**	**70.4**	**76.4**	**80.0**	**78.1**	**84.8**	**70.4**	**99.4**	**58.0**	**16.5**	**13.8**	**41.4**	2.4	8.6	**78.1**	**76.6**	**81.0**	**76.8**
Gemini-3.1-Pro	55.6	73.4	**95.6**	78.2	59.2	59.2	74.8	68.2	73.4	57.8	96.2	45.6	13.0	10.0	29.2	2.2	**10.6**	67.6	64.0	69.4	69.4
Claude-Opus-5	51.6	60.0	58.0	66.0	54.0	52.0	70.0	66.3	70.0	54.0	96.0	44.0	13.0	12.0	34.0	**4.0**	4.0	66.9	64.0	68.0	70.0
Qwen3.8-27B	**57.1**	**72.4**	**92.4**	72.0	**59.2**	**64.6**	**74.0**	**70.9**	**78.4**	**60.0**	**95.8**	**49.2**	15.9	11.4	32.6	2.8	16.8	**69.4**	**67.4**	**70.8**	**69.8**
Qwen3.5-9B	55.5	70.7	91.0	**74.0**	56.2	61.0	71.2	67.6	75.0	56.8	94.0	44.8	18.0	13.6	34.0	7.4	17.0	65.8	64.6	68.8	64.0
InternVL3.5-8B	51.7	53.5	43.0	71.0	45.0	47.4	61.4	58.3	59.6	45.0	90.2	38.6	35.2	24.0	41.2	20.4	**55.2**	59.7	56.8	62.6	59.8
Qwen2.5-VL-7B-Instruct	40.9	36.1	28.0	45.4	31.4	30.0	45.6	48.8	48.0	32.4	87.8	26.8	30.8	27.8	**73.2**	11.2	11.2	47.7	46.0	48.6	48.4
Tarsier2-7B	40.1	38.9	70.2	34.8	22.4	32.2	35.2	46.8	51.2	29.4	79.6	26.8	33.4	31.2	55.0	18.6	28.6	41.3	41.4	41.6	40.8
VideoLLaMA3-7B	40.0	30.7	52.0	37.2	15.0	25.0	24.4	42.9	43.0	25.8	81.4	21.4	**42.6**	**39.0**	52.6	**37.6**	41.2	43.7	43.6	44.4	43.0
VideoLLaMA2.1-7B	26.2	22.8	44.0	14.0	9.0	19.2	28.0	34.8	35.6	17.2	72.0	14.4	12.2	10.4	30.0	2.0	6.4	35.0	34.4	35.0	35.6
GPT-5.6-Terra	**64.4**	**71.7**	**67.6**	**74.8**	**69.4**	**70.6**	**76.0**	**74.7**	**81.6**	**76.6**	**84.6**	**56.0**	**33.8**	**28.6**	**52.2**	**20.8**	33.6	**77.4**	**75.6**	**79.0**	**77.6**
Gemini-3.1-Pro	52.1	55.6	34.2	67.4	56.4	53.0	67.0	57.4	65.8	59.8	61.0	42.8	29.0	22.4	40.2	18.4	**34.8**	66.6	64.4	68.0	67.6
Claude-Opus-5	43.7	52.3	50.0	56.0	48.0	50.0	58.0	60.6	66.0	56.0	82.0	38.0	0.6	0.0	0.0	0.0	0.0	61.2	58.0	62.0	64.0
Qwen3.8-27B	**52.4**	**53.1**	37.6	**59.4**	**54.4**	**50.8**	**63.4**	**59.7**	**68.0**	**59.0**	68.0	**43.8**	29.7	23.4	41.4	19.2	35.0	**67.0**	**65.2**	**67.8**	**67.8**
InternVL3.5-8B	52.2	53.0	**78.0**	57.0	40.4	39.6	50.0	57.2	58.4	47.6	88.4	34.2	**41.4**	30.8	46.8	**38.2**	**49.8**	57.0	55.6	58.2	57.2
Qwen3.5-9B	43.1	41.3	21.8	49.8	42.6	43.6	48.4	51.0	57.6	47.2	66.0	33.2	23.8	18.6	31.0	18.4	27.0	56.5	54.6	58.4	56.6
Qwen2.5-VL-7B-Instruct	32.9	25.6	35.0	26.4	19.8	17.4	29.8	43.2	41.2	23.0	92.4	16.4	22.6	22.2	41.2	13.4	13.4	40.2	39.0	40.2	41.4
Tarsier2-7B	30.9	12.6	12.6	14.2	9.2	16.4	10.8	37.2	37.0	15.0	83.8	13.0	40.1	**36.0**	**56.2**	34.4	34.0	33.8	33.4	33.2	34.8
VideoLLaMA2.1-7B	22.6	4.2	0.0	3.4	2.8	7.2	7.6	32.7	28.0	6.2	92.0	4.4	24.5	23.0	44.6	13.2	17.0	29.1	28.0	28.6	30.8
VideoLLaMA3-7B	21.5	5.9	0.8	8.8	5.0	7.8	7.4	34.1	28.6	8.4	**93.0**	6.4	13.9	12.4	17.2	13.6	12.2	32.0	31.0	31.8	33.4

Table 9: Results with ground-truth temporal-window hints on real-world abnormal and risk-only videos. Temporal localization is excluded because the annotated window is supplied. All values are percentages; "AVG" denotes the within-dimension mean.

Model	Score	Perception						Cognition					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	Per	Haz	Ovr
GPT-5.6-Terra	**74.5**	67.9	61.8	**82.6**	**52.6**	**74.0**	**68.6**	**77.1**	**88.8**	**71.0**	**98.6**	**50.2**	**78.4**	**78.4**	**79.2**	77.6
Claude-Opus-5	72.9	67.8	78.2	77.4	50.2	70.2	63.0	74.0	86.8	67.2	94.6	47.2	77.1	75.4	77.8	**77.8**
Gemini-3.1-Pro	69.4	**70.7**	**87.2**	81.0	51.6	69.6	64.0	68.7	83.0	54.0	94.6	43.0	68.8	65.6	70.4	70.2
Qwen3.5-9B	**69.8**	**70.9**	93.8	75.8	46.4	**68.4**	**70.2**	**70.2**	**83.8**	**60.2**	93.2	**43.6**	68.4	68.4	69.8	67.0
Qwen3.8-27B	68.0	64.8	74.0	72.6	**51.0**	65.6	60.8	69.8	81.2	59.8	96.2	42.2	**69.5**	**69.6**	**71.0**	**68.0**
InternVL3.5-8B	62.0	58.5	52.2	76.4	45.2	58.0	60.6	64.6	72.0	50.8	97.2	38.4	63.0	63.6	64.4	61.0
Qwen2.5-VL-7B-Instruct	55.4	55.2	68.2	67.8	35.8	54.8	49.0	55.9	62.8	35.2	95.2	30.2	55.1	54.4	57.2	53.8
VideoLLaMA2.1-7B	49.2	60.6	**99.4**	**77.4**	29.8	57.8	38.8	41.8	52.2	10.4	91.2	13.2	45.1	47.2	45.0	43.0
Tarsier2-7B	41.6	41.3	61.4	48.0	19.6	38.2	39.4	41.6	49.8	21.0	76.0	19.6	42.1	44.0	40.4	41.6
VideoLLaMA3-7B	31.2	20.5	30.8	24.2	10.0	21.4	16.4	40.7	37.6	14.2	**99.6**	11.2	32.4	32.6	33.0	31.6
GPT-5.6-Terra	**73.9**	**67.4**	**51.6**	**80.4**	**60.8**	**76.2**	**68.2**	**74.3**	**87.2**	**68.4**	**90.8**	**50.4**	**79.9**	**81.0**	**77.6**	**81.0**
Gemini-3.1-Pro	62.7	56.7	21.4	79.8	57.2	62.0	62.8	61.6	77.0	51.2	78.0	40.0	69.8	67.4	69.6	72.2
Claude-Opus-5	59.7	51.2	19.6	70.6	46.8	56.6	62.2	58.6	73.4	53.2	71.0	37.2	69.4	67.6	69.0	71.6
Qwen3.8-27B	**58.5**	48.6	16.4	66.0	49.6	55.0	56.0	**60.8**	71.6	52.2	81.0	38.2	**66.2**	**67.4**	**65.4**	**65.8**
InternVL3.5-8B	58.3	**55.0**	**59.8**	**71.6**	45.6	50.0	47.8	60.1	66.4	45.6	93.2	35.0	59.8	61.4	59.4	58.6
Qwen3.5-9B	57.2	48.3	3.0	70.4	**50.6**	**56.6**	**60.8**	59.0	**72.0**	**52.8**	72.6	**38.8**	64.2	65.4	63.4	64.0
Qwen2.5-VL-7B-Instruct	47.9	42.4	28.2	60.0	34.2	40.6	49.0	50.7	55.6	33.2	87.0	27.2	50.7	51.6	51.0	49.4
VideoLLaMA2.1-7B	38.1	36.9	0.6	65.8	27.2	41.6	49.2	36.6	46.2	11.4	77.4	11.6	40.9	43.8	39.2	39.6
Tarsier2-7B	36.4	35.4	28.2	52.0	18.2	33.6	44.8	34.6	44.0	19.0	58.2	17.0	39.1	41.4	37.4	38.6
VideoLLaMA3-7B	27.1	12.5	0.0	19.0	8.6	16.8	17.8	38.4	32.8	13.6	**97.8**	9.4	30.5	31.2	30.6	29.6

Evaluation with Ground-Truth Temporal Windows. We additionally evaluate models with the annotated risk interval supplied as a frame-level temporal hint. The interval is mapped to the nearest uniformly sampled frames and appended to the perception, cognition, and planning prompts; models still process the complete video rather than a temporally cropped clip. We therefore exclude Temporal from evaluation and renormalize the score as the equal-weight mean of Perception, Cognition, and Planning. Tables 9 and 10 report abnormal and risk-only results for real-world and generated videos, respectively. The supplied window improves most models on both settings, while risk-only videos remain more challenging because recognizing an emerging risk requires interpreting weaker visual evidence even after its temporal location is known.

Table 10: Results with ground-truth temporal-window hints on generated abnormal and risk-only videos. Temporal localization is excluded because the annotated window is supplied. All values are percentages; "AVG" denotes the within-dimension mean.

Model	Score	Perception						Cognition					Planning			
		AVG	Sta	Typ	Src	Act	Obj	AVG	Desc	Csq	Fact	Caus	AVG	Per	Haz	Ovr
GPT-5.6-Terra	**77.0**	**76.5**	78.8	**80.0**	**69.4**	**74.6**	**79.6**	**77.7**	**84.6**	**69.4**	**99.4**	**57.4**	**76.9**	**75.6**	**80.4**	74.8
Gemini-3.1-Pro	73.7	75.9	**95.4**	78.4	63.2	63.6	78.8	74.5	83.0	64.8	97.6	52.8	70.8	65.0	71.6	**75.6**
Claude-Opus-5	65.9	68.4	92.6	70.6	49.8	64.0	64.8	63.3	70.2	54.4	85.6	43.2	65.9	63.6	65.8	68.2
Qwen3.8-27B	**75.4**	75.9	92.8	**80.8**	**62.8**	**66.4**	76.4	**74.8**	84.8	**63.8**	97.8	**53.0**	**75.4**	**72.6**	**78.6**	**75.2**
Qwen3.5-9B	74.0	**76.4**	97.2	**80.8**	61.0	64.8	**78.2**	73.4	**85.2**	61.0	97.2	50.0	72.3	69.0	76.4	71.6
InternVL3.5-8B	63.6	61.4	56.6	78.6	52.2	53.2	66.4	65.4	69.6	51.2	**98.0**	42.8	64.0	63.2	66.6	62.4
Qwen2.5-VL-7B-Instruct	55.8	55.0	68.2	65.8	42.6	43.8	54.4	58.0	60.2	40.0	97.0	34.6	54.5	53.4	56.6	53.6
Tarsier2-7B	49.5	54.9	77.8	67.6	35.2	46.4	47.2	48.7	55.2	29.6	81.2	28.8	44.8	45.2	44.0	45.4
VideoLLaMA2.1-7B	48.2	58.2	**99.8**	69.4	33.0	48.2	40.8	43.0	50.0	21.4	80.2	20.2	43.5	44.6	42.6	43.2
VideoLLaMA3-7B	43.8	40.3	61.0	42.6	24.4	32.2	41.0	48.7	48.4	27.4	96.0	23.0	42.5	42.0	43.6	42.0
GPT-5.6-Terra	**75.4**	**74.0**	**80.0**	**76.2**	**68.2**	**66.8**	79.2	**76.1**	**84.0**	**75.2**	**89.0**	**56.2**	**75.9**	**75.2**	**78.4**	74.2
Gemini-3.1-Pro	68.2	64.9	44.4	73.2	65.2	61.8	**79.8**	66.0	77.0	71.8	63.0	52.4	73.6	70.4	74.4	76.2
Claude-Opus-5	67.2	63.1	49.0	71.4	56.8	66.6	71.8	64.6	74.6	70.6	61.2	52.2	73.8	70.8	74.0	**76.4**
Qwen3.8-27B	**66.9**	62.0	43.0	**72.6**	**63.0**	**57.8**	**73.2**	**66.7**	**76.4**	**69.8**	69.2	**51.4**	**72.1**	**68.8**	**74.2**	**73.2**
InternVL3.5-8B	64.3	**64.3**	83.8	71.0	52.2	51.4	63.2	66.0	70.8	57.8	92.8	42.8	62.7	62.4	64.0	61.6
Qwen3.5-9B	59.3	53.6	20.4	68.2	57.4	54.4	67.8	58.1	69.6	62.2	55.2	45.2	66.3	63.2	69.4	66.4
Qwen2.5-VL-7B-Instruct	46.4	39.7	38.2	49.0	36.4	35.2	39.6	51.8	53.2	38.8	87.0	28.4	47.7	47.2	48.8	47.0
Tarsier2-7B	39.8	41.1	57.2	50.4	25.6	34.0	38.0	37.0	44.8	23.4	59.0	20.8	41.2	41.8	40.0	41.8
VideoLLaMA2.1-7B	33.4	33.3	**99.8**	69.4	33.0	48.2	40.8	43.0	50.0	21.4	80.2	20.2	43.5	44.6	42.6	43.2
VideoLLaMA3-7B	43.8	40.3	61.0	42.6	24.4	32.2	41.0	48.7	48.4	27.4	**96.0**	23.0	42.5	42.0	43.6	42.0

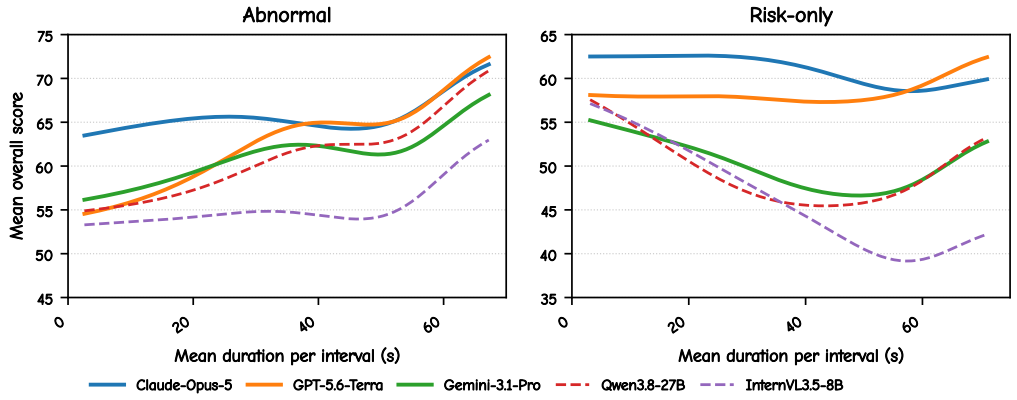

Figure 7: Weighted-smoothed mean overall score within fixed video-duration intervals for the three top proprietary and two top open-source models on real-world videos shorter than 90 seconds. Interval video counts are used as smoothing weights. The left and right panels show abnormal and risk-only videos, respectively. Solid and dashed lines denote proprietary and open-source models.

A.7 Further Analysis

Duration–Accuracy Relationship on Videos. To examine whether video duration affects model performance, we match each per-video overall score with its corresponding video duration and restrict the analysis to videos shorter than 90 seconds. We focus on the three top proprietary models (Claude-Opus-5 Anthropic (2026), GPT-5.6-Terra OpenAI (2026b), and Gemini-3.1-Pro Google DeepMind (2026)) and the two top open-source models (Qwen3.8-27B Qwen Team (2026) and InternVL3.5-8B Wang et al. (2025b)). The curves reveal distinct duration-dependent patterns for the two risk states. For abnormal videos, the score generally increases with duration, suggesting that longer videos provide more temporal context for interpreting a risk that has already occurred. Risk-only videos exhibit a different pattern: the score initially decreases as duration grows, likely because additional normal content dilutes the weak visual cues of an emerging risk. Beyond 60 seconds, however, the score partially recovers, possibly because a risk precursor that persists longer

provides more repeated evidence for recognition. Since video duration is also associated with source and content complexity, these findings should be interpreted as descriptive associations rather than causal effects. Figure 7 visualizes these opposing duration–accuracy trends across models.