AUDITING THE GROUND TRUTH: A FIELD-AWARE AUDIT OF VIDEO ANOMALY UNDERSTANDING BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language annotations in video anomaly understanding benchmarks guide model training and evaluation. Their quality must be assessed in relation to field responsibilities, temporal scopes, and granularity. We introduce a contract-aware audit framework that translates native field requirements into diagnostic questions and specifies the permitted evidence. Four tasks assess video quality, standalone text quality, native cross-text consistency, and video–text alignment. To determine whether a video–text pair meets multiple quality requirements simultaneously, we link video, standalone text, and alignment judgments to compute conditional joint usability. Our study includes 7,928 video entries and 56,103 video–text pairs across nine benchmarks. In the model-assisted audit, conditional joint usability ranges from 22.07% to 74.20% across benchmarks under field-macro aggregation. The predominant failure pattern is that video and text pass standalone checks, but their pairing fails alignment. To test whether audit-guided annotation revision improves performance, we revise CUVA descriptions and conduct controlled experiments with LVLMs, evaluating both description generation and multiple-choice question answering on VALU. The results show downstream benefits from revision, with effects varying by model and evaluation metric. Field-level audits further reveal quality differences within individual benchmarks. These diagnostics identify annotations requiring review and, together with native task coverage, guide the selection of benchmarks and fields suited to the evaluation goal and temporal granularity.

1 INTRODUCTION

Video anomaly detection (VAD) aims to identify anomalous events and localize them in time (Sultani et al., 2018; Zhu et al., 2026). With advances in vision–language models, video anomaly understanding (VAU) further requires natural-language descriptions of events and explanations of their causes and consequences (Du et al., 2024b; He et al., 2026). To support these understanding tasks, researchers reuse videos from existing datasets such as UCF-Crime and add language annotations in different forms and at different granularities to construct new benchmarks. As Figure 1 illustrates, UCF-CAP (Chatzikonstantinou et al., 2022), UCA (Yuan et al., 2024), and HIVAU-70K (Zhang et al., 2025a) share UCF-Crime as a video source, yet provide different annotation content and structures for these videos. These annotations serve as both training targets and evaluation references. Test-label errors can affect model comparisons (Northcutt et al., 2021), motivating a systematic audit of reference annotations and their video evidence.

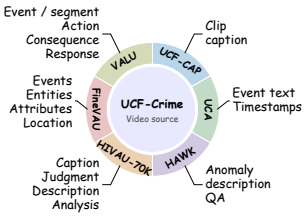


Figure 1: Diverse annotation fields.

To explore reference quality, we manually inspected benchmark videos and texts (Figure 2). One annotation described an obstruction as "complete," although a two-wheeler could still pass. Other cases included blur and occlusion obscuring video details; unresolved references or model limitation statements replacing event descriptions; and linked annotations disagreeing on whether an explosion occurred. These cases motivate a systematic audit of videos, texts, their alignment, and cross-text consistency to characterize quality concerns and their joint effects on video–text pair usability.

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.

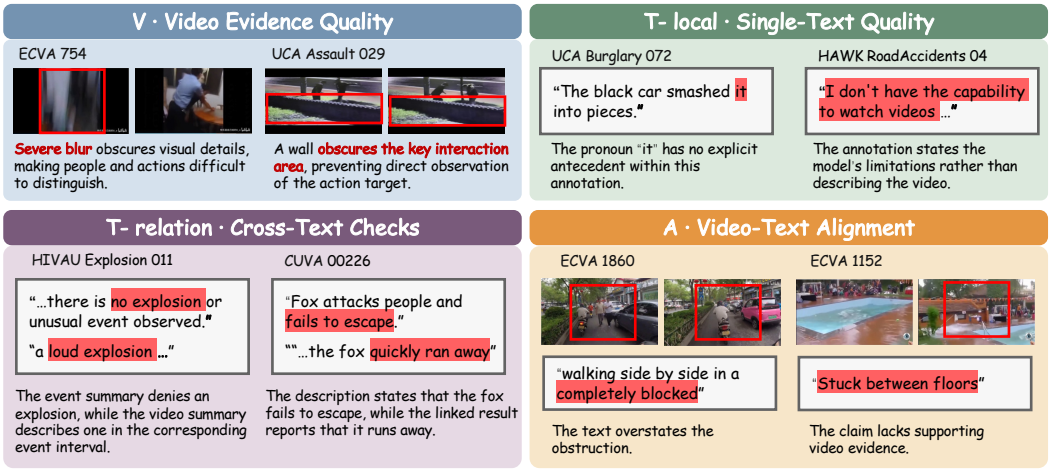


Figure 2: Quality concerns identified through manual inspection of videos, texts, cross-text consistency, and video–text alignment.

Conducting such an audit first requires specifying what different annotations are expected to convey. Event descriptions, causal explanations, and summaries in VAU benchmarks differ in semantic responsibility, temporal scope, and level of detail (Yuan et al., 2024; Du et al., 2024a; Zhang et al., 2025a; He et al., 2026). For example, a summary may omit secondary details, whereas the same omission in a description field that requires those details can constitute missing information. Similarly, two annotations describing different stages of an event need not contradict each other simply because their content differs. Assessing an annotation's accuracy, completeness, and consistency with other annotations therefore requires considering its field-specific requirements and corresponding video scope. Otherwise, an audit may flag legitimate differences as errors while overlooking genuine defects. Cross-benchmark auditing must therefore preserve native field requirements by specifying what each annotation should express, which video segments should be used to verify it, and which linked annotations require consistency checks.

Prior work examines benchmark quality through detection-label reliability, annotation consistency and richness, and semantic verification during data construction (Liu et al., 2025; Li et al., 2025; Long et al., 2026); other studies evaluate generated descriptions against structured references (Pereira et al., 2026; Zhang et al., 2026). These studies provide a foundation for quality assessment, but their findings do not establish whether language annotations across benchmarks meet their field-specific requirements or whether the same video–text pair jointly satisfies video, text, and alignment requirements. Answering these questions requires both assessing annotations under their native field definitions and linking multiple judgments for each pair. We therefore conduct a systematic audit of heterogeneous reference annotations that combines field-level diagnosis with joint usability analysis.

To this end, we introduce a contract-aware audit framework that translates native field requirements into explicit diagnostic questions. A *Field Contract* specifies what each field should express, which video evidence it may draw on, and which linked annotations require consistency checks. Four complementary tasks assess Video Evidence Quality (V), Standalone Text Quality (T-local), Native Cross-text Consistency (T-relation), and Video–Text Alignment (A). Each task receives only its permitted inputs and records question-level judgments, locating concerns within specific fields and criteria. We then link V, T-local, and A for each annotation to determine whether the pair meets all three sets of requirements. The conditional joint usability rate measures the proportion of assessable linked pairs that pass these checks; T-relation is analyzed separately. Together, these assessments locate quality concerns and identify alignment failures missed by standalone checks. The resulting diagnostics guide annotation review and, alongside native task coverage, inform benchmark selection.

We apply this framework to nine benchmarks to investigate reference quality, annotation revision, and benchmark selection through three research questions(**RQ**):

RQ1: Quality profiles. *What quality concerns occur across benchmarks, how are they distributed, and how do they jointly affect video–text pair usability?*

RQ2: Annotation improvement. *What revisions do the audit diagnostics suggest, and can training on revised annotations improve downstream performance?*

RQ3: Benchmark selection. *How can annotation quality and task coverage guide benchmark selection for different evaluation purposes?*

Our model-assisted audit of samples from nine benchmarks shows that high standalone video and text quality scores can coexist with alignment concerns. Under field-macro aggregation, most pairs that fail the joint checks within each benchmark pass both standalone checks but fail alignment. Field-level diagnostics identify annotations that make claims unsupported by the video, omit information required by their field, or state claims as facts when the video evidence is inconclusive, providing concrete targets for revision. Under matched training settings and a common test reference, fine-tuning on audit-guided revised annotations increases mean ROUGE-L across three random seeds from 23.59 to 25.95.Benchmark selection must consider both quality and task coverage: a high-scoring benchmark may lack required fields, while available fields may still contain unreliable annotations. Selection should therefore match field responsibilities to the target task and use audit diagnostics to guide review. Our contributions are threefold:

- **A field-aware audit framework.** We translate heterogeneous field requirements into diagnostic questions and link the answers to assess joint usability while preserving native field responsibilities.
- **Cross-benchmark quality findings.** We identify alignment concerns missed by standalone checks in samples from nine benchmarks and locate them within specific fields and criteria.
- **Audit-guided revision and use.** We provide experimental evidence of the downstream value of annotation revision and combine diagnostics with task coverage to guide benchmark selection.

2 Related Work

VAU benchmarks. Video anomaly benchmarks differ in both the content they ask models to express and the temporal granularity of their annotations. UCF-Crime provides video-level training labels and temporal test annotations for anomaly detection (Sultani et al., 2018). Language-annotated benchmarks make event content an explicit target: UCF-CAP pairs short clips with captions, while UCA associates event descriptions with temporal boundaries (Chatzikonstantinou et al., 2022; Yuan et al., 2024). The requested content extends beyond event description in CUVA and ECVA, which include causal accounts, and in HAWK, which provides anomaly descriptions and question–answer pairs (Du et al., 2024b;a; Tang et al., 2024). FineVAU structures event content into events, entities, and locations (Pereira et al., 2026). Alongside these differences in content, HIVAU-70K, introduced by Holmes-VAU, organizes annotations at clip, event, and video levels, while VALU provides descriptions and temporal boundaries at five semantic levels (Zhang et al., 2025a; He et al., 2026). Pistachio-VAU extends the video sources to synthetic content, with language fields that include generation requirements and video summaries (Li et al., 2025). Together, these designs give reference fields different responsibilities, temporal scopes, and relations to other annotations, which our audit retains when assessing their quality.

Quality assessment in VAU. Quality assessment in VAU examines detection labels, reference annotations, and model-generated descriptions. For detection labels, Liu et al. (2025) use independent re-annotations to show how annotated anomaly intervals affect VAD evaluation. For reference annotations, Pistachio uses multiple vision–language models to compare its video-level annotations with those of Holmes-VAU for consistency and richness (Li et al., 2025). For model-generated descriptions, FineVAU compares anomaly reports with structured event, entity, and location references (Pereira et al., 2026), while VCapsBench uses human-verified question–answer pairs to assess caption accuracy, inconsistency with reference answers, and coverage (Zhang et al., 2026). These output evaluations depend on reference content, whose quality is the focus of our study. We audit reference annotations across benchmarks under native field responsibilities and permitted video evidence scopes. Linking video, text, and alignment judgments locates specific quality concerns and reveals their joint effects on video–text pair usability.

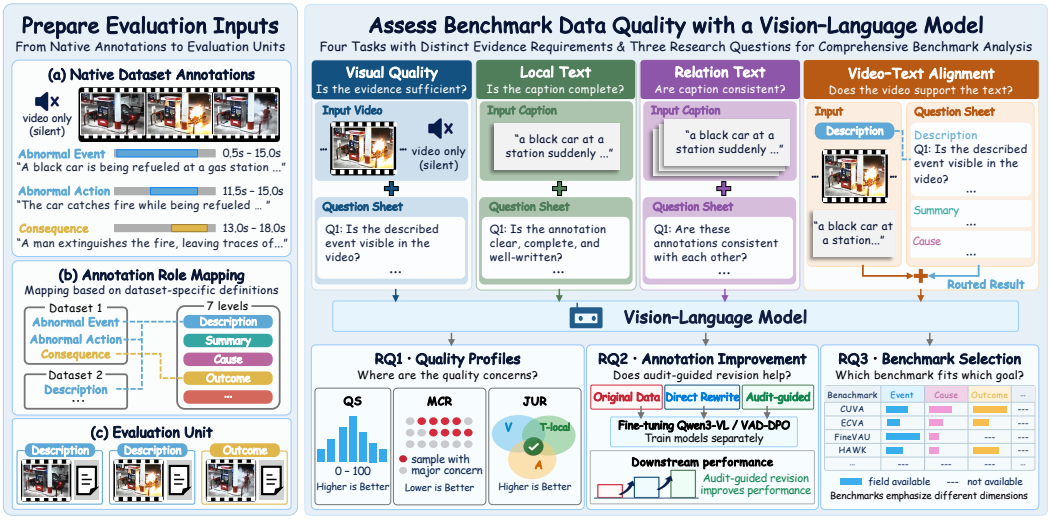


Figure 3: Audit framework for quality assessment, annotation revision, and benchmark selection.

3 Benchmark Audit Framework

3.1 Overview

We organize heterogeneous annotations into evaluation units using each benchmark's native annotation structure and field definitions. We then specify diagnostic questions and required evidence for four dimensions: video evidence, standalone text, cross-text relations, and video–text alignment. A large vision–language model (LVLM) conducts the audit using these questions and evidence. Figure 3 summarizes the framework and its applications.

3.2 Audit Dimensions

Before conducting the human content-quality audit, we first perform a unified data-validity check and exclude records from which valid evaluation inputs cannot be constructed (Appendix A).

Our manual inspection shows that, although the concrete manifestations of these issues vary across datasets and annotation types, they can be grouped into four categories according to where the failure occurs and what information relationship it affects: video-evidence issues, standalone-text issues, cross-text consistency issues, and video–text alignment issues (Appendix B). Based on this taxonomy, we characterize benchmark quality along four dimensions.

Video Evidence Quality (V). This dimension measures whether the video itself provides sufficiently usable evidence for supervision or evaluation; that is, whether the target content can be reliably observed and judged from the provided video.

Standalone Text Quality (T-local). This dimension assesses the quality of an individual textual annotation in isolation. Specifically, it asks whether the annotation correctly fulfills its intended semantic function without relying on the video or on other associated text fields.

Native Cross-Text Consistency (T-relation). This dimension evaluates whether multiple text fields that are explicitly linked by the source benchmark's native data structure remain mutually consistent. It captures the quality of relations among different language annotations within the same data instance.

Video–Text Alignment (A). This dimension evaluates cross-modal consistency between a textual annotation and its corresponding video evidence, namely whether the content expressed by the text is supported by the video evidence associated with that field.

These four dimensions capture distinct levels at which benchmark quality can fail. They are related but not interchangeable. For example, sufficient visual evidence does not imply that the associated text

annotation is correct; likewise, a text annotation may be linguistically clear and internally coherent while still being unsupported by the video.

3.3 Structured Audit Protocol

To make the audit quantitative, analyzable, and reproducible, we instantiate the four quality dimensions above as a set of structured diagnostic questions (Appendix C), and standardize their applicability, input evidence, execution procedure, and result aggregation.

Determining applicable diagnostic questions. Language fields across benchmarks are not semantically equivalent: they may differ in intended function, temporal scope, and information granularity. Based on dataset papers, official documentation, and released annotations, we specify a Field Contract for each native field included in the audit (Appendix D). The contract states what the field should express, which video evidence it may use, and which natively linked fields require consistency checks. We organize these semantic requirements into seven standardized sets (Appendix E). A field's semantic requirements determine which diagnostic questions apply to it: fields with the same requirements are evaluated under the same quality criteria, whereas fields with different requirements are assigned different question sets. For example, event-description fields are evaluated on the quality of their descriptions of observable events; *Cause* fields focus on causal explanations; and *Analysis* fields are allowed to provide inferential interpretations grounded in observable facts.

Determining input evidence. Guided by the field contracts, we predefine an independent evidence boundary for each audit task to prevent irrelevant context from influencing judgments, and expose only the information necessary for the corresponding decision. V receives only the full silent video. T-local receives only the target text and its semantic requirements. T-relation receives only an anonymized group of texts linked by relations defined in the dataset's native structure. A receives only the video range permitted by the field contract, the target text, and any explicitly allowed reference context. All evaluation units are processed independently; no task receives another task's output.

Audit execution. Once the applicable question set and input evidence have been determined, V, T-local, and T-relation directly answer their corresponding diagnostic questions. A requires finer-grained verification of video–text evidence and therefore introduces an additional atomic-decomposition step (Min et al., 2023). Specifically, we decompose the target text into independently verifiable atomic claims and identify its core claim. Each claim is then labeled, based on the permitted video evidence, as *supported*, *minor deviation*, *unverifiable*, or *contradicted*. The resulting claim-level assessments are subsequently used to answer the corresponding alignment diagnostic questions.

Scoring and aggregation. Applicable diagnostic questions receive ratings from 0 to 3. For each evaluation unit, we aggregate its applicable ratings and normalize the result to a 0–100 quality score (QS). Let $M_{u,d} = 1$ if unit u has any applicable rating of 0 or 1 in task d, indicating a major concern, and $M_{u,d} = 0$ otherwise. Units without applicable ratings are excluded from quality summaries. QS measures overall satisfaction of the task requirements, while the major-concern indicator retains critical failures that averaging can obscure.

Within each benchmark, V summaries give equal weight to distinct source videos with assessable outputs. T-local and A first average assessable units within each field, then give each field equal weight. T-relation follows the same two-stage aggregation over native text-relation types. For a fixed benchmark and task, let $\mathcal{C}$ contain the fields or relation types with assessable outputs, and let $\mathcal{U}_c$ contain the assessable units in group c. For a unit-level quantity z_u, define

$$\mathrm{Macro}(z) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} z_u. \tag{1}$$

Applying the corresponding aggregation to unit quality scores yields benchmark QS; applying it to major-concern indicators and multiplying by 100 yields the major-concern rate (MCR). Equation 1 gives each field or relation type equal weight regardless of its number of assessed units.

For joint usability, we link each annotation's T-local and A judgments to the V judgment of its source video. Let $g(u)$ identify the source video of annotation u. For annotations with assessable results in all three tasks, the joint-pass indicator is

$$J_u = (1 - M_{g(u),V})(1 - M_{u,T\text{-local}})(1 - M_{u,A}). \tag{2}$$

Thus, $J_u = 1$ exactly when all three tasks pass without a major concern. Conditional JUR is $100\,\mathrm{Macro}(J)$, using Equation 1 with $\mathcal{U}_c$ restricted to assessable linked annotations in field c and $\mathcal{C}$ restricted to fields with at least one such annotation. T-relation is reported separately from this pair-level joint check. Appendix F.1 provides the detailed aggregation and coverage rules.

3.4 Evaluator Validation and Selection

Improved video understanding does not necessarily yield more reliable evaluation judgments (Liu & Zhang, 2025). To assess whether candidate vision–language models can reliably execute the proposed evaluation protocol, we manually re-annotate 400 videos using the structured audit protocol described above. Two annotators independently evaluate each sample. A third annotator adjudicates disagreements to produce the human reference annotation.

We then apply four candidate vision–language models to exactly the same set of samples and compare their question-level outputs, as well as the resulting audit outcomes, against the human reference annotations. This evaluation measures how reliably each model can execute the proposed protocol. Detailed comparisons among the four candidate models are reported in Table 1. These results identify Gemini 3.7 Flash as the evaluator for the automated audit of nine benchmarks.

Table 1: Comparison of candidate evaluators against human reference annotations. F1 is the harmonic mean of the reported precision and recall.

					Concern detection ↑		
Candidate	Exact ↑	Weighted κ ↑	MAE ↓	MSE ↓	Precision	Recall	F1
Gemini 3.7 Flash	93.4%	0.814	0.085	0.143	0.912	0.995	0.952
Qwen 3.8 Max	89.2%	0.655	0.164	0.299	0.657	0.913	0.764
Doubao Seed 2.1 Pro	90.4%	0.680	0.154	0.304	0.626	0.974	0.762
Gemini 3.8 Flash	74.0%	0.625	0.287	0.364	0.722	0.932	0.814

4 Results and Discussion

4.1 RQ1: Quality profiles across tasks, fields, and linked pairs

Table 2 summarizes quality profiles across nine benchmarks, reporting QS, MCR, and conditional JUR under the definitions in Section 3.3.

Alignment has the highest major-concern rate in every benchmark. Under this audit protocol, A-MCR ranges from 22.38% to 77.84% and exceeds the rate of every other evaluated task within each benchmark. V, T-local, and T-relation MCR reach at most 6.13%, 6.00%, and 2.00%, respectively (Table 2). For quality scores, AQS ranges from 41.47 to 85.56, below VQS (93.97–99.48), T-local QS (98.37–99.88), and evaluated T-relation QS (99.44–100.00) (Table 2; Figure 4). These profiles identify video–text alignment as the most frequently flagged dimension of reference quality.

Alignment concern rates vary across fields within benchmarks. The benchmark-level profiles conceal differences across native fields. In ECVA, alignment MCR is highest for Effect (39.20%), followed by Key Sentences (31.48%), Cause (29.67%), and Description (26.62%). Other benchmarks show smaller differences: CUVA's Description, Cause, and Result have rates of 24.24%, 26.26%, and 26.77%, respectively. FineVAU's Normal and Abnormal descriptions have rates of 42.50% and 44.94% (Appendix H).

Factual-support concerns exceed coverage concerns in the examined fields. Criterion-level analysis further identifies the defects in ECVA's Key Sentences and HAWK's descriptions. In ECVA's Key Sentences, factual-support and coverage major-concern rates are 27.50% and 13.50%, respectively. In HAWK's descriptions, the corresponding rates are 74.05% and 42.49%, while certainty-calibration concerns affect 67.94%. These diagnostics identify factual support as a frequent concern and certainty calibration as an additional concern in HAWK.

Standalone quality does not determine joint usability. Conditional JUR range from 22.07% to 74.20% across the nine benchmarks (Table 2). For HAWK, the standalone T-local and V pass

Table 2: Quality scores (QS), major-concern rates (MCR, %), and conditional joint usability (JUR, %).

Benchmark	n	Pairs	V		T-local		T-relation		A		JUR
			QS	MCR	QS	MCR	QS	MCR	QS	MCR	
CUVA (2024b)	198	594	96.69	0.00	98.58	1.33	99.44	2.00	76.21	25.76	74.07
ECVA (2024a)	1989	7980	96.28	2.51	98.37	6.00	99.47	2.00	74.71	31.74	66.25
FineVAU (2026)	2009	23984	95.50	2.49	99.85	0.26	–	–	65.56	43.72	55.33
HAWK (2024)	786	786	98.39	1.53	98.82	1.91	–	–	41.47	74.55	25.06
HIVAU-70K (2025a)	1050	13844	96.61	4.48	98.93	3.45	99.87	0.67	79.23	26.14	69.47
Pistachio-VAU (2025)	192	1193	99.48	0.52	99.88	0.13	100.00	0.00	58.91	77.84	22.07
UCA (2024)	310	4331	93.97	6.13	99.77	0.62	–	–	72.53	28.49	69.80
UCF-CAP (2022)	396	396	94.50	1.52	99.73	2.02	–	–	85.56	25.00	72.98
VALU (2026)	998	2995	96.22	4.01	99.87	0.00	100.00	0.00	79.35	22.38	74.20

rates reach 98.09% and 98.47%, yet only 25.06% of pairs pass all three checks for text, video, and alignment. Of the 589 pairs that fail the joint check, 564 pass both standalone checks but fail alignment, accounting for 95.76%. Thus, passing separate video and text quality checks is insufficient to establish the usability of their pairing. Furthermore, average alignment quality cannot substitute for joint usability: UCF-CAP has a higher AQS than VALU , but a lower JUR.

Together, these results identify alignment as the most frequently flagged quality dimension across the nine benchmarks. Field and criterion analyses locate the concerns, while joint assessment reveals failures that standalone checks miss.

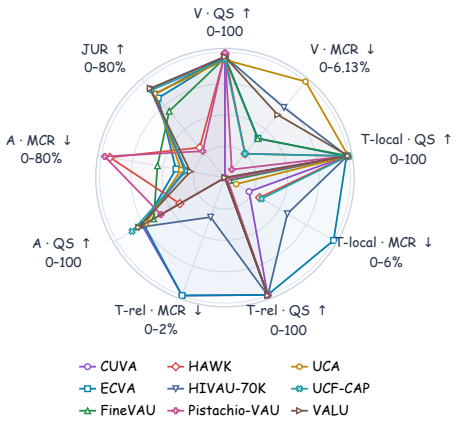


Figure 4: QS, MCR, and JUR profiles.

4.2 RQ2: Does audit-guided revision improve downstream performance?

RQ1 identifies quality concerns in annotations. RQ2 tests whether addressing these concerns improves downstream performance and whether explicit audit diagnostics add value beyond direct rewriting. We compare three training-label conditions: original annotations (A), direct LLM revision (B), and audit-guided revision (C). CUVA description generation measures agreement with revised references, while VALU-ADC (He et al., 2026) provides a complementary multiple-choice evaluation.

Matched comparisons of annotation revision. We evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) and the domain-adapted VAD-DPO (Zhang et al., 2025b). Conditions B and C both use Gemini 3.7 Flash, with C incorporating the audit diagnostics. The C–A comparison tests the downstream benefit of revision, whereas C–B examines the additional benefit of diagnostic guidance. Figure 5(a) illustrates how the revisions change specific claims rather than merely their phrasing. Condition C retains all 693 training descriptions: 354 are revised and 339 remain unchanged.

Training, evaluation, and seed variability. All conditions use the same 693 CUVA training examples, 32-frame inputs, and three training epochs. Within each model and seed, the starting checkpoint, prompts, training order, and hyperparameters are matched across A, B, and C. Validation loss on 78 examples selects the checkpoint. CUVA evaluation uses the same revised Description references for 198 held-out examples. VALU-ADC evaluates the trained checkpoints without additional task-specific training on the common set of 5,689 evaluable questions. Table 3 reports the mean and sample standard deviation across seeds 42, 123, and 2026. ROUGE-L (R-L; Lin & Och, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore F1 (BS; Zhang et al., 2020) are scaled by 100; VALU-ADC accuracy is reported as a percentage.

Audit-guided revision improves mean description scores. Condition C has higher mean scores than A on all three reported CUVA metrics for both models. For Qwen3-VL, ROUGE-L increases from 23.59 ± 0.41 to 25.95 ± 0.22, and METEOR increases from 9.35 ± 0.13 to 10.78 ± 0.16. For

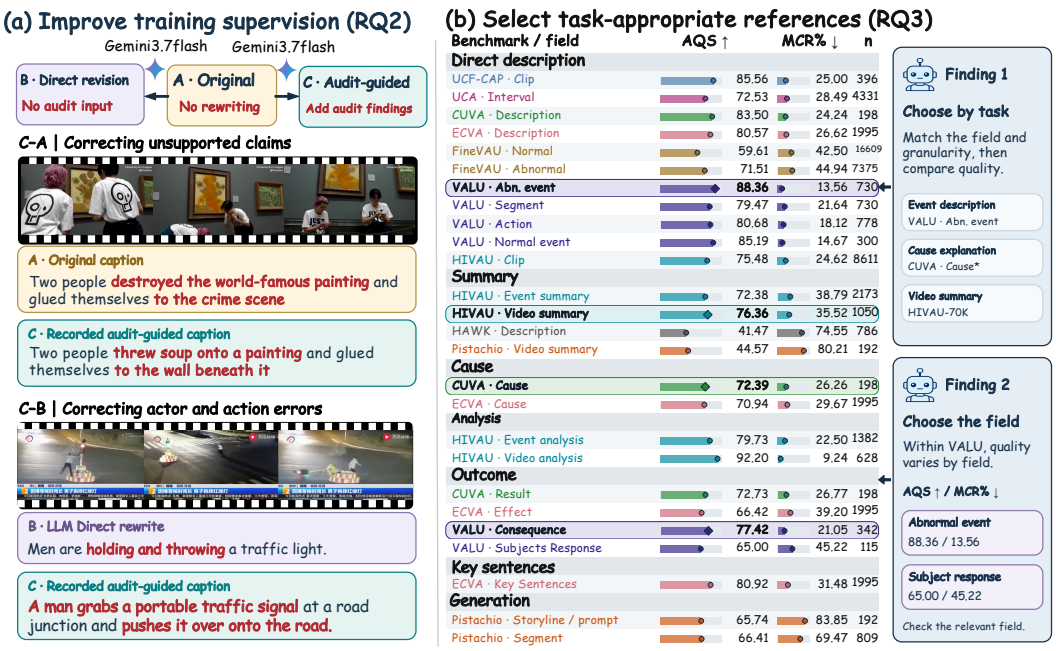


Figure 5: **Audit-guided annotation revision and benchmark selection.** (a) Revision examples comparing C with A and B. (b) Alignment quality (AQS ↑) and major-concern rate (MCR, % ↓).

Table 3: Downstream effects of revising CUVA training descriptions. R-L, METEOR, and BS evaluate CUVA descriptions; VALU-ADC reports accuracy (%). Results are mean ± sample standard deviation across seeds 42, 123, and 2026; higher is better.

	Qwen3-VL-8B-Instruct (Bai et al., 2025)				**VAD-DPO** (Zhang et al., 2025b)			
Training labels	R-L	METEOR	BS	VALU-ADC	R-L	METEOR	BS	VALU-ADC
Original (A)	23.59 ± 0.41	9.35 ± 0.13	89.33 ± 0.04	42.24 ± 0.32	20.41 ± 0.09	8.04 ± 0.09	88.37 ± 0.06	20.45 ± 0.42
Direct revision (B)	23.87 ± 0.14	9.23 ± 0.09	89.41 ± 0.02	42.94 ± 0.29	20.27 ± 0.15	8.35 ± 0.08	88.35 ± 0.08	20.87 ± 0.37
Audit-guided (C)	$\mathbf{25.95 \pm 0.22}$	$\mathbf{10.78 \pm 0.16}$	$\mathbf{89.51 \pm 0.02}$	$\mathbf{44.58 \pm 0.40}$	$\mathbf{20.68 \pm 0.19}$	$\mathbf{8.70 \pm 0.08}$	$\mathbf{88.68 \pm 0.04}$	$\mathbf{21.08 \pm 0.41}$

VAD-DPO, ROUGE-L increases from 20.41 ± 0.09 to 20.68 ± 0.19, while METEOR increases from 8.04 ± 0.09 to 8.70 ± 0.08. These comparisons show higher reference agreement after revision in both model settings, with smaller mean gains for VAD-DPO on ROUGE-L and METEOR.

Diagnostic guidance adds value beyond direct rewriting. Direct revision does not uniformly improve the original labels: Qwen3-VL's METEOR decreases from 9.35 to 9.23, and VAD-DPO's ROUGE-L and BERTScore decrease from 20.41 to 20.27 and from 88.37 to 88.35. In contrast, C has higher means than B on all three CUVA metrics in both models. Relative to B, C raises Qwen3-VL's ROUGE-L by 2.08 points and METEOR by 1.55 points; the corresponding gains for VAD-DPO are 0.41 and 0.35 points. With the revision model and downstream training settings matched, this pattern supports the added value of audit diagnostics over direct rewriting in these settings.

Benefits extend to multiple-choice evaluation. On VALU-ADC, Qwen3-VL reaches $44.58 \pm 0.40\%$ with C, compared with $42.24 \pm 0.32\%$ with A and $42.94 \pm 0.29\%$ with B: gains of 2.34 and 1.64 percentage points, respectively. VAD-DPO reaches $21.08 \pm 0.41\%$, compared with $20.45 \pm 0.42\%$ and $20.87 \pm 0.37\%$, yielding smaller gains of 0.63 and 0.21 percentage points. These results complement the CUVA reference-based evaluation because they use VALU's original multiple-choice answers without further training on that task. The reported standard deviations quantify between-seed variation; the smaller gains, particularly for VAD-DPO, should not be read as established statistical significance from these summaries alone.

Together, the comparisons support downstream benefits from annotation revision and an additional mean improvement from audit guidance over direct rewriting. The magnitude depends on the model and metric. The evidence concerns reference agreement and multiple-choice accuracy on the evaluated tasks, rather than a uniform improvement in every aspect of video understanding.

4.3 RQ3: WHICH BENCHMARKS SUIT DIFFERENT EVALUATION GOALS?

RQ3 uses the audit to help researchers choose reference annotations for different evaluation tasks. Figure 5(b) shows the annotation types available in each benchmark, together with their alignment quality scores (AQS) and major-concern rates (MCR), providing a basis for comparing candidate references. The displayed n counts eligible annotations in the study inputs. The complete field-level statistics and annotation counts are in Appendix H.

Select task-appropriate reference annotations. Different evaluation goals lead to different candidate fields. VALU's Abnormal_Event is a strong option for anomaly-event description; UCF-CAP and CUVA provide references for short-clip captioning and anomaly-interval description, respectively. For cause explanation, CUVA's Cause has modestly higher AQS and lower MCR than ECVA's. For video summarization, HIVAU-70K has the strongest observed profile among the three providers.

However, high quality in one field does not mean that all fields in the benchmark are equally reliable. For example, VALU's Abnormal_Event has AQS 88.36 and MCR 13.56%, whereas Subjects_Response, which describes how people react to an event, has AQS 65.00 and MCR 45.22%. This difference shows that VALU's strong event descriptions do not imply equally reliable descriptions of people's reactions. Benchmark selection should therefore identify the specific annotation type to be used and consider its quality and concern rate when judging its suitability as a reference for the task.

Combine complementary annotations for shared source videos. When evaluation covers multiple tasks, this selection principle can also guide the combination of benchmarks with overlapping source videos. As Figure 1 illustrates, UCF-CAP, UCA, and HIVAU-70K provide different forms of annotation for shared UCF-Crime videos. Researchers can draw on these resources to assemble references for multiple evaluation tasks. For example, an evaluation of both clip description and video summarization can select the corresponding fields from different benchmarks. Selection should begin by matching source videos and annotation time spans, then checking the chosen annotations against their field contracts on the shared subset. Field-level audit results can guide this review while each reference retains its own evidence scope and evaluation criteria. This strategy organizes complementary annotations by task, allowing shared-source resources to support evaluation at multiple semantic and temporal levels.

4.4 DISCUSSION

Together, the three research questions show that assessing annotation suitability requires checking whether its content is supported by the corresponding video evidence and fulfills its field requirements. Standalone video and text checks can miss these issues, while field-level diagnostics identify what needs revision. These diagnostics can therefore guide annotation revision, whose training value is supported by the downstream comparisons. Likewise, when selecting evaluation references, researchers can combine these quality diagnostics with task coverage to identify fields suited to the target task. By locating specific defects, the audit thus informs both the improvement and selection of existing annotations.

5 CONCLUSION

We introduce a field-aware framework for auditing video and language references across nine video anomaly understanding benchmarks. The audit reveals alignment failures missed by standalone video and text checks and locates them within specific fields and criteria. Controlled downstream experiments support the training value of audit-guided annotation revision. Combined with task coverage, these diagnostics also inform the selection of references for different evaluation goals. Together, the findings connect reference-quality assessment to annotation improvement and task-specific benchmark use.

Limitation. Our cross-benchmark quality audit covers only the test splits. The downstream revision experiments use CUVA training annotations; the effects of large-scale annotation revision across benchmarks remain untested. Future work will extend the audit to training splits and evaluate whether large-scale revision improves performance across a broader range of downstream tasks.

AI USE STATEMENT

We used Gemini and GPT-family models to assist with literature search, research discussions, feedback on methodology and experimental design, code development and debugging, and interpretation of experimental results. These tools also assisted with drafting, translation, language editing, and LaTeX formatting. As part of the research methodology, vision–language models performed automated annotation audits, and Gemini generated direct and audit-guided annotation revisions, as described in the paper. Evaluator validation and the scope of human review are reported in the methodology and limitations. The authors take responsibility for the final manuscript, code, experimental results, and conclusions.

ETHICS STATEMENT

This study examines existing video anomaly understanding benchmarks, whose videos may contain identifiable individuals and sensitive or violent events. Their use and any redistribution must respect the source datasets' licenses and applicable privacy requirements. Our audit assesses the quality of reference annotations and their supporting video evidence; it is not intended to make judgments about individuals depicted in the videos. Model-assisted judgments and annotation revisions may contain errors or biases. We therefore report evaluator validation and the scope of human review, and treat flagged concerns as candidates for review rather than definitive determinations. The findings are intended to support responsible benchmark improvement and selection.

REPRODUCIBILITY STATEMENT

We provide the detailed audit questions, scoring criteria, and prompts in the appendix, together with the native field definitions and their mapping to audit requirements (Appendices C, D, and E). The audit population and aggregation procedures are documented in Appendices F.3 and F.1. Section 4.2 describes the downstream comparisons, training settings, evaluation metrics, and three random seeds; Appendices G.4 and G.5 provide additional evaluation details. We include the code in the supplementary materials to support reproduction of the reported procedures.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Christos Chatzikonstantinou, Georgios Grigorios Valasidis, Konstantinos Stavridis, Georgios Malogiannis, Apostolos Axenopoulos, and Petros Daras. UCF-CAP, video captioning in the wild. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 1386–1390, 2022. doi: 10.1109/ICIP46576.2022.9898011.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, and Jun Liu. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2412.07183*, 2024a. URL https://arxiv.org/abs/2412.07183.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18793–18803, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Du_Uncovering_What_Why_and_How_A_Comprehensive_Benchmark_for_Causation_CVPR_2024_paper.html.

Yixiao He, Menghao Zhang, Haifeng Sun, Jing Wang, Kangheng Lin, Jinghan Wang, Chenye Xu, Pengfei Ren, Qi Qi, and Jingyu Wang. VALU: A benchmark for video anomaly temporal localization and understanding at multiple semantic levels. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 1252–1296, 2026. doi: 10.18653/v1/2026.acl-long.56. URL https://aclanthology.org/2026.acl-long.56/.

Jie Li, Hongyi Cai, Mingkang Dong, Muxin Pu, Shan You, Fei Wang, and Tao Huang. Pistachio: Towards synthetic, balanced, and long-form video anomaly benchmarks. *arXiv preprint arXiv:2511.19474*, 2025. URL https://arxiv.org/abs/2511.19474v6. Version 6, July 2026.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 605–612, 2004. doi: 10.3115/1218955.1219032. URL https://aclanthology.org/P04-1077/.

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dc4f891373d19087d1ddda33e81e00e4-Abstract-Conference.html.

Zihao Liu, Xiaoyu Wu, Wenna Li, and Linlin Yang. Rethinking metrics and benchmarks of video anomaly detection. *arXiv preprint arXiv:2505.19022*, 2025. URL https://arxiv.org/abs/2505.19022v1. Version 1.

Yuzhou Long, Haodong Zhang, Yunpeng Yang, Peng Wu, and Guansong Pang. AnomalyCraft-700K: Component-level controllable and verifiable synthetic anomalies for fine-grained video anomaly understanding. *arXiv preprint arXiv:2609.06978*, 2026. URL https://arxiv.org/abs/2609.06978.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html.

João Alexandre Cardeira Pereira, Vasco Lopes, João C. Neves, and David Semedo. FineVAU: A novel human-aligned benchmark for fine-grained video anomaly understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8403–8411, 2026. doi: 10.1609/aaai.v40i10.37790. URL https://ojs.aaai.org/index.php/AAAI/article/view/37790.

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.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. HAWK: Learning to understand open-world video anomalies. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4435. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22052–22061, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yuan_Towards_Surveillance_Video-and-Language_Understanding_New_Dataset_Baselines_and_Challenges_CVPR_2024_paper.html.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13843–13853, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.html.

Menghao Zhang, Huazheng Wang, Pengfei Ren, Kangheng Lin, Qi Qi, Haifeng Sun, Zirui Zhuang, Lei Zhang, Jianxin Liao, and Jingyu Wang. Do LVLMs truly understand video anomalies? revealing hallucination via co-occurrence patterns. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/99b419554537c66bf27e5eb7a74c7de4-Abstract-Conference.html.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. VCaps-Bench: A large-scale fine-grained benchmark for video caption quality evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12726–12734, 2026. doi: 10.1609/aaai.v40i15.38269. URL https://ojs.aaai.org/index.php/AAAI/article/view/38269.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yiyan Zhu, Menghao Zhang, Haifeng Sun, Pengfei Ren, Xianao Chu, Chenye Xu, Hong Tan, Jinghan Wang, Qi Qi, and Jingyu Wang. Alert-CLIP: Abnormality-aware latent-enhanced representation tuning of CLIP for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35545–35554, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhu_Alert-CLIP_Abnormality-aware_Latent-Enhanced_Representation_Tuning_of_CLIP_for_Video_Anomaly_CVPR_2026_paper.html.

A Data Preprocessing Results

Deterministic checks flag parsing errors, empty text, evidence intervals that exceed the video's real duration, relation-binding failures, and media errors before any model call; affected videos are excluded from all four tasks. Table 4 reports checked test inputs and invalid records excluded during preprocessing. FineVAU retains $2{,}230 - 221 = 2{,}009$ video entries, and VALU retains $1{,}019 - 21 = 998$, as reported in Table 2.

Table 4: Video exclusions from data validity checks.

Benchmark	Checked	Excluded
UCF-CAP	396	0
UCA	310	0
CUVA	200	2
ECVA	2,174	185
HAWK	786	0
HIVAU-70K	1,051	1
FineVAU	2,230	221
VALU	1,019	21
Pistachio-VAU	192	0

The exclusions reported in Table 4 are broken down below: UCF-CAP, UCA, HAWK, and Pistachio-VAU exclude none; CUVA excludes 2 of 200 (interval); ECVA excludes 185 of 2,174 (180 interval, 5 empty text); HIVAU-70K excludes 1 of 1,051 (empty text); FineVAU excludes 221 of 2,230; and VALU excludes 21 of 1,019. Counts are affected source video IDs; a video can trigger more than one issue type. FineVAU's parsing errors share one root cause: mixed seconds/clock/string timestamp encodings that the parser cannot resolve.

B Manual Review: Representative Annotation Errors

All review material in this appendix—case identification, video inspection, textual analysis, and judgments—comes entirely from human review, without model assistance. This review preceded the formulation of the audit dimensions and criteria. During manual inspection of the nine benchmarks, reviewers logged representative cases for each recurring problem pattern, grouped by the evidence scope the audit later formalizes as V, T-local (*Text-Local* below), T-relation (*Text Relation* below), and A (*Alignment* below). Each case reproduces the native annotation record verbatim, followed by a standardized explanation of the annotation, the recorded evidence or textual conflict, and the resulting concern. Pink highlights identify the relevant spans; for cross-text conflicts, both sides are marked. For omissions, highlighting identifies the incomplete annotation and any stated comparison detail, rather than text that is absent. Video-based explanations summarize the recorded manual review.

B.1 Text-Local

TL01: Unclear semantics or grammatical errors.

CUVA 00516

GT annotation.


```
[
  {
    "instruction": "Determine whether the video contains an anomalous event. Output 1 if an
        anomaly is present.",
    "visual_input": "00516.mp4",
    "output": 1,
    "task": "Detection",
    "ID": "test_18"
  },
  {
    "instruction": "Identify and categorize the anomaly using the provided categories.",
    "visual_input": "00516.mp4",
    "output": "[11-3,5]",
    "task": "Classification",
    "ID": "test_19"
  },
  {
    "instruction": "Locate the position of the anomalous segment in mmss format.",
    "visual_input": "00516.mp4",
    "output": "[0000,0251]",
    "task": "Timestamp",
    "ID": "test_20"
  },
  {
    "instruction": "Give a detailed description of the anomalous segment in the video.",
    "visual_input": "00516.mp4",
    "output": "Three people smoking on the roadside were stopped smoking in various ways",
```

```
  "task": "Description",
  "ID": "test_21"
 },
 {
  "instruction": "Give the root cause of the anomalies in the video in detail.",
  "visual_input": "00516.mp4",
  "output": "On the streets of India, there are women wearing red saris smoking on
      electric cars, a man with a flashy gold black coat holding a cigarette, and a man
      with a blue and white checkered shirt sitting on the roadside railing smoking.",
  "task": "Cause",
  "ID": "test_22"
 },
 {
  "instruction": "Give a detailed and organized account of the final outcome caused by the
       unusual events.",
  "visual_input": "00516.mp4",
  "output": "The woman smoking was cut off with scissors and splashed with water, the man
      with a flashy coat had his cigarette removed and stamped out on the ground, and the
       man sitting by the railing smoking was cut with scissors from behind",
  "task": "Result",
  "ID": "test_23"
 }
]
```


Explanation. The `Result` field states that "the woman smoking was cut off with scissors." Its grammar makes the woman, rather than her cigarette, the object being cut. Recovering the intended meaning requires the reader to repair the sentence, so the wording introduces semantic ambiguity instead of a clear account of the action.

UCA burglary 065

GT annotation.


```
{
 "Burglary065_x264": {
  "duration": 312.55,
  "timestamps": [
   [55.8, 69.2], [65.8, 89.2], [90.1, 110.3], [115.7, 124.6],
   [125.7, 144.6], [150.9, 160.8], [199.8, 206.3], [209.8, 213.3],
   [213.8, 216.3], [230.7, 275.0], [283.5, 290.5], [303.4, 308.2],
   [308.4, 312.2]
  ],
  "sentences": [
   "There is a figure shaking outside the gate of the small courtyard.",
   "Then the courtyard door was slowly opened, and an African-American man wearing a
       white hat and dark clothing walked into the small courtyard.",
   "The man and another African-American man who appeared at the door opened the
       courtyard door further.",
   "The two of them walked towards the house.",
   "At the same time, a silver car backed up and entered the small courtyard.",
   "Two more men got out of the car and walked inside one after another.",
   "Two men walked out of it,",
   "After loading some items in the trunk of the car,",
   "Close the car door and turn back.",
   "Four people walked out of it one after another.",
   "A pickup truck came to the entrance of the small courtyard from a distance.",
   "An African-American man wearing green camouflage uniform entered the courtyard and
       looked around.",
   "The man walked inside."
  ]
 }
}
```


Explanation. The annotation states, "Two men walked out of it," but supplies no antecedent for `it` within that sentence. Identifying the place being exited requires another annotation or external context. Because the unit is assessed as standalone text, its central referent is unresolved and the sentence does not communicate a self-contained event.

TL02: Description does not match the field's assigned responsibility.

ECVA 253

GT annotation.


```
{
  "video_id": "253",
  "classification": "Medical Abnormality-Nursing Event",
  "reason": "On-site first aid for burns and scalds.",
  "result": "Introduces the specific steps of on-site first aid for burns and scalds.",
  "moment": "[0213,0324]",
  "description": "The video provides a detailed introduction to the steps of on-site first
      aid for burns and scalds, emphasizing the correct first aid methods.",
  "key_sentences": "[\"emergency\", \"first aid\", \"burns\", \"scalds\", \"on-site
      treatment\"]"
}
```


Explanation. The `reason` field names the topic of first aid for burns, while `result` describes the presentation of first-aid steps. Neither supplies the information required by its assigned responsibility: an anomaly's cause for Cause, or its consequence for Outcome. Both fields therefore substitute a topic description for the expected semantic content.

ECVA 2144

GT annotation.


```
{
  "video_id": "2144",
  "classification": "Traffic Accidents-Motor",
  "reason": "On the daytime road, vehicles stop in the right lane waiting to pass, and
      three young people pass through the gaps between the vehicles and cross the road",
  "result": "When one of the young people walked into the left lane, they were hit by a
      vehicle coming from the left lane. It is very reasonable for pedestrians to cross the
       road and cause a car accident",
  "moment": "[0000,0013]",
  "description": "[It's reasonable for pedestrians to be killed when crossing the road]",
  "key_sentences": "[\"On the daytime road\",\"and three young people pass through the gaps
      \",\"When one of the young people walked into the left lane\",\" they were hit by a
      vehicle coming from the left lane\"]"
}
```


Explanation. The `description` field says that it is "reasonable for pedestrians to be killed when crossing the road." This is a value judgment rather than an account of the people, actions, or event progression in the segment. The field therefore fails its descriptive responsibility, regardless of whether its wording is grammatically interpretable.

TL03: Redundant content.

UCA Normal_Videos 547

GT annotation.


```
{
  "video_id": "Normal_Videos547_x264",
  "annotation_index": 189,
  "annotation_number": 190,
  "timestamp": [4725.2, 4784.2],
  "sentence": "The man in red is walking back and forth on the left side of the camera. A
      woman in a blue skirt enters the supermarket and walks to the left. The man in red
      is walking back and forth on the left side of the camera."
}
```


Explanation. The `sentence` field repeats the full statement about the man in red walking back and forth, with a sentence about a woman between the two copies. The repeated wording adds no distinct event or detail. This is redundancy within one annotation, rather than complementary information supplied by separate annotations.

ECVA 47

GT annotation.


```
{
  "video_id": "47",
  "classification": "Public Safety-Crimes",
  "reason": "The man discovered a thief attempting to steal something from his blue pickup
      truck.",
  "result": "The thief's car window was shattered, and the man's arm was cut. This incident
       garnered media coverage.",
  "moment": "[0036,0042],[0055,0101],[0113,0120]",
  "description": "[The thief quickly returned to his car, shut the door, and drove off.
      The man chased after him and punched through the thief's car window.],[The thief,
      using a flashlight, opened the tailgate of the man's blue pickup truck, intending to
      steal something.],[The thief quickly returned to his car, shut the door, and drove
      off. The man chased after him and punched through the thief's car window.]",
  "key_sentences": "[\"Nighttime theft\",\"Thief driving away\",\"Shattered glass\",\"Arm
      injury\"]"
}
```


Explanation. The `description` field repeats the same passage about the thief driving away and the man pursuing him and breaking the car window. The two copies surround a different passage about the attempted theft. Because the repeated passage contributes no additional information, the annotation contains duplicated content rather than a concise event description.

TL04: Internal statements contradict each other.

ECVA 2105

GT annotation.


```
{
  "video_id": "2105",
  "classification": "Traffic Accidents-Motor",
  "reason": "At the intersection, a white sedan turned left and collided with a straight
      ahead vehicle. After the collision, the white sedan continued to turn left and
      accelerated into the waiting black sedan on the roadside",
  "result": "The white sedan did not observe the road conditions when turning and avoided
      a car accident. After the accident, the white sedan made a mistake and continued to
      turn and drive, colliding with the normal waiting black tutorial",
  "moment": "[0000,0014]",
  "description": "[Consecutive traffic accident caused by turning vehicles not avoiding
      straight ahead]",
  "key_sentences": "[\"a white sedan turned left and collided with a straight ahead vehicle
      \",\"After the collision\",\"white sedan made a mistake\",\"colliding with the normal
       waiting black tutorial\"]"
}
```


Explanation. The `result` field first says that the sedan "avoided a car accident," then continues with "After the accident" and another collision. These clauses deny and presuppose the occurrence of the accident within the same annotation. The inconsistency is internal to the text and does not require comparison with another field or the video.

B.2 TEXT RELATION

TR01: Action, direction, state, or result conflicts.

ECVA 812

GT annotation.


```
{
  "video_id": "812",
  "classification": "Robbery-Shop",
  "reason": "CCTV captures the shocking moment thieves drive a Range Rover into the front
      of a west London jewellers before they rob the shop and attempt to flee. The incident
       happened in front of shocked customers at the store in Shepherds Bush last October.
      The three individuals used hammers to smash glass display cabinets before stuffing
      the jewels into bags and pockets. ",
```

```
  "result": " One of the robbers was caught and detained by members of the public before
       police arrived and arrested him. He is now serving a 10-year jail sentence after
       pleading guilty to robbery, dangerous driving, and possession of offensive weapon,
       criminal damage and receiving stolen goods.",
  "moment": "[0000,0025]",
  "description": "[Three robbers crashed their vehicles into a jewelry store, then smashed
       the glass with hammers, put the jewelry in bags, and escaped.]",
  "key_sentences": "[\"Jewelry store robbery\",\" vehicle collision\",\" hammer breaking
       counter\",\" filling jewelry and escaping\"]"
}
```


Explanation. The `description` field presents three robbers as escaping after the robbery, whereas `result` says that one robber was caught and detained by members of the public. The linked fields give incompatible outcomes for the same group and event. The concern is cross-text outcome consistency, rather than the grammatical quality of either field alone.

ECVA 782

GT annotation.


```
{
  "video_id": "782",
  "classification": "Robbery-Burglary",
  "reason": "Two women and a man were robbed at home by a group of robbers wearing white
       hoods.The robber grabbed one of the girls' hair and took her outside the house,
       ordering the other two to lie on the ground.",
  "result": "The owner of the house issued a command through surveillance, demanding that
       the robber leave his house, otherwise he would have to call the neighbors to come and
        help.",
  "moment": "[0000,0020][0040,0043]",
  "description": "[The robber made two people lie still on the ground and grabbed a girl's
       hair with their hands, dragging it into the room.][The robber made two people lie
       still on the ground and grabbed a girl's hair with their hands, dragging it into the
       room.]",
  "key_sentences": "[\"Home invasion robbery, grabbing a girl's hair, homeowner warning
       through surveillance, victim lying on the ground\"]"
}
```


Explanation. The `reason` field says the robber took the girl "outside the house," whereas `description` says she was dragged "into the room." Both refer to the same hair-pulling action but assign different destinations. The mismatch is therefore a cross-field conflict in the action's direction and endpoint, rather than a difference in descriptive detail.

TR02: Object identity, category, or attribute conflicts.

ECVA 2139

GT annotation.


```
{
  "video_id": "2139",
  "classification": "Traffic Accidents-Motor",
  "reason": "At a crossroads on the street at night, a silver gray SUV drives from left to
       right. However, the silver gray SUV passes the intersection when its signal light is
       still red",
  "result": "A silver gray sedan running a red light is responsible for all traffic
       accidents caused by a collision between a passing intersection and a passing vehicle
       ",
  "moment": "[0000,0013]",
  "description": "[Running a red light leads to a car accident]",
  "key_sentences": "[\"At a crossroads on the street at night\",\"a silver gray SUV drives
       from left to right\",\"when its signal light is still red\",\"is responsible for all
       traffic accidents\"]"
}
```


Explanation. The `reason` field identifies the offending vehicle as a "silver gray SUV," while `result` identifies it as a "silver gray sedan." The vehicle has the same role in the same traffic event, but its category changes between fields. This is an inconsistent entity attribute, rather than two compatible descriptions of different participants.

FineW3 ECVA 1

GT annotation.


```
{
  "video_id": "1",
  "source": "ECVA",
  "events": {
    "refined_events": [
      {
        "timestamp": [0.27, 0.28],
        "event_type": "Abnormal",
        "event_description": "The young man forcefully throws the white hubcap/wheel cover
             onto the hood/windshield area of the camera car, causing a visible impact and a
             loud thud.",
        "involved_entities": [
          "young man in orange t-shirt",
          "white hubcap/wheel cover",
          "camera car"
        ]
      }
    ]
  },
  "entities": [
    {
      "id": "silver hubcap",
      "category": "object",
      "color": "silver",
      "type": "hubcap"
    }
  ],
  "scene": {
    "environment": "city street intersection",
    "time_of_day": "Daytime",
    "lighting_conditions": "Bright/Well-lit",
    "crowd_density": "Few people",
    "salient_feature": "A person exiting a car in the middle of the road"
  }
}
```


Explanation. The `event_description` calls the thrown object a "white hubcap/wheel cover," while the linked `entities` inventory identifies a "silver hubcap" and records its color as silver. The descriptions refer to the same object but disagree on its color. The concern lies in the relationship between the event annotation and the entity inventory.

TR03: Quantity conflicts.

ECVA 924

GT annotation.


```
{
  "video_id": "000000198842",
  "split": "train",
  "scene_category": "commercial_anomaly_less",
  "anomaly_type": "explosion",
  "VAD": {
    "prompt": [
      "1. Panoramic fixed shot: Four stunt planes fly in formation against a clear blue sky,
           leaving trails of white smoke behind them as they perform synchronized maneuvers
          .",
      "2. Panoramic fixed shot: Suddenly, one of the planes veers sharply off course, its
          engine emitting a loud roar followed by visible flames shooting from the tail. The
           plane spirals downward, leaving a trail of black smoke.",
      "3. Panoramic fixed shot: The remaining three planes continue their flight, adjusting
          their formation to avoid the descending aircraft. The pilot of the malfunctioning
          plane attempts to regain control, but the situation worsens as the plane begins to
           spin uncontrollably."
    ]
  },
  "VAU": {
    "event_level": [
      "Four stunt planes fly in formation against a clear blue sky, leaving trails of white
          smoke behind them as they perform synchronized maneuvers.",
      "Suddenly, one of the planes veers sharply off course, its engine emitting a loud roar
           followed by visible flames shooting from the tail. The plane spirals downward,
          leaving a trail of black smoke.",
```

```
    "The remaining three planes continue their flight, adjusting their formation to avoid
        the descending aircraft. The pilot of the malfunctioning plane attempts to regain
        control, but the situation worsens as the plane begins to spin uncontrollably."
   ],
   "video_level": "The video shows a group of small aircraft performing an aerial display.
       Initially, four planes fly in formation, leaving white smoke trails. Later, two of
       the planes catch fire, emitting thick black smoke, while another plane continues to
        fly nearby."
  }
}
```


Explanation. The `VAD.prompt` and `VAU.event_level` describe one plane catching fire while the remaining three continue flying. The `VAU.video_level` summary instead states that two planes catch fire. The linked descriptions disagree on the number of affected aircraft, so the summary introduces a quantity conflict rather than merely compressing the event-level account.

TR04: Agent/patient conflicts.

ECVA 1288

GT annotation.


```
{
  "video_id": "1288",
  "classification": "Robbery-Other",
  "reason": "A woman accused a black man of stealing on a bus, and was later beaten by
      three black men who robbed her and threw her off the bus",
  "result": "A woman was accused of theft by a black man and was beaten and robbed of her
      phone by the man and two of his accomplices before leaving the bus",
  "moment": "[0032,0131]",
  "description": "[After being identified by a woman for theft, a black male thief
      assaulted the identified person and robbed her phone, then threw the identified
      person off the bus]",
  "key_sentences": "[\"identified by a woman for theft\",\"black male thief assaulted\",\"
      man robbed her phone\",\"threw the identified person off the bus\"]"
}
```


Explanation. The `reason` field says that a woman accused a man of theft, while `result` says that the man accused the woman. Both refer to the accusation preceding the same assault. Reversing who makes the accusation and who receives it changes the event's participant roles and creates an agent–patient conflict between fields.

Pistachio-VAU 000000071081

GT annotation.


```
{
  "video_id": "000000071081",
  "split": "train",
  "scene_category": "road_anomaly_more",
  "anomaly_type": "robbery",
  "VAD": {
   "prompt": [
    "1. Fixed wide-angle shot: A small convenience store (interior) is dimly lit, with a
        female cashier (short brown hair, wearing glasses) standing behind the counter. A
        male customer (tall, wearing a baseball cap) enters through the front door.",
    "2. Fixed wide-angle shot: The male customer approaches the counter and stands facing
        the cashier. He places his hand inside his jacket pocket.",
    "3. Fixed wide-angle shot: The male customer pulls out a handgun and points it
        forward at the cashier. He demands money loudly.",
    "4. Fixed wide-angle shot: The cashier, visibly shaken, opens the cash register drawer
         and hands over a stack of bills to the male customer.",
    "5. Fixed wide-angle shot: The male customer grabs the money and shoves the cashier
        backward with his free hand. The cashier stumbles but remains standing.",
    "6. Fixed wide-angle shot: The male customer turns and exits the store quickly through
         the front door, heading left.",
    "7. Fixed wide-angle shot: The cashier picks herself up from the floor and dials a
        phone, appearing distressed. She speaks urgently into the receiver.",
    "8. Fixed wide-angle shot: A police car (marked with blue lights) arrives outside the
        store, and two officers exit the vehicle, approaching the entrance. The cashier
        steps outside and speaks to them."
   ]
  },
```

```
  "VAU": {
    "event_level": [
      "A small convenience store (interior) is dimly lit, with a female cashier (short brown
           hair, wearing glasses) standing behind the counter. A male customer (tall,
          wearing a baseball cap) enters through the front door.",
      "The male customer approaches the counter and stands facing the cashier. He places his
           hand inside his jacket pocket.",
      "The male customer pulls out a handgun and points it forward at the cashier. He
          demands money loudly.",
      "The cashier, visibly shaken, opens the cash register drawer and hands over a stack of
           bills to the male customer.",
      "The male customer grabs the money and shoves the cashier backward with his free hand.
           The cashier stumbles but remains standing.",
      "The male customer turns and exits the store quickly through the front door, heading
          left.",
      "The cashier picks herself up from the floor and dials a phone, appearing distressed.
          She speaks urgently into the receiver.",
      "A police car (marked with blue lights) arrives outside the store, and two officers
          exit the vehicle, approaching the entrance. The cashier steps outside and speaks
          to them."
    ],
    "video_level": "In a convenience store, a man in a blue cap points a gun at another man,
         while a woman stands nearby. A police officer enters the store through the back
        door, approaching the situation."
  }
}
```


Explanation. The `VAD.prompt` and `VAU.event_level` have the gunman point his weapon at the female cashier. The `VAU.video_level` summary instead has him point it at another man, with a woman nearby. Changing the threatened participant from the cashier to the man creates a cross-text action-target conflict, rather than a harmless difference in detail.

TR05: Scene, environment, or spatial-relation conflicts.

ECVA 875

GT annotation.


```
{
  "video_id": "875",
  "classification": "Socialunrest-Shooting",
  "reason": "Four police officers were inside a home trying to serve a warrant.",
  "result": "Four police officers were killed while serving a warrant. Four other officers
       were injured.",
  "moment": "[0016,0065]",
  "description": "[Four police officers were trying to execute an arrest warrant at a house.
        They were waiting outside with guns ready to arrest the criminals, but they were
       alert and resisted by the criminals.]",
  "key_sentences": "[\"execute an arrest warrant\",\"waiting outside with guns\",\"resisted
        by the criminals\",\"police officers were killed\"]"
}
```


Explanation. The `reason` field places the officers "inside a home" while serving a warrant. The `description` and `key_sentences` instead place them waiting outside during the same stage. These locations cannot describe the same spatial situation, so the linked fields conflict on the officers' position rather than supplying complementary scene information.

ECVA 2128

GT annotation.


```
{
  "video_id": "2128",
  "classification": "Traffic Accidents-Motor",
  "reason": "On a densely packed highway, vehicles move one after another, with very narrow
       gaps between vehicles. A motorcycle follows behind a black SUV",
  "result": "When the motorcycle is parked in front of the black SUV, there is no time to
       stop and the SUV collides with the rear end. The motorcycle is mainly responsible for
        traffic accidents because it does not maintain a good distance between the cars",
  "moment": "[0000,0019]",
  "description": "[Motorcycle and SUV rear end collision]",
  "key_sentences": "[\"ehicles move one after another\",\"with very narrow gaps between
       vehicles\",\"in front of the black SUV\",\"The motorcycle is mainly responsible for
       traffic\"]"
```

```
}
```


Explanation. The `reason` field places the motorcycle behind the black SUV, while `result` and `key_sentences` place it in front of that vehicle. Both descriptions concern the same pair of vehicles in the collision account. Reversing their relative positions changes the spatial relation, creating a cross-field conflict that affects the event interpretation.

TR06: Sequencing, duration, or stage-relation conflicts.

ECVA 1305

GT annotation.


```
{
  "video_id": "1305",
  "classification": "Pedestrain Incidents-Jaywalking",
  "reason": "A bearded man wearing a black coat and khaki pants with a duck tongue cap
       crossed the road. He waved to stop passing vehicles, then jumped in front of a car
       and crouched down, standing on his hands upside down, then stood up and walked away",
  "result": "He waved to stop passing vehicles, made a strange inverted gesture, and then
       left.A bearded man wearing a black coat and khaki pants with a duck tongue cap
       crossed the road.",
  "moment": "[0000,0033]",
  "description": "[A man crossed the road, stopped a car, and made strange movements before
        leaving]",
  "key_sentences": "[\"crossed the road\",\"stop passing vehicles\",\" made a strange
       inverted gesture\",\"wearing a black coat and khaki pants\"]"
}
```


Explanation. The `reason` field narrates crossing the road before stopping a car, whereas `result` places the road-crossing statement after the stopping and departure account. The review records this as a sequencing inconsistency between linked fields. The highlighted passages identify the competing narrative orders; the concern is their temporal organization rather than differing event vocabulary.

B.3 VIDEO

V01: Insufficient visual clarity.

V01-1: VALU `ECVA/1158.mp4`

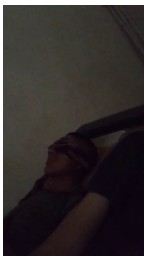
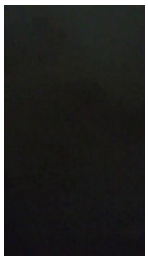
Video frames.




Explanation. The nighttime recording contains strong motion blur and smeared highlights. In the frames at 2 and 8 seconds, the edges of nearby objects, scattered furniture, and surrounding structures are poorly resolved. This loss of visual detail limits reliable inspection of the scene.

V01-2: HIVAU-70K `YANmwpkuWRo(segment1)`

Video frames.

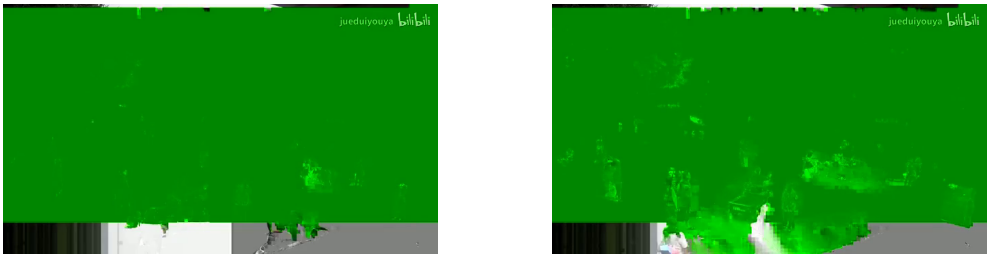



Explanation. The recording is severely underexposed. The frame at 10 seconds retains only dim outlines, while the frame at 42 seconds is nearly black, with little recoverable scene detail. The lack of visible structure makes the depicted content difficult to assess.

V02: Media integrity defects.

V02-1: ECVA `1525`

Video frames.



Explanation. The frames at 20.5 and 22 seconds are dominated by a large green region and corrupted image blocks. Most of the underlying surveillance scene is lost or fragmented. The defect prevents observation of the affected content during this part of the video.

V02-2: HIVAU-70K `Burglary024_x264`

Video frames.



Explanation. The surveillance scene is visible at 83.8 seconds, but the frame at 84 seconds is almost entirely replaced by a gray, block-patterned image. This brief corruption interrupts the visual record and removes the scene information available immediately beforehand.

V03: Critical visual occlusion.

V03-1: VALU `Explosion010_x264_clip_1.mp4`

Video frames.



Explanation. Dense foreground trees conceal the area behind the roadside woodland. The frames at 5 and 12 seconds show the tree barrier and smoke rising from behind it, while the underlying event region remains hidden. The visible aftermath does not reveal the activity taking place behind the trees.

V03-2: VALU `ECVA/877_clip_1.mp4`

Video frames.



Explanation. A broad artificial blur strip covers the lower central part of the view in the frames at 25 and 30 seconds. The overlay conceals the ground-level subject and activity region while the surrounding room remains visible. It removes the visual evidence needed to inspect that part of the scene.

B.4 ALIGNMENT

A01: Unsupported people or objects.
People, objects, identities, counts, or attributes asserted by the text are not established by the video.

A01-1: HAWK `ruiSuGHJf6Q_004479_video.mp4 description`

Video frames.





GT annotation.

```
{
  "description": "While attempting to overtake a truck in the opposite lane, the driver
      misjudges the distance and speed, resulting in a head-on collision with the truck."
}
```

Explanation. The description identifies both the overtaken vehicle and the collision counterpart as trucks. The frames instead show passenger vehicles: a dark sedan passes across the camera view and collides with a light-colored passenger car. The visible vehicle bodies do not support the stated truck identities.

A01-2: CUVA `00243 Result`

Video frames.





GT annotation.

```
{
  "visual_input": "00243.mp4",
  "output": "Two robbers flipped into the counter and used their own pockets to take away
      the jewelry inside. One of the robbers was wearing a reflective vest, while the other
       was wearing a fire safety helmet",
  "task": "Result",
  "ID": "test_389"
}
```

Explanation. The Result annotation identifies one participant by a fire safety helmet. In the surveillance footage, that participant wears a broad-brimmed hat. The second view during the counter raid shows the same headwear. No visible features establish the more specific fire-safety identification; that attribute is unsupported, even though the taking of items is visible.

A02: Unsupported scene or spatial information.
The asserted environment, location, direction, or spatial relation does not match the visible scene.

A02-1: HAWK `Eq7_uD2yN5Y_001004_video.mp4 description`

Video frames.





GT annotation.

```
{
  "description": "As the driver attempts to avoid the white teddy bear sitting in the
      middle of the street, they swerve to the side but accidentally clip the side of the
      bus, causing minor damage to the current vehicle's side mirror and fender."
}
```

Explanation. The white teddy bear remains fixed at the bottom of the view on the recording vehicle's dashboard while the road and surrounding vehicles move behind it. It is not an obstacle in the street. The caption therefore assigns a visible object to the wrong spatial location; the toy's existence does not establish its claimed position.

A02-2: HAWK `7uTMZ5-aWKA_000590_video.mp4 description`

Video frames.





GT annotation.

```
{
  "description": "While driving down the highway in the heavy rain, the car's visibility is
       reduced due to the wet windshield. The white truck in front of the car suddenly
      brakes, causing the car to rear-end the truck, resulting in damage to both vehicles."
}
```

Explanation. The caption describes heavy rain and a wet windshield. The inspected driving sequence shows clear views of lane markings and vehicles through the windshield, with no visible heavy rainfall or rain-obscured glass. The later haze appears when the vehicle approaches the roadside. The asserted adverse-weather scene is not supported by these observations.

A03: Unsupported actions or processes.
The stated action or event process is unestablished or differs from the visible behavior.

A03-1: HAWK `hfwb3mZZ0fs_004369_video.mp4` `description`

Video frames.



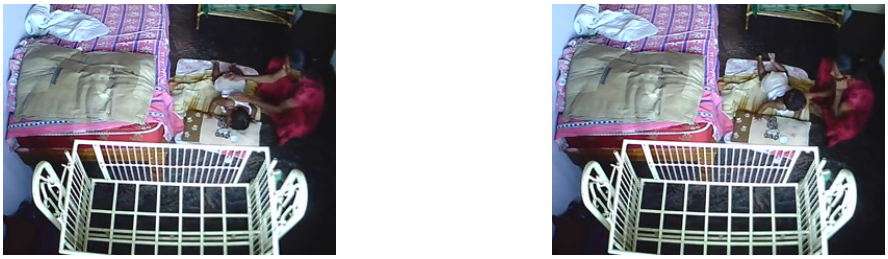
GT annotation.

```
{
  "description": "While driving down the city street, the white bus suddenly swerves to
      avoid an opening car door from a parked car and ends up colliding with the side
      mirror of the car adjacent to it, causing damage to both vehicles."
}
```

Explanation. The bus loses control and ends up sideways across the roadway, with no substantial collision visible. No parked-car door opens in its path, and no impact with an adjacent car mirror is shown. The description invents a door-avoidance and collision sequence that is not supported by the footage.

A03-2: UCF-CAP `Abuse042_x264_0028` `caption`

Video frames.



GT annotation.

```
{
  "video_id": "vid875",
  "caption": "a woman abuses a baby"
}
```

Explanation. The complete short clip shows a woman seated beside an infant, rubbing or handling the infant's torso and legs in a manner consistent with routine care. No strike, forceful shaking, or other clearly abusive action is visible in this interval. The categorical action label is therefore not established by the scoped footage.

A04: Unsupported outcomes or consequences.
The text adds or overstates an outcome, damage, injury, reaction, or final disposition beyond the visible evidence.

A04-1: VALU `RoadAccidents010_x264.mp4` `Abnormal_Consequence`

Video frames.




GT annotation.


```
{
  "subset": "Abnormal_Consequence",
  "video": "RoadAccidents010_x264.mp4",
  "source": "UCF",
  "description": "The red bus stops and the white car is destroyed.",
  "timestamps": [9.0, 17.6]
}
```


Explanation. The consequence interval shows the bus stopped against the car, whose visible roof, rear window, and much of the body remain intact. Contact and possible damage do not establish the stronger claim that the car is destroyed. The highlighted outcome exceeds what the visible aftermath supports.

A04-2: CUVA `00584` `Result`

Video frames.




GT annotation.


```
{
  "visual_input": "00584.mp4",
  "output": "After passing the curve, due to not slowing down, the angle of entry was too
       large and the speed was too fast. The front of the car could not be pulled up in time
        to enter the straight lane, causing the vehicle to collide with the roadside
       guardrail and injuring the onlookers on the roadside.",
  "task": "Result",
  "ID": "test_641"
}
```


Explanation. The vehicle strikes the roadside area and nearby people move away, then approach or look toward the crash. The footage establishes the impact and the bystanders' reactions, but does not resolve a bodily injury to an onlooker. The stated injury outcome is unestablished, rather than disproved merely because no wound is visible.

A05: Unsupported temporal information.
The text misstates temporal location, onset, duration, or the order of established stages. A missing action alone is classified under A03, not as a reversed order.

A05-1: UCA `Abuse037_x264 sentence`

Video frames.



GT annotation.

```
{
  "timestamps": [40.4, 47.6],
  "sentences": [
    "The gray dog pulled out the tail of the dog that had just been run over by the car,
        then stayed for a few minutes and left."
  ]
}
```

Explanation. The sentence assigns a stay of several minutes to the dog before it leaves. The native sentence interval spans only 7.2 seconds, and the continuous surveillance footage shows the dog moving and briefly investigating the ground within that interval. No minutes-long stay is established by the scoped video. The highlighted error is the stated duration, independently of the other claims about the dogs.

A05-2: HIVAU-70K `Normal_Videos_365_x264 description`

Video frames.



GT annotation.

```
{
  "events_summary_split": {
    "description": "The video shows a normal sequence of events, where a man with a backpack
         sits in the upper left corner, a staff member sits in front of a computer in the
        upper right corner, the man leaves, and then another man in the lower right corner
        gets up, takes an order, and returns to his seat, with all events unfolding in a
        routine and expected manner."
  }
}
```

Explanation. The lower-right man stands and walks across the room around 18–30 seconds, returning to his seat around 36 seconds, while the upper-left man remains seated. The upper-left man rises and leaves his seat only around 53–55 seconds. Thus, the lower-right movement precedes the upper-left departure from the seat, contrary to the explicit “and then ordering. Times here are relative to the inspected interval, not the original full video.

A06: Unsupported explanations or inferences.
A cause, intention, responsibility, or analytic explanation goes beyond the evidence supporting the visible event. Overconfidence is a cross-cutting property, not a separate category.

A06-1: HAWK `bI9JzQSrEGo_005170_video.mp4 description`

Video frames.




GT annotation.

```
{
  "description": "As the red double-decker bus continues down the street, a sudden brake
      failure causes it to rear-end the white car in front, ultimately pushing it into the
      truck ahead, leading to a three-vehicle collision."
}
```

Explanation. The clip shows a red bus alongside moving traffic, a white car crossing into the camera view, and an abrupt interaction near the recording vehicle. It supplies no observation of the bus's braking mechanism failing. Naming brake failure as the specific cause goes beyond the external traffic motion; that mechanical explanation cannot be established from the pictured event.

A06-2: ECVA `26 Cause`

Video frames.





GT annotation.

```
{
  "video_id": 26,
  "classification": "Public Safety-Crimes",
  "reason": "A man, intoxicated, was unable to control his actions."
}
```

Explanation. The surveillance sequence shows a man approaching and toppling a bird sculpture, then falling with it. These actions do not independently establish intoxication or an inability to control his behavior. The Cause field converts visible conduct into a definite physiological and behavioral explanation without the evidence needed to support it.

A07: Necessary video content omitted.
The text omits an independently necessary event, stage, or detail that is visible within the field's required coverage. Correcting a stated false fact is not counted again as an omission.

A07-1: UCF-CAP `Shooting040_x264_0000 caption`

Video frames.




GT annotation.


```
{
  "video_id": "vid1710",
  "caption": "three police officers are searching a car"
}
```


Explanation. The initial scene includes police activity beside a car. Near the end of the same clip, the dark-shirted man abruptly lunges toward the person facing him and a physical struggle begins. A caption limited to searching the car omits this distinct, salient escalation, which the complete-clip description is required to cover. The highlight marks the incomplete account, not nonexistent text for the missing event.

A07-2: UCF-CAP `Fighting046_x264_0006 caption`

Video frames.





GT annotation.


```
{
  "video_id": "vid1439",
  "caption": "a group of children standing inside an internet cafe"
}
```


Explanation. The opening group is followed by people rushing toward the foreground and other individuals advancing through the aisle carrying long sticks. Describing only children standing in the cafe leaves out the clip's salient change from gathering to flight and confrontation. This missing event stage is necessary to characterize the observed activity, rather than an optional background detail.

C Evaluation Questions and Scoring Rubric

Each release is assessed under four separately queried tasks. V receives only the silent source video; T-local receives one text and its field responsibility; T-relation receives only the anonymized text pair(s) permitted by the benchmark's native relation contract (App. D lists which benchmarks have one); A receives the video evidence scoped to the field, the target text, and any permitted reference context. No task ever receives the full video together with every field's text at once.

V.

1. **Media visibility.** Is overall image quality clear, degraded, severely degraded, or unusable?
   - 3: clear, only minor compression or softening.
   - 2: a noticeable overall degradation or a locally severe defect with real impact.
   - 1: pervasive or persistent severe blur, noise, exposure loss, or compression artifacting.
   - 0: essentially unusable.
2. **Media integrity.** Is there real corruption that removes or freezes content?
   - 3: no corruption beyond minor compression or sampling artifacts.
   - 2: a localized, real defect with a clear information gap.
   - 1: frequent or persistent corruption with substantial information loss.
   - 0: mostly unusable.
3. **View occlusion.** Does occlusion prevent adequate observation of the key subject, target, or interaction of the main activity?
   - 3: no qualifying gap (ordinary background/corner occlusion and brief mutual overlap are allowed).
   - 2: a limited, minor gap while the key target and interaction remain mostly observable.
   - 1: the key stage, subject, target, or interaction region is persistently or repeatedly obscured.
   - 0: the main activity is essentially unobservable.

No V question is inapplicable.

T-local. Receives only the current text and its configured field responsibility, without video, neighboring text, or other fields.

1. **Clarity.** Do grammar, word choice, and sentence structure convey an unambiguous meaning?
   - 3: clear (minor non-native phrasing or recoverable ellipsis allowed).
   - 2: a local defect but the main proposition is determinable.
   - 1: a fragment or structural ambiguity makes the core action/relation hard to determine.
   - 0: essentially unparseable.
2. **Reference clarity.** Are pronouns, demonstratives, and elisions resolvable within this text alone?
   - 3: clear or no reference to resolve.
   - 2: a minor local ambiguity, main participants still clear.
   - 1: a key participant, object, source, or destination is unresolvable within this text, or multiple candidate antecedents affect the main meaning.
   - 0: key references are unresolvable throughout.
3. **Information fit.** Does the text supply the semantic content its configured field responsibility requires?
   - 3: the configured responsibility is fulfilled.
   - 2: a minor irrelevant addition or slight deviation, main purpose preserved.
   - 1: required core content is missing or the text substantially deviates from the responsibility, though some usable content remains.

- 0: essentially no usable content for this responsibility.

4. **Redundancy.** Does the text contain non-informative internal repetition?
   - 3: no material redundancy.
   - 2: one instance or limited local repetition, main content still clear.
   - 1: repetition dominates or clearly impedes reading.
   - 0: extreme repetition leaves essentially no usable narrative.
5. **Internal consistency.** Do the text's explicit statements avoid mutually exclusive claims?
   - 3: no clear conflict.
   - 2: a minor, non-central self-contradiction.
   - 1: a conflict between important statements that harms the main reading.
   - 0: core statements are directly mutually exclusive.

T-relation. Compares only the anchor–target pairs declared by the benchmark's native relation contract; UCF-CAP, UCA, and HAWK have no such contract and receive a structural N/A for this task.

1. **Claim compatibility.** Do the paired texts conflict on the action, direction, state, or result of the same activity?
   - 3: no confirmed conflict.
   - 2: a minor local conflict.
   - 1: an important action/state/result is incompatible.
   - 0: the shared event's main meaning is directly contradicted.
2. **Entity consistency.** Do the paired texts conflict on the identity, category, or comparable attributes of the same object?
   - 3: no confirmed conflict (hyponym/hypernym generalization allowed).
   - 2: a minor attribute conflict.
   - 1: a key identity/category/attribute conflict harms the main reading.
   - 0: the core object identity is directly contradicted.
3. **Quantity consistency.** Do the paired texts conflict on an explicit count for the same complete group?
   - 3: no confirmed conflict (subset/superset/step counts can coexist).
   - 2: a minor conflict within the same complete count.
   - 1: a key total is substantively inconsistent though the shared main event still holds.
   - 0: a core quantity constraint is directly contradicted.
4. **Role consistency.** Do the paired texts conflict on who performs or receives the same action?
   - 3: no confirmed conflict.
   - 2: a minor role conflict on a secondary action.
   - 1: the agent or patient of a key action is substantively inconsistent.
   - 0: the core meaning is determined by this relation and is reversed or negated.
5. **Scene consistency.** Do the paired texts conflict on the scene, environment, or spatial relation of the same stage?
   - 3: no confirmed conflict (different stages, or part/whole scope, can coexist).
   - 2: a minor scene/location conflict.
   - 1: a key location/environment/spatial relation is incompatible, though shared content remains.
   - 0: the core scene relation is directly contradicted.
6. **Temporal consistency.** Do the paired texts conflict on the order, duration, or stage relation of the same activity?
   - 3: the stated temporal relation can hold jointly.
   - 2: a minor local temporal conflict.

- 1: a key ordering/duration/stage relation is substantively incompatible, though shared content remains.
- 0: the core temporal structure is directly contradicted.
- **null**: no comparable temporal relation exists between the two texts.

A. A first extracts atomic claims from the target text under its field responsibility and rates each against the scoped video evidence (established / minor issue / not established / contradicted) — an unscored step. A final unscored step reports whether video quality or text ambiguity limited the assessment. The scored questions below are either answered directly by the model or derived programmatically from the rated claims; certainty-calibration questions are always derived, never independently judged.

Direct description (routes UCF-CAP, UCA, CUVA, ECVA, HIVAU-70K clip captions, FineVAU, and VALU's direct fields):

1. **Factual support.** Are the stated people, objects, actions, states, and relations established by the current video?
   - 3: all assessable claims established, no material issue.
   - 2: no core claim unestablished or contradicted, only a minor or non-core issue.
   - 1: a core claim is unestablished, or a partial core contradiction still leaves substantive alignment.
   - 0: a main-meaning core contradiction changes the field into another event or the opposite relation.
2. **Scope/localization support.** Does the text's explicit commitment to a time or segment position match locatable video evidence?
   - 3: the committed position or boundary matches locatable evidence.
   - 2: a specific, minor localization deviation.
   - 1: the required localization cannot be established.
   - 0: the visible location explicitly conflicts with the commitment.
   - **N/A**: the field is not configured to require localization.
3. **Critical omission.** Does the description retain the necessary visible content the field explicitly requires coverage of? (Only enabled when the field requires coverage.)
   - 3: required content is expressed.
   - 2: a necessary secondary point is missing, main understanding intact.
   - 1: a major necessary point is missing, harming understanding.
   - 0: the omission makes the expressed main event substantially different or the opposite.
4. **Certainty calibration.** Is the field's certainty wording commensurate with the evidence? (Program-derived.)
   - 3: no overconfidence, including hedged wording appropriately preserved for an unrefuted explanation.
   - 2: a minor or non-core instance of overconfidence.
   - 1: an unestablished core claim is asserted with certainty, or a partial core contradiction.
   - 0: a main-meaning core contradiction.

Summary (routes HAWK's description, HIVAU-70K's event/video summaries, Pistachio-VAU's video summary):

1. **Factual support.** Are the stated people, actions, events, and results established by the field-specified video range?
   - 3: all assessable claims established, no material issue.
   - 2: no core claim unestablished or contradicted, only a minor or non-core issue.
   - 1: a core claim is unestablished, or a partial core contradiction still leaves substantive alignment.
   - 0: a main-meaning core contradiction.

2. **Coverage.** Does the summary cover the main events and key stages the field responsibility requires?
   - 3: required content is expressed.
   - 2: a necessary secondary point is missing, main understanding intact.
   - 1: a major necessary point is missing, harming understanding.
   - 0: the omission makes the expressed main event substantially different or the opposite.
3. **Temporal order.** For stages required by the text or field and established by the video, does their order match the stated commitment?
   - 3: the comparable stages match the commitment.
   - 2: a minor deviation in a stated interval or adjacency, while the required order still holds.
   - 1: both stages are present but their relative order cannot be determined.
   - 0: the video shows an order incompatible with an explicit commitment.
   - **N/A**: no explicit order is required, or no comparable established stage pair exists.
4. **Certainty calibration.** Is the field's certainty wording commensurate with the evidence? (Program-derived.)
   - 3: no overconfidence.
   - 2: a minor or non-core instance of overconfidence.
   - 1: an unestablished core claim is asserted with certainty, or a partial core contradiction.
   - 0: a main-meaning core contradiction.

Cause (routes CUVA's and ECVA's Cause fields):

1. **Content support.** Are the background, trigger, behavior, and specific explanation supported by the corresponding video?
   - 3: all assessable claims established, no material issue.
   - 2: no core claim unestablished or contradicted, only a minor or non-core issue.
   - 1: a core claim is unestablished, or a partial core contradiction still leaves substantive alignment.
   - 0: a main-meaning core contradiction.

   A visible background or trigger alone can fulfill the responsibility; the complete causal chain need not be shown.
2. **Certainty calibration.** Is the field's certainty wording commensurate with the evidence? (Program-derived.)
   - 3: no overconfidence.
   - 2: a minor or non-core instance of overconfidence.
   - 1: an unestablished core claim is asserted with certainty, or a partial core contradiction.
   - 0: a main-meaning core contradiction.

Analysis (routes HIVAU-70K's event/video analysis):

1. **Premise/argument support.** Do the specific premises in the analysis, and their support for its conclusion, have video grounding?
   - 3: the necessary premises and the explanation both hold, no material issue.
   - 2: the core holds; a minor explanation gap or non-core deviation remains.
   - 1: a core premise or the explanation is unestablished, or only a general normative principle is given without concrete video grounding; also covers a partial core contradiction that retains substantive support.
   - 0: a core premise or conclusion contradiction breaks the main analytic meaning.
2. **Certainty calibration.** Is the field's certainty wording commensurate with the evidence? (Program-derived.)
   - 3: no overconfidence.

- 2: a minor or non-core instance of overconfidence.
- 1: an unestablished core claim is asserted with certainty, or a partial core contradiction.
- 0: a main-meaning core contradiction.

Outcome (routes CUVA's Result, ECVA's Effect, VALU's Consequence/Response fields):

1. **Factual support.** Are the object, nature, and degree of the result, impact, or subject's reaction supported by the corresponding video?
   - 3: all assessable claims established, no material issue.
   - 2: no core claim unestablished or contradicted, only a minor or non-core issue.
   - 1: a core claim is unestablished, or a partial core contradiction still leaves substantive alignment.
   - 0: a main-meaning core contradiction.

   A visible result state alone can constitute support; the causing action or its formation process need not be re-shown.
2. **Certainty calibration.** Is the field's certainty wording commensurate with the evidence? (Program-derived.)
   - 3: no overconfidence.
   - 2: a minor or non-core instance of overconfidence.
   - 1: an unestablished core claim is asserted with certainty, or a partial core contradiction.
   - 0: a main-meaning core contradiction.

Key sentences (routes ECVA's Key Sentences):

1. **Factual support.** Are the stated people, objects, actions, states, and relations in the key factual sentences established by the current video?
   - 3: all assessable claims established, no material issue.
   - 2: no core claim unestablished or contradicted, only a minor or non-core issue.
   - 1: a core claim is unestablished, or a partial core contradiction still leaves substantive alignment.
   - 0: a main-meaning core contradiction.
2. **Scope/localization support.** Does the text's explicit commitment to a time or segment position match locatable video evidence?
   - 3: the committed position or boundary matches locatable evidence.
   - 2: a specific, minor localization deviation.
   - 1: the required localization cannot be established.
   - 0: the visible location explicitly conflicts with the commitment.
   - **N/A**: the field is not configured to require localization.
3. **Coverage.** Does the set of key factual sentences cover the main event points the field responsibility requires?
   - 3: required content is expressed.
   - 2: a necessary secondary point is missing, main understanding intact.
   - 1: a major necessary point is missing, harming understanding.
   - 0: the omission makes the expressed main event substantially different or the opposite.

Generation (routes Pistachio-VAU's storyline/prompt and segment requirements; its post-generation video summary follows the Summary rubric):

1. **Requirement coverage.** Does the generated video faithfully realize and cover the people, actions, events, and stages the text explicitly requires? (Direction is text $\rightarrow$ video: every explicit requirement is extracted and checked.)
   - 3: all explicit requirements are realized, no material issue.
   - 2: no core requirement unrealized or contradicted, only a minor or non-core issue.

- 1: a core requirement is unrealized, or a partial core contradiction still leaves substantive alignment.
- 0: a main-meaning core contradiction.

2. **Unrequested content.** Does content the video actually adds independently interfere with or change the expression of the generation requirement?
   - 3: no interference from unrequested content.
   - 2: a minor, clear interference.
   - 1: unrequested content substantially changes understanding of the main storyline.
   - 0: unrequested content dominates the video, replacing the required main line.
3. **Stage order.** For stages the text or field requires and the video establishes, does their order match the stated commitment?
   - 3: the comparable stages match the commitment.
   - 2: a minor deviation in a stated interval or adjacency, while the required order still holds.
   - 1: both stages are present but their relative order cannot be determined.
   - 0: the video shows an order incompatible with an explicit commitment.
   - **N/A**: no stage order is required.

D Dataset Native Field Specification

Native fields verified directly against each benchmark's released annotation files (JSON, JSONL, XLSX, or Parquet).

UCF-CAP.

- `vidID`: release-internal clip identifier.
- `caption`: single free-text sentence describing the clip.
- `youtube_mapping`: maps `vidID` to the released clip file.

UCA.

- `sentences[i]`: visible-event description for interval i.
- `timestamps[i]`: start/end time (s, 0.1s precision) paired with sentence i.
- `duration`: declared video duration.

CUVA.

- `Detection`: binary anomaly flag.
- `Classification`: hierarchical anomaly category.
- `Timestamp`: one or more MMSS-coded anomaly intervals.
- `Description`: observable events within the anomaly interval(s).
- `Cause`: explanation of why the anomaly occurred.
- `Result`: observable consequence of the anomaly.

ECVA.

- `A1 Classification`: anomaly category label.
- `A2 reason` (Cause): cause of the anomaly.
- `A2 result` (Effect): consequence of the anomaly.
- `A3 moment`: one or more [MMSS,MMSS] intervals.
- `A3 description`: observable events within the moment interval(s).
- `A4 key sentences`: 4–6 short, independent factual statements.

HAWK.

- `description`: mixed description with explanatory inference over the full video.
- `QA`: open-world question/answer pair.

HIVAU-70K.

- `clips_caption`: action/state directly visible within a short clip.
- `events_summary_split.description`: summary of an event's stages.
- `events_summary_split.judgement`: normal/anomalous judgment and category.
- `events_summary_split.analysis`: justification for the judgment.
- `video_summary_split.description`: summary of the full video.
- `video_summary_split.judgement`: video-level judgment and category.
- `video_summary_split.analysis`: justification for the video-level judgment.

FineVAU.

- `event_description`: event description within a resolved timestamp interval.
- `timestamp`: interval in a mixed seconds/clock/string encoding.
- `event_type`: event category (e.g., Abnormal, Normal, Arrival, . . .).
- `involved_entities`: entities participating in the event.
- `entities`: video-level entity inventory.
- `scene`: video-level scene/context description.

VALU.

- `description`: layered description (Abnormal_Event / Normal_Event / Abnormal_Segment / Abnormal_Action / Abnormal_Consequence / Subjects_Response).
- `timestamps`: interval paired with each layered description.

Pistachio-VAU.

- `event_storyline` / `generation_prompt`: full generation requirement for the synthesized video.
- `video_summary`: summary of the synthesized video.
- `vad_prompt.segment`: ordered, officially numbered generation segment.

E Field Responsibility Mapping

The seven standard field responsibilities are listed below. Under each responsibility, the entries identify the datasets and their corresponding native annotation fields.

Direct description.

- **UCF-CAP**: `caption`.
- **UCA**: `sentences[i]`.
- **CUVA**: `Description`.
- **ECVA**: `Description`.
- **HIVAU-70K**: `clips_caption`.
- **FineVAU**: `event_description` (all event types).
- **VALU**: `Abnormal_Event`, `Normal_Event`, `Abnormal_Segment`, `Abnormal_Action`.

Summary.

- **HIVAU-70K**: `events_summary_split.description`, `video_summary_split.description`.
- **Pistachio-VAU**: `video_summary`.

Cause.

- **CUVA**: `Cause`.
- **ECVA**: `Cause`.

Analysis.

- **HIVAU-70K**: `events_summary_split[*].analysis`, `video_summary_split.analysis`.

Outcome.

- **CUVA**: `Result`.
- **ECVA**: `Effect`.
- **VALU**: `Abnormal_Consequence`, `Subjects_Response`.

Key sentences.

- **ECVA**: `Key Sentences`.

Generation.

- **Pistachio-VAU**: `event_storyline` / `generation_prompt`, `vad_prompt.segment`.

F Benchmark Aggregation and Audit Execution Details

F.1 Benchmark-level aggregation and coverage

Quality profiles (RQ1). For RQ1, our primary summary gives each annotation field equal weight, so that its contribution to the benchmark profile does not depend on its number of annotations. We first average units within each field for T-local and A, or within each native relation type for T-relation, and then average these means. For release b and task $d \in \{T\text{-local}, T\text{-relation}, A\}$, let $R_{b,d}$ contain the applicable fields or routes with assessable outputs, and let $X_{b,c,d}$ contain the assessable units in field or route c. The field-macro average of a unit-level quantity z_x is

$$\overline{z}_{b,d}^{\text{field macro}} = \frac{1}{|R_{b,d}|} \sum_{c \in R_{b,d}} \frac{1}{|X_{b,c,d}|} \sum_{x \in X_{b,c,d}} z_x. \tag{3}$$

For V, we instead average over unique source videos with assessable outputs, giving each source equal weight within a benchmark (Appendix G.2). Applying the corresponding aggregation to quality scores yields the benchmark quality score; applying it to major concern indicators and multiplying by 100 yields MCR. These rates describe source videos for V, annotations for T-local and A, and relations for T-relation. Each benchmark retains its native fields and criteria; sharing a 0–100 scale does not make different task scores interchangeable.

Concern concentration within RQ1. To locate concerns within these profiles, we retain the native field or relation type and the criterion responsible for each flagged judgment. Within a field, a criterion-specific rate divides the affected units by the units for which that criterion is assessable. A unit may be flagged on several criteria, so diagnostic counts need not sum to the number of affected annotations. These records distinguish the requirements associated with concerns without inferring their causes from aggregate scores.

Coverage. Quality summaries are conditional on valid, applicable outputs from the declared analysis population. Each evaluation unit contributes its final validated result, with source identities retained within benchmarks.

F.2 Joint usability and overlapping concerns

Within RQ1, we ask how often an annotation has a source video that passes V, a text that passes T-local, and field-specific visual support that passes A. The joint usability rate (JUR) measures the proportion meeting all three conditions. Let $g(u)$ identify the source video of annotation u, and let $\mathcal{L}_b$ contain annotations with assessable V, T-local, and A results linked through that source. Suppressing the fixed threshold $\tau = 1$, define

$$J_u = \big(1 - M_{g(u),V}\big)\big(1 - M_{u,T\text{-local}}\big)\big(1 - M_{u,A}\big), \qquad \mathrm{JUR}_{b,\mathrm{conditional}} = 100\ \mathrm{Aggregate}_{u \in \mathcal{L}_b}\, J_u. \tag{4}$$

The primary aggregation gives each field equal weight over its assessable linked annotations, as in Equation 3. A source-level V result may enter several annotation pairs, while marginal VQS counts that source once. T-relation remains a separate property of related texts and is not a fourth requirement in JUR.

This joint criterion combines whole-video quality with annotation-level quality and alignment. A major V concern affects every linked annotation, even when its local evidence lies elsewhere in the video. Conversely, a passing whole-video V result does not establish the quality of every local interval. Local defect records aid interpretation, but do not create a separate segment-level V score. JUR is therefore interpreted under this source-level requirement, rather than as a test of local evidence quality.

JUR summarizes how often all three requirements are met; the remaining patterns show how they fail. The concern triple $(M_V, M_{T\text{-local}}, M_A)$ defines eight mutually exclusive patterns, including the all-clear pattern. Their frequencies separate isolated concerns from co-occurring failures. Within the same linked population and using identical weights,

$$\mathrm{JUR}_{\mathrm{conditional}}/100 = 1 - P(M_V \vee M_{T\text{-local}} \vee M_A). \tag{5}$$

Marginal concern rates do not determine this joint quantity. Multiplying three marginal pass rates would assume independence; retaining the native links measures their overlap directly.

F.3 Benchmark scope and audit population

We apply the framework to nine language-annotated benchmarks (Table 5). The evaluation units are source videos for V, individual language annotations for T-local and A, and permitted native relation groups for T-relation. VALU uses the TDG release; FineVAU is named FineW3 in the implementation. Closed-form detection labels and auxiliary QA/MCQ fields are outside the RQ1 language-field audit. VALU-ADC multiple-choice questions are used separately for downstream evaluation in RQ2 (Appendix G.5).

Study scope and completion. The study population follows Table 2: 7,928 retained video entries and 56,103 video–text pairs across nine benchmarks after the data checks in Appendix A. Table 5 reports the same per-benchmark input counts together with the audited fields and video scopes. These counts describe the included releases and inputs, rather than the number of independent source videos across benchmarks.

The benchmark-level analysis follows the study population and results in Table 2. Field-level quality is reported in Appendix H, with native field responsibilities and evidence scopes retained throughout the analysis. The same linked population and field-macro weights are used for JUR and the decomposition of joint failures.

Source identities are retained within benchmarks; shared content across separately packaged benchmarks is not counted as independent evidence in interpreting the results.

Native relation routing. CUVA compares Description with Cause and Result; ECVA compares Description with Cause and Effect. Neither route compares Cause directly with Result or Effect. VALU compares a parent event description with its bound lower-level descriptions. HIVAU-70K

Table 5: Benchmark input counts and scope, using the same counts as Table 2. "Rel." indicates that T-relation is evaluated; "–" marks benchmarks excluded from T-relation evaluation. Video scopes refer to A inputs; V always uses the whole source video.

Release	Videos	Pairs	A video scope	Audited fields	Rel.
UCF-CAP	396	396	official clip	anomalous-clip caption	–
UCA	310	4,331	event interval	timestamped event sentence	–
CUVA	198	594	event/context intervals	description, cause, result	yes
ECVA	1,989	7,980	event/full video	description, cause, effect, key sentences	yes
HAWK	786	786	full video	anomaly description	–
HIVAU-70K	1,050	13,844	clip/event/video	clip caption, event/video summary, description, analysis	yes
FineVAU	2,009	23,984	event interval	normal and abnormal event descriptions	–
VALU	998	2,995	native timed interval	descriptions at native semantic levels	yes
Pistachio-VAU	192	1,193	segment/video	storyline, summary, ordered segments	yes
Total	7,928	56,103			

compares event summaries with their clips, video summaries with their events, and descriptions with their associated analyses. Pistachio-VAU compares its generation storyline with the corresponding segment requirements, excluding the post-generation summary from T-relation. These parent–child routes do not create sibling comparisons. T-relation evaluation covers CUVA, ECVA, VALU, HIVAU-70K, and Pistachio-VAU.

F.4 Execution and evidence provenance

V and A use silent video sampled locally at 4 fps, preserving the source spatial resolution and the task's specified temporal scope. The initial H.264 encoding uses CRF 18; V uses the slow encoding preset. V covers the full source timeline, whereas A receives only the scope authorized for the annotation. The API request also specifies 4 fps and high media resolution. T-local and T-relation receive text only. Each evaluation unit is assessed in a separate request containing its applicable question set, so judgments from one task are not supplied as evidence to another.

Responses pass structural and protocol-specific validation before scoring. For A, validation checks the fidelity of extracted claim spans and the consistency of evidence judgments before deterministic score derivation. Bounded retries handle unsuccessful requests. Explicit size rejections or repeated transport failures can trigger stronger compression while retaining the time range, spatial resolution, and frame rate. Consequently, visual quality is assessed on the actual submitted video, which may differ in compression from the released file.

The frozen records retain unit and source identities, field or relation routes, question and prompt versions, input-media metadata, responses, validation outcomes, and replacement history. These records distinguish an assessable low-quality input from a failed request or an inapplicable question. The analysis uses the question banks and execution snapshot that produced the retained outputs.

F.5 Analysis and interpretation

RQ1 combines task-level quality profiles, field and criterion diagnostics, and linked V/T-local/A concern patterns. Only pairs with three assessable results enter the joint decomposition; T-relation is analyzed separately. RQ2 uses these diagnostics to revise CUVA training annotations and compares original, direct-revision, and audit-guided training conditions on CUVA description generation and VALU-ADC multiple-choice accuracy. RQ3 combines native-field coverage with the audit profiles to formulate task-specific use guidance. Appendix G.4 reports the complete downstream comparison, and Appendix G.5 describes the multiple-choice evaluation.

All primary summaries use the weighting in Section F.1: source means for V, field-macro means for T-local and A, and relation-type macro means for T-relation. Conditional JUR and its concern-pattern decomposition use identical field-macro weights over linked pairs. We report descriptive comparisons.

G Coverage and Downstream Evaluation Details

G.1 Coverage accounting

The audit distinguishes source-video groups, task outputs, and linked annotation pairs. Completion is recorded for the configured inputs of each retained source. Quality aggregation uses outputs with applicable ratings, and JUR uses linked V, T-local, and A results as defined in Appendix F.1. Completed-group counts, valid-output counts, and paired denominators describe successive stages of the analysis.

G.2 Aggregation and interpretation

The primary summaries use source means for V, field-macro means for T-local and A, and relation-type macro means for T-relation, as defined in Section F.1. Conditional JUR and the concern pattern decomposition use the same field-macro weights over linked pairs. A major concern requires at least one applicable rating of 0 or 1. Inapplicable questions are excluded from scoring.

G.3 Interpretation of representative cases

Representative cases illustrate how a field's responsibility, authorized video scope, and native text relations affect a quality judgment. Cross-text comparisons follow the linked event and temporal scope; video-quality judgments consider the full source video. Aggregate findings are reported in Section 4.

G.4 Complete downstream evaluation results

CIDEr-D follows Vedantam et al. (2015). Table 6 extends the main-text comparison with CIDEr-D and retains VALU-ADC accuracy. Both models use the same 693 training videos, 78 validation examples, and 198 held-out CUVA videos. In the audit-guided condition, 354 training descriptions are revised and 339 retained. Scores are means over seeds 42, 123, and 2026. The description metrics use the common revised CUVA references; VALU-ADC uses the original multiple-choice answers. The shared values follow Table 3.

Table 6: Complete downstream results on CUVA description generation and VALU-ADC. Scores are three-seed means; higher is better. R-L, METEOR, and BS are scaled by 100, CIDEr-D retains its original scale, and VALU-ADC is accuracy in percent.

Model	Training labels	R-L	METEOR	CIDEr-D	BS	VALU-ADC
Qwen3-VL-8B-Instruct	Original (A)	23.59	9.35	0.455	89.33	42.24
	Direct revision (B)	23.87	9.23	0.443	89.41	42.94
	Audit-guided (C)	**25.95**	**10.78**	**0.474**	**89.51**	**44.58**
VAD-DPO	Original (A)	20.41	8.04	0.309	88.37	20.45
	Direct revision (B)	20.27	8.35	0.344	88.35	20.87
	Audit-guided (C)	**20.68**	**8.70**	**0.371**	**88.68**	**21.08**

G.5 VALU-ADC multiple-choice evaluation

Evaluation set. The original VALU-ADC input contains 5,695 questions associated with 1,019 videos. One video cannot be decoded, so its six questions are excluded from every condition. Evaluation uses the same 5,689 questions from 1,018 videos. These questions form a separate evaluation population from the VALU description fields audited in RQ1. We preserve the original question content, non-null options, option order, and answer labels.

Inputs and inference. The CUVA-trained checkpoints are evaluated without additional training on VALU-ADC. Each input supplies 32 uniformly sampled video frames, the question, and its answer options. The reference answer and the source and task metadata are not given to the model. Normal-event and abnormal-event items use the same event-description question wording. Decoding

is greedy, with at most 16 generated tokens. The evaluation procedure uses a common prompt format and answer parser across annotation conditions.

Video preparation. Nonzero timestamp origins in 143 videos are subtracted before frame sampling. Sequential decoding recovers 112 videos for which random access fails. The remaining undecodable video is the common exclusion noted above. These steps preserve the intended video timeline and do not change the questions or reference answers.

Scoring and aggregation. A response is accepted only if it identifies one available option letter; parentheses, a final period, and an `Answer:` prefix are permitted. Unparseable responses count as incorrect and remain in the denominator. For model m, annotation condition g, and training seed s, accuracy is

$$\mathrm{Acc}_{m,g,s} = \frac{100}{5{,}689} \sum_{i=1}^{5{,}689} \mathbf{1}\{\widehat{y}_{m,g,s,i} = y_i\}. \tag{6}$$

Tables 3 and 6 report the arithmetic mean across training seeds 42, 123, and 2026. Thus, the comparison measures correct option selection using VALU's original labels, complementing agreement with the revised CUVA description references.

H TASK-SPECIFIC REFERENCE QUALITY BY NATIVE FIELD

Tables 7 and 8 report alignment quality for all 26 field routes across the nine benchmarks included in Table 2. The study includes 7,928 video entries and 56,103 video–text pairs. AQS is the mean normalized alignment score; MCR is the percentage of annotations with at least one applicable rating of 0 or 1. Each row retains its native responsibility and evidence scope. n counts eligible annotations in the study inputs, as in Figure 5(b). FineVAU's Normal and Abnormal are routes of event_description. UCA's sentence is the audit route for sentences[i]; ECVA's field names use the role aliases in Appendix D. FineW3 is reported as FineVAU. These tables support comparisons conditional on task and granularity; T-local and T-relation are reported separately in Table 2.

Table 7: Direct descriptions and summaries. AQS is higher-is-better; MCR (%) is lower-is-better.

Benchmark	Native field / route	n	AQS	MCR (%)
Direct description				
UCF-CAP	caption	396	85.56	25.00
UCA	sentence	4331	72.53	28.49
CUVA	Description	198	83.50	24.24
ECVA	Description	1995	80.57	26.62
FineVAU	Normal	16609	59.61	42.50
FineVAU	Abnormal	7375	71.51	44.94
VALU	Abnormal_Event	730	88.36	13.56
VALU	Abnormal_Segment	730	79.47	21.64
VALU	Abnormal_Action	778	80.68	18.12
VALU	Normal_Event	300	85.19	14.67
HIVAU-70K	clips_caption	8611	75.48	24.62
Summary				
HIVAU-70K	events_summary_split.description	2173	72.38	38.79
HIVAU-70K	video_summary_split.description	1050	76.36	35.52
HAWK	description	786	41.47	74.55
Pistachio-VAU	video_summary	192	44.57	80.21

Table 8: Causes, analyses, outcomes, key sentences, and generation requirements. AQS is higher-is-better; MCR (%) is lower-is-better.

Benchmark	Native field / route	n	AQS	MCR (%)
Cause				
CUVA	Cause	198	72.39	26.26
ECVA	Cause	1995	70.94	29.67
Analysis				
HIVAU-70K	events_summary_split[*].analysis	1382	79.73	22.50
HIVAU-70K	video_summary_split.analysis	628	92.20	9.24
Outcome				
CUVA	Result	198	72.73	26.77
ECVA	Effect	1995	66.42	39.20
VALU	Abnormal_Consequence	342	77.42	21.05
VALU	Subjects_Response	115	65.00	45.22
Key sentences				
ECVA	Key Sentences	1995	80.92	31.48
Generation				
Pistachio-VAU	event_storyline/generation_prompt	192	65.74	83.85
Pistachio-VAU	vad_prompt.segment	809	66.41	69.47