UNIAR: A UNIFIED MULTI-MODAL BENCHMARK FOR EVIDENCE-GROUNDED ANOMALY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly reasoning is a core perceptual and cognitive ability in humans, who can identify anomalies across distinct modalities and provide corresponding evidence. However, a unified benchmark for rigorously evaluating this ability in Multimodal Large Language Models (MLLMs) across video, image, and text remains lacking. To this end, we present UniAR, a unified benchmark for evaluating the anomaly reasoning capabilities of MLLMs across video, image, and text. Rather than assuming identical anomaly semantics across modalities or requiring multi-modal fusion within each instance, UniAR profiles MLLMs' anomaly reasoning through a shared set of reasoning tasks and human-normalized performance, enabling more comparable analysis of modality-dependent capabilities across video, image, and text. UniAR contains 40,199 high-quality question-answer pairs covering four shared reasoning challenges: dual anomaly discrimination, context understanding, fine-grained anomaly classification, and holistic anomaly analysis. Beyond standard multiple-choice question (MCQ) evaluation, UniAR assesses the evidence supporting model predictions, enabling more reliable and interpretable evaluation of anomaly understanding. Evaluations of over 20 MLLMs reveal substantial gaps from human performance and pronounced modality biases in human-normalized performance, particularly in evidence-grounded anomaly discrimination, while further suggesting precise anomaly localization as a key bottleneck. We hope UniAR can facilitate research on generalizable and reliable anomaly reasoning.

1 INTRODUCTION

Anomaly detection aims to identify observations that deviate from contextual normality. With the emergence of multimodal large language models (MLLMs), anomaly understanding is no longer limited to detecting whether an anomaly exists, but extends to reasoning about its context, fine-grained semantics, and supporting evidence Yang et al. (2023); Wang & Zhao (2026); Gu et al. (2024); Iino et al. (2024). We refer to this broader capability as *anomaly reasoning (AR)*. A key challenge, however, is that anomalies are inherently heterogeneous across modalities: they may appear as abnormal events in videos, defective regions in images, or anomalous spans in text, each with distinct semantics and evidence structures. While these manifestations differ, they raise a common set of reasoning requirements, from recognizing abnormality to understanding and explaining the anomalous phenomenon. This motivates a fundamental question: *Can MLLMs consistently perform anomaly reasoning across video, image, and text through a shared set of reasoning tasks?*

Despite substantial progress of MLLMs on general-purpose multimodal benchmarks Li et al. (2024); Liu et al. (2024); Yue et al. (2024), comprehensive evaluation of anomaly reasoning across modalities remains lacking. Existing benchmarks typically focus on individual modalities, including video Zhang et al. (2024); Chen et al. (2025b); Zhao et al. (2025), image Jiang et al. (2024); Fan et al. (2025); Cheng et al. (2025), and text Liu et al. (2022); Yang et al. (2025b); Cao et al. (2025), and remain limited in several respects:

- *Fragmented cross-modal evaluation.* Existing benchmarks evaluate anomaly reasoning separately across video, image, and text, with heterogeneous task formulations and evaluation protocols. This fragmentation hinders systematic assessment of cross-modal consistency, while differences in task and data difficulty make raw performance across modalities difficult to compare.

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.

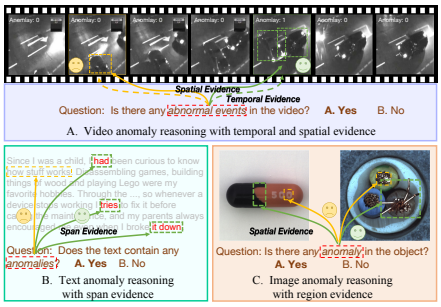


Figure 1: Unlike existing MCQ-based benchmarks, UniAR evaluates multi-modal anomaly reasoning with evidence grounding, enabling more reliable and interpretable assessment.

Table 1: Benchmark comparison across MCQ-based video (*V.*), image (*I.*), and text (*T.*) modalities between existing AR benchmarks and our UniAR. "-" and "E.G." denotes nonexistence and evidence grounding, respectively.

Benchmark	Modality	QAs	Interval Anno.	Region Anno.	Span Anno.	E.G. Eval.
UCVL Chen et al. (2025b)	*V.*	3,000	✓	✗	-	✗
HIVAU-70k Zhang et al. (2025)	*V.*	∼70,000	✗	✗	-	✗
VAD-Instruct50k Zhang et al. (2024)	*V.*	∼50,000	✗	✗	-	✗
SmartHome-Bench Zhao et al. (2025)	*V.*	3,609	✗	✗	-	✗
MMAD Jiang et al. (2024)	*I.*	39,672	-	✗	-	✗
AnomVerse Jiang et al. (2026)	*I.*	-	-	✓	-	✗
MANTA Fan et al. (2025)	*I.*	∼2,000	-	✓	-	✗
AnomalyCoT Cheng et al. (2025)	*I.*	37,565	-	✓	-	✗
Hades Liu et al. (2022)	*T.*	-	-	-	✓	✗
AD-NLP Bejan et al. (2023)	*T.*	-	-	-	✗	✗
AD-LLM Yang et al. (2025b)	*T.*	-	-	-	✗	✗
Tad-bench Cao et al. (2025)	*T.*	-	-	-	✗	✗
UniAR	*V.+I.+T.*	40,199	✓	✓	✓	✓

- *Limited task coverage of anomaly reasoning.* Existing benchmarks mainly focus on discrimination, with limited coverage of contextual understanding, fine-grained classification, and holistic analysis.
- *Insufficient fine-grained evidence annotations.* Existing benchmarks lack fine-grained event-, region-, and span-level annotations needed to ground anomaly reasoning across modalities.
- *Unreliable evaluation.* MCQ-based evaluation overlooks the assessment of supporting evidence, making it difficult to distinguish genuine reasoning from superficial guessing Chen et al. (2024b).

Addressing these limitations requires a benchmark that enables comparable evaluation of heterogeneous anomaly reasoning across modalities. We identify three key challenges. (i) *Shared reasoning tasks with modality-specific evidence.* Anomaly reasoning should be evaluated through common reasoning tasks—from anomaly discrimination to contextual and fine-grained understanding—while preserving modality-specific evidence structures. (ii) *Comparable cross-modal profiling.* Since task difficulty varies across modalities, human performance provides a reference for normalizing model scores and enabling more comparable cross-modal profiling. (iii) *Reliable evidence-grounded evaluation.* Beyond answer correctness, evaluation should assess whether predictions are supported by appropriate modality-specific anomaly evidence. To this end, we propose **UniAR**, a unified benchmark for profiling MLLMs' anomaly reasoning capabilities across video, image, and text, as shown in Figure 1. UniAR re-annotates and substantially expands $11,110$ samples from seven publicly available datasets, yielding $40,199$ high-quality question-answer pairs, $1,005$ event intervals, $10,575$ bounding boxes, and $42,941$ span evidence annotations. It organizes heterogeneous anomaly scenarios into four shared coarse-to-fine reasoning tasks: dual (global and referring) anomaly discrimination, context understanding, fine-grained anomaly classification, and holistic anomaly analysis. To support cross-modal profiling, we further compare model performance against modality-specific human references, providing human-normalized measures of anomaly reasoning capability. Beyond conventional answer-based evaluation, UniAR introduces an evidence-grounded metric that jointly assesses prediction correctness and supporting anomaly evidence. Evaluations of over 20 open- and closed-source MLLMs reveal substantial gaps from human performance and pronounced modality-dependent disparities in human-normalized performance, particularly for evidence-grounded anomaly discrimination. We further observe substantial performance improvements when anomaly locations are provided, suggesting precise anomaly localization as an important bottleneck for grounded anomaly reasoning. In brief, our contributions are threefold:

- **Unified Cross-Modal Anomaly Reasoning Benchmark.** We introduce UniAR, a unified benchmark for systematically profiling MLLMs' anomaly reasoning capabilities across video, image, and text. UniAR organizes heterogeneous anomalies into four shared reasoning tasks while preserving modality-specific anomaly manifestations and evidence structures.
- **Evidence-Grounded and Human-Normalized Evaluation.** We introduce an evidence-grounded AR evaluation beyond answer correctness that jointly assesses predictions and their supporting temporal, spatial, or textual evidence to separate grounded reasoning from superficial guessing. We also normalize performance against human references for better cross-modal comparability.
- **Comprehensive Profiling of MLLMs.** We evaluate over 20 open- and closed-source MLLMs, revealing substantial human–model gaps and pronounced modality-dependent disparities in anomaly reasoning. Further analysis highlights evidence-grounded anomaly discrimination and precise anomaly localization as major challenges for current MLLMs.

2 Related Work

Multi-modal Large Language Models. Large language models (LLMs) such as Qwen series Bai et al. (2023a); An Yang et al. (2024); Yang et al. (2025a), Llama series Touvron et al. (2023a;b); Grattafiori et al. (2024); Adcock et al. (2026) and GLM series Du et al. (2022); Glm et al. (2024) serve as the fundamental backbone of modern MLLMs. Built upon these text-only LLMs, vision-language models (VLMs) further extend model capabilities to the visual domain, with representative open-source implementations including InternVL series Chen et al. (2024c); Zhu et al. (2025); Wang et al. (2025), QwenVL series Bai et al. (2023b); Wang et al. (2024a); Bai et al. (2025a), MiniCPM-V series Yao et al. (2024); Yu et al. (2025), and Eagle series Li et al. (2025a;b); Chen et al. (2025a). Beyond this, advanced omni-modal models achieve unified perception and reasoning across modalities, covering open-source Ola Liu et al. (2025) as well as mainstream closed-source proprietary systems including GPT series OpenAI (2026); Singh et al. (2025), Gemini series DeepMind (2025; 2026), and Claude series Anthropic (2025; 2026).

Anomaly Reasoning Benchmarks. Existing anomaly reasoning benchmarks are mostly limited to unimodal evaluation and lack in-depth, multi-dimensional assessment tasks.

In video modality, existing benchmarks are primarily confined to sample-level anomaly discrimination. For example, HIVAU-70k Zhang et al. (2025) provides a large-scale video anomaly MCQ dataset but mainly evaluates whether a video clip contains an anomaly, lacking temporal context understanding, fine-grained anomaly classification, and holistic analysis. Other video benchmarks, such as UCVL Chen et al. (2025b), VAD-Instruct50k Zhang et al. (2024), and SmartHome-Bench Zhao et al. (2025), face similar limitations: despite their large scale, they focus on anomaly discrimination rather than systematic anomaly reasoning. In the text modality, comprehensive benchmarks for anomaly reasoning remain largely underexplored. To the best of our knowledge, there is still a notable absence of MCQ-based datasets designed for intra-text anomaly reasoning. Most existing benchmarks focus exclusively on sample-level anomaly discrimination, while failing to support contextual comprehension or in-depth reasoning analysis. For instance, Hades Liu et al. (2022) only targets hallucinated token detection within text and lacks evaluation tasks that cover diverse anomaly types. Other widely adopted datasets, such as AD-NLP Bejan et al. (2023), AD-LLM Yang et al. (2025b), and Tad-bench Cao et al. (2025), primarily concentrate on semantic-rare text classification. These datasets overlook intra-text anomaly reasoning and similarly fall short of providing comprehensive task-level evaluation. In contrast, although image-modality anomaly reasoning benchmarks are relatively advanced, they still lack fine-grained task evaluation. MMAD Jiang et al. (2024) designs various tasks but lacks instance-level discrimination and classification. AnomalyCoT Cheng et al. (2025) offers a discrimination-only benchmark with large-scale MCQs and Chain-of-Thought (CoT) annotations. AnomVerse Jiang et al. (2026) and MANTA Fan et al. (2025) explore spatial or contextual understanding but lack task completeness and evaluation depth.

A fundamental limitation common to all unimodal anomaly reasoning benchmarks is the absence of evidence-grounded, reliable evaluation protocols. This creates a gap between multiple-choice performance metrics and authentic anomaly reasoning capability. As a result, high scores on existing benchmarks do not guarantee effective reasoning in real-world anomaly scenarios. Table 1 presents a comprehensive comparison between existing benchmarks and our proposed UniAR.

3 Unified Anomaly Reasoning Benchmark

3.1 Unified Task Formulation

We formulate anomaly reasoning as identifying deviations from contextual normality, understanding the anomalous phenomenon, and grounding the judgment in modality-specific evidence. Given an input $\mathcal{X}_m$ from modality $m \in \{\text{vid}, \text{img}, \text{txt}\}$, we represent its structural elements as $\mathcal{E}_m = \{e_i\}_{i=1}^{N}$ and the anomalous subset as $\mathcal{A}_m \subseteq \mathcal{E}_m$. Specifically,

- **Video:** $\mathcal{E}_{\text{vid}} = \{(t_i, r_i)\}_{i=1}^{N}$, where t_i and r_i denote a temporal interval and its spatial region; $\mathcal{A}_{\text{vid}}$ denotes anomalous spatio-temporal events.
- **Image:** $\mathcal{E}_{\text{img}} = \{\text{region}_i\}_{i=1}^{N}$, where $\mathcal{A}_{\text{img}}$ denotes anomalous regions.
- **Text:** $\mathcal{E}_{\text{txt}} = \{\text{span}_i\}_{i=1}^{N}$, where $\mathcal{A}_{\text{txt}}$ denotes anomalous spans.

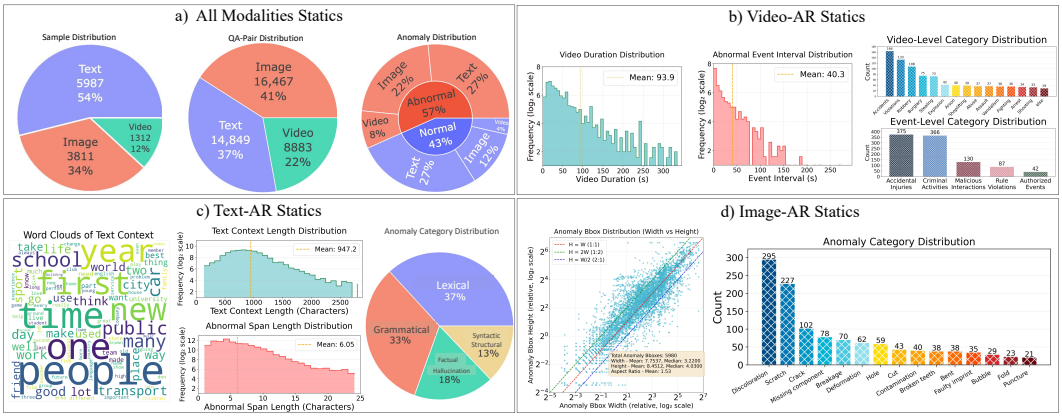


Figure 2: Comprehensive statistics of the proposed UniAR across video, text, and image modalities. a) Overall distribution of samples, QA pairs, and anomaly status across modalities. b) Video modality statistics, including video duration, abnormal event intervals, and category distributions at both video- and event-levels. c) Text modality statistics, featuring a word cloud of contexts, text length distributions, and anomaly category breakdowns. d) Image modality statistics, covering anomaly bounding box distributions (width/height) and fine-grained anomaly category distribution.

A sample is anomalous if $\mathcal{A}_m \neq \emptyset$, with $\mathcal{A}_m$ providing the supporting evidence for the anomaly judgment. Importantly, UniAR does not assume that anomalies across modalities share the same semantics, manifestations, or underlying representations. Instead, the *unification* lies at the level of reasoning requirements: despite their different forms, anomaly reasoning across video, image, and text can be evaluated through a common task schema covering anomaly discrimination, context understanding, fine-grained classification, and holistic analysis. We instantiate these reasoning dimensions as dual anomaly discrimination (DAD), context understanding (CU), fine-grained anomaly classification (FAC), and holistic anomaly analysis (HAA), respectively as detailed in Sec. 3.4.

This formulation provides a unified basis for organizing modality-specific anomaly reasoning within a consistent task and evaluation framework. By instantiating the same reasoning dimensions across video, image, and text, UniAR aligns heterogeneous anomaly scenarios at the level of reasoning requirements while preserving the anomaly semantics and evidence structures specific to each modality. Accordingly, each modality constitutes a distinct instantiation of the unified task formulation, rather than an independent evaluation setting with separate task definitions and protocols. This common formulation allows the same MLLM to be systematically examined along consistent reasoning dimensions across modalities, providing a coherent basis for profiling modality-dependent anomaly reasoning capabilities. Combined with human-normalized performance and evidence-grounded evaluation, UniAR further supports more comparable and reliable cross-modal profiling while respecting the intrinsic heterogeneity of anomaly reasoning across modalities.

3.2 Benchmark Overview

Figure 2 a) summarizes overall statistics across modalities, including samples, QA pairs, and anomaly distribution. Text accounts for 54% of samples and 37% of QA pairs, followed by image and video. QA pairs are relatively balanced, with images slightly leading (41%). Abnormal instances (57%) slightly exceed normal ones (43%), with text contributing the most in both categories and video the least. The dataset exhibits a reasonably balanced cross-modal composition.

Video-AR Track. Temporal statistics of video modality data reveal a long-tailed duration distribution with an average length of 93.9s, while abnormal events occur sparsely with a mean interval of 40.3s. At the video level, *Accidents*, *Violations*, and *Robbery* are the most frequent anomaly types. At a finer granularity, *Accidental Injuries* and *Criminal Activities* dominate, covering a wide range of real-world anomaly scenarios. We fine-grainedly re-annotate videos from three public datasets: UCF-Crime Sultani et al. (2018), TAD Xu et al. (2024), and ShanghaiTech Liu et al. (2018).

Image-AR Track. We provide detailed statistics over 5980 annotated anomaly regions. The bounding box distribution shows a mean width of 7.75, height of 8.45, and an average aspect ratio of

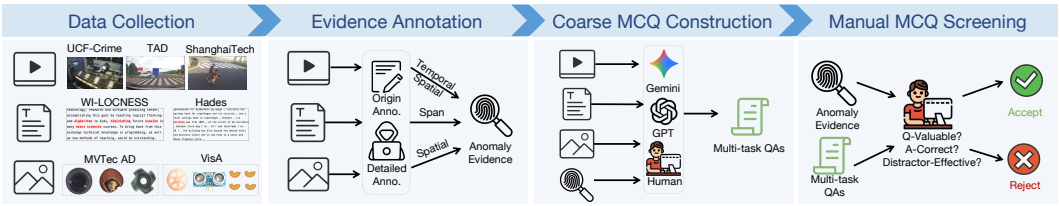


Figure 3: Construction pipeline of cross-modal MCQs in UniAR. We first manually obtain anomalous evidence on the source dataset, then generate coarse question–answer pairs using closed-source foundation models, and finally refine the full evaluation dataset via careful human verification.

1.53 (all in relative proportions), covering diverse anomaly scales and shapes. In terms of category distribution, *Discoloration* (295) and *Scratch* (227) are the most frequent types, followed by *Crack*, *Missing Component*, and other defect categories, ensuring comprehensive coverage of real-world visual anomaly scenarios. The initial image anomaly samples are derived from the widely used comprehensive datasets MVTec AD Bergmann et al. (2019) and VisA Zou et al. (2022).

Text-AR Track. The context length follows a long-tailed distribution, averaging 947.2 characters and ranging from short passages to long documents. In contrast, anomalous spans are typically short, with a mean length of 6.05 characters, indicating localized anomalies. Regarding category distribution, Lexical Error accounts for the largest proportion (37%), followed by Grammatical Error (33%). Factual Hallucination (18%) and Syntactic Structural Error (13%) further enrich the coverage of real-world textual anomalies. We source anomalous text data from the WI-LOCNESS corpus Bryant et al. (2019), which provides lexical, grammatical, and syntactic structural errors. We additionally use Hades Liu et al. (2022) to incorporate factual hallucinations.

3.3 MCQ Construction

Multiple-choice questions (MCQs) provide a standardized, objective, and unambiguous evaluation framework for multi-modal anomaly reasoning in MLLMs, closely mirroring real-world diagnostic scenarios. Following prior works Chen et al. (2024a); Li et al. (2024); Jiang et al. (2024); Cheng et al. (2025), we therefore adopt MCQs as our primary evaluation format. As shown in Figure 3, we construct multi-modal MCQs via a four-stage pipeline: (1) Data Collection, gathering raw samples from public video, text, and image anomaly detection benchmarks for cross-modal coverage and task diversity; (2) Evidence Annotation, performing fine-grained temporal, region, and span-level annotation to extract fine-grained anomaly evidence as the ground truth for MCQ generation; (3) Coarse MCQ Construction, generating initial multi-task QA pairs with large multimodal models and human annotators; (4) Manual MCQ Screening, rigorously filtering low-quality pairs by question value, answer correctness, and distractor effectiveness. This end-to-end pipeline ensures high-quality, consistent, and evidence-supported MCQs for comprehensive evaluation of MLLMs' cross-modal anomaly reasoning abilities. Details of the base dataset, MCQ construction pipeline, quality control, and prompt design for validated data generation are provided in Appendix A.

3.4 Unified Task Definition

Based on the unified task formulation above, UniAR establishes a unified evaluation framework for multi-modal anomaly reasoning, encompassing four core tasks across video, image, and text modalities as shown in Figure 4, designed to comprehensively assess the capabilities of MLLMs in full-spectrum anomaly reasoning.

- **Dual Anomaly Discrimination (*DAD*).** We adopt two anomaly binary classification settings: one is the conventional global setting (*glob.*), where the model judges anomalies solely based on the raw input; the other is the referring setting (*ref.*), which requires fine-grained discrimination on designated temporal/regional/span segments of the input. This dual-mode design enables comprehensive evaluation of both anomaly *detection* and *verification* capabilities across all modalities.
- **Context Understanding (*CU*).** Contextual information encodes critical reasoning logic for anomaly judgment, making context understanding a core capability of holistic anomaly reasoning. This task evaluates the model's ability to capture key contextual background cues. Specifically, it covers the timing (when) and location (where) of video events, the topical context of the textual claim, and the object descriptions, as well as imaging conditions where the anomaly occurs in images.

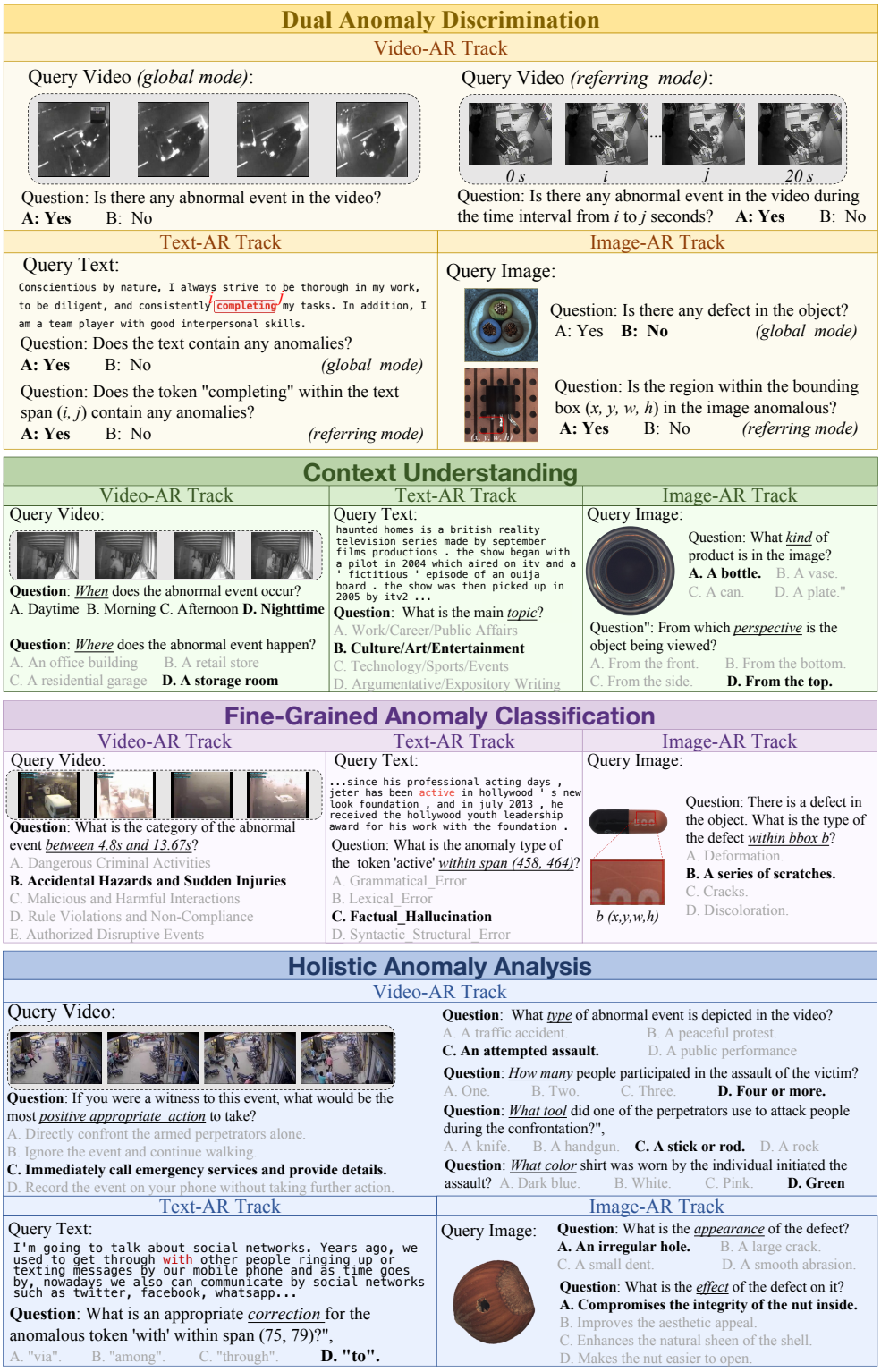


Figure 4: MCQ examples of four cross-modal anomaly reasoning tasks for full-spectrum evaluation.

- **Fine-Grained Anomaly Classification (*FAC*).** This task enables fine-grained anomaly classification across modalities: *event-level* for videos, *span-level* for texts, and *box-level* for images.
- **Holistic Anomaly Analysis (*HAA*).** This task challenges in-depth cross-modal anomaly reasoning capabilities. It requires models to conduct comprehensive reasoning on in-depth anomaly details.

Table 2: Performance comparison of both proprietary and open-source MLLMs in Video-AR track.

	Scale	Success (%)	*DAD*				*FAC*		*CU*	*HAA*
			glob.acc.	prom.acc.	trim.acc.	acc.@evi.	prom.acc.	trim.acc.	glob.acc.	glob.acc.
Human (5min)	-	100.0	96.4	98.3	98.5	95.7	92.2	92.8	99.5	96.2
Gemini-3-Flash DeepMind (2025)	-	100.0	**87.3**	87.0	88.0	64.6	**71.4**	**71.0**	**97.0**	**90.2**
Gemini-2-Flash DeepMind (2024)	-	100.0	84.1	83.5	84.5	60.2	67.8	66.9	95.3	87.5
GPT-5.4-mini OpenAI (2026)	-	100.0	86.9	84.3	86.0	49.7	51.4	52.2	89.9	82.2
GPT-5.1 OpenAI (2025b)	-	100.0	86.8	86.0	**88.1**	**67.3**	70.7	70.8	94.8	88.4
GPT-5-Nano OpenAI (2025a)	-	99.9	75.1	66.3	75.1	51.5	50.7	50.2	91.0	78.5
Video-LLaVA Lin et al. (2024)	7B	100.0	72.6	68.3	71.2	32.9	32.0	38.3	69.2	66.6
Ola Liu et al. (2025)	7B	99.1	78.9	74.9	79.6	20.0	45.7	41.9	90.4	84.4
Qwen2-VL Wang et al. (2024b)	7B	99.9	82.5	84.6	80.4	47.3	50.2	52.5	92.4	82.4
Qwen2.5-VL Bai et al. (2025b)	7B	99.2	79.4	77.0	78.2	53.6	53.3	49.4	93.5	83.9
Eagle2.5 Chen et al. (2025a)	8B	99.6	86.4	**88.7**	85.6	63.9	60.2	57.8	94.2	87.7
InternVL2.5 Chen et al. (2024c)	8B	99.1	83.9	81.9	83.1	33.4	35.5	36.8	90.1	84.1
InternVL3 Zhu et al. (2025)	8B	98.2	64.7	65.1	64.7	49.2	55.8	51.0	91.9	83.5
InternVL-3.5 Wang et al. (2025)	8B	99.9	75.1	57.4	67.7	27.2	41.6	42.7	58.6	55.5
Qwen3-VL Bai et al. (2025a)	8B	99.7	85.1	84.2	83.8	51.1	52.2	51.6	93.8	84.9
MiniCPM-V-4.5 Yu et al. (2025)	9B	99.6	80.6	76.7	82.8	25.6	55.4	61.0	94.3	86.8

For video, this includes understanding the *cause, process, and outcome* of abnormal events, as well as the *count, appearance, and behavior* of involved participants, and inferring appropriate responses to such events. For textual anomalies, it demands *correct revision* suggestions. For visual anomalies, it asks models to identify potential *causes* or corresponding *impacts*.

3.5 Cross-Modal Evaluation.

To thoroughly evaluate the core anomaly discrimination ability of anomaly reasoning, we design two evaluation settings together with dedicated metrics. In the global setting, we adopt multi-choice accuracy (*glob.acc.*). In the referring setting, we employ two strategies: (1) prompt injection, which embeds spatio-temporal clues into input prompts (*prom.acc.*); and (2) data trimming, which truncates raw data into clue-aligned segments (*trim.acc.*). Both strategies support temporal clues, whereas only prompt injection is used for region and text span clues due to stronger contextual coherence. Adopting the normal-prior discrimination paradigm, which requires explicit evidence exclusively for anomalous rather than normal cases, we prompt MLLMs to output binary predictions along with supporting evidence and rectify unsubstantiated anomalous predictions to normal. We propose a novel evidence-aware metric *acc.@evi.* for credible evaluation: an anomalous sample is regarded as correctly identified only when the model delivers accurate binary classification and provides modality-specific valid evidence, including intervals, bounding boxes, and text spans. This standard enables interpretable, evidence-grounded anomaly evaluation across all modalities. For fine-grained anomaly classification with multiple anomalous instances, we adopt the referring setting with *prom.acc.* and *trim.acc.*. For contextual comprehension and holistic anomaly analysis, we consistently use *glob.acc.*. We adopt a model-shared input pipeline to mitigate interference from confounding factors in end-to-end reasoning-centric anomaly evaluation. We parse the output evidence and standardize it into 1D/2D masks, where IoU is adopted to quantify the support strength of evidence across modalities. Detailed evaluation protocols, including *input-compatibility design*, *output-parsing pipelines*, and *modality-specific IoU threshold settings for evidence*, are elaborated in Appendix B.

4 Experiments

Settings. To systematically assess the performance of leading MLLMs in this specific domain, we evaluate a diverse set of commercial closed-source and open-source models. For commercial closed-source models, we test the GPT-5 Singh et al. (2025) series, Gemini-3-Flash DeepMind (2025), and Claude-Haiku-4.5 Anthropic (2025). For open-source models, we evaluate the InternVL series Chen et al. (2024c); Zhu et al. (2025); Wang et al. (2025), Qwen series Wang et al. (2024b); Bai et al. (2025b;a); Yang et al. (2025a); Xu et al. (2025a), DeepSeek series Wu et al. (2024); DeepSeek-AI (2025), MiniCPM series Yu et al. (2025); OpenBMB (2026), GLM series Glm et al. (2024), alongside Ola Liu et al. (2025) and Eagle2.5 Chen et al. (2025a). We implement a unified MLLM evaluation framework based on VLMEvalKit Duan et al. (2024). For video, we sample 32 frames for models without native video support and use original videos otherwise. In the Video-AR track, the first abnormal event serves as the referring input, and an additional frame is selected for evidence bbox prediction. For Image-AR, a randomly selected bbox annotation is injected into the

Table 3: Performance comparison of both proprietary and open-source MLLMs in Image-AR track.

	Scale	Success (%)	*DAD*			*FAC*	*CU*	*HAA*
			glob.acc.	prom.acc.	acc.@evi.	prom.acc.	glob.acc.	glob.acc.
Human (3min)	-	100.0	95.2	96.8	94.6	85.0	96.1	94.2
Claude-Haiku-4.5 Anthropic (2025)	-	98.6	67.6	47.3	27.6	59.3	84.4	79.9
Gemini-3-Flash DeepMind (2025)	-	100.0	**83.4**	**91.2**	**66.1**	**78.8**	89.7	87.1
GPT-5.4-mini OpenAI (2026)	-	100.0	71.4	60.3	35.1	59.9	84.2	75.0
GPT-5.1 OpenAI (2025b)	-	100.0	73.6	76.5	42.7	65.8	88.0	79.8
GPT-5-Nano OpenAI (2025a)	-	100.0	69.9	66.2	39.8	58.5	82.6	75.3
DeepSeek-VL2 Wu et al. (2024)	∼3B	100.0	74.8	68.8	36.4	55.5	86.9	77.3
Ola Liu et al. (2025)	7B	100.0	76.5	79.5	57.4	65.9	87.6	79.9
Qwen2-VL Wang et al. (2024b)	7B	100.0	66.8	74.3	42.0	64.3	89.1	79.2
Qwen2.5-VL Bai et al. (2025b)	7B	100.0	73.3	60.3	40.3	61.0	87.4	79.8
Eagle2.5 Chen et al. (2025a)	8B	100.0	66.2	50.2	43.6	60.2	86.9	82.6
InternVL2.5 Chen et al. (2024c)	8B	100.0	72.8	71.8	37.3	59.1	85.5	78.8
InternVL3 Zhu et al. (2025)	8B	100.0	75.4	56.3	42.0	68.6	84.2	81.2
InternVL3.5 Wang et al. (2025)	8B	100.0	74.6	52.2	31.6	54.7	82.0	73.0
Qwen3-VL Bai et al. (2025a)	8B	100.0	76.4	77.7	53.3	69.1	88.2	81.5
MiniCPM-V-4.5 Yu et al. (2025)	9B	100.0	79.2	72.4	26.4	74.3	**91.3**	**88.5**

prompt for referring discrimination. For Text-AR, we evaluate *glob.acc.* and *acc.@evi.* solely on WI-LOCNESS with full span annotations. For all other metrics in referring discrimination, fine-grained classification, and anomaly analysis tasks, both text data from WI-LOCNESS and Hades (with partial span annotations) are utilized. We employ global prompts (*e.g.*, full-duration interval, image-size bbox, text-length span) as referring cues for normal samples.

Human Evaluation We perform a rigorous *human evaluation* under the unified protocol on 200 randomly sampled instances per track (1538 MCQs on average). Three annotators with multimodal domain expertise conduct the assessment to ensure reliable judgments. To maintain evaluation consistency, they complete MCQs for each video/image/text sample within 5/3/5 minutes, respectively. The averaged annotator accuracy is treated as the human-level performance baseline for each task.

Main Result. In Video-AR Track (Table 2), the closed-source Gemini-3-Flash achieves the best performance: *DAD glob.acc.* 87.3, *trim.acc.* 88.0, *acc.@evi* 64.6. The open-source SOTA is Eagle2.5 (*DAD glob.acc.* 86.4, *prom.acc.* 88.7, *acc.@evi* 63.9). Still, the best model lags far behind human performance (*DAD glob.acc.* 96.4, *trim.acc.* 98.5, *acc.@evi* 95.7) and also underperforms on *FAC/CU/HAA* sub-tasks. In Image-AR Track (Table 3), Gemini-3-Flash again leads among closed-source models: *DAD glob.acc.* 83.4, *prom.acc.* 91.2, *acc.@evi* 66.1, while also outperforming rivals on other sub-tasks. Among open-source models, MiniCPM-V-4.5 attains *DAD glob.acc.* 79.2, *prom.acc.* 72.4, and Ola achieves *acc.@evi* 57.4. The top model still shows a clear gap from human performance (*DAD glob.acc.* 95.2, *prom.acc.* 96.8, *acc.@evi* 94.6), trailing humans on remaining metrics. In Text-AR Track (Table 4), Gemini-3-Flash is the best closed-source model: *DAD glob.acc.* 81.4, *prom.acc.* 47.5, *acc.@evi* 65.7. Ola achieves SOTA *prom.acc.* 73.8. The strongest model still falls short of human performance (*DAD glob.acc.* 87.9, *prom.acc.* 91.7, *acc.@evi* 86.5) and and remains far behind on *FAC*, *CU* and *HAA* tasks by 29.4, 20.2 and 18.8, respectively.

Modality bias of MLLMs. We report human-normalized scores across three discrimination tasks in Figure 5. Compared with humans' full-modality anomaly reasoning capability, existing models exhibit obvious modal bias. On *acc.@evi.*, GPT-5.4-mini performs well on Video-AR (51.9) and excels in Text-AR (75.0) but lags markedly in Image-AR (37.1). In contrast, Ola excels in Image-AR (60.7), performs moderately on Text-AR (45.0), but underperforms greatly on Video-AR (20.9). We attribute the modality bias to imbalanced pre-training data and architectural inductive biases.

Credible Evaluation. We compare *acc.@evi.* with alternative evaluation metrics (binary *acc.* and evidence *IoU*) by ranking three MLLMs on *full* anomalous image samples. Human preference scores ($0, 1, ..., 5$, normalized sum to $0 \sim 100$) of model response are collected by 3 annotators based on prediction persuasiveness. Result in Table 5 shows that the proposed *acc.@evi.* yields rankings and relative distances far closer to human preference, confirming its superiority as a credible metric.

Anomaly Discrimination with and without Reference. Although global and referring anomaly discrimination are fundamentally distinct, referring discrimination acts as a sufficient prerequisite for global discrimination under our setting with full-duration/full-region normal prompts. As shown in Figure 6, referring discrimination outperforms its global counterpart across several MLLMs in the Video and Image AR tracks. This indicates precise clues guide MLLMs toward valid anomaly manifestations, revealing that anomaly locus proposal is the key bottleneck for anomaly discrimination.

Table 4: Performance comparison of both proprietary and open-source MLLMs in Text-AR track.

	Scale	Success (%)	*DAD*			*FAC*	*CU*	*HAA*
			glob.acc.	prom.acc.	acc.@evi.	prom.acc.	glob.acc.	glob.acc.
Human (5min)	-	100.0	87.9	91.7	86.5	84.2	91.4	93.7
Claude-Haiku-4.5 Anthropic (2025)	-	100.0	73.7	52.2	60.8	46.9	59.4	63.3
Gemini-3-Flash DeepMind (2025)	-	100.0	**81.4**	47.5	**65.7**	**54.8**	**71.2**	**74.9**
GPT-5.4-mini OpenAI (2026)	-	100.0	68.1	52.4	64.9	45.9	70.9	46.8
GPT-5.1 OpenAI (2025b)	-	100.0	65.3	50.7	63.1	50.6	71.5	51.9
GPT-5-nano OpenAI (2025a)	-	100.0	72.2	33.3	63.6	45.6	64.9	68.6
DeepSeek-R1-Distill Guo et al. (2025)	7B	100.0	57.6	49.8	38.4	42.7	50.0	63.8
Ola Liu et al. (2025)	7B	100.0	60.7	**73.8**	38.9	31.0	54.0	54.6
Qwen2.5 Xu et al. (2025b)	7B	100.0	55.6	70.0	29.6	28.4	38.9	36.6
Llama-3.1 AI (2024)	8B	100.0	63.7	56.5	26.7	25.7	34.3	39.7
Qwen3 Xu et al. (2025c)	8B	100.0	70.8	64.7	40.0	31.9	29.9	41.6
GLM-4 Glm et al. (2024)	9B	100.0	68.9	50.6	32.1	21.9	33.3	29.9

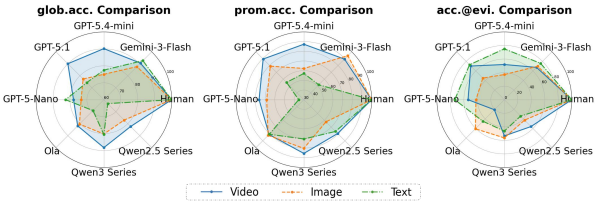


Figure 5: Modal bias analysis in anomaly discrimination tasks (We use human-normalized score, human is 100).

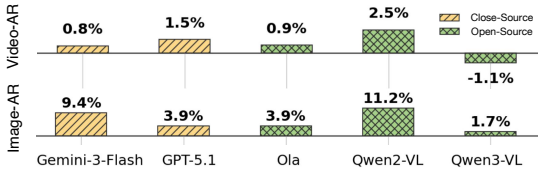


Figure 6: Discrimination performance gains of *referring* setting over *global* setting on Video/Image-AR tracks.

Table 5: Metrics ranking (Tab. 3) with human preference in Image-AR track.

Evaluation	Ola	InternVL3.5	Qwen3-VL
acc.	#2 $_{76.5}$	#6 $_{74.6}$	#3 $_{76.4}$
IoU	#3 $_{19.9}$	#1 $_{20.7}$	#2 $_{20.3}$
acc.@evi.	#1 $_{57.4}$	#9 $_{31.6}$	#2 $_{53.3}$
human	#1 $_{71.4}$	#8 $_{54.8}$	#2 $_{68.3}$

Table 6: Impact of augmented prompt on *glob.acc./acc.@evi.* in Video-AR track.

Variants	Ola	Eagle2.5	Qwen3-VL
Origin	79.8/**2.75**	86.2/66.4	83.5/**53.4**
CoT	80.7/1.83	87.2/68.3	69.7/47.6
Context	**86.2**/**2.75**	**88.1**/71.7	**91.7**/53.3
CoT+Context	83.5/**2.75**	87.2/**72.1**	87.2/52.8

Reasoning with Augmented Prompt. To investigate the impacts of prompt augmentation on anomaly reasoning, we evaluate *glob.acc.* and *acc.@evi.* on 109 video samples from the Video-AR track. We adopt two strategies: Chain-of-Thought (CoT) Kojima et al. (2022); Wang et al. (2023), guided by the prompt "*Let's think step by step*"; and context injection, which incorporates non-anomaly-leaky background and event descriptions derived from Gemini-3-Flash via the prompt "*Describe the video in detail*". We also evaluate their combination for further comparison. As demonstrated in Table 6, all augmented prompts lead to consistent improvements in *glob.acc.*, yet their gains on *acc.@evi.* remain unstable. It reveals that linguistic reasoning enhancement effectively facilitates global anomaly discrimination but lacks the granularity required for vision-dependent evidence grounding. Collectively, these findings provide preliminary insights into training-free optimization for cross-modal anomaly reasoning, while identifying general and robust prompt enhancement as a direction for future research.

5 Conclusion

In this work, we present UniAR, a unified benchmark for systematically profiling MLLMs' anomaly reasoning capabilities across video, image, and text. UniAR organizes heterogeneous anomaly scenarios under four shared reasoning tasks while preserving modality-specific anomaly semantics and evidence structures. It comprises *40,199* high-quality MCQs and introduces evidence-grounded evaluation and human-normalized performance for more reliable and comparable cross-modal profiling. Extensive evaluation of over *20* MLLMs reveals substantial human–model gaps and pronounced modality-dependent disparities, particularly in evidence-grounded anomaly discrimination, while highlighting precise anomaly localization as an important challenge for current MLLMs. We hope UniAR provides a unified and reliable foundation for advancing anomaly reasoning across modalities.

AI USE STATEMENT

We used AI-assisted tools to improve the clarity, grammar, and overall presentation of the manuscript. These tools were used solely for language editing and polishing purposes and were not used to generate research ideas, conduct experiments, analyze results, or draw scientific conclusions. All scientific content and claims were developed, reviewed, and verified by the authors, who take full responsibility for the content of this manuscript.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. The manuscript provides detailed descriptions of the proposed method, experimental settings, and implementation details necessary to reproduce the reported results. Upon publication, we will publicly release all relevant resources associated with this work, including the datasets, source code, and model weights, to facilitate independent reproduction and further research.

REFERENCES

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, et al. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*, 2026.

Meta AI. Introducing meta llama 3.1, jul 2024. URL https://ai.meta.com/blog/meta-llama-3-1/.

Binyuan Hui Bo Zheng Bowen Yu An Yang, Baosong Yang et al. Qwen2 technical report. *https://arxiv.org/abs/2407.10671v4*, 2024.

Anthropic. Introducing claude haiku 4.5. *https://www.anthropic.com/news/claude-haiku-4-5*, 2025.

Anthropic. Claude models overview. https://platform.claude.com/docs/en/about-claude/models/overview, 2026. Accessed: 2026-03-31.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.

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*, 2025a.

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. *https://arxiv.org/abs/2502.13923*, 2025b.

Matei Bejan, Andrei Manolache, and Marius Popescu. Ad-nlp: A benchmark for anomaly detection in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10766–10778, 2023.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. The bea-2019 shared task on grammatical error correction. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pp. 52–75, 2019.

Yang Cao, Sikun Yang, Chen Li, Haolong Xiang, Lianyong Qi, Bo Liu, Rongsheng Li, and Ming Liu. Tad-bench: A comprehensive benchmark for embedding-based text anomaly detection. *arXiv preprint arXiv:2501.11960*, 2025.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024a.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024b.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *NeurIPS*, 2025a.

Haoran Chen, Dong Yi, Moyan Cao, Chensen Huang, Guibo Zhu, and Jinqiao Wang. Ucvl: a benchmark for crime surveillance video analysis with large models. In *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*, pp. 2051–2057. IEEE, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Jiaxi Cheng, Yuliang Xu, Shoupeng Wang, Ma Tao, Yuchen He, Jinghe Zhang, Sihang Cai, Jiawei Zhen, Jingyi Jia, Yao Wan, et al. Anomalycot: A multi-scenario chain-of-thought dataset for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, dec 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Google DeepMind. Gemini 3 flash: Intelligence in a flash. *https://deepmind.google/models/gemini/flash/*, 2025.

Google DeepMind. Gemini 3.1 pro. https://deepmind.google/models/gemini/pro/, 2026. Accessed: 2026-03-31.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25518–25527, 2025.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378, 1971.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 1932–1940, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shota Iino, Hideki Nomoto, Takashi Fukui, Yohei Yagisawa, Sayaka Ishizawa, Takayuki Hirose, and Yasutaka Michiura. Towards explainable anomaly detection in safety-critical systems: Employing fram and spectrm in international space station telemetry. *International Journal of Prognostics and Health Management*, 15(3), 2024.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2024.

Yuxin Jiang, Wei Luo, Hui Zhang, Qiyu Chen, Haiming Yao, Weiming Shen, and Yunkang Cao. Anomagic: Crossmodal prompt-driven zero-shot anomaly generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5485–5493, 2026.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025a.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737, 2022.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025.

OpenAI. Introducing gpt-5, aug 2025a. URL https://openai.com/index/gpt-5/.

OpenAI. Gpt-5.1: A smarter, more conversational chatgpt, nov 2025b. URL https://openai.com/index/gpt-5-1/.

OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026. Accessed: 2026-03-31.

OpenBMB. Minicpm-o 4.5: A gemini 2.5 flash level mllm for vision, speech, and full-duplex multimodal live streaming on your phone. https://github.com/OpenBMB/MiniCPM-o, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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*, 2024b.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yu Wang and Shengjie Zhao. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. *arXiv preprint arXiv:2603.00550*, 2026.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025c.

Yajun Xu, Huan Hu, Chuwen Huang, Yibing Nan, Yuyao Liu, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. Tad: A large-scale benchmark for traffic accidents detection from video surveillance. *IEEE Access*, 13:2018–2033, 2024.

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*, 2025a.

Fan Yang, Jiacen Xu, Chunlin Xiong, Zhou Li, and Kehuan Zhang. {PROGRAPHER}: An anomaly detection system based on provenance graph embedding. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 4355–4372, 2023.

Tiankai Yang, Yi Nian, Li Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A Rossi, Kaize Ding, et al. Ad-llm: Benchmarking large language models for anomaly detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1524–1547, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*, 2024.

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 computer vision and pattern recognition conference*, pp. 13843–13853, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3975–3985, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Benchmark Details

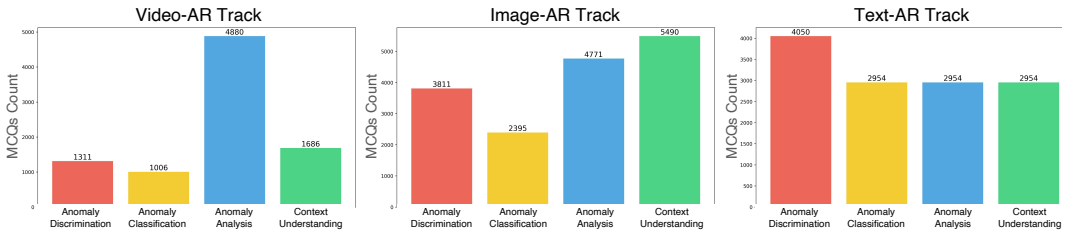


Figure 7: Statistics of MCQs across three modalities and four tasks. For the anomaly discrimination task, the original MCQs are converted into three distinct modes: global, referring, and evidence-based.

Table 7: Dataset details: Our benchmark contains diverse datasets in dimensions of scenes, objects, anomaly sample sizes, and resolutions.

Dataset	Modality	$\|\mathcal{C}\|$	N.S	A.S.	MCQ-As	Content Size	Remark
UCF-Crime	Video	13	147	718	6816	3.47∼1355.19s	Surveillance Videos in Diverse Scenarios
TAD	Video	7	248	189	1979	1.40∼1000.00s	Road Surveillance
ShanghaiTech	Video	3	0	9	88	8.03∼20.03s	Campus Surveillance
MVTec AD	Image	15	444	1226	7042	$w_{min}, w_{max}, h_{min}, h_{max}$ 700,1024,700,1024	Industrial Obj. & Texture Anomalies
VisA	Image	12	943	1198	9425	$w_{min}, w_{max}, h_{min}, h_{max}$ 1274,1562,960,1176	Industrial Obj. Anomalies
WI-LOCNESS	Text	1	1096	1684	7832	140∼9089	Grammatical Error Correction
Hades	Text	1	1937	1270	7017	172∼2121	Hallucinated text tokens

A.1 Dataset Constitution

To evaluate multi-modal anomaly reasoning capabilities across video, image, and text domains, we construct our benchmark based on seven publicly available datasets covering diverse anomaly scenarios. The detailed statistics are summarized in Table 7.

Video Datasets. UCF-Crime is a large-scale surveillance video dataset containing 13 anomaly categories (*e.g.*, *abuse, assault, burglary, shooting*) with 147 normal and 718 abnormal videos. The video duration ranges from 3.47 s to 1355.19 s, covering diverse real-world surveillance scenarios. TAD is a road surveillance dataset with 7 traffic anomaly categories, containing 248 normal and 189 abnormal videos. The duration ranges from 1.40 s to 1000.00 s, focusing on urban traffic anomaly events such as accidents, illegal turns, and pedestrian violations. ShanghaiTech is a campus surveillance dataset with 3 anomaly types, including 9 abnormal videos (no normal videos in the test split). Videos are short (8.03–20.03 s) and depict crowd-related anomalies in campus environments. The entire Video-AR Track annotation process takes three annotators approximately six days, with each annotator processing around 500 MCQs per day.

Image Datasets. MVTec AD is a standard industrial anomaly detection dataset covering 15 object and texture categories (*e.g.*, bottle, cable, carpet). We use 444 normal and $1,226$ anomalous images from this dataset, with resolutions ranging from 700×700 to $1,024 \times 1,024$. VisA is an industrial anomaly dataset including 12 object categories, with 943 normal and $1,198$ anomalous images. The resolutions range from 1274×960 to $1,562 \times 1,176$, capturing fine-grained manufacturing defects. The entire Image-AR Track annotation process takes two annotators approximately four days, with each annotator processing around $2,000$ MCQs per day.

Text Datasets. WI-LOCNESS is a widely used dataset for grammatical error correction, with $1,096$ normal and $1,684$ anomalous texts. The text length ranges from 140 to $9,089$ characters, providing diverse grammatical error patterns (*e.g.*, tense, article, and spelling mistakes) that resemble LLM hallucinations. Hades is a dataset focusing on text hallucination detection, with $1,937$ normal and $1,270$ hallucinated texts. The length ranges from 172 to $2,121$ characters, containing explicit factual inconsistencies and hallucinated tokens that align with our text anomaly reasoning task. The entire Image-AR Track annotation process takes three annotators approximately five days, with each annotator processing around $1,000$ MCQs per day.

A.2 Task Distribution

The distribution of MCQs across three modalities (Video-AR, Image-AR, and Text-AR) and four core tasks is visualized in Figure 7. Across all tracks, the dataset comprehensively covers Anomaly Discrimination, Anomaly Classification, Anomaly Analysis, and Context Understanding, with varying MCQ counts reflecting the task-specific complexity and modality-specific data availability. Notably, for the Anomaly Discrimination task, the original MCQs are further structured into three distinct

settings: global, referring, and evidence-based reasoning. The statistics reveal a consistent pattern where anomaly-related tasks account for the majority of samples in each modality, while context understanding questions are designed to evaluate the models' ability to reason about the background and causal dependencies of anomalies. This multi-task, multi-modal design ensures that the benchmark provides a holistic evaluation of models' capabilities in anomaly reasoning.

A.3 Annotation Quality Control and Human Baselines

Following established practices in large-scale multimodal benchmarks, UniAR employs an LLM-assisted MCQ generation pipeline followed by rigorous human verification. Starting from 68,735 automatically generated MCQs, we manually filter 28,536 samples (41.5%) to obtain 40,199 high-quality QA pairs. Among the retained samples, 21,382 (53.2%) undergo further refinement to correct ambiguous question stems (21.7%), sensitive information leakage (7.5%), incorrect answers (34.2%), and low-quality distractors (36.6%). This multi-stage process ensures that UniAR-Bench is a carefully curated resource rather than a simple aggregation of generated question–answer pairs.

Annotation consistency is evaluated on $1,000$ randomly sampled QA pairs (300 video, 300 image, and 400 text). Three independent annotators achieve an overall Fleiss' κ Fleiss (1971) of 0.88, indicating substantial cross-modal agreement as shown in Table 8.

Table 8: Inter-annotator agreement on UniAR QA pairs.

Modality	# Samples	Annotation Type	Metric	Agreement
Video	300	QA correctness	Fleiss' κ	0.88
Image	300	QA correctness	Fleiss' κ	0.91
Text	400	QA correctness	Fleiss' κ	0.86
Overall	1000	QA correctness	Fleiss' κ	0.88

We further report human baseline performance (mean $\pm$ standard deviation) obtained from three annotators in Table 9. The small variance confirms high consistency of the UniAR annotations.

Table 9: Human baseline performance (mean $\pm$ std) across UniAR tracks.

Track	$\text{DAD}_{\text{glob.}}$	$\text{DAD}_{\text{prom.}}$	$\text{DAD}_{\text{trim.}}$	$\text{DAD}_{\text{acc.@evi.}}$	$\text{FAC}_{\text{prom.}}$	$\text{FAC}_{\text{trim.}}$	$\text{CU}_{\text{glob.}}$	$\text{HAA}_{\text{glob.}}$
Video	96.4 ± 1.2	98.3 ± 0.8	98.5 ± 0.7	95.7 ± 1.4	92.2 ± 2.1	92.8 ± 1.9	99.5 ± 0.5	96.2 ± 1.3
Image	95.2 ± 1.5	96.8 ± 1.2	–	94.6 ± 1.8	85.0 ± 2.8	–	96.1 ± 1.7	94.2 ± 1.9
Text	87.9 ± 2.3	91.7 ± 2.0	–	86.5 ± 2.5	84.2 ± 2.7	–	91.4 ± 2.1	93.7 ± 1.8

A.4 Modality-Input-Blinded Evaluation

To verify that the evaluations of UniAR rely on genuine anomaly evidence rather than linguistic cues or dataset artifacts, we perform modality-input-blinded experiments using Gemini-3-Flash. The anomaly reasoning performance comparison of w/w.o. modality input is presented in Table 10. In this blinded setup, the modality containing anomalies is removed while the rest of the textual context is preserved. We observe substantial performance degradation across all tracks, which confirms that the model depends on authentic multimodal anomaly signals.

A.5 Robustness and Human Alignment of *acc.@evi.*

To further assess whether *acc.@evi.* aligns with human judgment, we conduct human preference studies on the Video-AR and Text-AR tracks. Since manually evaluating and ranking all model outputs on the full benchmark is prohibitively expensive, we report results on randomly sampled MCQs from 100 video samples and 500 text samples, respectively. For each track, we select representative MLLMs and compare the rankings derived from automatic metrics against the human preference ranking. Table 11 summarizes the results. Here, smaller values indicate stronger alignment with human judgment. On both tracks, *acc.@evi.* achieves substantially smaller Σ_{diff} values than plain accuracy and vanilla IoU metrics, indicating that the rankings produced by *acc.@evi.* closely

Table 10: Modality-input-blinded evaluation results. Removing the anomaly-bearing modality causes large performance degradation.

Modality Setting	DAD	Δ	FAC	Δ	CU	Δ	HAA	Δ
Video-Input	87.3	–	71.4	–	97.0	–	90.2	–
Video-Blinded	51.6	↓35.7	24.6	↓46.8	30.5	↓66.5	34.5	↓55.7
Image-Input	83.4	–	78.8	–	89.7	–	87.1	–
Image-Blinded	51.8	↓31.6	27.9	↓50.9	24.6	↓65.1	28.4	↓58.7
Text-Input	81.4	–	54.8	–	71.2	–	74.9	–
Text-Blinded	50.9	↓30.5	23.3	↓31.5	18.7	↓52.5	26.2	↓48.7

match human preferences. Together with the human-standardized evaluation on the Image-AR track reported in the main paper, these results confirm that *acc.@evi.* consistently aligns with human judgments of anomaly evidence across all three modalities.

Table 11: Human preference alignment of *acc.@evi.* on Video-AR and Text-AR tracks. Lower rank numbers indicate better performance.

Track	Evaluation	Ola	InternVL3.5	Qwen3-VL	Σ_{diff}
Video	acc.	7	10	2	5
	Temporal IoU	6	7	5	6
	acc.@evi.	10	8	2	**2**
	human	9	8	3	-
Track	Evaluation	Ola	Qwen3	Llama-3.1	Σ_{diff}
Text	acc.	4	1	3	5
	Span IoU	3	2	4	4
	acc.@evi.	2	1	6	**1**
	human	2	1	5	-

A.6 Mitigation of Potential Biases and Contamination

To reduce possible model-family bias arising from LLM-assisted MCQ generation, we employ multiple closed-source models for initial question synthesis, enforce a standardized cross-modal task formulation that constrains question format and evaluation criteria, and apply extensive human filtering and refinement to remove unnatural patterns or model-specific artifacts. Empirical results show no systematic advantage for models belonging to the same families as the generators.

Regarding potential benchmark contamination from public training data, UniAR mitigates the risk through its unified cross-modal task framework, extensive newly introduced modality-specific annotations, and the evidence-grounded evaluation metric. These design choices require genuine anomaly reasoning beyond simple pattern memorization. We observe no consistent performance advantages for models that may have been exposed to related public data, further supporting the reliability of the benchmark.

A.7 MCQ Generation Prompt

MCQs serve as a unified evaluation paradigm in our MLLM assessment. For the text modality, we adopt the datasets and annotations for global anomaly discrimination, global anomaly classification, and analysis from MMAD Jiang et al. (2024). On this basis, we further supplement bounding box annotations of anomalous regions, along with refined referential anomaly discrimination and fine-grained anomaly classification labels. We only retain local abnormal defects to evaluate the reference-free anomaly reasoning capability of the MLLMs. We explicitly incorporate the instruction

"improve the misleadingness of distractors" into the MCQ generation prompt to increase task difficulty. For the open-ended reasoning HAA task requiring textual revision suggestions, we add the constraint *"generate revision proposals grounded in specific context and topic background"* to the prompt. For video anomaly analysis, we further embed the guideline *"put forward positive and constructive operational suggestions by considering scenario context, participant characteristics, and potential event impacts"*. These prompt designs ensure the outputs cover diverse valid solutions rather than being limited to a single standard answer. Moreover, manual inspection is conducted to verify **question difficulty**, as well as the **correctness and diversity of responses**, so as to guarantee the validity of the constructed MCQs.

We use the following prompt to generate rough MCQs for videos based on Gemini-2.0-Flash.

```
"""
You are given a video (only visual information, no audio).  Your task is to generate high-quality multiple-
choice questions (**four-choice, single answer**) about ABNORMAL EVENTS in the video. You need to generate
questions following these guidelines:
1. TWO questions about the REAL TIME and LOCATION where the abnormal event occurred in real life:
- DO NOT use any video timestamps such as 0:01 or 0:05 marks.
- Describe ONLY real-world time: daytime, morning, noon, afternoon, evening, night, dawn, etc.
- If a clock, watch, or text showing the exact real time is visible in the frames, use that exact time.
- If NO real time information can be inferred from the frames, **DO NOT generate any time-related questions**.
2. TWO questions about the number, appearance, and other key identifying characteristics of the participants (
perpetrator and victim) or related entities.
3. ONE question about the type of abnormal event that occurred in the video. Use a general term such as theft,
 robbery, traffic accident, arson, or violation of regulations.
4. TWO questions about the event process: the cause, process, result, and key details of the abnormal event.
5. ONE question about the correct and positive actions to take if you are involved in or witness such events.
Put forward positive and constructive operational suggestions by considering scenario context, participant
characteristics, and potential event impacts.
Requirements:
- All questions must be based solely on the VISUAL content of the video.
- All questions must be related to ABNORMAL events in the video.
- Each question must be clear, specific, and objectively verifiable.
- Each question has 4 candidates labeled A, B, C, D. Improve the misleadingness of distractors.
- Output strictly in the JSON format shown below.
Output format (follow exactly):
{"questions": [{
        "question": "Your question about the video",
        "answer": "The correct text answer",
        "candidates": ["Option A", "Option B", "Option C", "Option D"],
        "right_answer": "A" or "B" or "C" or "D"
    }, ...]}
Return the JSON enclosed in <json></json> tags.
"""
```

We use the following prompt to generate rough MCQs for texts based on GPT-5.4.

```
### NON-NEGOTIABLE RULES
1. Output ONLY a valid JSON object (no extra text, no markdown, no placeholders).
2. "anomaly_type": The anomaly type of the token "{first_anomaly_token}" within the span position index {
first_span}, with four categories represented by A/B/C/D:
A: Grammatical_Error B: Lexical_Error
C: Factual_Hallucination D: Syntactic_Structural_Error
3. "text_field": The topic/subject of the given text, with six categories represented by A/B/C/D/E/F:
A: History & Biography  B: Culture/Art/Entertainment
C: Society/Education/Daily Life D: Technology/Sports/Events
E: Work/Career/Public Affairs F: Argumentative/Expository Writing
4. "edits": An appropriate Correction_String for the anomalous token "{first_anomaly_token}" within the span
position index {first_span}. Generate revision proposals grounded in specific context and topic background.
    - The correction must be a short English string that replaces the anomalous token without introducing
    grammatical, lexical, or factual errors in context.
    - If the correction is to delete the token, simply return an empty string "".
5. "confident": The confidence score for the well-formedness of the full text after replacing the anomalous
token "{first_anomaly_token}" at span index {first_span} with the above Correction_String.
6. "cands": For the anomaly token "{first_anomaly_token}" within span start/end index {first_span}, provide 3
different INCORRECT English strings at that location. Improve the misleadingness of distractors.
   - Each string should be **absolutely** grammatically/lexically/factually **error** in the context when used
    to replace tokens within span {first_span}.
   - Each string should be **semantically** similar to "{first_anomaly_token}" but MUST different.
   - Format: List of strings.
7. JSON must be fully closed and parsable.
### INPUT
Claim: {item['claim']}
anomaly_token: {first_anomaly_token}
anomaly_spans: {first_span}
### OUTPUT ONLY THIS JSON (fill in the actual values):
{{"anomaly_type": Option,
"text_field": Option,
"edits": Correction_String,
"confident": Confidence_Score,
"cands": [...]}}
"""
```

A.8 Data Structure

To facilitate standardized and unified evaluation across diverse modalities and tasks, we adopt a unified annotation data structure for evaluation data spanning different modalities. This unified framework mainly consists of three core components: basic sample information, fine-grained evidence annotations, and multi-choice question (MCQ) items. An annotation example from the Video-AR track is presented in the following content. Notably, the data annotation structures for image and text modalities are designed analogously to ensure cross-modal consistency and scalable evaluation.


```
{
    basic_info:{
        "video_path": "01_Accident_001.mp4",
        "id": 7698,
        "fps": 24.9777,
        "anomaly": "1",
        "video_duration": 14.933333,
        "total_frames": 373
        },
    "evidences": {
        "temporal": [
            {
                "labels": ["Accidental Hazards and Sudden Injuries"],
                "time_range": {"start": 66, "end": 192}
            }
        ],
        "spatial": [{"frame": 66, "time": 2.64, "position": {"x": 47.15, "y": 27.33, "width": 10.14, "height":
         18.33}},
            ...
        {"frame": 129, "time": 5.16, "position": {"x": 41.04, "y": 21.33, "width": 8.19, "height": 12.67}}]
    },
    "QAs": [{"question": "Is there any abnormal event in the video?",
             "answer": "Yes",
             "candidates": ["A. Yes", "B. No"],
             "right_answer": "A",
             "task": "Video Anomaly Discrimination"},
               ...
             {"question": "If you were a witness to this traffic accident, what is the most appropriate
             immediate action to take?",
             "candidates": [
                    "A. Drive past quickly to avoid involvement.",
                    "B. Immediately confront the driver of the white car.",
                    "C. Ensure personal safety, then call emergency services and offer assistance if safe to
                    do so.",
                    "D. Record the incident on your phone without taking any further action."
                ],
             "right_answer": "C",
             "type": "Anomaly Event Analysis"}]
}
```


B Evaluation Details

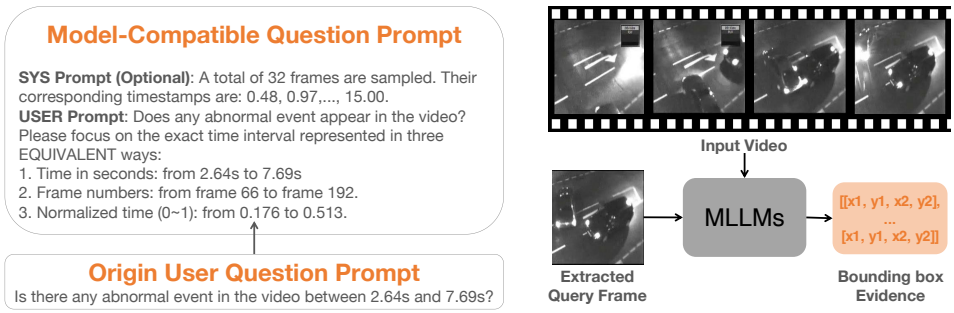


Figure 8: Video anomaly reasoning-centric evaluation. Left: In addition to adopting trimmed event clips as input, we convert referring prompts and temporal cues into multiple equivalent formats, ensuring referential information is complete and universally model-compatible. Right: Unlike temporal interval evidence, spatial evidence is frame-bound. Thus, we explicitly supply key query frames as additional input, guaranteeing full access to all necessary information during reasoning.

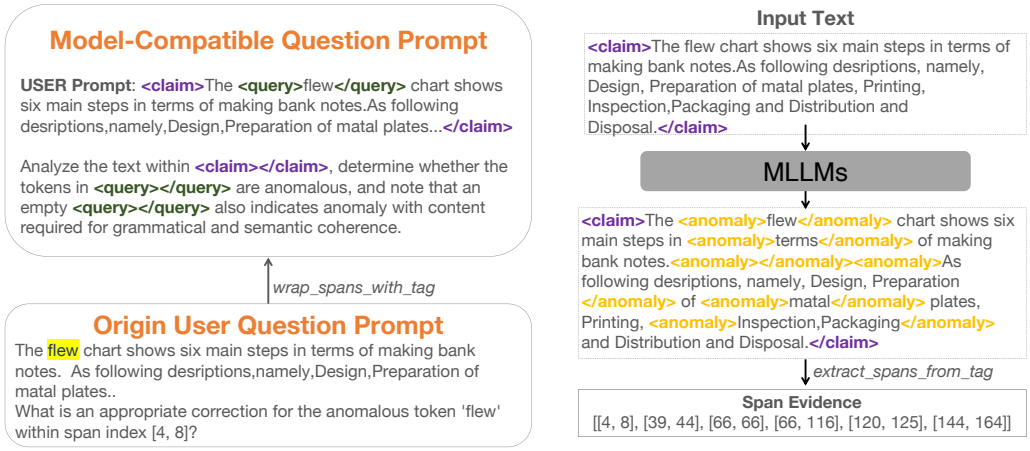


Figure 9: Text anomaly reasoning-centric evaluation. Instead of relying on span indices, which MLLMs process inefficiently, we adopt customized <tag>/</tag> markers to supply model-friendly textual positional cues. In the referring setting, we employ <query>/</query> to demarcate target text spans (left). Meanwhile, we require MLLMs to enclose anomalous content in <anomaly>/</anomaly> tags to explicitly mark abnormal words. Anomaly spans are subsequently extracted via post-processing (right).

B.1 Anomaly Reasoning-Centric Evaluation

To fairly and comprehensively evaluate the cross-modal anomaly reasoning ability of MLLMs, we accommodate their inherent prompt compatibility across modalities. In referring discrimination and evidence generation, models must accurately understand referential cues and generate outputs following standard formats. Both referential comprehension and format adherence greatly influence end-to-end anomaly reasoning performance. For the image modality, we use unified aspect-ratio normalized bounding box coordinates in top-left and bottom-right format. This lightweight design enjoys broad compatibility and native support among mainstream MLLMs.

For video data, we adopt two strategies to guarantee sufficient referential information for all models: pre-trimming temporal referential content (*trim.acc.*), and embedding precise timestamps and frame indices into question prompts. This provides complete temporal details, including exact timestamps, frame numbers, and relative temporal proportions. For evidence generation, key frames containing participant and entity positional information are fed into the model along with raw video inputs, thereby reducing intermediate processing and reasoning costs. The evaluation pipeline for video anomaly reasoning with compatible input is visualized in Figure 8.

For text anomaly reasoning, we propose a tag-based paradigm to provide concise, model-friendly positional cues for stable anomaly localization. Instead of inefficient span indices, we adopt custom markup tags to structure textual position information. We use ¡query¿¡/query¿ tags to highlight target spans in user prompts for easier referential understanding, and require models to enclose abnormal content within ¡anomaly¿¡/anomaly¿ tags to explicitly mark boundaries. Standardized post-processing then parses tagged outputs to extract anomaly spans, simplifying result collection and eliminating extra reasoning overhead from ambiguous text indexing. The text anomaly reasoning evaluation framework is shown in Figure 9.

B.2 Metric Details

Interval and BBox IoU In the video and image modalities, we instruct the model to directly output predicted interval and bounding box lists, which are subsequently converted into the *mask* representation. We then compute the one- and two-dimensional *mask IoU* against the ground truth.

Span IoU Predicted spans are extracted from model outputs according to ¡anomaly¿¡/anomaly¿ tags. Each string is sequentially matched against the original claim text to obtain character-level spans. Empty strings are ignored for simplicity. Finally, we compute the one-dimensional IoU as the ratio of the total overlap length to the total union length of the predicted and ground truth spans.

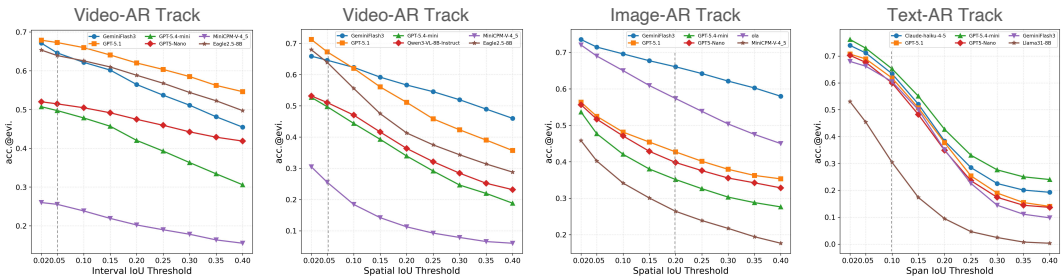


Figure 10: Ablation studies on *IoU* thresholds of the *acc.@evi.* for anomaly discrimination across Video-AR Track, Image-AR Track, and Text-AR Track. Results demonstrate that varying interval, spatial, and span evidence thresholds exert negligible influence on the relative performance ranking of MLLMs. The gray dashed line indicates our empirically adopted threshold in the main experiments.

We study how the threshold, which quantifies the proximity between model-generated evidence and ground truth across modalities, affects our proposed reliable anomaly discrimination metric *acc.@evi.*. We conduct evaluations over six models to examine how spatial-temporal IoU (video), spatial IoU (image), and span IoU (text) influence *acc.@evi.*. As shown in Figure 10, *acc.@evi.* gradually drops as the threshold increases. Within a reasonable threshold range, all models retain consistent relative rankings and comparable performance gaps, which implies that the evaluation results are not sensitive to threshold selection. In our implementation, thresholds are chosen according to the strictness of evidence matching required by each task. Specifically, we set both temporal and spatial IoU thresholds to 0.05 for Video-AR; the spatial IoU threshold to 0.2 for Image-AR; and the span IoU threshold to 0.1 for Text-AR. The IoU threshold defines the stringency of evidence matching for *acc.@evi.*: given a fixed threshold, *acc.@evi.* assesses whether anomaly predictions are backed by sufficiently grounded evidence, supporting fair comparisons among MLLMs. We select relatively low thresholds to accommodate varied output-parsing abilities of different MLLMs and mitigate biases introduced by non-reasoning factors, such as formatting and parsing failures. Crucially, our emphasis is on the robustness of relative model rankings, rather than absolute performance under a single threshold. Even when the IoU threshold varies across a wide range for temporal, spatial, and span evidence in all modalities, *acc.@evi.* yields largely consistent MLLM rankings, verifying its robustness to threshold configurations.

B.3 Task Prompts

We use the following task prompts in global, referring, and evidence-based settings in Video-Track.

```
 "glob_acc": (
    "You will be provided with sampled frames from a video, along with a binary-choice question that includes
    a question and two answer options.\n Your task is to analyze the provided frames, infer the most plausible
     answer based on the visual information.\n If the video does not provide enough information, infer the
    answer based on the options available and still provide a result. Therefore, In all cases, an answer must
    be given.\n Only output the answer in the following format:\n\n"
    '{"result": "option"}'
    'The "option" is the uppercase letter corresponding to your answer.\n\n'
)
```

```
"ref_prom": (
    "You will be provided with sampled frames from a video, along with a binary-choice question that includes
    a question and two answer options.\n Your task is to analyze the provided frames, infer the most plausible
     answer based on the visual information.\n If the video does not provide enough information, infer the
    answer based on the options available and still provide a result. Therefore, In all cases, an answer must
    be given.\n Only output the answer in the following format:\n\n"
    '{"result": "option"}'
    "The 'option' is the uppercase letter corresponding to your answer.\n\n")
```

```
"ref_trim": (
    "You will be provided with sampled frames from a video, along with a binary-choice question that includes
    a question and two answer options.\n Your task is to analyze the provided frames, infer the most plausible
     answer based on the visual information.\n If the video does not provide enough information, infer the
    answer based on the options available and still provide a result. Therefore, In all cases, an answer must
    be given.\n Only output the answer in the following format:\n\n"
    '{"result": "option"}'
    "The 'option' is the uppercase letter corresponding to your answer.\n\n"
)
```

```
"st-iou": """
    You will be provided with uniformly sampled frames from a video and their timestamps, along with a frame
    image (the last image) extracted from the original video. Your task is to answer a binary-choice (yes/no)
    question on anomaly discrimination with explicit visual evidence. If your answer is "B. No", you MUST
    output ONLY: {"intervals": [], "bboxes": []}

    If your answer is "A. Yes", you MUST output both of the following:
    - Time intervals of abnormal events
    - Bounding boxes of related or main participants/entities in given frame image
    You MUST provide both, not just one.

    Intervals Definition:
    - MUST cover the FULL duration of the abnormal event
    - Each interval is [start, end], 0 = video start, 1 = video end
    - Values of start and end MUST be normalized between 0 and 1, NEVER output start and end greater than 1
    Intervals Constraints:
    - Do NOT output overly short or fragmented intervals
    - NEVER exceed 4 intervals under ANY circumstance
    If too many intervals:
    - Merge overlapping intervals
    - Merge adjacent or very close intervals
    - Prefer longer continuous intervals over fragmented ones
    - Keep the most representative abnormal segments only

    Bounding-box  Definition:
    - Each bbox is [x1, y1, x2, y2]
    - Coordinates are normalized between 0 and 1
    - (x1, y1) = top-left, (x2, y2) = bottom-right
    Bounding-box Constraints:
    - MUST satisfy: 0 < x1 < x2 < 1 and 0 < y1 < y2 < 1
    - Boxes must correspond to anomaly-relevant participants/entities
    - Do NOT output redundant or duplicate boxes
    - NEVER exceed 5 bounding boxes under ANY circumstance
    If too many bounding boxes:
    - Select ONLY the most important entities
    - Prioritize entities directly involved in the abnormal event
    - Remove less relevant or overlapping boxes

    Output Format: Output ONLY the JSON. Do NOT include any explanation or extra text.
    Format:
    {"intervals": [[start1, end1], [start2, end2], ...],
    "bboxes": [[x1, y1, x2, y2], ...]}
    - Number of intervals <= 4
    - Number of bounding boxes <= 5
    - All interval values satisfy 0 < start < end < 1
    - All bbox values satisfy 0 < x1 < x2 < 1 and 0 < y1 < y2 < 1
    """
```

We use the following task prompts in global, referring, and evidence-based settings in Image-Track.

```
"glob_acc": (
    "You will be provided with an image, along with a "
    "binary-choice question that includes a question and two answer options.\n"
    "Your task is to analyze the provided image, infer the most plausible "
    "answer based on the visual information.\n"
    "If the image does not provide enough information, infer the answer based "
    "on the options available and still provide a result. "
    "Therefore, In all cases, an answer must be given.\n"
    "Only output the answer in the following format:\n\n"
    '{"result": "option"}'
    'The "option" is the uppercase letter corresponding to your answer.\n\n'
    )
```

```
"ref_acc_prom": (
    "You will be provided with an image, along with a "
    "binary-choice question that includes a question and two answer options.\n"
    "Your task is to analyze the provided image, infer the most plausible "
    "answer based on the visual information.\n"
    "If the image does not provide enough information, infer the answer based "
    "on the options available and still provide a result. "
    "Therefore, In all cases, an answer must be given.\n"
    "Only output the answer in the following format:\n\n"
    '{"result": "option"}'
    "The 'option' is the uppercase letter corresponding to your answer.\n\n"
    )
```

```
"s-iou": """
    You will be provided with an image. And your task is to answer a binary-choice (yes/no) question on
    anomaly discrimination with explicit visual evidence.
    If your answer is "B. No", you MUST output ONLY:
    {"bboxes": []}

    If your answer is "A. Yes", you MUST output both Bounding boxes of related/main anomalous regions in given
     image.

    Bounding-box  Definition:
    - Each bbox is [x1, y1, x2, y2]
```

```
    - Coordinates are normalized between 0 and 1
    - (x1, y1) = top-left, (x2, y2) = bottom-right

    Bounding-box Constraints:
    - MUST satisfy: 0 < x1 < x2 < 1 and 0 < y1 < y2 < 1
    - Boxes must correspond to anomaly-relevant participants/entities
    - Do NOT output redundant or duplicate boxes
    - NEVER exceed 5 bounding boxes under ANY circumstance

    If too many bounding boxes:
    - Select ONLY the most important entities
    - Prioritize entities directly involved in the abnormal event
    - Remove less relevant or overlapping boxes

    Output Format: Output ONLY the JSON. Do NOT include any explanation or extra text.

    Format:
    {"bboxes": [[x1, y1, x2, y2], ...]}
    - Number of bounding boxes <= 5
    - All bbox values satisfy 0 < x1 < x2 < 1 and 0 < y1 < y2 < 1
    """
```

We use the following prompt to evaluate MLLMs with global, referring, and evidence-based settings in Text-Track.

```
"glob_acc":
    """Define Text Anomalies: Lexical Error, Grammatical Error, Syntactic Structural Error, and contextual
    Factual Hallucination. You are provided with an text wrapped with <claim></claim>, along with a binary-
    choice question that includes a question and two answer options.\n" Your task is to analyze the provided
    text, infer the most plausible answer based on the textual information.\n If the text does not provide
    enough information, infer the answer based on the options available and still provide a result. Therefore,
     In all cases, an answer must be given.\n Only output the answer in the following format:\n\n {"result": "
    option"} The "option" is the uppercase letter corresponding to your answer.\n\n"""
```

```
"spec_acc_prom":
    """Define Text Anomalies: Lexical Error, Grammatical Error, Syntactic Structural Error, and contextual
    Factual Hallucination. You are provided with an text wrapped with <claim></claim>, along with a  binary-
    choice question that includes a question and two answer options.\n Your task is to analyze the provided
    text, infer the most plausible  answer based on the textual information.\n If the text does not provide
    enough information, infer the answer based on the options available and still provide a result. Therefore,
     In all cases, an answer must be given.\n Only output the answer in the following format:\n\n"
    '{"result": "option"} '
    "The 'option' is the uppercase letter corresponding to your answer.\n\n"""
```

```
"s-iou":
    """You are provided with text wrapped by <claim></claim>.
    Your task: perform binary anomaly discrimination judgment, and strictly output in the required JSON format
     only.
    Anomaly definition: Lexical Error, Grammatical Error, Syntactic Structural Error, Contextual Factual
    Hallucination.
    YOU MUST DETECT ALL ANOMALIES AND ADD <anomaly> TAGS.
    IF ANOMALY EXISTS:
        - YOU DO NOT ADD <anomaly> TAGS, YOUR OUTPUT IS INVALID.
        - YOU MUST NOT OUTPUT THE TEXT WITHOUT TAGS.
        - YOU MUST WRAP EVERY ANOMALOUS WORD WITH <anomaly> AND </anomaly>
        - For **missing content / incomplete context anomalies at a position**: directly INSERT empty tag <
        anomaly></anomaly> at THAT missing position.
        - OUTPUT: {"spans": "full input text WITH <anomaly> WRAPPED AROUND EVERY ERROR"}
    ELSE:
        - OUTPUT: {"spans": "full text"}
    ## STRICT NON-NEGOTIABLE MANDATORY RULES
    1. ONLY process content inside <claim></claim>. Ignore all outside content.
    2. NEVER output partial text, NEVER truncate text, NEVER output A.Yes/B.No directly.
    3. Do NOT modify any normal words, do NOT rewrite sentences, only ADD <anomaly> tags on abnormal parts.
    4. Do NOT output any explanation, comment, extra text, reasoning, answer words. Only output the single
    JSON object.

    ## CORRECT OUTPUT EXAMPLES (MUST FOLLOW THIS FORMAT)
    Example 1 (has anomaly):
    {"spans": "My parents are very <anomaly>support</anomaly>."}

    Example 2 (no anomaly):
    {"spans": ""}

    Now detect the text inside <claim></claim>."""
```

C CASE STUDY

C.1 VIDEO-AR TRACK

GLOBAL ANOMALY DISCRIMINATION (GLOB.ACC.)

Good case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: A



Full Video

Bad case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: B



Full Video

REFERRING ANOMALY DISCRIMINATION (PROM.ACC.)

Good case

Question:
Does any abnormal event appear in the video? Please focus on the exact time interval: from 0.0s to 352.9s.
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: A



Full Video

Bad case

Question:
Does any abnormal event appear in the video? Please focus on the exact time interval: from 12.0s to 18.8s.
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: B



Full Video

REFERRING ANOMALY DISCRIMINATION (TRIM.ACC.)

Good case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: A



Trimmed Video

Bad case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: A
Gemini-3-Flash: B



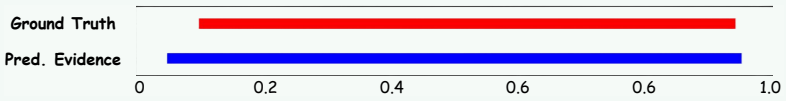
Trimmed Video

EVIDENCE-BASED ANOMALY DISCRIMINATION (ACC.@EVI.)

Good case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: Red Intervals and Bboxes
Gemini-3-Flash: Blue Intervals and Bboxes
Interval IoU: 0.91 ¿ $thres_t = 0.05$
Bbox IoU: 0.38 ¿ $thres_s = 0.05$ –¿ A. Yes (Supported)

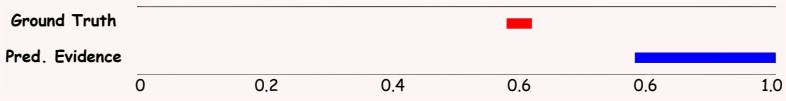




Bad case

Question:
Is there any abnormal event in the video?
Choices:
A. Yes B. No
Ground Truth: Red Intervals and Bboxes
Gemini-3-Flash: Blue Intervals and Bboxes
Interval IoU: 0 ¡ $thres_t = 0.05$
Bbox IoU: 0.001 ¡ $thres_s = 0.05$ –¿ B. No (Unsupported)





ANOMALY CLASSIFICATION (PROM.ACC.)

Good case

Question:
What is the category of the abnormal event? Please focus on the exact time interval: from 9.5s to 13.9s (Time in seconds:).
Choices:
A. Dangerous Criminal Activities
B. Accidental Hazards and Sudden Injuries
C. Malicious and Harmful Interactions
D. Rule Violations and Non-Compliance
E. Authorized Disruptive Events
Ground Truth: D
Gemini-3-Flash: D



Full Video

Bad case

Question:
What is the category of the abnormal event? Please focus on the exact time interval: from 0.5s to 26.0s (Time in seconds:).
Choices:
A. Dangerous Criminal Activities
B. Accidental Hazards and Sudden Injuries
C. Malicious and Harmful Interactions
D. Rule Violations and Non-Compliance
E. Authorized Disruptive Events
Ground Truth: D
Gemini-3-Flash: A
Task: Event-level Anomaly Classification



Full Video

ANOMALY CLASSIFICATION (TRIM.ACC.)

Good case

Question:
What is the category of the abnormal event in the trimmed video?
Choices:
A. Dangerous Criminal Activities
B. Accidental Hazards and Sudden Injuries
C. Malicious and Harmful Interactions
D. Rule Violations and Non-Compliance
E. Authorized Disruptive Events
Ground Truth: B
Gemini-3-Flash: B



Trimmed Video

Bad case

Question:
What is the category of the abnormal event in the trimmed video?
Choices:
A. Dangerous Criminal Activities
B. Accidental Hazards and Sudden Injuries
C. Malicious and Harmful Interactions
D. Rule Violations and Non-Compliance
E. Authorized Disruptive Events
Ground Truth: C
Gemini-3-Flash: A



Trimmed Video

CONTEXT UNDERSTANDING (GLOB.ACC.)

Good case

Question:
Where does the multi-car accident primarily occur?
Choices:
A. On an open highway
B. Inside an underpass/tunnel
C. At a busy intersection above ground
D. In a parking lot
Ground Truth: B
Gemini-3-Flash: B
Task: Context Understanding



Full Video

Bad case

Question:
According to the visible timestamp, at what real-world time of day did the abnormal event occur?
Choices:
A. Morning
B. Afternoon
C. Evening
D. Night
Ground Truth: D
Gemini-3-Flash: A
Task: Context Understanding



Full Video

Anomaly Analysis (glob.acc.)

Good case

Question:
What does the perpetrator initially do with the shopping cart?
Choices:
A. He leaves it unattended.
B. He pushes it down an aisle and stops.
C. He loads it with many items.
D. He uses it to block an exit.
Ground Truth: B
Gemini-3-Flash: B



Full Video

Bad case

Question:
How many individuals were directly involved in the theft at the counter?
Choices:
A. One B. Two
C. Three D. Four or more
Ground Truth: A
Gemini-3-Flash: B



Full Video

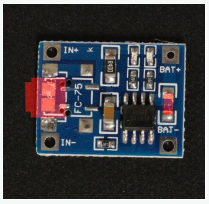
C.2 Image-AR Track

Global Anomaly Discrimination (glob.acc.)

Good Case

Question: Is there any defect in the object?
Choices:
A. Yes
B. No

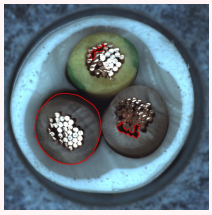
Ground Truth: A
Gemini-3-Flash: A



Bad Case

Question: Is there any defect in the object?
Choices:
A. Yes
B. No

Ground Truth: A
Gemini-3-Flash: B

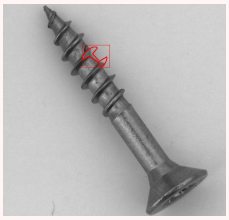


REFERRING ANOMALY DISCRIMINATION (PROM.ACC.)

Bad Case

Question: Is there any anomaly within the bounding box [[0.34, 0.26, 0.43, 0.34]] in the image? Bounding box format: (x1, y1, x2, y2) are normalized to [0, 1], where (x1, y1) denotes the top-left corner and (x2, y2) denotes the bottom-right corner.
Choices:
A. Yes
B. No

Ground Truth: A
Gemini-3-Flash: B

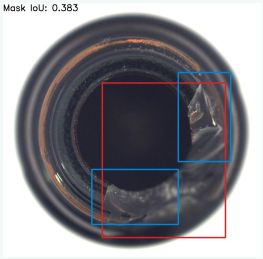


EVIDENCE BASED ANOMALY DISCRIMINATION (ACC.@EVI.)

Good Case

Question: Is there any defect in the object?
Choices:
A. Yes.
B. No.

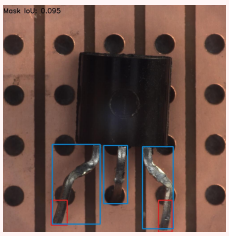
Ground Truth: Yes. Red Box
Gemini-3-Flash: Yes. Blue Box
Bbox IoU: 0.4304 –¿ A. Yes (Supported)



Bad Case

Question: Is there any defect in the object?
Choices:
A. Yes.
B. No.

Ground Truth: Yes. Red Box
Gemini-3-Flash: Yes. Blue Box
Bbox IoU: 0.095 –¿ B. No (Unsupported)

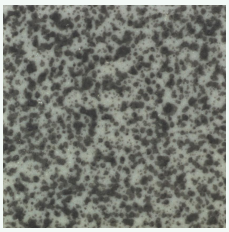


CONTEXT UNDERSTANDING (GLOB.ACC.)

Good Case

Question: What kind of product is in the image?
Choices:
A. A fabric swatch.
B. A digital screen.
C. A tile.
D. A piece of paper.

Ground Truth: C
Gemini-3-Flash: C

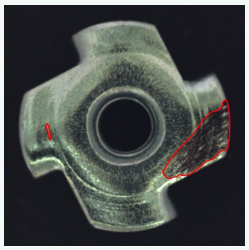


ANOMALY ANALYSIS (GLOB.ACC.)

Good Case

Question: There is a defect in the object. What effect does the defect have?
Choices:
A. Makes it easier to tighten.
B. Changes the nut's color.
C. Increases the nut's strength.
D. Affects the uniformity of the metal's finish.

Ground Truth: D
Gemini-3-Flash: D



C.3 TEXT-AR TRACK

GLOBAL ANOMALY DISCRIMINATION (GLOB.ACC.)

Good Case

Question: Does the text contain any anomalies?
Choices:
A. Yes B. No

Ground Truth: A
Gemini-3-Flash: A

`plane, train, and boat are the ways to travel. I think the boat is the best way to travel and I have three reasons for my idea. First of all, boat moves across the sea and the sea is very beautiful. For instance, when we move on the sea, we can see many types of animals, such as birds, turtles, and fishes. Next, the boat usually is`

Bad Case

Question: Does the text contain any anomalies?
Choices:
A. Yes B. No

Ground Truth: A
Gemini-3-Flash: B

`I write to you to express my interest to apply for the weekend worker job that was posted on your cafe door. I believe I am a good fit for this task since I had previous experience working in similar jobs. I am good at dealing with people. I demonstrated courteous and professional attitude and`

REFERRING ANOMALY DISCRIMINATION (PROM.ACC.)

Good Case

Question: Are the word tokens wrapped by ¡query¿ and ¡/query¿ anomalous?
Choices:
A. Yes B. No

Ground Truth: A
Gemini-3-Flash: A

<claim>I´m Sara and i´m 24 years old. I´m interested in this job because at the moment I haven´t got a job. I´m a <query> **worker**</query> person and I like children. In the work I´m a responsibility person, I never last arriving. Also I always work with children and I have a teachercertificate. I´m a friendlier person and I have a good curriculum....</claim>

Bad Case

Question: Are the word tokens wrapped by ¡query¿ and ¡/query¿ anomalous?
Choices:
A. Yes B. No

Ground Truth: A
Gemini-3-Flash: B

<claim>Many people still consider television only as a way to relax, <query>get</query> entertained and stop thinking about all the problems that might worry them. It is also important, though, not to underestimate all the educational possibilities that it gives you, as sometimes you can learn more from a documentary than from a school book. </claim>

Evidence-based Anomaly Discrimination (acc.@evi.)

Good Case

Question: Does the text contain any anomalies?
Choices:
A. Yes B. No

Ground Truth: A (Span in red and green)

Gemini-3-Flash: A (iou=0.37) –¿ A (Supported)

Evidence Span in blue and green

In my birthday's day, I directed go to school this morning. At this point, someone invites sing to me , after school. they touched to me because there are a lot of people had planned a birthday celebrate to me together . After finish to the home , due to I was so tired that I fell asleep and fell down . The next day , I was reluctanted to rise and class . Suddenly, I thought of my gift was gone in the class , the students said he / she didn't take ^, but I doubt them , so we quarrel. I came home from school, only to found a gift on the sofa . loffer an apology out of harm students, but fortunately he didn't got angry for me

Bad Case

Question: Does the text contain any anomalies?
Choices:
A. Yes B. No

Ground Truth: A (Span in red and green)

Gemini-3-Flash: A (iou=0.03) –¿ B (Unsupported)

Evidence Span in blue and green

Belarus is not a popular tourist destination nowadays^ but a lot of people would be surprised to find ^ real medieval atmosphere in Grodno,a western city -of the country. Founded in the 12th century, it came through all^ important events of Eastern Europe history, from^ Tatar invasions till the World War II^ which was a real disaster. Two ancient castles remind us of the times of kings and nobility, quiet pedestrian streets show us the life as^ it could be in the 1800s. Grodno is situated on the crossroads of great trading ways that's why one could hear Polish, Jewish, German,^ Lithuanian there. Its multiethnic population was the reason to organise and held the Festival of national cultures every two years. It's worth visiting because traditions of the ancient Belarusian, Tatar, Russian, Jewish^ shown alive make us feel history itself. City of different religions, different nationalities and \"talking stones\" is waiting for you.

Context Understanding (glob.acc.)

Good Case

Question: What is the main topic of this text?
Choices:
A. History & Biography
B. Culture/Art/Entertainment
C. Society/Education/Daily Life
D. Technology/Sports/Events
E. Work/Career/Public Affairs
F. Argumentative/Expository Writing
Ground Truth: C
Gemini-3-Flash: C

I felt nervous when the phone rang. I was sitting calmly ina cafe, when I heard it ringing. Firstly I thought it could be one of my best friends as we really liked to have lunch together. But I saw unknown phone number on a display and felt a little bit nervous. What if it were my boss or clients and possibly something dreadful had happened to my job. It went on ringing and with great fear I clicked on the green button. Luckily it was my colleague Jane who wanted to drink a cp of fresh coffee with me

ANOMALY ANALYSIS (GLOB.ACC.)

Bad Case

Question: What is an appropriate correction for the anomalous token " as" within span index [172, 175]?
Choices:
A. for resolving legal cases and disputes
B. because resolve legal cases and disputes
C. to resolve legal cases and disputes
D. "" (delete it)

Ground Truth: D
Gemini-3-Flash: A

The Jirga is a 'traditional system' of justice, which is actively practiced in, Khyber Pakhtunkhwa , Baluchistan as well as in districts of Sindh bordering with Baluchistan as to resolve legal cases and disputes. The system, which was in practice since the British rule, was modified under the 'Special Provision Ordinance (Criminal Provisions 11) of 1968' . Besides dispensing justice and dealing with judicial cases, the Jirga also covers all important aspects of monetary, ...

Good Case

Question: What is an appropriate correction for the anomalous token "for the public transport there is a future" within span index [13, 55]?
Choices:
A. there is a future for public transport
B. for the public transport there will be futures
C. for public transport there are a future
D. for the public transport it has a future
Ground Truth: A
Gemini-3-Flash: A

I think that for the public transport there is a future, because over time the gasoline prices and the value of the reparations of the vehicles increase. Like in other things the prices increase, from the aliments to clothes and home supplies, gas, water, eletricity.. Then the people will choose the public transport in the future.

D LIMITATION

Despite this work's contributions, several limitations exist, pointing to future research directions.

A key limitation of our dataset is that initial MCQs were mainly generated by Gemini and GPT series models. These models may embed their inherent reasoning biases into annotations, even though manual verification mitigated this issue to some extent. Future work should explore diverse MCQ generation sources to enhance dataset robustness and generalizability. Another limitation concerns the evaluation of model reasoning performance, particularly evidence output. Requiring models to produce format-compliant, reasoning-reflective evidence is challenging for some models, especially in cross-modal scenarios. Although model-compatible pre- and post-processing were applied, models' limited instruction execution capabilities still cause discrepancies between evaluation results and true performance. Addressing this requires improving models' instruction understanding/output control and enhancing the evaluation framework's compatibility.

E COMPUTING RESOURCES

For open-source models, all experiments are conducted uniformly on an NVIDIA RTX A6000 GPU with 48 GB of memory. For closed-source models, we adopt the third-party API with 32 processes.