MAVEN: A Multi-stage Agentic Annotation Pipeline for Video Reasoning Tasks

Han Zhang Wanting Jiang Tomasz Kornuta Tian Zheng Vidya Murali
NVIDIA

Abstract

Training Vision Language Models (VLMs) for video event reasoning requires high-quality structured annotations capturing not only what happened, but when, where, why, and with what consequence, at a scale manual labelling cannot support. We present MAVEN *(Multi-stage Agentic Video Event aNnotation), a multi-stage agentic pipeline that turns raw videos into multi-task training data with Chain-of-Thought (CoT) reasoning traces, organized around a designated* Event of Focus*. At its core,* MAVEN *synthesizes a* Multi-Scale Spatio-Temporal Event Description *(MSTED) from three complementary caption levels; this explicit intermediate serves as the sole input to downstream Q&A generation across multiple task formats. Crucially,* MAVEN *supports* agent-driven domain adaptation*: given a new video dataset and target question examples, the agent redesigns all prompts top-down without manual re-engineering. A hierarchical refinement loop further classifies annotation errors against a taxonomy, traces root causes to the originating pipeline stage, and applies targeted edits that rewrite prompts or modify the pipeline structure itself, iteratively improving data quality. We apply* MAVEN *to label over 5,300 traffic videos and fine-tune Cosmos-Reason2-8B on the resulting data. On a private CCTV evaluation set, fine-tuning surpasses both Gemini 2.5 Pro and 3.1 Flash, including a* $+38.8$*-point gain in MCQ accuracy over zero-shot. On AccidentBench, CCTV-only training lifts Cosmos-Reason2 by* $+10.7$ *MCQ points and matches Gemini 2.5 Pro despite seeing no dashcam videos; adding agent-adapted dashcam annotations narrows the gap to Gemini 3.1 Flash, and RL post-training pushes overall performance past both Gemini baselines. Qualitative results on warehouse surveillance and public safety videos further show the agentic workflow readily adapts the pipeline to new domains.*

1. Introduction

Training Vision Language Models (VLMs) for video event reasoning poses a common analytical demand: understanding *who* was involved, *what* sequence of events unfolded, *where* and *when* key moments occurred, and *why* the event happened. This structured causal reasoning is critical for intelligent transportation systems, workplace safety monitoring, and physical AI, yet remains precisely where recent VLMs fall short despite strong general video understanding.

The central challenge is training data. Structured chain-of-thought (CoT) annotations for video events (descriptions capturing temporal dynamics, spatial relationships, causal factors, and multi-step reasoning) are expensive to produce at scale. Existing auto-labeling approaches [3, 13, 23] either rely on flat single-pass video descriptions that permanently lose fine-grained detail, or are constrained to fixed taxonomies and domain-specific sensor inputs, limiting generalizability. None produce a reusable intermediate event representation that can ground diverse downstream task formats from a single annotation pass.

We present MAVEN (Multi-stage Agentic Video Event aNnotation), a multi-stage agentic pipeline that addresses this gap. We define the *Event of Focus* (EoF) as any notable event in video, whether anomalous or routine within its domain, that the pipeline should characterize and generate training data around. The core design principle is to construct the most complete structured representation of the scene and Event of Focus *before* generating any downstream annotation. MAVEN proceeds in three stages: (1) three-level video captioning capturing global context, dense timestamped events, and fine-grained chunk-level detail; (2) synthesis of a *Multi-Scale Spatio-Temporal Event Description* (MSTED) that consolidates all caption levels into a structured characterization of the Event of Focus; and (3) generation of multi-task CoT Q&A (MCQ, binary verification, and open-ended) grounded solely on the MSTED. Because all downstream annotations derive from the MSTED rather than the raw video, the Q&A generator cannot hallucinate details absent from the structured representation, and human reviewers can audit the MSTED as a natural verification checkpoint before large-scale Q&A generation.

Crucially, MAVEN supports *agent-driven domain adaptation*: given target benchmark question examples and a new video domain description, an agent consultation workflow redesigns prompts top-down across all pipeline stages, adapting the pipeline to new video domains, camera views, event types, and question styles automatically without manual re-

engineering.

Beyond one-shot adaptation, MAVEN supports *hierarchical pipeline refinement* from human feedback. When reviewers identify systematic annotation issues, the agent classifies errors against a structured taxonomy, traces each root cause through the pipeline hierarchy to the originating stage, and applies targeted fixes: rewriting prompts for gaps the current configuration misses, or inserting new pipeline stages for structural limitations that prompt changes alone cannot address. This distinguishes MAVEN from static pipelines that degrade silently when applied to challenging video distributions.

We apply MAVEN to label over 5,300 traffic videos (3,841 CCTV and 1,500 dashcam) and fine-tune Cosmos-Reason2 [18] on the resulting data. On our private CCTV evaluation set, fine-tuning yields $+38.8$, $+35.0$, and $+24.1$ point improvements in MCQ accuracy, verification accuracy, and open-ended score over zero-shot, respectively, surpassing both Gemini 2.5 Pro and Gemini 3.1 Flash [8] on all three metrics. On the public AccidentBench [11] benchmark, our CCTV-trained model, which has never seen dashcam footage during training, matches Gemini 2.5 Pro, demonstrating that the structured CoT reasoning capability induced by the pipeline is *generalizable* rather than domain-specific. Adding agent-adapted dashcam annotations narrows the gap to Gemini 3.1 Flash, and RL post-training pushes overall performance past both Gemini baselines while CCTV evaluation set performance remains stable. We additionally demonstrate qualitative generalization of the agentic pipeline to warehouse surveillance and public safety domains.

Contributions:

- **Pipeline.** MAVEN: a multi-stage agentic pipeline producing structured MSTED descriptions and multi-task CoT Q&A from raw videos, structured around Events of Focus with an explicit intermediate representation that avoids the information loss of single-pass approaches.
- **Agentic domain adaptation.** An agent consultation workflow, packaged as a single Agent Skill [1], that adapts the pipeline to new domains and question styles given only a domain description and example questions, requiring no manual prompt engineering.
- **Hierarchical refinement.** A structured three-stage refinement process (error taxonomy classification, root cause tracing through the pipeline hierarchy, and targeted configuration edits) that distinguishes prompt gaps from structural limitations and resolves each appropriately.
- **Dataset.** 3,841 CCTV and 1,500 dashcam traffic videos labeled with diverse CoT training data across three task formats; qualitative demonstrations on warehouse and public safety domains.
- **Results.** Empirical evidence that structured intermediate representations enable domain-general reasoning: CCTV-only training matches Gemini 2.5 Pro on dashcam benchmarks, and agent-adapted dashcam training with RL post-training exceeds both Gemini baselines without degrading in-domain performance.

2. Related Work

2.1. Video Anomaly Datasets and Benchmarks

Prior video anomaly datasets largely target frame- or clip-level classification rather than structured reasoning. UCF-Crime [21] established the weakly-supervised surveillance paradigm with 1,900 videos across 13 categories, while CADP [20] provides spatio-temporal annotations. None include the causal reasoning chains (why the accident occurred, what behaviors contributed, what followed) needed to train reasoning VLMs.

Recent VLM-oriented benchmarks have begun to fill this gap. AccidentBench [11] provides $\sim$19,000 human-annotated MCQ pairs stratified by difficulty and reasoning type, including temporal reasoning, spatial reasoning, and intent goal reasoning; its land split is our primary public benchmark. SurveillanceVQA-589K [16] shows the scale achievable with AI-assisted labeling for open-ended Q&A. These works evaluate reasoning but do not themselves generate training data; MAVEN addresses this gap by producing diverse CoT Q&A grounded in structured event representations.

2.2. Auto-labeling and Training Data for Video VLMs

General-purpose video-language models [4, 22] show strong video understanding, but structured causal reasoning (fault attribution, temporal localization, consequence prediction) remains weak without targeted fine-tuning. The bottleneck is data: chain-of-thought annotation at scale requires either prohibitive human effort or automated pipelines that preserve fine-grained detail. The dominant approach uses stronger models to label reasoning data for weaker, deployable ones. Cosmos-Reason1 [3] compresses each video into a single global description before synthesizing Q&A, which is scalable but lossy. VAD-Reasoning [13] concatenates per-frame captions and prompts an LLM for anomaly explanations, missing events between sampled frames. Alpamayo-R1 [23] targets ego-centric autonomous driving within a closed taxonomy and requires proprietary sensor metadata unavailable in general surveillance, while VLM-AutoDrive [6] post-trains VLMs on dashcam safety-critical events within a fixed task formulation. Closest to our data pipeline, LongVILA-R1 [7] chunks long videos, captions each chunk, and uses an LLM to consolidate the chunk captions into CoT training data for long-video reasoning. Though not directly comparable, these designs inspired MAVEN's multi-scale captioning and its emphasis on an explicit intermediate representation.

MAVEN differs from these approaches structurally in two

ways. First, rather than flat chunk captions alone, MAVEN produces a *hierarchical* three-level decomposition that is mutually corrective across scales. Second, rather than consolidating directly into Q&A, MAVEN synthesizes an explicit structured intermediate (the MSTED) that serves as a verification checkpoint and the sole input to downstream Q&A generation, avoiding irrecoverable information loss while enabling multiple task formats from a single annotation effort.

2.3. Agentic Pipelines for Video Understanding

LLM-based agents are an emerging tool for both video analysis and data curation. At inference time, several frameworks apply agentic reasoning to anomaly detection: QVAD [5] treats VLM-LLM interaction as a dynamic dialogue, PANDA [26] deploys a planning-and-reflection "AI Engineer" for VAD, Follow the Rules [25] translates normality definitions into textual rules, and VERA [27] optimizes guiding questions offline. These improve how models reason about anomalies at test time but do not address the upstream problem of generating structured training data.

At data-generation time, Colon-Bench [12] introduces a multi-stage agentic pipeline for dense colonoscopy annotation, integrating temporal proposals, tracking, AI confirmation, and human review; we share their orchestration principle but target structured reasoning annotations rather than spatial detection. Pipeline-level prompt optimization has also been explored outside video: single-stage optimizers such as DSPy [14] and OPRO [24] address individual tasks, while multi-stage optimizers like MIPRO [19] treat the pipeline as a black box, unable to trace errors to their originating stage or distinguish prompt deficiencies from model limitations. MAVEN's agent, by contrast, performs backward inference from target tasks to derive stage-level requirements and modifies pipeline *structure* (not just prompt text) based on human feedback. Unlike inference-time agents (QVAD, PANDA), which operate on fixed models, our agent operates at *pipeline design time*; unlike Colon-Bench's fixed pipeline, the MAVEN agent reconfigures top-down for each new domain without manual prompt engineering, making domain adaptation a first-class capability.

3. Method

3.1. The MAVEN Pipeline

MAVEN (Figure 1) transforms raw videos into structured CoT training data organized around *Events of Focus* (EoF), defined as any notable events in the video whether anomalous or routine, through three sequential stages. An EoF may be anomalous (*e.g.*, a traffic accident in smart-city surveillance, or a worker safety incident in warehouse monitoring) or routine (*e.g.*, a pedestrian crossing at a signalized intersection, or a pick-and-place operation at a workstation); what matters is that it is the salient event the pipeline characterizes and generates training data around. For each video, the EoF is selected automatically as the most salient event surfaced during Stage 1 captioning; when a source dataset provides an explicit event label (*e.g.*, accident class), that label is used to anchor the EoF.

The pipeline operates on a configuration (prompts and structure) produced by the top-down adaptation workflow (Section 3.2). When the hierarchical refinement loop (Section 3.3) identifies a structural limitation, it can insert new stages or modify existing ones without altering the core three-stage design. The key design principle is to construct the most complete structured representation of the scene and event *before* generating any downstream annotation, factoring the challenges of video understanding, event synthesis, and task generation into distinct stages with explicit intermediate representations.

Stage 1: Three-level video captioning. A single-pass video caption cannot simultaneously capture global scene context, precise event timing, and fine-grained local detail.

We generate three complementary levels using a video VLM (*e.g.*, Gemini 3.1 Pro): (i) *Global caption*: a holistic description capturing scene layout, weather, time of day, and pre-event baseline behavior, establishing the context against which a notable behavior is recognized; (ii) *Dense caption*: a temporally grounded event-level description pairing each major event with start and end timestamps, providing the causal chain and timing needed for reasoning; (iii) *Chunk captions*: fine-grained captions over short video segments (5–30s depending on video length), recovering subtle behaviors, small objects, and brief decisive moments that a global caption of a long clip would miss. These three levels are designed to be mutually correcting: global captions provide context that disambiguates vague chunk-level descriptions, chunk captions recover details that dense captions underspecify, and dense captions impose temporal structure that organizes chunk-level observations.

Stage 2: MSTED synthesis. An LLM (*e.g.*, Gemini 3.1 Flash) consolidates all three caption levels into the *Multi-Scale Spatio-Temporal Event Description* (MSTED), consisting of three parts: (1) *Holistic Scene Description*: environment, weather, time of day, scene layout, and pre-event baseline behavior; (2) *Temporal and Spatial Localization*: a chronological narrative of the event's progression with precise start/end timestamps and spatial region; (3) *Event of Focus Description*: a structured characterization of event category, temporal and spatial properties, root cause, and consequences (open-ended, with no predefined taxonomy).

The MSTED serves two critical roles. First, it acts as a **verification checkpoint**: because the MSTED explicitly characterizes all salient properties of the event in a structured form, human annotators or automated validation can review

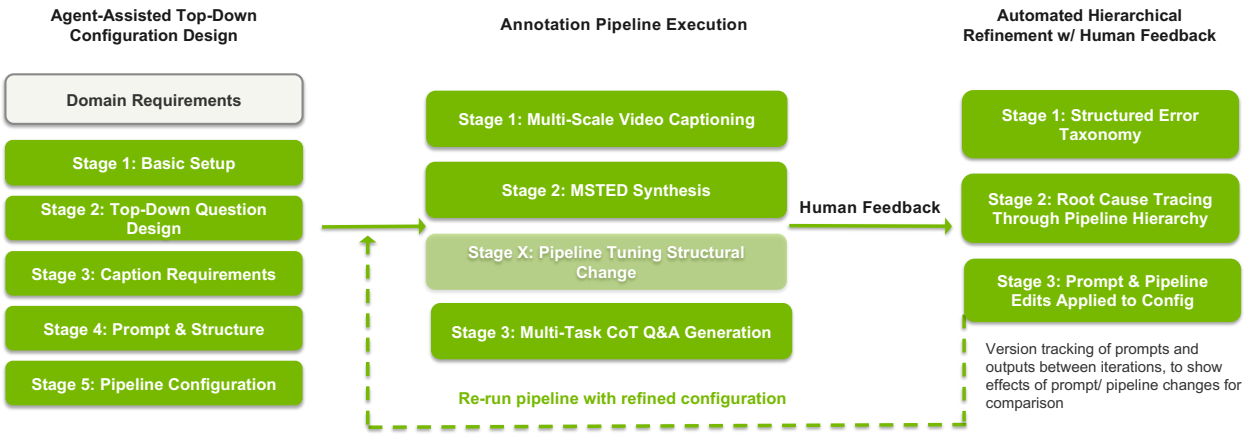

Figure 1. The MAVEN pipeline, organized into three components: agent-assisted top-down configuration design (left), annotation pipeline execution (center), and hierarchical pipeline refinement with human feedback (right).

it for completeness and accuracy before Q&A generation, preventing error propagation into the training data. Second, it is the **sole input** to Stage 3: Q&A generators never see the raw video or original captions, only the MSTED. This factorization ensures that all generated questions are answerable from explicitly captured information; the model cannot hallucinate details absent from the structured representation.

Stage 3: Multi-task CoT Q&A generation. A second LLM pass takes the MSTED as sole context and generates three task formats, each with an explicit reasoning trace: **MCQ** (4-option, with step-by-step reasoning on timestamps and spatial locations); **binary verification** (yes/no questions with supporting reasoning); **open-ended QA** (free-form questions requiring causal or descriptive reasoning). The task types and question formats are fully configurable via prompts; no architectural changes are required.

3.2. Agent-Driven Domain Adaptation

Adapting the pipeline to a new domain (*e.g.*, from CCTV to dashcam footage) or new task types typically requires manual prompt re-engineering at each stage, a time-consuming process requiring domain expertise. MAVEN automates this by packaging the pipeline as a single Agent Skill [1]. Following recent work [15] showing that a multi-agent committee can be compiled into an equivalent single-agent-with-skills system at substantially lower token and latency cost, we adopt a single-orchestrator design. The skill is instantiated via the `opencode` CLI harness[1] backed by Claude Opus 4.6 [2] with file-read, file-write, and bash tool access. The Agent Skill abstraction is model- and harness-agnostic, so the same skill can be executed under Claude Code, Codex, or any other harness that supports Agent Skills.

The agent reads the base pipeline configuration files, the domain description, and the target question examples as context, and writes back updated prompts and any structural edits. Given (1) the base pipeline configuration, (2) a target domain description, and (3) desired question types, the agent performs *backward inference*: what temporal granularity, spatial relationships, and causal depth must the MSTED capture to make those questions answerable? From this analysis, it derives a per-stage must-capture checklist, ensuring that each captioning stage produces what downstream stages will need, then rewrites all prompts to satisfy these requirements, adjusting for domain-specific visual characteristics, camera perspectives, and event semantics.

For example, when adapting from CCTV to dashcam footage targeting AccidentBench [11], the agent adjusted captioning prompts for ego-vehicle perspective and motion dynamics, and strengthened Q&A generation to emphasize intent attribution and temporal reasoning, question types that dominate the target benchmark. The result is a complete domain-adapted pipeline configuration that can be applied to new video batches without further human intervention.

3.3. Hierarchical Pipeline Refinement with Human Feedback

Beyond one-shot domain adaptation, MAVEN supports iterative pipeline refinement through natural-language human feedback. When reviewers identify systematic issues in the generated annotations, the agent diagnoses root causes and updates the pipeline configuration accordingly, modifying not just prompt text but pipeline *structure*. This refinement proceeds through three structured stages.

Stage 1: Structured Error Taxonomy. Given human feedback and sampled annotation outputs, the agent classifies each discrepancy against a fixed error taxonomy: *misinformation*, *hallucination*, *missing information*, *temporal error*, *spatial error*, and *attribution error*. This structured

[1] https://opencode.ai

classification prevents over-diagnosis: not every output error requires a pipeline change.

Stage 2: Root Cause Tracing Through Pipeline Hierarchy. For each identified error, the agent traces responsibility back through the pipeline hierarchy, from Q&A output through the MSTED to the specific captioning level where the error originates. It classifies the root cause as either a *prompt gap* (the information was capturable but the prompt failed to elicit it, fixable by prompt modification) or a *system limitation* (the information cannot be recovered from the existing pipeline stages, requiring structural intervention).

Stage 3: Prompt & Pipeline Edits Applied to Configuration. Prompt gaps are resolved by rewriting the relevant prompt; system limitations trigger a structural change, inserting a new stage or modifying an existing one. For instance, when a reviewer noted that uniform-length chunking sometimes splits events across chunk boundaries and yields inaccurate chunk captions, Stage 2 diagnosed this as a system limitation and Stage 3 introduced an additional captioning stage supplementing the chunk captions: *event-centered highlight chunks*. An LLM first identifies the key event timestamp from the existing captions, then a variable-duration segment is extracted around it for targeted re-captioning, producing more accurate descriptions of the event itself and improving MSTED temporal fidelity.

3.4. Training Dataset

CCTV corpus. We apply MAVEN to 3,841 open-source traffic videos from roadside CCTV cameras: 808 accident videos and 3,033 normal traffic videos. Each video is labeled with all three task formats, yielding 3,841 MCQ, 7,682 binary verification, and 3,841 open-ended QA samples with full CoT reasoning traces.

Dashcam corpus (agent-adapted). Using the agent-adapted pipeline described in Section 3.2, we label 1,500 dashcam collision videos from the Nexar dataset [17]: 750 accident videos and 750 normal videos. The agent-redesigned prompts target AccidentBench-style questions, producing 11,200 MCQ samples with matching difficulty gradations and task type coverage.

Evaluation sets. We evaluate on two held-out sets: (i) *Private CCTV evaluation set*: 80 traffic CCTV videos sourced from YouTube with human-labeled MSTEDs and verified Q&A, totaling 80 MCQ, 160 binary verification, and 80 open-ended samples. Focused on causal reasoning: fault attribution, root cause identification, and consequence characterization. (ii) *AccidentBench* [11] (land split): 1,630 videos with 17,069 human-annotated MCQ at three difficulty levels (easy, medium, hard) and three task types (temporal, spatial, intent). Short videos (1,500) are exclusively dashcam views, while medium-length (58) and long (70) videos include a mixture of dashcam and CCTV footage.

3.5. Training Protocol

We fine-tune Cosmos-Reason2-8B (CR2) [18] on MAVEN-labeled data following the now-standard SFT-then-RL protocol adopted in recent video and multimodal reasoning work [3, 7, 9, 13, 23]: supervised fine-tuning (SFT) followed by reinforcement learning (RL).

SFT. We fine-tune the full model for 3 epochs with a learning rate of $1e{-}5$, global batch size 512, using $8\times$ A100 GPUs. Video frames are sampled at rate 2 up to 128 frames per video. Each training example contains the sampled video frames, the question, and the full CoT & answer; loss is computed on the entire CoT and answer tokens so the model learns to mimic the style and depth of the generated reasoning traces before outputting a final answer across all three task formats.

RL (DAPO). On top of the intermediate SFT checkpoint after 1 epoch, we apply DAPO [28] for 1000 steps with learning rate $1e{-}6$. For each prompt we sample $n{=}16$, with a prompt batch of 256 and mini-batch size 2. The policy objective uses a symmetric clip range $\epsilon_{\text{low}}{=}\epsilon_{\text{high}}{=}0.2$ and KL coefficient $\beta = 0.01$. The reward is a weighted sum of (i) format correctness ($w_{\text{format}} = 0.2$), enforcing the reasoning-trace schema with thinking tags, and (ii) answer accuracy ($w_{\text{acc}} = 1$), exact match for MCQ and binary verification. RL is applied only to MCQ and binary verification; open-ended performance is evaluated but not RL-optimized.

4. Experiments

4.1. Experimental Setup

Following the training protocol in Section 3.5, we report three model variants of CR2: **+ CCTV SFT** (SFT on CCTV data only), **+ Dashcam SFT** (SFT on CCTV + agent-adapted dashcam data), and **+ RL** (RL post-training on CCTV + dashcam data, initialized from the + Dashcam SFT intermediate checkpoint). We compare these against zero-shot CR2, Gemini 2.5 Pro, and Gemini 3.1 Flash [8], all evaluated with the same prompts. Evaluation is conducted on the private CCTV evaluation set and the AccidentBench land split (Section 3.4).

4.2. Private CCTV Evaluation Set

Table 1 shows results on our private CCTV evaluation set. + CCTV SFT yields dramatic improvements over zero-shot CR2 across all three task formats ($+38.8$ MCQ, $+35.0$ verification, $+24.1$ open-ended points) and surpasses both Gemini 2.5 Pro and Gemini 3.1 Flash on all three metrics. The open-ended gap is particularly striking: zero-shot CR2 and Gemini 2.5 Pro both score below 16% on rescaled BertScore F1, likely because their default answer style diverges from

Model	Training	MCQ	Verif.	Open
Gemini 2.5 Pro	0-shot	82.50	76.25	15.60
Gemini 3.1 Flash	0-shot	80.00	70.63	23.20
CR2	0-shot	47.50	50.00	15.37
+ CCTV SFT	SFT	86.25	**85.00**	**39.45**
+ Dashcam SFT	SFT	86.25	83.75	**39.47**
+ RL	SFT+RL	**88.75**	81.25	37.29

Table 1. Results on the private CCTV evaluation set (80 videos). MCQ and verification are reported as accuracy (%); open-ended as rescaled BertScore F1 (%).

the reference answers generated by our pipeline, whereas fine-tuning on MAVEN data aligns the model's output distribution with the target answer format.

Adding agent-adapted dashcam data (+ Dashcam SFT) keeps CCTV performance near-stable (Verif dips by 1.25 points; MCQ and open-ended remain essentially unchanged) despite the domain shift, indicating that the pipeline-generated labels are complementary rather than conflicting. + RL further improves MCQ accuracy ($86.25 \rightarrow 88.75$) with slight decreases in verification and open-ended, consistent with the RL reward signal targeting only MCQ and verification tasks. We note that the open-ended score is rescaled BertScore F1 against pipeline-generated reference answers, which rewards lexical overlap with the training distribution; it should be read as a directional signal rather than an absolute reasoning-quality measure.

4.3. AccidentBench

Table 2 shows MCQ accuracy on AccidentBench land split across all nine video-length-by-difficulty cells, with columns grouped by video length.

+ CCTV SFT lifts the backbone across all video lengths. + CCTV SFT improves over zero-shot CR2 by $+10.7$ points overall ($29.9 \rightarrow 40.6$), with consistent gains on every length bin: Short $32.4 \rightarrow 42.4$, Medium $34.2 \rightarrow 42.4$, and Long $23.1 \rightarrow 36.9$. This demonstrates that the training signal from structured CoT reasoning over CCTV events transfers to the Cosmos-Reason2 backbone well beyond the training distribution, including the all-dashcam Short bin.

Domain generalization without dashcam data. + CCTV SFT matches Gemini 2.5 Pro overall (40.6% vs. 40.3%) and approaches Gemini 3.1 Flash (42.8%) despite seeing no dashcam videos during training. On long videos it exceeds Gemini 2.5 Pro by 4.4 points and approaches Gemini 3.1 Flash; on medium videos it matches Gemini 2.5 Pro and trails Gemini 3.1 Flash by only 2.0 points. On short videos, it still trails both Gemini baselines, reflecting the residual domain gap that CCTV-only training cannot fully cover and motivating the agent-adapted dashcam corpus next.

Agent-adapted dashcam data closes the short-video gap. Adding dashcam labels generated by the agent-adapted pipeline drives the largest improvement precisely where it is needed: Short-Avg rises from 42.4 to 47.9, now surpassing both Gemini baselines (46.9 and 46.4), with gains across all difficulty levels. Overall accuracy rises to 42.0%, approaching Gemini 3.1 Flash; medium- and long-video performance remains comparable to + CCTV SFT, indicating that the dashcam labels add short-video capability without displacing the CCTV-domain reasoning already learned.

RL further amplifies reasoning on short videos and hard questions. + RL achieves the highest overall accuracy at 44.2%, exceeding both Gemini 2.5 Pro ($+3.9$) and Gemini 3.1 Flash ($+1.4$). The gains concentrate on short videos and on the hard-level questions (Short-Hard: $26.1 \rightarrow 37.7$; Medium-Hard: $34.5 \rightarrow 38.3$), consistent with the RL reward targeting MCQ accuracy on questions requiring multi-step causal reasoning.

4.4. Cross-Domain Pipeline Generalization

To demonstrate that the agentic consultation workflow generalizes beyond traffic, we apply MAVEN to two additional domains by providing only a new domain description to the agent. *Public safety*: crowded-scene footage of behavioral anomalies (*e.g.*, fights, crowd surges); the agent adjusts for dense multi-person tracking, role attribution, and intent-level reasoning. *Warehouse surveillance*: indoor CCTV of worker safety and security anomalies (*e.g.*, falls, unsafe zone violations, unauthorized access, theft); the agent redesigns prompts for overhead camera perspective, occlusion from shelving, and safety- and security-rule-grounded behavior analysis. In both cases the pipeline produces structured MSTED outputs and Q&A samples without manual prompt engineering or domain expertise.

Figure 2 shows two representative outputs. The *public safety* example captures a complex altercation in a pedestrian underpass: two loiterers ambush a passing couple, but the intended victim counter-attacks and knocks the initial aggressor unconscious, turning a 2-on-2 ambush into a 1-on-2 defensive counter-attack. This is a challenging multi-actor temporal reasoning case, requiring the pipeline to recover both the role reversal and the causal chain connecting the successive actions, all of which the agent-adapted MSTED correctly identifies. The *warehouse surveillance* example captures a suspicious criminal case: an individual enters an enclosed office, systematically searches two separate workstations, scans the room for observers, picks up a white object and conceals it under their shirt, then exits. The MSTED captures the suspicious-behavior pattern, the deliberate concealment action, and the access-legitimacy reasoning, illustrating that the agent-adapted pipeline recovers fine-grained intent attribution from multiple short actions without any manual prompt engineering.

Model	Training	Short				Medium				Long				Overall
		E	M	H	Avg	E	M	H	Avg	E	M	H	Avg	
Gemini 2.5 Pro	0-shot	63.0	42.8	34.8	46.9	**54.7**	33.9	35.8	41.5	46.0	32.7	18.7	32.5	40.3
Gemini 3.1 Flash	0-shot	63.1	45.1	31.0	46.4	51.5	**42.7**	**38.9**	44.4	48.0	**40.0**	**25.3**	**37.8**	42.8
CR2	0-shot	43.6	32.4	21.3	32.4	40.0	33.0	29.6	34.2	30.3	24.3	14.7	23.1	29.9
+ CCTV SFT	SFT	60.4	40.8	26.1	42.4	52.3	40.5	34.5	42.4	50.2	37.5	23.0	36.9	40.6
+ Dashcam SFT	SFT	66.6	46.9	30.2	47.9	47.7	42.1	35.0	41.6	**52.6**	35.2	21.3	36.4	42.0
+ RL	SFT+RL	**68.4**	**50.3**	**37.7**	**52.1**	53.0	42.6	38.3	**44.6**	51.0	35.4	21.6	36.0	**44.2**

Table 2. AccidentBench land split MCQ accuracy (%). Columns are grouped by video length (Short/Medium/Long); nested columns report difficulty (E/M/H = Easy/Medium/Hard). Short videos are exclusively dashcam; medium and long videos mix dashcam and CCTV footage.

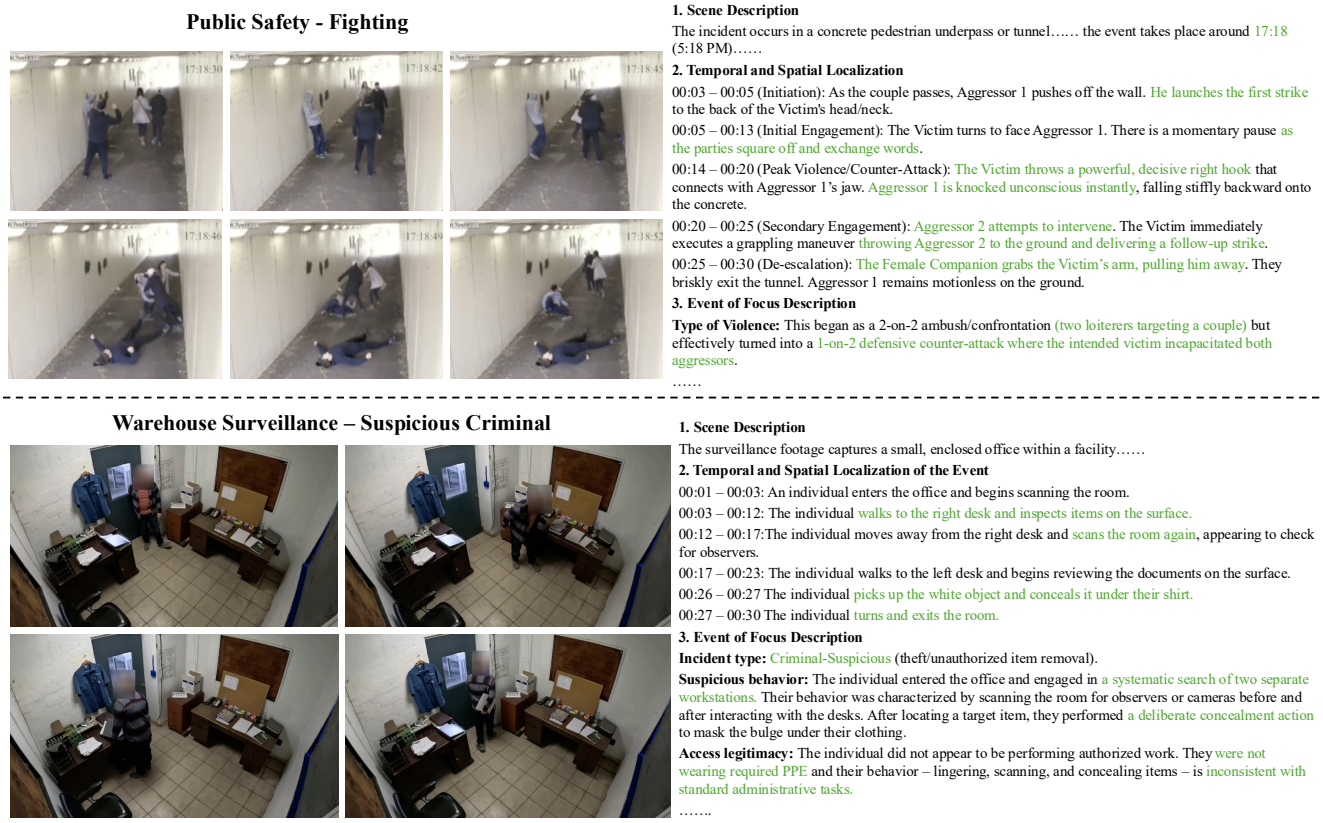

Figure 2. Qualitative MAVEN outputs on two generalized domains. **Top (Public Safety):** a complex altercation in a pedestrian underpass, where an ambush turns into a defensive counter-attack and the initial aggressor is knocked unconscious by the intended victim. **Bottom (Warehouse Surveillance):** a suspicious-criminal case where an individual systematically searches two workstations, conceals a white object under their shirt, and exits. Highlighted spans in the MSTED mark the key information that the agent-adapted pipeline captures.

4.5. Ablations

Does the structured intermediate representation matter? To isolate the contribution of three-level captioning and MSTED synthesis, we compare MAVEN against a *flat* baseline that generates CoT Q&A directly from a single global caption of the same 3,841 CCTV videos, without the dense/chunk caption decomposition and without the MSTED intermediate representation. All other training details are held constant.

Table 3 reports the comparison. MAVEN outperforms the single-pass captioning baseline on all three task formats, with gains of +6.25 MCQ, +11.25 verification, and +3.50 open-ended points. The single-pass baseline itself sits at roughly the Gemini baseline level on MCQ and verification (Table 1), which indicates that a standard caption-then-generate recipe effectively distills the teachers' zero-shot performance. MAVEN's three-level captioning and MSTED capture and organize event information that a single global

Training CoT generation	MCQ	Verif.	Open
Single-pass Captioning	80.00	73.75	35.95
MAVEN	**86.25**	**85.00**	**39.45**

Table 3. Ablation on the CCTV evaluation set: flat single-pass captioning vs. MAVEN (three-level captioning + MSTED). Both variants trained on the same CCTV videos with identical SFT setup.

caption cannot convey; the resulting gain over the single-pass baseline is attributable to pipeline structure rather than to a stronger generator, and allows the 8B student to exceed the Gemini baselines themselves.

4.6. Discussion

On distillation and baseline comparison. Our pipeline uses Gemini 3.1 Pro as the VLM for three-level captioning and Gemini 3.1 Flash as the LLM for MSTED synthesis and CoT Q&A generation, so on the surface the training labels are distilled from Gemini-class models. A natural concern is that the fine-tuned Cosmos-Reason2-8B model simply inherits its teacher's behavior. The ablation in Section 4.5 addresses this at the data-generation level: the same teacher models combined with a single-pass pipeline fall well short of MAVEN, localizing the gain to pipeline structure rather than to the teacher. Evaluation-time evidence further shows that the + RL variant reaches 44.2% overall on AccidentBench, exceeding Gemini 3.1 Flash on an 8B backbone; on the private CCTV evaluation set it also outperforms both Gemini baselines across all three tasks. Therefore, the MSTED structured representation together with DAPO post-training extracts a reasoning signal from the generated data that goes beyond surface-level imitation of the generator.

Structured representations enable cross-domain transfer. Surprisingly, CCTV-only training matches Gemini 2.5 Pro and approaches Gemini 3.1 Flash on AccidentBench, a dashcam-centric benchmark. We attribute this to the MSTED intermediate representation: by forcing the model to reason over structured event characterizations (temporal bounds, spatial locations, causal factors) rather than raw visual features, the training signal captures reasoning patterns that are *domain-invariant*.

Agentic adaptation amplifies generalization. The agent-redesigned prompts for the dashcam corpus target AccidentBench question style and difficulty gradations directly, producing training data whose distribution better matches the evaluation benchmark. The progressive improvements from + CCTV SFT to + Dashcam SFT to + RL demonstrate that data quality (from agent prompt design) and post-training methodology contribute additively, ultimately surpassing both Gemini baselines while maintaining CCTV evaluation set accuracy.

RL disproportionately benefits hard reasoning. The progressive improvement from + CCTV SFT to + RL is not uniform across difficulty levels: Easy improves by $+3.1$ points, Medium by $+3.2$, and Hard by $+4.6$. RL post-training with answer-accuracy rewards particularly strengthens the model on questions requiring multi-step causal inference (root cause identification, intent attribution, and counterfactual reasoning), which dominate the Hard split.

5. Conclusion

We presented MAVEN, a multi-stage agentic pipeline that transforms raw videos into structured CoT training data organized around a designated *Event of Focus*. By synthesizing three complementary caption levels into an explicit MSTED intermediate before generating any Q&A, MAVEN avoids the irrecoverable information loss of single-pass auto-labeling. An agentic consultation workflow and a hierarchical refinement loop together enable top-down domain adaptation and error-driven pipeline evolution without manual re-engineering.

Fine-tuning Cosmos-Reason2-8B on MAVEN-labeled CCTV data yields a $+38.8$ MCQ-point gain on our private evaluation set and matches Gemini 2.5 Pro on AccidentBench despite seeing no dashcam videos, indicating that the induced reasoning capability is generalizable rather than domain-specific. Adding agent-adapted dashcam labels and DAPO post-training pushes the model past both Gemini baselines overall, while CCTV performance remains stable. Qualitative results on warehouse and public safety domains further show that the agentic workflow readily adapts the pipeline given only a domain description.

Limitations and future work. MAVEN currently relies on Gemini-class models for captioning and synthesis, and training gains are validated only on Cosmos-Reason2-8B; a natural next step is cross-backbone validation on other open video-language models (*e.g.*, Qwen-VL [4] and Nemotron series [10]) to confirm that the observed gains transfer beyond a single model. Other planned work includes data-scale ablations on the number of generated questions per video, component-level ablations isolating the contribution of MSTED synthesis, top-down configuration, and hierarchical refinement with quantitative results, as well as cross-domain benchmarks for warehouse and public safety.

At present, MSTED quality is assessed via spot-checking, and each domain adaptation takes 2–4 rounds of human review over 10–20 samples (roughly 1–2 hours). Human reviewers excel at the detailed diagnosis that automated metrics cannot surface, but this dependence limits the scale at which the refinement loop can operate. Our longer-horizon goal is *fully automated closed-loop self-improvement*: using downstream evaluation signals to drive hierarchical refinement without human input. Such a system would automatically attribute errors to their originating pipeline stage, distinguish prompt gaps from system limitations, and apply targeted prompt or structural fixes, progressing toward a truly *self-optimizing* pipeline.

References

[1] Anthropic. Agent skills. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview, 2025. 2, 4

[2] Anthropic. Introducing claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026. 4

[3] Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 1, 2, 5

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 8

[5] Lokman Bekit, Hamza Karim, Nghia T Nguyen, and Yasin Yilmaz. Qvad: A question-centric agentic framework for efficient and training-free video anomaly detection. *arXiv preprint arXiv:2604.03040*, 2026. 3

[6] Mohammad Qazim Bhat, Yufan Huang, Niket Agarwal, Hao Wang, Michael Woods, John Kenyon, Tsung-Yi Lin, Xiaodong Yang, Ming-Yu Liu, and Kevin Xie. Vlm-autodrive: Post-training vision-language models for safety-critical autonomous driving events. *arXiv preprint arXiv:2603.18178*, 2026. 2

[7] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025. 2, 5

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 5

[9] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[10] Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Guo Chen, et al. Nvidia nemotron nano v2 vl. *arXiv preprint arXiv:2511.03929*, 2025. 8

[11] Shangding Gu, Xiaohan Wang, Donghao Ying, Haoyu Zhao, Runing Yang, Ming Jin, Boyi Li, Marco Pavone, Serena Yeung-Levy, Jun Wang, et al. Accidentbench: Benchmarking multimodal understanding and reasoning in vehicle accidents and beyond. *arXiv preprint arXiv:2509.26636*, 2025. 2, 4, 5

[12] Abdullah Hamdi, Changchun Yang, and Xin Gao. Colonbench: An agentic workflow for scalable dense lesion annotation in full-procedure colonoscopy videos. *arXiv preprint arXiv:2603.25645*, 2026. 3

[13] Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *arXiv preprint arXiv:2505.19877*, 2025. 1, 2, 5

[14] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[15] Xiaoxiao Li. When single-agent with skills replace multi-agent systems and when they fail. *arXiv preprint arXiv:2601.04748*, 2026. 4

[16] Bo Liu, Pengfei Qiao, Minhan Ma, Xuange Zhang, Yinan Tang, Peng Xu, Kun Liu, and Tongtong Yuan. Surveillancevqa-589k: A benchmark for comprehensive surveillance video-language understanding with large models. *arXiv preprint arXiv:2505.12589*, 2025. 2

[17] Daniel Moura, Shizhan Zhu, and Orly Zvitia. Nexar dashcam collision prediction dataset and challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2583–2591, 2025. 5

[18] NVIDIA. Cosmos-reason2-8b. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2025. 2, 5

[19] Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 3

[20] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9, 2018. 2

[21] 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*, pages 6479–6488, 2018. 2

[22] 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*, 2025. 2

[23] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025. 1, 2, 5

[24] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[25] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pages 304–322. Springer, 2024. 3

[26] Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *arXiv preprint arXiv:2509.26386*, 2025. 3

[27] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688, 2025. 3

[28] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 5