← Back to **Author Console** (/group?id=AAAI.org/2026/Conference/Authors#your-submissions)

FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding

(/pdf?id=pAFkeg8U8M

Joao Alexandre Cardeira Pereira (/profile?id=~Joao_Alexandre_Cardeira_Pereira1), Vasco Lopes (/profile?id=~Vasco_Lopes1), João C. Neves (/profile?id=~Jo%C3%A3o_C._Neves1), David Semedo (/profile?id=~David_Semedo1)

Published: 07 Nov 2025, Last Modified: 07 Nov 2025 AAAI-26 Poster
Conference, Area Chairs, Senior Program Committee, Program Committee, Publication Chairs, Authors Revisions (/revisions?id=pAFkeg8U8M) BibTeX

Serve As Reviewer: Joao Alexandre Cardeira Pereira (/profile?id=~Joao_Alexandre_Cardeira_Pereira1)
Keywords: Video Anomaly Understanding, Large Vision-Language Models, Surveillance
Primary Keyword: CV: Video Understanding & Activity Analysis
TL;DR: We introduce FineVAU, a benchmark that addresses the gap for human-aligned evaluation in Video Anomaly Understanding through a novel metric and dataset structured into Events, Entities and Location, according to human perception of anomalies.
Secondary Keywords: CV: Language and Vision, CV: Large Vision Models, CV: Scene Analysis & Understanding
Abstract:
Video Anomaly Understanding (VAU) is a novel task focused on describing unusual occurrences in videos. Despite its growing interest, the evaluation of VAU remains an open challenge. Existing benchmarks rely on n-gram-based metrics (e.g., BLEU, ROUGE-L) or LLM-based evaluation. The first fails to capture the rich, free-form, and visually grounded nature of LVLM responses, while the latte focuses on assessing language quality over factual relevance, often leading to subjective judgments misaligned with human perception. In this work, we address this issue by proposing FineVAU, a ne benchmark for VAU that shifts the focus towards rich, fine-grained and domain-specific understanding of anomalous videos. We formulate VAU as a three-fold problem, with the goal of comprehensiv understanding key descriptive elements of anomalies in video: events (What), participating entities (Who) and location (Where). Our benchmark introduces a) FV-Score, a novel, human-aligned evaluat metric that assesses the presence of critical visual elements in LVLM answers, providing interpretable, fine-grained feedback; and b) FineW^3, a novel, comprehensive dataset curated through a structu and fully automatic procedure that augments existing human annotations with high quality, fine-grained visual information. Human evaluation reveals that our proposed metric has a superior alignm with human perception of anomalies in comparison to current approaches. Detailed experiments on FineVAU unveil critical limitations in LVLM's ability to perceive anomalous events that require spati and fine-grained temporal understanding, despite strong performance on coarse grain, static information, and events that typically comprise strong visual cues.

Country Of Institutions: Portugal
Supplementary Material: zip (/attachment?id=pAFkeg8U8M&name=supplementary_material)
Profile Policy Agreement: I confirm that all authors have up-to-date OpenReview profiles, including their current position, institution-affliated email address, and DBLP URL. I understand that submissions with incomplete author profiles will be subject to desk rejection.
Submission Number: 27872

Filter by reply type... Filter by author... Search keywords... Sort: Newest First
Everyone Program Chairs Submission27872... Submission27872... Submission27872... Submission27872... Submission27872... Submission27872...
Submission27872... Submission27872... Submission27872... Submission27872... Submission27872...
9 / 9 replies shov

Add: Withdrawal Ethics Chair Author Comme

Paper Decision

Decision by Program Chairs 07 Nov 2025 at 18:53 (modified: 07 Nov 2025 at 23:35) Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=O88NRBw6xU)

Decision: Poster

Add: Ethics Chair Author Comment

Phase 2 AC Recommendation by Area Chairs

Phase 2 AC Recommendation by Area Chairs 31 Oct 2025 at 15:59 (modified: 07 Nov 2025 at 22:36)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=Ay7QbTtmE7)

Metareview:
While this paper is interesting in its own right, with a new dataset for fine-grained video anomaly understanding, it raises numerous issues, as pointed out by the reviewers and SPC, including design, evaluation, and a lack of methodological detail for reproducibility. As such, the AC agrees with the SPC's and reviewers' recommendations.

Acceptance Recommendation: This paper is in the bottom 25% of papers presented at a top tier venue like AAAI. (Weak accept recommendation.)
Confidence: 5: The AC is absolutely certain

Add: Ethics Chair Author Comment

Phase 2 SPC Recommendation by Senior Program Committee pZ7E

Phase 2 SPC Recommendation by Senior Program Committee pZ7E 23 Oct 2025 at 15:53 (modified: 08 Nov 2025 at 00:30)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=pUjBVrv0Ef)

Metareview:
The reviewers describe this work as well-motivated and clearly written, introducing a human-aligned benchmark for fine-grained video anomaly understanding. They appreciate the organization and annotation design, and note that the dataset is comprehensive and covers diverse scenarios. However, the reviews repeatedly point out that the paper's technical novelty is limited, and that it focuses mainly on dataset construction rather than methodological innovation. The rebuttal provides additional details about annotation consistency and evaluation, but reviewers agree that it does not change the overall assessment. The consensus is that the paper is valuable as a dataset resource but does not meet the bar for the AAAI main track.

Acceptance Recommendation: This paper is slightly below papers presented at a top tier venue like AAAI. (Weak reject recommendation.)
Confidence: 5: The SPC is absolutely certain

Add: **Ethics Chair Author Comment**

Review of "FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding"

Official Review by Program Committee WwU2 05 Oct 2025 at 14:37 (modified: 07 Nov 2025 at 21:13)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee WwU2, Authors Revisions (/revisions?id=cYKfqy1Ts9)

Review:

This paper introduces a new benchmark aimed at advancing the evaluation of video anomaly understanding. The main contributions are: (1) FV-Score, a novel evaluation metric that measures the presence of critical visual elements (What, Who, Where) in LVLM responses, providing interpretable and human-aligned feedback; and (2) FineW3, a new dataset that enhances existing human annotations with fine-grained, automatically generated visual information. Experimental results and human studies demonstrate that FV-Score correlates more strongly with human judgment and effectively reveals key weaknesses of current LVLMs in modeling fine-grained spatial and temporal anomaly understanding.

Pros:

1. The paper addresses the objective evaluation of video anomaly understanding, an important yet underexplored problem.
2. The proposed FV-Score is novel, interpretable, and well-aligned with human perception.
3. The construction of FineW3 through an automated fine-grained annotation pipeline is well-motivated and valuable for future research.

Cons / Suggestions:

1. It would be helpful to clarify what specific design aspects of the proposed evaluation method are tailored to video anomaly understanding, and how the task fundamentally differs from related areas such as video captioning or event detection.
2. While FV-Score appears promising, the paper would benefit from ablation studies or correlation analyses with existing metrics to better contextualize and validate its advantages.

Rating: 6: Marginally above acceptance threshold
Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

Review of FineVAU

Official Review by Program Committee oTJ9 05 Oct 2025 at 03:48 (modified: 07 Nov 2025 at 21:13)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee oTJ9, Authors Revisions (/revisions?id=sTw6AU9y8x)

Review:

Summary

- This paper proposes FienVAU, a benchmark tailored towards Video Anomaly Understanding with fine-grained human-aligned evaluation of VLM outputs. The paper claims that the classic metrics based on lexical overlap, such as n-gram or LLM-as-a-judge based metrics, are not sufficient in capturing such details and instead evaluate anomaly understanding through three structured dimensions - namely, what, who and where. The authors construct a new dataset (FineW^3) and propose FV-score and conduct a human correlation study.

Strengths

- The motivation that current evaluation methods may not be sufficient for anomaly understanding is clear.
- The densely annotated dataset (FineW^3) will be a useful benchmark.
- Human study reveals that FV-score moderately correlates better than standard metrics.

Weaknesses

- The most critical weakness is that the coefficients are virtually tuned on the test set. This is an unfair comparison to other methods. In particular, Tabel 3 shows that the proposed FV score does not yield the best correlation metrics for different choices of the parameters, further undermining the validity of the proposed method. Unless the authors can justify a principled way to tune the weighting parameters, there should at least be a separation of the validation and test sets, on which the parameters are tuned and then validated (compared with other methods).
- The sensitivity of the results on the parameters also raise critical concerns on whether the results will generalize to different datasets, or different human judgement.
- The reported correlation is marginally higher, especially when the human-aligned claim is supported by one correlation study only on 60 videos, which is a small portion of the entire dataset.

Questions

- Is the FV-score robust to prompt ablations or paraphrases?
- What are possible failure scenarios of using FV-score?

Rating: 4: Ok but not good enough - rejection
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

Benchmark for VA

Official Review by Program Committee CbQN 29 Aug 2025 at 10:49 (modified: 07 Nov 2025 at 21:13)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee CbQN, Authors Revisions (/revisions?id=Dpn2sCQcsx)

Review:

Paper Summary This paper introduces FineVAU, a benchmark for fine-grained video anomaly understanding (VAU). The key idea is to move beyond surface-level metrics and adopt a structured evaluation (What, Who, Where) that aligns with human perception. The proposed contributions are: (1) FineW3 dataset, built via an LVLM-assisted pipeline for anomaly annotation, and (2) FV-Score, an interpretable evaluation metric that achieves higher correlation with human judgment than existing baselines. The paper is timely and relevant, addressing an important gap in anomaly evaluation. The experimental results are convincing and highlight weaknesses of current LVLMs. However, there are concerns: 1.Reliance on LVLMs for dataset construction could propagate biases and errors, yet the paper provides limited error analysis. 2.The dependency on Gemini for FineVAU-Judge raises questions about generalizability. 3.Human evaluation scale is relatively small compared to dataset size. 4.While diagnostic value is clear, practical applicability for improving anomaly models is under-explored. Overall, the paper makes a solid and meaningful contribution, but some aspects (dataset validation, broader human study, cross-LLM robustness) could be improved. Paper Strengths 1.Addresses a critical and underexplored problem in VAU evaluation. 2.Strong conceptual and methodological contributions. 3.Provides new dataset and metric with demonstrated human alignment. 4.Clear writing, strong motivation, and thorough experimental support. 5.Likely to become a reference benchmark in the field. Paper Weaknesses 1.FineW3 relies heavily on LLM-assisted annotation, which may introduce noise and bias, and lacks sufficient error analysis and comparative experiments. 2.The human evaluation scale is too small to ensure the robustness of the conclusions. 3.The generalization of different LLMs as judges is lacking experimental verification, which may affect the universality of the metric. 4.The discussion of its practical application value (such as how to help the model improve VAU performance) is insufficient, and remains mainly at the evaluation level. 5.The lack of statistical analysis of annotation consistency makes it impossible to fully guarantee the reliability of FineW3 data.

Rating: 3: Clear rejection

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

FineVAU is a benchmark with potential, but it still requires experimental validation.

Official Review by Program Committee 3AJC 27 Aug 2025 at 16:15 (modified: 07 Nov 2025 at 21:13)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee 3AJC, Authors Revisions (/revisions?id=lpGGSk5E6A)

Review:

Conclusion

This paper introduces a new benchmark in VAU called FineVAU, which focus on fine-grained and domain-specific understanding. To formulate this task, they split the anomalies to What, Who and Where, then create FV-Score and a new dataset FineW^3 with around 1.5k videos and 17k events. Experiments reveal that state-of-the-art models struggle on FineW^3.

Strengths

1. This paper create a novel large dataset in VAU with fine-grained annotation.
2. The proposed FV-Score is structured, interpretable, and highly aligned with human perception.

Weakness

1. The paper does not specify which LVLM was used for Fine^3 dataset annotation.
2. The paper does not mention quality control for the dataset. Since the dataset was annotated by LVLM based on existing datasets, manual quality verification is necessary.
3. In table 4, the paper lacks results from closed-source SOTA models (e.g. GPT4v, Gemini) and also lacks testing on larger version closed-source models (e.g. Qwen2.5-VL-72B, InternVL2.5-78B).
4. In table 2, the paper lacks comparison with embedding-based metrics (e.g. BERTScore, BLEURT, MoverScore). This type of metric has been published for quite some time. These metrics are also missing from the VAU Evaluation section of the Related Work.
5. The contribution of the dataset is limited to evaluation purposes in this paper. The authors only explore the potential benefits for training, lacking concrete experiments to further demonstrate the dataset's practical utility.

Rating: 4: Ok but not good enough - rejection
Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

AI Review

AI Review by Program Committee AI 27 Aug 2025 at 07:48 (modified: 10 Oct 2025 at 20:07)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee AI, Authors Revisions (/revisions?id=NApHQLeAKa)

Review:

Title: FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding

Synopsis of the paper The paper introduces FineVAU, a benchmark for evaluating video anomaly understanding that decomposes evaluation into three human-aligned dimensions: events (What), entities (Who), and location (Where). It proposes FV-Score, an LLM-judged metric that performs element-wise semantic membership scoring over these dimensions, aiming for interpretable, anomaly-specific assessment. The authors also curate FineW3, a structured dataset produced by automatically augmenting existing anomaly annotations with detailed entity attributes and scene information. Experiments report higher human alignment for FV-Score than n-gram and existing LLM-judge metrics, and show that current vision-language models struggle on fine-grained, temporally localized anomalies.

Summary of Review This work tackles a substantive evaluation gap in video anomaly understanding by shifting from lexical or fluency-oriented scoring to visually grounded, anomaly-specific element detection. The What–Who–Where decomposition is well-motivated and, coupled with FV-Score, yields interpretable diagnostics of model strengths and failures. The paper's core ideas are promising and supported by a human-correlation study and multi-model evaluation, but several technical and reporting issues limit rigor and reproducibility: notation errors in the scoring formulation, missing definitions for attribute scoring and normalization, evaluator–generator coupling in the human study, and under-specified dataset augmentation and long-video evaluation protocols. These issues are correctable and do not undermine the main contribution, but addressing them would materially strengthen the paper's validity and utility. With these fixes and broader comparisons to semantic/factuality-aware baselines, the benchmark can be a useful resource for the community.

Strengths

- Problem framing and contribution scope
  - The paper clearly articulates why n-gram overlap and generic LLM judges are ill-suited for anomaly evaluation, and grounds its design in human perception of anomalies via What–Who–Where. This is a well-motivated shift from language quality to visually grounded, anomaly-specific assessment.
  - The positioning relative to UCA (Yuan et al., 2024), CUVA/ECVA (Du et al., 2024), HAWK (Tang et al., 2024), and Holmes-VAU (Zhang et al., 2025) is apt: prior efforts often emphasize causal narratives or QA quality, but under-specify entity/scene grounding and rely on synthetic annotations that may hallucinate facts.
- Metric design and interpretability
  - FV-Score formalizes evaluation as element-wise semantic membership against G_{what}, G_{who}, and G_{where}, with simple scales (binary for Who/Where, ternary for What). This yields interpretable diagnostics (which elements were missed or wrong) and is more actionable than holistic, subjective scales common in LLM-based judges.
- Dataset structuring and coverage
  - FineW3 leverages a two-stage LVLM-assisted pipeline—event decomposition with entity linking and subsequent entity/scene augmentation—to produce structured labels. Reported statistics (thousands of videos, tens of thousands of events/entities/attributes) indicate challenging density and diversity, especially for long surveillance videos.
- Empirical evidence and insights
  - Human-correlation study: Across multiple agreement measures, FV-Score surpasses n-gram metrics and existing LLM judges, supporting the claim of improved human alignment.
  - Ablation on weights suggests human preferences align strongly with accurate entity identification (higher correlation when λ_{who} is larger), indicating value in emphasizing "Who" alongside "What."
  - Multi-model analysis reveals consistent patterns: stronger performance on static/coarse information (scene attributes, broad entity classes) and weak performance on fine-grained event detection and per-entity attributes, with higher accuracy when anomalies exhibit strong visual cues (e.g., explosions, arson). The "normalcy bias" observation—abnormal events being described as normal—is an important, actionable finding echoed by broader video evaluations (Fang et al., 2024; Liu et al., 2024; Pătrăucean et al., 2023).
- Relevance of comparisons
  - The paper benchmarks against both n-gram and contemporary LLM-judge baselines, including ECVA's AnomEval (Du et al., 2024) and VAU-EVAL (Zhu et al., 2025b), which is appropriate given the claims.

Weaknesses

- Metric formalization and notation errors
  - In Eq. (1), the membership function is written as $m_\theta(g_i, G_{dim})$, but it should depend on the generated report R (e.g., $m_\theta(g_i, R)$), consistent with the FineVAU-Judge description that uses the triplet R, G, P. As written, the equation obscures the dependence on R and conflicts with the judging mechanism.
  - The text states "two membership degrees" for What, yet the defined scale is ternary $(0, 1/2, 1)$. This is a presentation error; it should refer to two scales (binary vs. ternary), not "two

degrees."

- Formalism–results mismatch and missing definitions
  - Attributes are formally subsumed under G_{who}, but Table 4 reports separate "Entity" and "Attribute" columns. There is no explicit definition of an attribute-specific scoring function or a precise rule for partitioning entity versus attribute correctness within $J_{who}(R)$.
- Normalization and aggregation are under-specified
  - The paper reports percentages for Event/Entity/Attribute/Location and an "All" column, but does not define how $J_{dim}(R)$ is normalized (e.g., dividing by $|G_{dim}|$), how partial credit for What is handled in "accuracy," or whether "All" is a simple average or a weighted aggregation using λ's. The λ values used for the main tables are not specified alongside the results.
- Evaluator–generator coupling and limited human-study reporting
  - Gemini serves as both the generator of candidate reports (Gemini 2.5-flash) and the judge. This can introduce style/model-family alignment bias and inflate observed correlations. The study size is modest and results lack confidence intervals or hypothesis tests. There is also a procedural inconsistency: the paper states "8 human experts rank three video reports per video" but also that "each video is evaluated by three different experts."
- Dataset augmentation details and quality control
  - The LVLM(s) used for augmentation (model/version), input sampling (frames or clips, stride), decoding parameters, and any safety/consistency filters are not specified. Since Stage 1 explicitly "complements annotations by identifying unmentioned events or objects," there is a risk of injecting hallucinated content. No quantitative quality audit (e.g., human spot-check accuracy, inter-annotator agreement on augmented elements) is provided.
- Long-video evaluation protocol is under-specified
  - Many videos are long, yet the paper does not report frame/clip sampling strategies, context length limits per model, temporal stride, or coverage policy. These choices strongly influence event detection difficulty and may partially explain the low event scores.
- Baseline breadth for human-alignment claims
  - The comparisons omit common semantic/factuality-aware metrics such as BERTScore (Zhang et al., 2020), CLIPScore (Hessel et al., 2021), and robust rule-based LLM judging setups like AutoEval-Video (Chen et al., 2024). Given the paper's claim of superior human alignment, including such baselines would substantially strengthen the evidence.
- Minor clarity and consistency issues
  - Typos and naming inconsistencies that hinder replication: "Imms-eval" should be "LMMs-Eval" (Zhang et al., 2024a); "HIVAU" is better standardized as "Holmes-VAU"; inconsistent "FI" vs. "FI" abbreviations in Table 1; inconsistent "FineW3" vs. "FineW^3" naming; the text mentions "MovieChat" without a citation. Note that Video-ChatGPT (Maaz et al., 2023) is already cited in the references.
- Interpretation nuance on ablations
  - The paper emphasizes the centrality of the What dimension, yet Table 3 shows the strongest correlation with human judgment when λ_{who} is larger. The review agrees with the empirical observation (humans value entity correctness), but the paper's narrative should reconcile this with the stated centrality of What.

Suggestions for Improvement

- Fix and unify the formalism
  - Correct Eq. (1) to $J_{dim}(R) = \sum_{g_i \in G_{dim}} m_\theta(g_i, R)$, and unambiguously define m_θ as the judge-based membership scorer that consumes (g, R, P) and returns scores per the defined scales.
  - Explicitly define how attributes are scored. Either: (a) introduce a fourth dimension G_{attr} with its own $J_{attr}(R)$ and report it explicitly; or (b) clarify how "Entity" and "Attribute" are separated and aggregated within $J_{who}(R)$.
  - Specify normalization to convert $J_{dim}(R)$ to percentages, detail how partial credits for What are treated in "accuracy," define how "All" is computed (macro vs. micro average; with or without λ weights), and state the λ values used for each reported table.
- Strengthen evaluator independence and human-study rigor
  - Decouple the generator from the judge: include candidate reports from multiple LVLMs (e.g., InternVL, LLaVA variants, Qwen-VL) in the human-correlation study; and/or use multiple, diverse judges to assess whether FV-Score's correlation persists across judges.
  - Report confidence intervals and statistical tests for correlation differences. Clarify the exact annotation protocol: number of experts, per-video assignments, total judgments, and any adjudication or tie-breaking procedures.
- Improve dataset augmentation transparency and quality assurance
  - Document the augmentation LVLM(s), input sampling (frames per second, clip length, stride), decoding temperature/top-k/p, and any safety or consistency filters. Provide a quantitative audit: human spot-check accuracy for added events/entities/attributes, inter-annotator agreement on a held-out subset, and an estimate of false-additions when "complementing" events.
  - Include a complete, end-to-end annotation example showing the seed annotation, Stage 1 decomposition and entity linking, Stage 2 attribute/scene augmentation, and the final G_{what}, G_{who}, G_{where}.
- Elaborate long-video evaluation protocol
  - Detail the per-model input budget (number of frames/tokens), temporal sampling strategy (fps, stride), whether sliding windows or temporal samplers (cf. Holmes-VAU) are used, and how reports R are generated for long videos (single pass vs. segment aggregation). Discuss how these choices impact event recall and normalcy bias.
- Expand baseline coverage and error analysis
  - Add semantic/factuality-aware baselines: BERTScore (Zhang et al., 2020), CLIPScore (Hessel et al., 2021), and rule-based LLM judging (AutoEval-Video; Chen et al., 2024). Also consider reporting alignment against the factuality metric proposed by Liu and Wan (2023) for video captioning hallucinations.
  - Substantiate "normalcy bias" with controlled metrics (e.g., per-class precision/recall, confusion between abnormal vs. normal event types, and analysis of missed short-window anomalies).
  - Where feasible, include cross-dataset evaluation (e.g., a subset from Holmes-VAU) to examine whether FV-Score and the What–Who–Where schema remain robust on long, sparse anomaly scenarios.
- Clarify narrative around ablation findings
  - Reconcile the stated centrality of What with the empirical finding that up-weighting Who yields higher human correlation. One option is to present a rationale that entity correctness is a strong proxy for perceived report quality in anomaly scenarios, and to consider reporting both "event-centric" and "entity-centric" scores for different use cases.
- Presentation and consistency
  - Standardize names ("Holmes-VAU"), fix typos ("LMMs-Eval"), ensure consistent abbreviations across tables, and reconcile "FineW3" vs. "FineW^3." Add the missing citation for MovieChat where referenced alongside Video-ChatGPT.

References

- Bai, S., et al. (2025). Qwen2.5-VL technical report. arXiv:2502.13923.
- Banerjee, S., & Lavie, A. (2005). 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.
- Chen, X., Lin, Y., Zhang, Y., & Huang, W. (2024). AutoEval-Video: An automatic benchmark for assessing large vision-language models in open-ended video QA. In Computer Vision – ECCV 2024 (LNCS 15126, pp. 179–195). Springer.
- Dong, H., et al. (2024). Benchmarking and improving detail image caption. arXiv:2405.19092.
- Du, H., Nan, G., Qian, J., Wu, W., Deng, W., Mu, H., Chen, Z., Mao, P., Tao, X., & Liu, J. (2024). Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. arXiv:2412.07183.
- Du, H., Zhang, S., Xie, B., Nan, G., Zhang, J., Xu, J., Liu, H., Leng, S., Liu, J., Fan, H., et al. (2024). 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).
- Fang, X., Mao, K., Duan, H., Zhao, X., Li, Y., Lin, D., & Chen, K. (2024). MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. In Advances in Neural Information Processing Systems 37 (Datasets & Benchmarks Track).

- Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. arXiv preprint.
- Li, B., et al. (2024). LLaVA-OneVision: Easy visual task transfer. arXiv:2408.03326.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out.
- Liu, H., & Wan, X. (2023). Models see hallucinations: Evaluating the factuality in video captioning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 11807–11823). Association for Computational Linguistics.
- Liu, B., Qiao, P., Ma, M., Zhang, X., Tang, Y., Xu, P., Liu, K., & Yuan, T. (2025). SurveillanceVQA-589K: A benchmark for comprehensive surveillance video-language understanding with large models. arXiv:2505.12589.
- Liu, H., et al. (2023). Visual instruction tuning. In Advances in Neural Information Processing Systems.
- Liu, Y., Ma, Z., Qi, Z., Wu, Y., Shan, Y., & Chen, C.-W. (2024). E.T. Bench: Towards open-ended event-level video-language understanding. In Advances in Neural Information Processing Systems 37 (Datasets & Benchmarks Track).
- Maaz, M., et al. (2023). Video-ChatGPT: Towards detailed video understanding via large vision and language models. arXiv:2306.05424.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.
- Pătrăucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., ... Carreira, J. (2023). Perception Test: A diagnostic benchmark for multimodal video models. In Advances in Neural Information Processing Systems 36 (Datasets & Benchmarks Track).
- Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- Ye, M., Liu, W., & He, P. (2025). VERA: Explainable video anomaly detection via verbalized learning of vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8679–8688).
- Yuan, T., Zhang, X., & Liu, K. (2024). 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).
- Zhang, B., et al. (2025). VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. arXiv:2501.13106.
- Zhang, H., Xu, X., Wang, X., Zuo, J., Huang, X., Gao, C., Zhang, S., Yu, L., & Sang, N. (2025). 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).
- Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., & Liu, Z. (2024). LMMs-Eval: Reality check on the evaluation of large multimodal models. arXiv preprint.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations.
- Zhang, Y., et al. (2024). Video instruction tuning with synthetic data. arXiv:2410.02713.
- Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. (2025). InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479.
- Zhu, L., Chen, Q., Shen, X., & Cun, X. (2025). VAU-R1: Advancing video anomaly understanding via reinforcement fine-tuning. arXiv:2505.23504.

Add: **Ethics Chair Author Comment** **Author AI Review Evaluation**

Official Review of "FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding"

Official Review by Program Committee 55f7 21 Aug 2025 at 09:04 (modified: 07 Nov 2025 at 21:13)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee 55f7, Authors Revisions (/revisions?id=PT35TUcNvK)

Review:

UPDATE: The author choose not to rebuttal, so I will keep my score.

This paper introduces FineVAU, a new benchmark for Video Anomaly Understanding (VAU) aimed at addressing the shortcomings of existing evaluation methods. The authors propose a three-fold evaluation framework focusing on "What" (events), "Who" (entities), and "Where" (location). Central to this benchmark are two key contributions: 1) the FineW³ dataset, created through a fully automated LVLM-assisted pipeline that augments existing annotations, and 2) the FV-Score, a new metric that leverages an LLM-judge (FineVAU-Judge) to assess model-generated reports against the structured ground truth. The paper conducts experiments on several open-source LVLMs, revealing their significant weaknesses in understanding fine-grained temporal events despite competence in recognizing static scenes.

Strengths:

1. Important Problem: The paper correctly identifies a critical gap in VAU research. Existing metrics like n-gram-based scores are ill-suited for the free-form outputs of LVLMs, and current LLM-based evaluations often lack the necessary granularity for anomaly-specific understanding. The goal of creating a more human-aligned, interpretable, and fine-grained benchmark is timely and valuable.
2. Intuitive Framework: The decomposition of anomaly understanding into "What, Who, Where" is conceptually simple, interpretable, and aligns well with how humans might structure a description of an event. This structured approach provides more actionable feedback than a single holistic score.
3. Interesting Findings: The experimental results, particularly the stark contrast between models' ability to describe static location information (61.3% mean accuracy) and their profound failure to identify events (12.2% mean accuracy), is a significant finding. It highlights a critical disconnect in current LVLMs' capabilities and provides a clear direction for future research.

Weaknesses:

While the paper's direction is promising, the current manuscript suffers from several major issues related to methodological transparency, dataset validity, and the scope of the evaluation that undermine the confidence in its contributions.

1. Critical Questions on FineW³ Dataset Quality: The paper's cornerstone is the FineW³ dataset, yet its generation process raises significant red flags. The authors state it is created via a "structured and fully automatic procedure" using an LVLM. However, the well-known issue of LVLM "hallucination" is not adequately addressed.
   - While the FV-Score metric underwent a human correlation study, the paper provides no evidence of human verification for the FineW³ dataset annotations themselves. Without a human-in-the-loop validation or at least a thorough report on a sample-based audit (e.g., accuracy, recall, F1-scores against human experts), the quality of this foundational dataset remains questionable.
   - How can we be sure that the dataset is not riddled with systematic errors or biases from the backbone LVLM used for annotation? The paper does not specify which LVLM was used in this "fully automated" annotation pipeline, a critical detail for assessing potential bias.
2. Lack of Methodological Detail for Reproducibility: The paper omits a crucial detail about the experimental setup: how video data is processed and presented to the LVLMs.
   - Are the models analyzing the entire video, sampled clips, or a set of keyframes? If sampling is used, what is the strategy (e.g., sampling rate, uniform vs. importance-based sampling)?
   - The strategy for temporal data representation can impact an LVLM's performance, especially its ability to capture events that unfold over time and its susceptibility to hallucination. This lack of clarity hinders the reproducibility of the reported results.
3. Unaddressed Ambiguity from Source Data: The FineW³ dataset is derived entirely from existing datasets. These source datasets often contain categories with significant semantic overlap (e.g., `abuse` vs. `assault` vs. `fighting`; `stealing` vs. `shoplifting`). The paper does not discuss how this ambiguity in the source data might propagate into the fine-grained `What`, `Who`, and `Where` annotations. Does the automated pipeline attempt to disambiguate these cases, or does it risk creating noisy or inconsistent labels? The potential impact of this "inherited ambiguity" on the evaluation's reliability is a notable limitation.
4. Reproducibility of the FV-Score Judge: The authors use a frontier LLM (Gemini) as the FineVAU-Judge. While powerful, relying on a closed-source, API-based model for a core benchmark component harms long-term reproducibility. Have the authors considered the feasibility of using a powerful open-source LLM (e.g., Qwen) as the judge? A discussion on the trade-off between peak performance and community accessibility would strengthen the paper.

5. Insufficient Qualitative Analysis: The paper reports a strikingly low average accuracy of 12.2% on the "Event" dimension. This is a key result, yet the analysis remains superficial. The paper would be much stronger if it included a few qualitative examples of these failures. For instance, are the models completely silent about the anomaly, or are they misinterpreting it (e.g., describing a "robbery" as "customers browsing")? Such examples are vital for understanding the nature of the models' failures.
6. Limited Scope of Model Evaluation: For a paper proposing a new benchmark intended to push the frontiers of VAU, the evaluation is limited to a handful of open-source models. The exclusion of state-of-the-art closed-source models like GPT-4o, Claude 4 or Gemini is a missed opportunity. Including these models would provide a more complete picture of the current SOTA, better contextualize the performance of the open-source models, and more powerfully demonstrate the challenging nature of the FineVAU benchmark.

Questions:

1. Regarding FineW^3 Quality: Could you please elaborate on the quality control measures for the automated annotation pipeline? Was there any human verification conducted? If so, what were the resulting accuracy, recall, or F1 scores? If not, how do you assure the community of the dataset's quality against LVLM hallucinations? Which specific LVLM was used as the backbone for this pipeline?
2. Video Input Strategy: How are the video frames specifically processed and fed into the LVLMs for evaluation? Is it based on keyframe sampling, segments, or the full video? What was the sampling strategy, and have you analyzed how different strategies might affect performance?
3. Inherited Ambiguity: How does your framework handle the potential for semantic ambiguity inherited from the source datasets (e.g., the overlap between `assault` and `fighting`)? Do you believe this could affect the reliability of the FineW^3 annotations and the FV-Score?
4. FV-Score Judge: Did you consider using a high-performing open-source LLM as the FineVAU-Judge to improve reproducibility? What are your thoughts on the trade-off between using a frontier model like Gemini-flash versus a more accessible one?
5. Qualitative Failure Analysis: Could you provide some concrete, qualitative examples from your experiments that illustrate *why* LVLMs perform so poorly on the "Event" dimension?
6. Human Evaluator Disagreement: In your human evaluation study for FV-Score, the 32% of samples with disagreement is non-trivial. Did you analyze these cases? Do they concentrate on specific types of videos or anomalies that might suggest limitations of the "What, Who, Where" framework?
7. Model Scope: What was the reasoning for excluding leading closed-source models like ChatGPT, Claude, and Grok from your evaluation? Including them would seem essential for establishing a comprehensive SOTA on your new benchmark.

Rating: 5: Marginally below acceptance threshold
Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Add: **Ethics Chair Author Comment** **Author Review Evaluation**