Can We Defend Against AI-Generated Video Attacks on Real-World Crisis Events? A Systematic Evaluation of Detectors, Generators and Social Dissemination

Website: RA-Bench **Dataset:** Hugging Face **Code:** GitHub

Shuo Liang[*,1,2] • Yixing Ma[*,1,3] • Pengfei Zhou[*,§,1]
Zhenglin Wan[1] • Xingyan Chen[3] • Zihan Mei[5] • Manting Li[3] • Feihan Chen[6] • Zhiwen Wang[7] • Bin Xu[8]
Haotian Zhang[6] • Jiajun Song[6] • Shiya Su[9] • Run Liu[10] • Zhenghang Ni[2] • Yifa Yu[11]
Jintao Hong[2] • Bolong Feng[10] • Yifei Liu[3] • Zirui Zhang[12] • Jingxuan Zhang[1] • Songlin Zhao[14]
Yifan Bai[13] • Kang Tan[15] • Yizhe Liu[13] • Junhao Du[13] • Yongtao Ge[19] • Zhaopan Xv[16] • Xinyuan Zhang[16]
Mengru Ma[2] • Chunhua Shen[17] • Wei Wang[†,4] • Yang You[†,1] • Zheng Zhu[†,18] • Kaipeng Zhang[†,19] • Wangbo Zhao[†,4]

[*]Equal contribution [§]**Project lead:** Pengfei Zhou (zpf4wp@outlook.com)
[†]**Corresponding authors:** Wangbo Zhao (wangbo.zhao96@gmail.com) Kaipeng Zhang (kaipeng.zhang@shanda.com) Zheng Zhu (zhengzhu@ieee.org)
Yang You (yangyou@nus.edu.sg) Wei Wang (weiwa@cse.ust.hk)

[1]National University of Singapore • [2]Xidian University • [3]University of California, Berkeley • [4]The Hong Kong University of Science and Technology
[8]InfRec, Cardinal AI Lab • [7]Monash University • [6]Renmin University of China • [5]Arizona State University • [14]University of Illinois Urbana–Champaign
[11]Stanford University • [17]Zhejiang University • [15]Carnegie Mellon University • [12]ETH Zurich • [10]Shanghai Jiao Tong University
[13]Xi'an Jiaotong University • [18]GigaAI • [9]University of Wisconsin–Madison • [19]Alaya Lab • [16]Independent Researcher

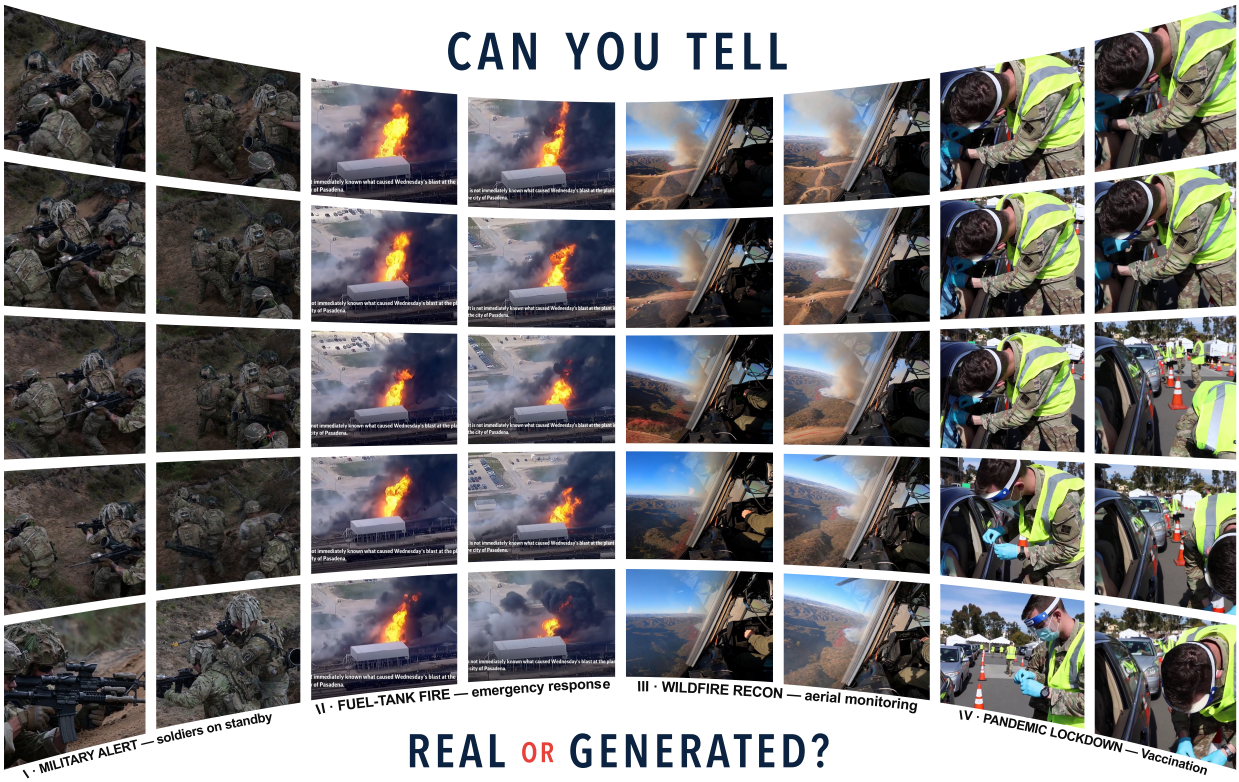


Figure 1 | Which is which? Real and AI-generated clips from four crisis scenarios are paired side by side; identify the synthetic example in each pair before consulting the answer key on the following page.

Answer key → page 2

Contents

1 Introduction **3**

2 Related Work **4**

3 RA-Bench **6**

3.1 Source Data Collection . . . 6
3.2 Automated Preprocessing . . . 7
3.3 Manual Review . . . 7
3.4 Retained Clip Postprocessing . . . 7
3.5 Paired Video Generation . . . 7

4 Experiments and Analysis **8**

4.1 How Well Do Current Detectors Generalize? . . . 9
4.1.1 Traditional Detectors Do Not Transfer Reliably to RA-Bench . . . 9
4.1.2 Zero-Shot Multimodal Models Remain Unreliable Across Prompts and Sources . . . 12
4.1.3 Fine-Tuned MLLMs Exhibit Protocol Dependence and Class Bias . . . 14
4.1.4 Summary of Detector Generalization . . . 15
4.2 What Makes Generated Videos Hard to Detect? . . . 16
4.2.1 Generation Quality Does Not Uniformly Determine Detection Difficulty . . . 16
4.2.2 Real-Image Conditioning Affects Detector Families Differently . . . 17
4.2.3 Detection Results Vary Little Across Sampling Seeds . . . 19
4.2.4 Summary of Generation Properties . . . 20
4.3 Human Perception and Social Dissemination . . . 20
4.3.1 Human Recognition Varies Sharply Across Generation Sources . . . 20
4.3.2 RA-Bench-HumanProof Remains Difficult Across Detector Families . . . 21
4.3.3 Social Dissemination Simulation Further Weakens Detection . . . 23
4.3.4 Summary of Human Perception and Social Dissemination . . . 24

5 Discussion **24**

6 Future Directions **25**

7 Conclusion & References **26**

A RA-Bench Construction and Data Documentation **32**

A.1 L2 Taxonomy and Source Distribution . . . 32
A.2 Source Inventory and Rights Basis . . . 33
A.3 Automated Preprocessing Details . . . 33
A.4 Manual Review Details . . . 34
A.5 Postprocessing Details . . . 34
A.6 Generation Details . . . 35
A.7 Qualitative Real-versus-Generated Examples . . . 40

B Detector Evaluation Protocols and Extended Results **41**

B.1 Detector Evaluation Protocol and Model Inputs . . . 41
B.2 Traditional-Detector Reference Transfer and Operating Points . . . 43
B.3 Zero-Shot Multimodal Model Protocol and Extended Analysis . . . 45
B.4 Fine-Tuned MLLM Protocol Sensitivity and Class-Conditional Behavior . . . 49

C Generation Factors and Robustness Analysis **51**

C.1 Generation Quality and Detectability . . . 51
C.2 Generation Settings and Detectability . . . 53
C.3 Stability Across Generation Seeds . . . 55

D Human Evaluation and Social Dissemination **57**

D.1 Human Evaluation Protocol and Source-Level Recognition . . . 57
D.2 RA-Bench-HumanProof Construction and Detector Evaluation . . . 58
D.3 RA-Bench-LastMile Protocol and Complete Results . . . 61

Figure 1 answer key (from page 1; left/right): I Generated/Real; II Real/Generated; III Real/Generated; IV Generated/Real.

Abstract
Recent video generators can fabricate realistic depictions of wars, disasters, public emergencies, and other real-world crises, creating substantial risks of misinformation. Existing benchmarks, however, provide limited evidence on detector and generator behavior in such settings, including how detectability varies with generation conditions, how people perceive generated videos, and whether detectors remain reliable during social dissemination. To address this gap, we introduce **RA-Bench**, a benchmark for AI-generated video detection that uses **R**eal videos as **A**nchors. RA-Bench contains 17,886 videos, comprising 1,830 real-video anchors across 10 social-risk categories and 16,056 generated clips from four open-source and five closed-source generators. Based on RA-Bench, we organize our evaluation along three dimensions. We first assess detector generalization across seven traditional detectors, ten zero-shot multimodal models under three review settings, and two MLLMs specifically fine-tuned on AI-generated video detection. Across these methods, none of the three detector families generalizes consistently across RA-Bench instances. We then examine how detectability varies with generation quality, conditioning information, and sampling seeds. These analyses show that generation properties affect detector families differently, while source-level detection patterns remain stable across seeds. Finally, we study human authenticity judgments and detector reliability during social dissemination. We find that videos that mislead people are also difficult for current detectors, and that social dissemination makes detection harder. Together, these findings show that current methods struggle to detect realistic AI-generated videos, highlighting the need for detectors robust to evolving video generators.

1. Introduction

Recent video generation models (Brooks et al., 2024; van den Oord and Roman, 2024; Team Seedance et al., 2026) can produce increasingly realistic clips with coherent appearance and motion. By reducing the time and effort required to create high-quality videos, these models can substantially improve video-production efficiency. However, these advances can also create new societal risks. For instance, generated videos depicting real-world crises, such as wars, disasters, and public-health emergencies, may deceive viewers and provoke public panic. The examples in Figure 1 highlight the challenge of distinguishing generated content from real videos.

In response to these risks, numerous benchmarks (Ma et al., 2025; Chen et al., 2024; He et al., 2024) have been developed to evaluate the ability of existing detectors (Wang et al., 2020; Ojha et al., 2023; Ma et al., 2025; Chen et al., 2024; Internò et al., 2025) to identify realistic AI-generated videos. Recent benchmarks have broadened generator coverage, increased dataset scale, and introduced more comprehensive evaluation protocols (Ma et al., 2025; Chen et al., 2024; Ni et al., 2026; Ma et al., 2026). Meanwhile, other studies have extended the task beyond binary classification by examining forensic explanations, world-simulation settings, and source backtracking (Wen et al., 2025a; Chen et al., 2025a; Liao et al., 2026).

However, as shown in Table 1, existing studies primarily focus on general video content, with limited systematic analysis of detector performance and generator behavior in the context of real-world crises and other important socially consequential events. Therefore, *it remains unclear whether current detectors are sufficiently reliable to protect society from the threats posed by increasingly realistic video generation models and potential misuses.*

To tackle this problem, we introduce **RA-Bench**, a benchmark for AI-generated video detection that uses **R**eal videos as **A**nchors. RA-Bench comprises 1,830 real-video anchors drawn from 675 publicly available source videos and spans 10 social-risk categories and 44 subcategories. To simulate a plausible misinformation scenario in which real image is used to generate subsequent events, we condition image-to-video (I2V) generators on the first frame of each anchor. Given the same first frame and text prompt, four open-source and five closed-source generators produce 16,056 generated clips, resulting in a total of 17,886 videos.

Based on RA-Bench, we perform a systematic evaluation along three dimensions: *1. Detector General-*

ization. We evaluate three detector families: seven traditional detectors, ten zero-shot multimodal model under three review settings, and two MLLMs specifically fine-tuned for AI-generated video detection. *2. Generation Properties.* We investigate how video detectability varies with generation quality, conditioning information, and sampling seeds. Specifically, we examine whether higher perceptual quality makes generated videos harder to detect, how the amount of real-image conditioning affects detector behavior, and whether source-level detection patterns remain stable across generation seeds. *3. Human Behavior.* We examine perceptual judgments of authenticity and detector behavior during social dissemination. We conduct a human study to identify generated videos that viewers find difficult to distinguish from real videos and specifically evaluate detectors on these human-deceptive cases, which constitute **RA-Bench-HumanProof**. We further introduce **RA-Bench-LastMile**, a social dissemination simulation for evaluating detector robustness.

Our experiments reveal three main findings:

- **None of the three detector families generalizes consistently across RA-Bench sources.** Traditional detectors fall from public-reference AUCs of 67.6–98.6% to source-level means of 43.9–57.3% on RA-Bench, and their rankings change across generators. Scaling zero-shot multimodal models does not remove their sensitivity to prompts and generation sources. Replacing timestamps with frame indices reduces Skyra to 54.4–54.9% mean BAcc, while BusterX++ achieves only 4.1–9.1% FakeR (Section 4.1).
- **Generation properties affect detector families differently, with limited variation across seeds.** A 50-point Condition Fidelity increase is associated with a 14.4-point decrease in Gemini Diagnostic FakeR, whereas dynamic content increases Gemini's evidence for the generated class but leaves the traditional-detector mean nearly unchanged. Across T2V, first-frame I2V, and first+last-frame I2V, the seven-detector mean AUC changes from 33.4% to 50.9% and 44.6%, whereas the average FakeR across fine-tuned MLLMs decreases from 70.5% to 42.7% and 28.3%. (Section 4.2).
- **Videos that mislead people are also difficult for current detectors, and social dissemination weakens detection further.** Reviewers identify 68.6% of open-source videos as generated, but only 52.9% of closed-source videos, with Seedance2.0 and Kling falling to 40.7% and 45.1%. We use 633 generated videos that are labeled *Real* by all five reviewers to form **RA-Bench-HumanProof**. On this subset, Gemini Binary and Diagnostic reach only 54.7% and 54.5% BAcc, while the seven traditional detectors average 47.5% AUC. Separately, the Full condition of the social dissemination simulation reduces mean FakeR across the five fine-tuned configurations from 46.0% to 1.4% (Section 4.3).

In conclusion, our findings demonstrate that current methods still cannot reliably detect realistic AI-generated videos. This limitation is especially consequential when generated content depicts real-world crises, highlighting the need for detection systems that remain effective as generation models evolve and videos circulate in the real world.

2. Related Work

Benchmarks for AI-Generated Video Detection. As video generation models become more realistic, benchmarking whether existing detectors can accurately identify AI-generated videos has become increasingly important. Early efforts, such as GVF (Ma et al., 2025) and GenVideo (Chen et al., 2024), pair real videos with generated counterparts and evaluate cross-generator transfer. Subsequent benchmarks further expand the scale, generator coverage, and evaluation protocols, including GenVidBench (Ni et al., 2026) and AIGVDBench (Ma et al., 2026). Recent benchmarks further extend evaluation beyond binary classification to MLLM-based explanation (Wen et al., 2025a),

Benchmark	Real-event grounding	Social-risk taxonomy	Real-event-conditioned I2V	Human-deceptive challenge set	Social dissemination simulation
GVF (Ma et al., 2025)	✗	✗	✗	✗	✗
GenVideo (Chen et al., 2024)	✗	✗	✗	✗	○
GenVidBench (Ni et al., 2026)	✗	✗	○	✗	✗
GenBuster-Bench (Wen et al., 2025a)	✗	✗	✗	✗	○
GenWorld (Chen et al., 2025a)	✗	✗	○	✗	✗
AIGVDBench (Ma et al., 2026)	✗	✗	✗	✗	✗
Chameleon (Liao et al., 2026)	○	✗	○	✗	○
RA-Bench	✓	✓	✓	✓	✓

Table 1 | Comparison of benchmark designs for AI-generated video detection. We compare how existing benchmarks handle real-event sources, social-risk categories, real-event-conditioned generation, human deception, and social dissemination simulations during dataset construction and evaluation. RA-Bench also includes recent open- and closed-source video generators and representative detector families.
Criteria. A checkmark, circle, and cross denote full, partial, and absent coverage, respectively. Full coverage requires traceable footage from identifiable socially consequential events, a hierarchical social-risk taxonomy, I2V conditioned on a matched event-source frame, a challenge set selected through blind human authenticity judgments, and a social dissemination simulation. For the final column, partial coverage denotes isolated social dissemination operations rather than a complete social dissemination simulation. Dataset scale, generator coverage, explanations, and source attribution are outside scope.

world-simulation settings (Chen et al., 2025a), and source backtracking (Liao et al., 2026). Despite this progress, existing benchmarks remain limited in their coverage of misuse scenarios involving real-world crises and other socially consequential events. To address this gap, we introduce RA-Bench and use it to systematically assess the robustness and limitations of existing detectors in these settings.

Realistic video generation models. In recent years, video generation models have rapidly evolved from an emerging research topic into a technology with significant real-world and societal impact. Improving video realism has long been a central objective in this field, from early approaches extended from image generation models, such as Stable Video Diffusion (Blattmann et al., 2023) and VideoCrafter (Chen et al., 2023), to recent advanced models (Bar-Tal et al., 2024; Polyak et al., 2024; Kong et al., 2024; Brooks et al., 2024; van den Oord and Roman, 2024; Team Wan, 2025; HaCohen et al., 2025; Team Seedance et al., 2026; Cloudflare, 2026), which have achieved substantial improvements in narrative coherence and visual consistency. However, these advances in generation quality have raised concerns about potential misuse and negative impacts on society. This motivates us to investigate whether existing detection methods can reliably identify increasingly realistic AI-generated videos, which can further push forward responsible video generation methods.

Detectors for AI-generated videos. Inspired by the detector taxonomy used in AIGVDBench (Ma et al., 2026), we broadly divide existing AI-generated video detectors into traditional discriminative approaches and MLLM-based approaches. Traditional approaches include video classification models, generated-image detection models, and generated-video detection models. Video classification models (Feichtenhofer et al., 2019; Bertasius et al., 2021; Liu et al., 2022; Tong et al., 2022) directly treat detection as a binary video classification problem. Generated-image detectors (Wang et al., 2020; Ojha et al., 2023; Tan et al., 2024; Chen et al., 2025b) mainly focus on image-level artifacts, while generated-video detectors (Ma et al., 2025; He et al., 2024; Chen et al., 2024; Internò et al., 2025) further capture temporal evidence across frames. Beyond these traditional approaches, MLLM-based detectors (Wen et al., 2025a; Li et al., 2025) have gained increasing attention because they can provide detailed explanations beyond binary classification results. In this work, we evaluate representative detector families on real-world crises and other socially consequential events, and derive insights to guide their future design.

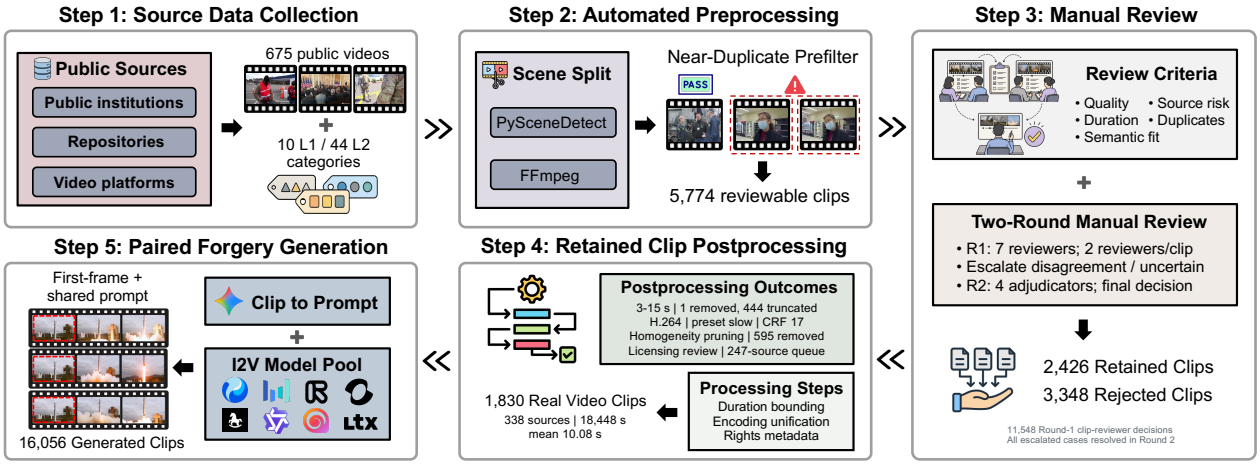


Figure 2 | Overview of the five-stage RA-Bench construction pipeline. We collect public videos from real-world crises and other socially consequential events, segment them into scene-level clips and screen near-duplicates, conduct two-round manual review, and apply release-oriented postprocessing to form the real anchor set. Each retained anchor is then paired with image-to-video (I2V) generated counterparts using first-frame conditioning and a shared prompting pipeline.

3. RA-Bench

Unlike prior benchmarks (Ma et al., 2026; Wen et al., 2025a), which collect seed video data from general-purpose web video corpora such as OpenVid-1M (Nan et al., 2025), RA-Bench focuses on real videos from real-world crises and other socially consequential events, where forged videos can pose public risk. We first collect public source videos and organize them into social-risk categories (Section 3.1). Then, each source video is segmented into scene-level clips and screened for near-duplicates (Section 3.2). The resulting clips are filtered through two rounds of human review (Section 3.3), after which postprocessing standardizes the retained clips into the final anchor set (Section 3.4). Each real clip is then paired with image-to-video generated clips that form the generated side of the benchmark (Section 3.5). The resulting RA-Bench consists of 1,830 real-video clips and 16,056 generated clips. An overview of the construction pipeline is shown in Figure 2, and the benchmark composition is summarized in Figure 3.

3.1. Source Data Collection

To prepare source data for constructing the benchmark, we first collect 675 videos depicting real-world crises and other socially consequential events from public platforms. We then group these videos into 10 social-risk categories (L1) and 44 subcategories (L2). The L1 categories are weather and natural disasters, war and armed conflict, politics and governance, public safety, accidents and infrastructure failures, economic and social panic, public health, technology, space and exploration, and large public events. Details of the L2 categories are provided in Appendix A.1. We do not impose a uniform per-category quota, since equal-size quotas would distort the natural prevalence of social-risk scenes and move the sample away from realistic deployment conditions. The L1 taxonomy and the released category distribution are shown in Figure 3(a,b).

Among the collected video sources, government and public-institution accounts form the largest group, while open-license repositories (*e.g.* Wikimedia Commons) and public video platforms (*e.g.* YouTube) account for most of the remaining sources. Appendix A.2 lists the sources used for data collection. Since licensing conditions vary across sources, we record the redistribution rights for each

source and use this information in the rights check described in Section 3.4.

3.2. Automated Preprocessing

We segment each source video into scene-level clips for later review and image-to-video generation. Specifically, we detect scene cuts with PySceneDetect[1] and split the videos at these cuts using FFmpeg[2], producing 5,774 reviewable clips from 675 source videos. Since adjacent scenes in a source video tend to share similar framing and content, we run a near-duplicate prefilter on the clip pool. We adopt a ResNet-18 (He et al., 2016) pretrained on ImageNet-1K (Russakovsky et al., 2015) to encode each clip and compare cosine similarities between neighboring clips from the same source video. A pair is flagged when both frame- and clip-level similarities exceed the specified thresholds, and flagged pairs are shown to reviewers as duplicate warnings during manual review (Section 3.3). More details of the preprocessing are provided in Appendix A.3.

3.3. Manual Review

We conduct a two-round manual review of the 5,774 clips obtained after preprocessing. In Round 1, seven volunteers review the clips using predefined evaluation criteria for visual quality, semantic fit to the assigned subcategory, duration suitability, duplicate content, and source-related risks. Each clip is independently assessed by two reviewers, who assign one of three actions: *retain*, *reject*, or *uncertain*. Clips with reviewer disagreement and those marked as uncertain are routed to Round 2. In Round 2, another four volunteers serve as adjudicators and make the final decision for each routed clip. After this process, 2,426 clips are retained as the standard sample pool and 3,348 are rejected. Further details of the manual review process are provided in Appendix A.4.

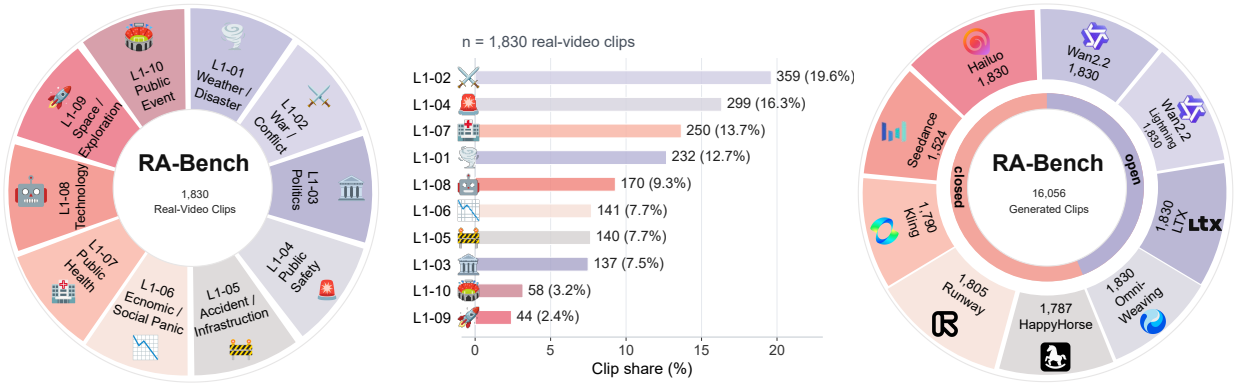
3.4. Retained Clip Postprocessing

We postprocess the 2,426 retained clips to obtain the final real video set, using duration bounding, encoding unification, homogeneity pruning, and licensing review. Specifically, we first bound each clip to a 3–15 seconds window, discarding clips shorter than 3 seconds and truncating longer ones, thereby maintaining a moderate clip duration that balances sufficient information and acceptable cost for detection models. We then re-encode every clip with H.264 and later apply the same encoding to the generated clips. This reduces the risk that detectors rely on codec configurations as label cues. Subsequently, we remove redundant clips extracted from each source video to reduce near-duplicates from the same event. Finally, we review source-level licensing conditions and record the resulting redistribution rights as metadata for each clip. After these operations, the final set contains 1,830 clips from 338 unique source videos, totaling 18,448 s (about 5.1 hours) with a mean duration of 10.08 s. The category composition of the released set is shown in Figure 3(b). Further postprocessing details are provided in Appendix A.5.

3.5. Paired Video Generation

We pair each real clip with counterparts generated by image-to-video (I2V) models using a unified generation pipeline, as shown in Figure 4. Under this design, each real clip and its generated counterparts share the same scene semantics and are grounded in real-world crises and other socially consequential events.

[1] https://www.scenedetect.com/docs/latest/
[2] https://ffmpeg.org/ffmpeg.html



(a) Real-event taxonomy **(b) Category distribution** **(c) Generation sources**

Figure 3 | Overview of RA-Bench. (a) The L1 real-event taxonomy used to organize the 1,830 real-video clips into 10 broad social-risk domains. The finer-grained L2 taxonomy is provided in Appendix A.1. (b) The clip distribution across the L1 domains. (c) The image-to-video generation sources used to construct the generated-video set, grouped into open-source models and closed-source API providers; numbers indicate the generated clips included from each source.

We first caption each real clip with Gemini-3.1-Pro-Preview (Google, 2026), following structured video-captioning practice (Ju et al., 2024; Luo et al., 2025), and convert the caption into a prompt shared across all generators (Appendix A.6). We then condition each generator on the obtained caption and the first frame of the real clip. This design yields visually plausible generated clips, increases benchmark difficulty, and reflects a common social media manipulation scenario in which a still image from a real-world crisis or another socially consequential event is used to fabricate a video that may pose public risk. Finally, we set the generated video length proportional to the duration of the corresponding real clip while constraining it to a 2–8 s range. Each generator maps this target to a supported duration setting so that every generated clip remains within the range supported by most existing video generation models, as detailed in Appendix A.6. This dynamic duration setting helps assess whether detection models learn duration-based shortcuts.

After generation, we re-encode every generated clip using the same H.264 codec as the real clips. We preserve each generator's native resolution and frame rate, and remove audio from all clips. We generate clips with four open-source and five closed-source generators, as shown in Figure 3(c). The four open-source generators cover all 1,830 anchors. We submit the same anchors to each closed-source provider, but provider-specific content-safety filters reject some requests, resulting in smaller paired subsets. In total, RA-Bench contains 16,056 generated clips.

4. Experiments and Analysis

We evaluate current methods for detecting AI-generated video on RA-Bench and analyze the factors associated with detection difficulty. Our experiments cover all four open-source and five closed-source generators in RA-Bench; the generators and paired clip counts are summarized in Figure 3(c). Each open-source generator is evaluated on 1,830 generated clips and their matched real anchors. For closed-source generators, provider-side moderation can reduce coverage, so each is evaluated on the returned generated clips and their corresponding real anchors. To examine sensitivity to generation duration, the detector-side experiments in Section 4.1 additionally include a fixed-duration Wan2.2 variant as a control for dynamic-duration generation. We mark this auxiliary control with an asterisk and exclude it from all benchmark-level averages. For continuous fake scores, we report paired AUC and TPR@5%FPR; for discrete decisions, we report balanced accuracy (BAcc), macro-F1, and fake

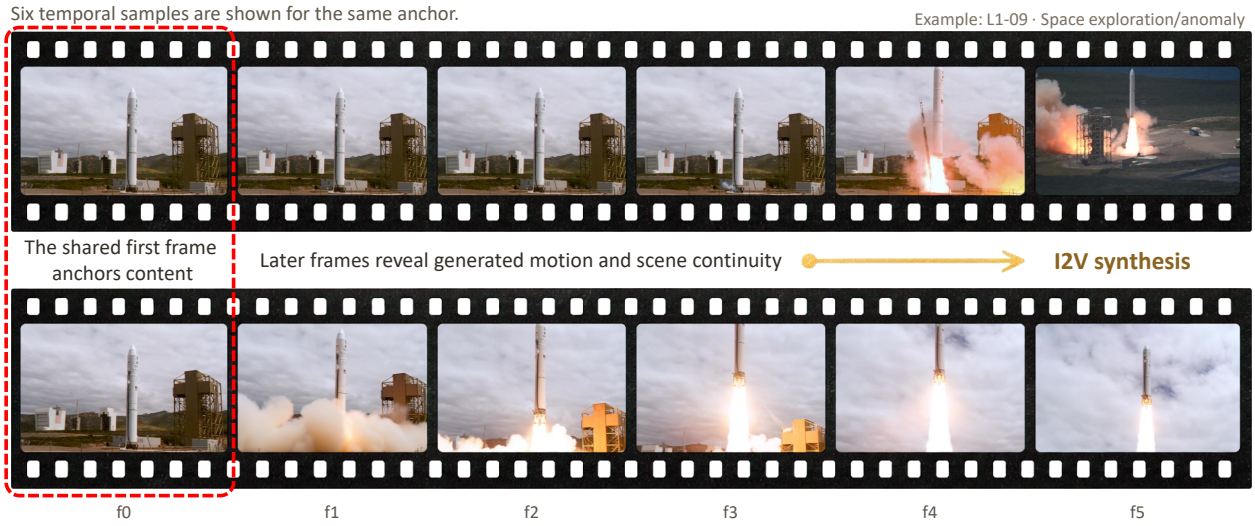


Figure 4 | A real anchor and its generated counterpart for a representative RA-Bench crisis event. The top filmstrip shows the real space-launch anchor, and the bottom shows the generated video, each represented by six uniformly sampled frames (`f0`–`f5`). The red dashed box marks the shared first frame used to condition the generator. The generated video continues from this frame, preserving much of the scene layout, capture style, and lighting while synthesizing new motion as the rocket leaves the pad. This example uses the dynamic-duration Wan2.2 setting. Additional examples across crisis categories and generators are provided in Figures 12 and 13.

recall (FakeR). Our analysis has three parts. We first evaluate whether traditional detectors, zero-shot multimodal models, and MLLMs fine-tuned for AI-generated video detection generalize across generation sources (Section 4.1). We then examine how generation quality, conditioning information, and sampling seeds relate to detectability (Section 4.2). Finally, we compare detector behavior with human authenticity judgments and evaluate detector robustness during social dissemination (Section 4.3). Evaluation implementation details are summarized in Appendix B.1; prompt templates and complete source-wise results are provided in the corresponding appendices.

4.1. How Well Do Current Detectors Generalize?

4.1.1. Traditional Detectors Do Not Transfer Reliably to RA-Bench

Table 2 reports results for seven traditional detectors across the nine RA-Bench generation sources and the fixed-duration Wan2.2 control. For context, the public-reference column reports AIGVDBench LTX-I2V results for six detectors (Ma et al., 2026) and the VidProM result for ReStraV (Internò et al., 2025); these AUCs range from 67.6% to 98.6%. On RA-Bench, the seven-detector mean falls to 50.9–57.3% across the open-source generators and 43.9–54.0% across the closed-source generators. For 26 of the 63 detector–source pairs, AUC is below 50%, indicating that generated videos receive lower fake scores than real anchors in the same paired evaluation subset more often than the reverse.

Public-reference rankings also fail to transfer to RA-Bench. Among the six detectors evaluated on the same AIGVDBench LTX-I2V reference, the public ranking has a Spearman correlation of only 0.26 with the ranking by mean AUC across the nine RA-Bench sources. UnivFD moves from second to sixth, whereas NPR moves from sixth to third. The leading detector also changes across generators: DeCoF and ForgeLens each rank first on four sources, while ReStraV ranks first on Kling. The loss is therefore not a uniform decrease from the public references; it changes which detector appears strongest.

Table 2 | Traditional detector performance across the nine RA-Bench generation sources We report paired AUC and TPR at 5% FPR (T@5%) in %. T@5% is generated-video recall at an operating point with 5% FPR on the matched real videos. The public-reference column reports AIGVDBench LTX-I2V AUC for six detectors and VidProM AUC for ReStraV; these are contextual references rather than matched-domain baselines. Wan2.2 fixed* is an auxiliary control and is excluded from the Open average. Spearman correlations use only the six detectors that share the AIGVDBench reference.

		Public	Open-source settings					Open	Closed-source generators					Closed
Detector	Metric	High ref.	Wan2.2 dyn.	Wan2.2 fixed*	Wan2.2 Light.	LTX	Omni Weav.	avg.	Happy Horse	Runway	Kling	Seedance 2.0	Hailuo	avg.
Image / frame-level detectors														
CNNSpot	AUC	81.4	35.5	34.6	43.8	64.0	45.8	47.3	39.2	31.9	45.4	40.4	53.9	42.2
	T@5%	–	2.3	1.7	3.0	11.7	3.5	5.1	2.8	1.6	3.9	2.8	5.8	3.4
NPR	AUC	67.6	50.6	49.9	55.2	58.6	50.9	53.8	54.6	52.3	44.3	48.7	54.6	50.9
	T@5%	–	3.8	4.0	4.6	4.5	5.6	4.6	4.5	3.4	3.2	3.9	5.7	4.2
UnivFD	AUC	89.9	32.7	33.4	43.0	51.2	39.9	41.7	43.0	38.4	39.1	35.0	44.0	39.9
	T@5%	–	0.4	0.5	2.1	0.9	0.5	1.0	1.7	0.5	1.1	0.6	1.3	1.1
ForgeLens	AUC	92.9	60.0	61.0	**69.6**	**64.1**	63.0	**64.2**	**64.9**	**59.5**	59.2	48.1	61.7	58.7
	T@5%	–	19.0	18.7	25.2	11.5	17.6	18.3	18.7	8.0	9.6	3.1	8.9	9.7
Video / temporal-level detectors														
DeCoF	AUC	81.5	**62.5**	**63.2**	56.7	62.5	**72.1**	63.4	58.5	55.2	57.2	**60.6**	**63.2**	**58.9**
	T@5%	–	3.9	4.7	2.8	2.3	6.7	3.9	1.8	1.8	0.5	1.0	4.8	2.0
D3	AUC	77.7	55.4	55.7	49.1	58.0	52.9	53.8	28.0	24.4	28.9	50.4	44.0	35.1
	T@5%	–	6.4	5.5	2.1	6.2	3.3	4.5	6.3	1.8	6.2	10.6	1.9	5.3
ReStraV	AUC	98.6	59.3	51.2	62.4	42.5	68.3	58.1	55.7	45.8	**67.4**	49.3	56.4	54.9
	T@5%	–	8.5	6.7	10.1	4.6	15.0	9.6	8.6	2.7	0.0	6.0	8.0	5.0
7-det. mean AUC		84.2	50.9	49.9	54.2	57.3	56.1	54.6	49.1	43.9	48.8	47.5	54.0	48.7
7-det. mean T@5%		–	6.3	6.0	7.1	6.0	7.5	6.7	6.4	2.8	3.5	4.0	5.2	4.4
Spearman vs ref.		–	0.14	0.14	0.20	0.31	0.14	0.20	0.54	0.54	0.54	−0.37	0.31	0.31

The failure pattern also differs across detectors and sources. D3 obtains 53.8% mean AUC across the open-source generators but only 35.1% across the closed-source generators. CNNSpot and UnivFD fall below 50% AUC on seven and eight of the nine RA-Bench sources, respectively, despite public-reference AUCs of 81.4% and 89.9%. Performance remains weak when false positives are constrained: at 5% FPR, the seven-detector mean identifies only 6.0–7.5% of generated videos from the open-source generators and 2.8–6.4% from the closed-source generators. Figure 5 visualizes the gaps from the public references and the crossings among source profiles. Reference alignment, rank transfer, and additional operating points at 1% FPR and 95% TPR are reported in Appendix B.2.

Temporal reallocation yields only modest gains. One possible explanation for the weak results is that sparse uniform sampling misses brief local inconsistencies. We test this explanation for five sparse-frame detectors by comparing Uniform-8 with Global–Local-8 under the same eight-frame budget. Global–Local-8 combines four uniformly spaced frames with four consecutive frames centered on the strongest temporal-change response; detector weights, visual preprocessing, and score aggregation remain fixed. The routing procedure is described in Appendix B.1.

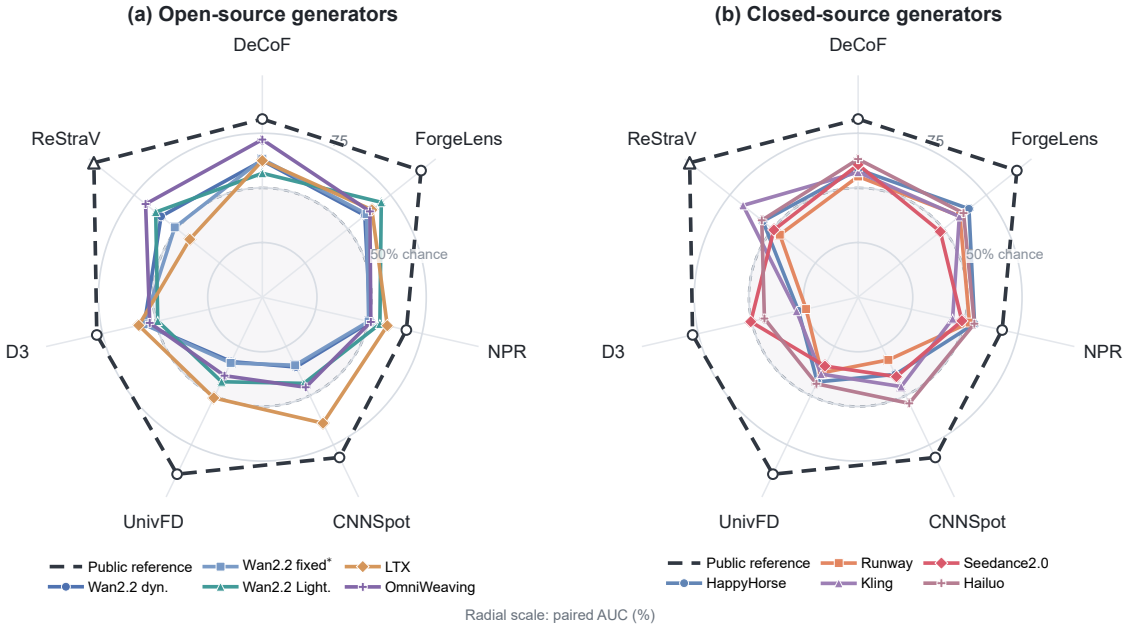


Figure 5 | Source-specific AUC profiles for the seven traditional detectors. Each spoke represents one detector, and each polygon represents one generation setting. The outer dashed contour shows the public-reference result for each detector (AIGVDBench LTX-I2V for six detectors and VidProM for ReStraV), while the faint inner ring marks 50% AUC. (a) The four open-source generators in RA-Bench and the fixed-duration Wan2.2 control. (b) The five closed-source generators in RA-Bench.

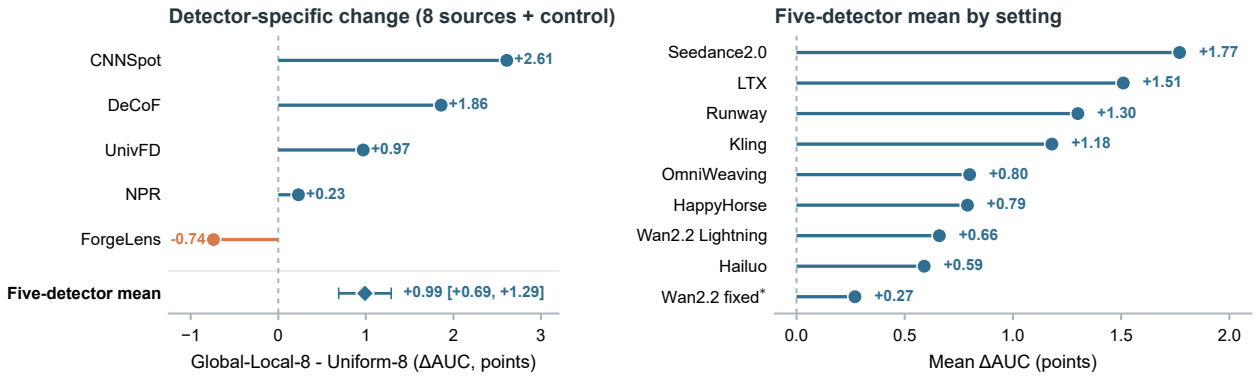


Figure 6 | Effect of temporal allocation under a fixed eight-frame budget. (a) AUC change from Uniform-8 to Global–Local-8 for each of five sparse-frame detectors, averaged over eight fixed-length RA-Bench sources and the fixed-duration Wan2.2 control; the diamond and interval show the five-detector mean and source-block bootstrap 95% confidence interval. (b) The corresponding five-detector mean for each evaluated setting. The ablation excludes dynamic-duration Wan2.2. Positive values favor Global–Local-8.

Across the eight fixed-length RA-Bench sources and the fixed-duration Wan2.2 control included in this ablation, Global–Local-8 increases mean AUC by only 0.99 points over Uniform-8 (source-block bootstrap 95% CI, [0.69, 1.29]; Figure 6). Four of the five detectors improve, with the largest gains for CNNSpot (+2.61 points) and DeCoF (+1.86 points), while ForgeLens decreases by 0.74 points. The detector-averaged change is positive for all nine evaluated settings but ranges from only +0.27 to +1.77 points. Reallocating the same frames recovers some local temporal evidence, but it does not

change the central result: high public-reference AUC neither translates into reliable detection nor identifies a consistently strong detector across RA-Bench sources.

Takeaway. Traditional detector AUC falls from 67.6–98.6% in public references to source-level means of 43.9–57.3% on RA-Bench, with different detectors leading on different generators.

4.1.2. Zero-Shot Multimodal Models Remain Unreliable Across Prompts and Sources

Table 3 evaluates representative zero-shot multimodal models under Binary, Diagnostic, and Rating prompts; their output formats and complete templates are provided in Appendix B.3. Overall detection performance remains limited. Qwen3.5-27B obtains 53.2/46.7 BAcc/macro-F1 under Binary and 51.3/37.2 under Diagnostic, while Qwen3.5-122B-A10B reaches 53.0/45.7 and 54.8/50.8, respectively. Gemini-3.1-Pro-Preview is the strongest evaluated model, reaching 63.4/62.9 under Binary, 63.1/62.5 under Diagnostic, and 63.6/63.0 AUC/verdict macro-F1 under Rating. Its Binary BAcc nevertheless ranges from 54.3 on Seedance2.0 to 74.5 on LTX. No evaluated model therefore combines strong overall performance with stable behavior across prompt formats and RA-Bench sources.

Table 3 | Source-specific zero-shot multimodal model results on the nine RA-Bench generation sources. Binary and Diagnostic cells report BAcc/macro-F1, while Rating cells report paired AUC/verdict macro-F1 (top/bottom). Values are in %. Open and Closed are source-equal averages over the four open-source and five closed-source RA-Bench generators, respectively; Wan2.2 fixed* is excluded from the Open average.

		Open-source settings					Open	Closed-source generators					Closed
Model	**Prompt**	**Wan2.2 dyn.**	**Wan2.2 fixed***	**Wan2.2 Light.**	**LTX**	**Omni Weav.**	**avg.**	**Happy Horse**	**Runway**	**Kling**	**Seedance 2.0**	**Hailuo**	**avg.**
Discrete classification prompts (BAcc/macro-F1)													
Qwen3.5-27B	Binary	51.3 43.8	51.2 43.6	55.3 49.9	59.7 56.2	54.8 49.2	55.3 49.8	51.6 44.2	55.3 50.0	50.0 41.4	49.8 41.7	51.4 43.9	51.6 44.2
	Diagnostic	50.9 36.5	50.7 36.1	51.8 38.4	53.3 41.2	51.7 38.2	51.9 38.6	50.8 36.4	50.9 36.5	50.5 35.6	50.2 35.0	51.1 36.9	50.7 36.1
Qwen3.5-122B-A10B	Binary	51.6 43.5	50.7 41.9	52.7 45.3	58.3 53.8	54.1 47.5	54.2 47.5	51.8 43.8	50.2 41.1	53.4 46.4	52.4 44.9	52.3 44.6	52.0 44.2
	Diagnostic	54.1 50.2	53.9 49.8	55.7 52.3	65.6 64.7	60.2 58.2	58.9 56.3	51.6 46.5	51.7 46.6	52.1 47.1	49.7 44.6	52.2 47.5	51.5 46.4
Qwen3.7-Plus (thinking)	Binary	53.7 45.8	53.6 45.7	55.5 48.8	64.6 61.8	54.8 47.7	57.2 51.0	53.7 45.9	54.9 47.8	53.9 46.2	51.2 41.5	52.9 44.5	53.3 45.2
	Diagnostic	54.2 47.7	53.6 46.7	55.7 50.0	64.5 62.1	57.2 52.2	57.9 53.0	53.4 46.4	54.0 47.5	54.2 47.6	51.9 43.9	52.9 45.6	53.3 46.2
Gemini-3.1-Pro-Preview	Binary	63.4 63.1	65.8 65.7	61.6 61.1	74.5 74.5	66.6 66.5	66.5 66.3	62.5 62.1	67.8 67.7	60.1 59.5	54.3 52.6	59.9 59.2	60.9 60.2
	Diagnostic	63.8 63.4	63.7 63.3	61.8 61.2	74.5 74.5	65.1 64.8	66.3 65.9	61.5 60.9	67.5 67.3	60.2 59.4	53.8 51.6	60.2 59.4	60.6 59.7
GPT-5.5	Binary	50.8 35.7	51.4 37.0	51.5 37.1	56.9 47.6	50.9 35.9	52.5 39.1	51.7 37.5	51.6 37.5	51.2 36.5	50.6 35.2	51.1 36.3	51.2 36.6
	Diagnostic	50.7 35.1	51.0 35.9	51.1 36.0	54.7 43.2	50.7 35.2	51.8 37.4	51.0 35.9	50.7 35.3	51.0 35.8	50.4 34.7	50.8 35.5	50.8 35.4
Continuous rating prompt with explicit verdict (AUC/macro-F1)													
Qwen3.5-27B	Rating	52.1 34.5	52.0 34.3	53.4 35.4	53.4 37.6	50.1 34.6	52.3 35.5	50.9 34.8	53.0 35.0	49.6 34.4	47.7 33.7	50.4 35.0	50.3 34.6
Qwen3.5-122B-A10B	Rating	56.3 55.8	56.0 55.6	57.2 57.2	65.8 64.3	60.4 59.5	59.9 59.2	53.2 53.4	57.1 56.4	54.2 54.2	45.4 46.8	52.7 52.8	52.5 52.7
Qwen3.7-Plus (thinking)	Rating	57.8 48.5	57.3 47.7	59.6 51.6	69.3 62.7	63.8 50.8	62.6 53.4	55.1 46.9	56.6 50.7	58.6 49.5	51.0 43.3	55.6 46.0	55.4 47.3
Gemini-3.1-Pro-Preview	Rating	62.5 64.8	63.6 65.7	62.2 61.2	77.2 75.8	64.5 65.2	66.6 66.7	63.0 60.8	67.7 67.7	59.8 60.1	54.4 51.2	61.4 60.3	61.2 60.0
GPT-5.5	Rating	59.3 35.5	60.2 36.0	61.6 36.2	69.7 45.1	66.6 35.2	64.3 38.0	60.8 36.8	64.5 36.6	61.4 36.3	53.8 34.5	61.2 35.7	60.3 36.0

Changing the prompt format can reverse a model's class preference even when the visual input and decision task remain fixed. Qwen3.5-0.8B increases from 19.7% FakeR under Binary to 97.8% under Diagnostic and 100.0% under Rating. Qwen3.5-4B moves from 85.2% to 5.5% and 21.9%. Gemini's FakeR varies by only 2.7 points and remains at least 51.8% across the three prompts. GPT-5.5 also changes little across prompts, but its FakeR remains 2.9–4.4% while RealR exceeds 99.2%. Prompt stability can therefore reflect a persistent class preference rather than reliable detection (Figure 15).

Requiring structured outputs does not consistently improve detection. Under Rating, Gemini-3.1-Pro-Preview obtains 63.6/63.0 paired AUC/verdict macro-F1, and Qwen3.5-122B-A10B obtains 55.8/55.6. GPT-5.5 instead reaches 62.1/36.9. Its continuous score ranks generated videos above their matched real anchors better than chance, but its explicit verdict predicts almost every input as *Real*. The Diagnostic aspect scores can also become uninformative. Qwen3.5-0.8B and Qwen3.5-9B repeat one value across all five aspects in 99.7% and 99.8% of complete outputs, respectively. Gemini repeats one value across all five aspects in 55.2%. Even for Qwen3.5-122B-A10B, the five aspect AUCs remain 54.5–54.8%, while its Diagnostic verdict reaches 54.8% BAcc. Structured prompting therefore provides neither consistently stronger class separation nor reliable aspect-level evidence. Full class-conditional, rating, and diagnostic analyses are provided in Appendix B.3.

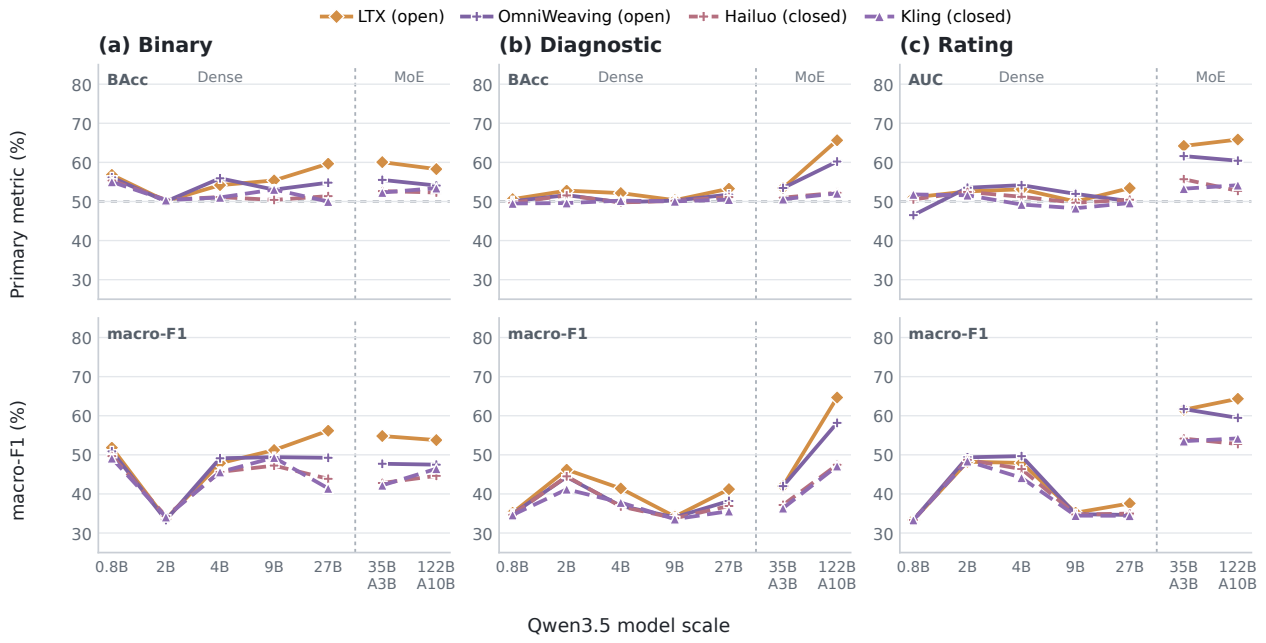


Figure 7 | Qwen3.5 scaling across prompts and generation sources. The four columns correspond to representative RA-Bench generators. The top row reports BAcc for Binary and Diagnostic and paired AUC for Rating; the bottom row reports macro-F1 from each prompt's explicit Overall Verdict. Horizontal dashed lines in the top row mark the 50% reference level. The vertical dotted line separates dense and MoE variants, whose trajectories are drawn separately. Solid and dashed lines denote open-source and closed-source generators, respectively.

Increasing Qwen3.5 scale does not produce consistent gains across prompts and generators. From 0.8B to 122B-A10B, Rating AUC rises from 51.0 to 65.8 on LTX but only from 50.5 to 52.7 on Hailuo. Over the same endpoints, macro-F1 changes from 49.7 to 44.6 for Binary on Hailuo, 35.2 to 64.7 for Diagnostic on LTX, and 33.3 to 64.3 for Rating on LTX. At 122B-A10B, the Binary BAcc, Diagnostic BAcc, and Rating AUC values for Hailuo and Kling remain within 4.2 points of the 50% reference, whereas LTX reaches 65.6% BAcc under Diagnostic and 65.8% AUC under Rating. Scaling therefore

improves selected source–prompt pairs but does not produce a reliably accurate detector across RA-Bench.

Takeaway. Scaling zero-shot multimodal models does not remove their sensitivity to prompt format or generation source.

4.1.3. Fine-Tuned MLLMs Exhibit Protocol Dependence and Class Bias

Table 4 | Fine-tuned MLLMs across the nine RA-Bench generation sources. Skyra prompts receive the same 16 frames; BusterX++ (Wen et al., 2025b) uses its released pipeline. Each source is evaluated on matched real–generated pairs, with BAcc, FakeR, and macro-F1 reported in %. The RealR values beside each model name are measured on the full set of 1,830 real anchors and are listed in official-timestamp/frame-index order for Skyra. Open and Closed are source-equal averages over the four open-source and five closed-source RA-Bench generators, respectively; Wan2.2 fixed* is excluded from Open.

		Open-source settings					Open	Closed-source generators					Closed
Prompt	**Metric**	**Wan2.2 dyn.**	**Wan2.2 fixed***	**Wan2.2 Light.**	**LTX**	**Omni Weav.**	**avg.**	**Happy Horse**	**Runway**	**Kling**	**Seed ance2.0**	**Hailuo**	**avg.**
Skyra-SFT (RealR: 87.4 / 55.4)													
Official timestamp	**BAcc**	60.6	91.9	59.9	74.7	51.9	61.8	80.2	80.3	80.1	73.5	54.9	73.8
	FakeR	33.8	96.4	32.5	61.9	16.4	36.1	73.3	73.2	72.8	59.6	22.5	60.3
	macro-F1	57.5	91.9	56.7	74.2	45.0	58.4	80.1	80.2	80.0	73.0	49.6	72.6
Frame index	**BAcc**	60.2	59.1	58.7	31.1	50.9	50.2	59.4	61.0	59.1	52.8	56.0	57.7
	FakeR	65.1	62.8	62.0	6.7	46.3	45.0	63.9	66.6	62.5	52.4	56.6	60.4
	macro-F1	60.1	59.0	58.7	26.7	50.8	49.1	59.3	60.9	59.1	52.8	56.0	57.6
Skyra-RL (RealR: 84.0 / 49.5)													
Official timestamp	**BAcc**	61.1	90.8	60.7	76.9	52.1	62.7	81.2	81.1	81.4	74.8	56.1	74.9
	FakeR	38.2	97.7	37.3	69.8	20.2	41.4	78.7	78.3	78.9	65.9	28.3	66.0
	macro-F1	58.9	90.8	58.4	76.8	46.7	60.2	81.2	81.1	81.4	74.6	52.4	74.1
Frame index	**BAcc**	60.0	59.0	58.6	29.7	53.0	50.3	59.7	61.3	60.6	54.5	57.1	58.6
	FakeR	70.5	68.4	67.7	9.9	56.4	51.1	70.5	73.2	71.5	61.5	64.6	68.3
	macro-F1	59.6	58.6	58.3	26.8	52.9	49.4	59.2	60.7	60.1	54.3	56.9	58.2
BusterX++ (RealR: 93.7)													
Released pipeline	**BAcc**	49.8	50.2	51.4	48.9	49.9	50.0	50.1	50.2	49.7	49.7	49.9	49.9
	FakeR	6.0	6.7	9.1	4.1	6.1	6.3	6.5	6.8	5.6	6.4	6.1	6.3
	macro-F1	37.9	38.6	40.8	36.1	38.0	38.2	38.4	38.6	37.6	38.1	38.0	38.1

Under the official-timestamp prompt, Skyra-SFT and Skyra-RL reach source-equal mean BAcc values of 68.5% and 69.5%, respectively, on RA-Bench (Table 4). These means do not indicate consistent cross-source detection: FakeR ranges from 16.4% to 73.3% for Skyra-SFT and from 20.2% to 78.9% for Skyra-RL. BusterX++ exhibits a different failure mode. Its FakeR remains between 4.1% and 9.1%, while its macro-F1 remains between 36.1% and 40.8%. Together with its 93.0–93.8% real recall, this pattern indicates a strong preference for the *Real* class rather than reliable recognition of generated videos. The evaluated fine-tuned systems therefore exhibit distinct weaknesses: source-sensitive detection for Skyra and a strong class bias for BusterX++.

The fixed-duration control reveals an additional anomaly in Skyra's official-timestamp results. Wan2.2 dynamic and its fixed-duration control use the same generator, anchors, prompts, and first-frame conditioning, with duration policy as the controlled difference. Fixing the duration at approximately 5 seconds increases FakeR from 33.8% to 96.4% for Skyra-SFT and from 38.2% to 97.7% for Skyra-RL. A related pattern appears within RA-Bench. Across the eight RA-Bench generation sources that include clips ending at exactly 5.00 seconds and clips with other final timestamps, the former receive

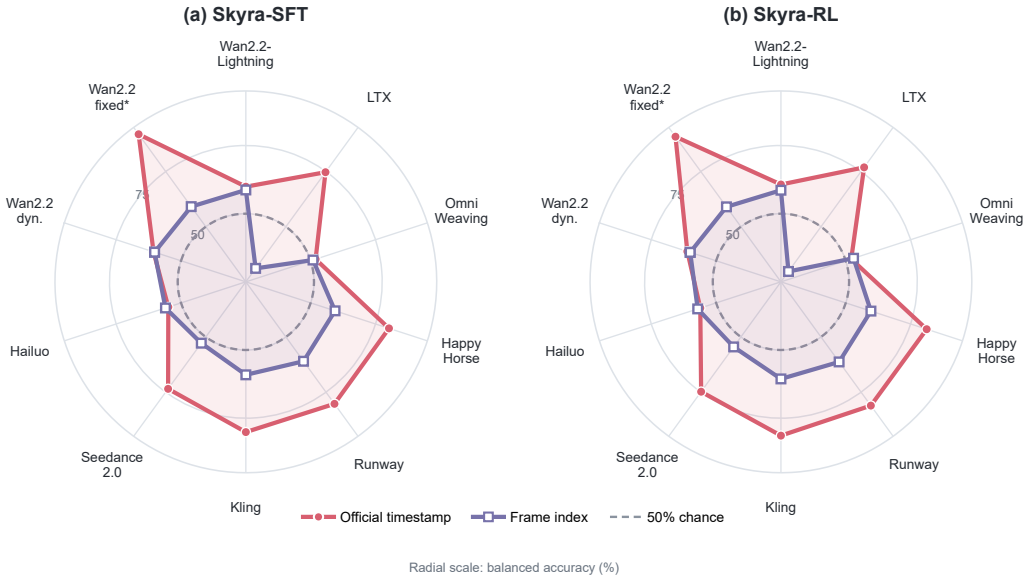


Figure 8 | Skyra performance changes with temporal-tag representation. The official-timestamp and frame-index prompts receive the same 16 frames. Each radar reports per-source BAcc across the nine RA-Bench generators and Wan2.2 fixed*, an auxiliary control excluded from benchmark-level averages; the dashed ring marks 50%.

substantially higher FakeR; the source-equal gaps are 48.6 and 43.4 points, respectively. These comparisons reveal a strong association between the displayed temporal grid and Skyra's predictions, but they do not determine whether the difference arises from duration-dependent visual content or from the temporal labels themselves.

To distinguish these explanations, we replace absolute timestamps with frame indices while preserving the same 16 visual frames and their order. This intervention reduces the exact-5-second FakeR gap from 48.6 to 1.5 points for Skyra-SFT and from 43.4 to 1.4 points for Skyra-RL. Mean BAcc also falls from 68.5% to 54.4% and from 69.5% to 54.9%, respectively. Because the visual input is unchanged, the collapse of the temporal-grid gap identifies a strong sensitivity to temporal-label representation. An audit of the released ViF metadata (Li et al., 2025) further shows that an exact-5-second final timestamp is correlated with the class label, providing a plausible source of this protocol prior (Appendix B.4).

Removing timestamps does not recover stable content-based detection. Under frame indices, per-source BAcc still spans 31.1–61.0% for Skyra-SFT and 29.7–61.3% for Skyra-RL (Figure 8). Temporal labels therefore explain an important component of Skyra's behavior, but not all variation across generation sources. RL fine-tuning removes neither the protocol dependence nor the remaining source-specific failures.

Takeaway. For the fine-tuned MLLMs, replacing timestamps with frame indices lowers Skyra to 54.4–54.9% mean BAcc, while BusterX++ reaches only 4.1–9.1% FakeR.

4.1.4. Summary of Detector Generalization

Across all three detector families, performance fails to transfer consistently across RA-Bench sources. Traditional detector AUC falls from 67.6–98.6% in public references to source-level means of 43.9–57.3% on RA-Bench, with different detectors leading on different generators. Scaling parameters of zero-shot multimodal models does not remove their sensitivity to prompt format or generation source.

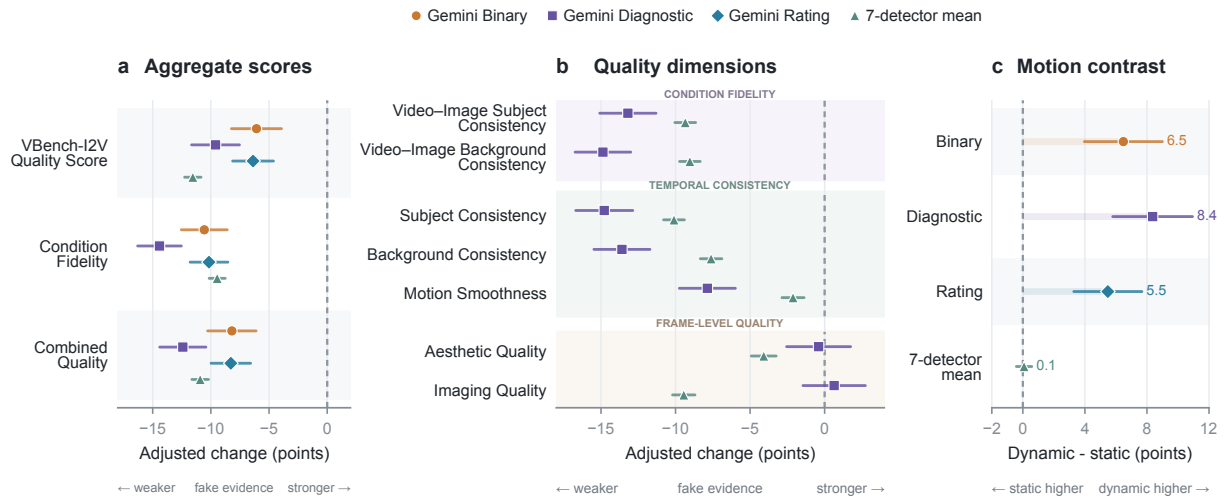


Figure 9 | Generation quality and fake-side detectability on RA-Bench. (a) Associations of the VBench-I2V Quality Score, Condition Fidelity, and Combined Quality with three Gemini-3.1-Pro-Preview outputs and the mean fake-score percentile of seven traditional detectors. (b) Dimension-level associations for Gemini Diagnostic and the traditional-detector mean; complete prompt-wise results are reported in Appendix C.1. (c) Dynamic-minus-static contrasts after adjustment for generation source and matched real-video anchor. Panels (a) and (b) report adjusted changes associated with an interquartile increase within each source and Dynamic Degree group. Negative values indicate weaker fake evidence at higher quality, while positive values in panel (c) indicate stronger fake evidence for dynamic clips. Traditional scores are percentile-normalized within detector and source before averaging. Error bars denote 95% confidence intervals based on standard errors clustered by real-video anchor.

For the fine-tuned MLLMs, replacing timestamps with frame indices lowers Skyra to 54.4–54.9% mean BAcc, while BusterX++ reaches only 4.1–9.1% FakeR. Public benchmark performance, larger model scale, and task-specific fine-tuning therefore do not by themselves ensure generalization across generation sources.

4.2. What Makes Generated Videos Hard to Detect?

4.2.1. Generation Quality Does Not Uniformly Determine Detection Difficulty

We evaluate all 16,056 generated clips in RA-Bench using the released VBench++ I2V evaluators (Huang et al., 2025). We report three aggregate scores: the official VBench-I2V Quality Score, which combines six dimensions of temporal consistency, motion, and frame-wise quality; Condition Fidelity, defined as the mean of the normalized Video–Image Subject and Background Consistency scores; and Combined Quality, their equal-weight mean. We omit Camera Motion because RA-Bench prompts do not specify the controlled camera-motion labels required by that evaluator. To separate clip-level quality variation from differences among generation sources, we estimate associations within each source. We further stratify clips by Dynamic Degree because low-motion clips can receive high temporal-consistency scores simply by changing little. For each continuous quality score, we report the adjusted change in fake-side detector output associated with an interquartile increase, giving each source equal weight. These estimates describe associations within RA-Bench rather than causal effects. Figure 9 presents the results, with full scoring and statistical details provided in Appendix C.1.

Across the three Gemini outputs and the traditional-detector mean in Figure 9(a), higher values

of all three aggregate scores are associated with weaker fake evidence. For Gemini Diagnostic, an interquartile increase in the VBench-I2V Quality Score, Condition Fidelity, and Combined Quality is associated with decreases of 9.6, 14.4, and 12.4 percentage points in fake recall, respectively. Condition Fidelity also has the largest negative association for Gemini Binary and Rating. The traditional-detector mean decreases by 9.5–11.5 percentile points across the three scores, with its largest change associated with the VBench-I2V Quality Score. No single aggregate score therefore characterizes detection difficulty consistently across detector families.

The dimension-level results explain this difference. For Gemini Diagnostic, higher Video–Image Subject and Background Consistency and higher within-video Subject and Background Consistency are each associated with 13.2–14.9 percentage-point decreases in fake recall. Motion Smoothness shows a smaller decrease of 7.9 percentage points, while Aesthetic Quality and Imaging Quality show no clear association with the Diagnostic verdict. The traditional-detector mean follows a different pattern: Subject Consistency and Imaging Quality show the largest decreases, at 10.1 and 9.4 percentile points, respectively, while Motion Smoothness changes it by only −2.1 points. Gemini Diagnostic is thus most closely associated with Condition Fidelity and temporal consistency, whereas the traditional-detector mean also varies strongly with frame-level Imaging Quality.

Dynamic Degree further shows why generation quality cannot be treated as a single axis. After adjustment for generation source and matched real-video anchor, dynamic clips receive 5.5–8.4 points more fake evidence than static clips across the three Gemini outputs. The traditional-detector mean changes by only 0.1 percentile points, and its 95% confidence interval spans zero. Because Dynamic Degree contributes positively to the VBench-I2V Quality Score, this association partly offsets the negative associations of its consistency dimensions and helps explain why the Quality Score has a weaker negative association with Gemini than Condition Fidelity. Taken together, Condition Fidelity and temporal consistency are associated with weaker fake evidence, whereas dynamic clips are associated with stronger Gemini fake evidence and no clear change in the traditional-detector mean. Detection difficulty therefore depends on both the quality dimension and the detector family being evaluated.

Takeaway. Within sources, stronger Condition Fidelity and temporal consistency are associated with weaker evidence for the generated class, whereas dynamic content strengthens Gemini's fake evidence but leaves the traditional-detector mean nearly unchanged.

4.2.2. Real-Image Conditioning Affects Detector Families Differently

RA-Bench conditions each generated video on the first frame of its matched real-video anchor. To examine detector behavior under different amounts of real-image conditioning, we use Wan2.2 to generate the same 1,830 anchor-derived prompts under three settings: T2V, first-frame I2V, and first+last-frame I2V. T2V receives only the prompt, whereas the two I2V settings additionally receive the matched first frame or the matched first and last frames. We report seed-0 results for the seven traditional detectors and five settings of MLLMs fine-tuned for AI-generated video detection: official-timestamp and frame-index prompts for Skyra-SFT and Skyra-RL, and the released BusterX++ pipeline. Table 5 reports the absolute results, while Figure 10 shows the changes between adjacent generation settings. Complete classification metrics and cross-seed results are provided in Appendix C.2.

From T2V to first-frame I2V, the seven-detector mean AUC increases from 33.4% to 50.9%, with higher AUC for five of the seven traditional detectors. Over the same comparison, the mean FakeR across the five fine-tuned MLLM settings decreases from 70.5% to 42.7%, and every setting shows a decrease of 21.0–32.2 points. The two detector families therefore respond in opposite directions to first-frame conditioning.

Table 5 | Detection under three Wan2.2 generation settings. We compare T2V, first-frame I2V (the RA-Bench setting), and first+last-frame I2V on the same 1,830 anchor-derived prompts at seed 0. Traditional detectors report AUC with T@5% (TPR at 5% FPR) in gray below; Skyra-SFT, Skyra-RL, and BusterX++ report FakeR. All values are percentages. BusterX++ abstentions are counted as incorrect.

Detector / configuration	Metric	T2V	First frame (I2V; RA-Bench)	First+last frames (I2V)
Traditional detectors				
CNNSpot	AUC	15.9	35.5	40.5
	T@5%	0.2	2.4	3.1
NPR	AUC	30.7	50.6	44.4
	T@5%	0.9	3.8	3.0
UnivFD	AUC	16.3	32.7	34.5
	T@5%	0.1	0.4	0.6
ForgeLens	AUC	24.2	60.0	44.8
	T@5%	1.0	19.0	6.8
DeCoF	AUC	66.6	62.5	68.5
	T@5%	1.5	3.9	6.8
D3	AUC	20.2	55.4	29.8
	T@5%	2.0	6.4	3.5
ReStraV	AUC	59.6	59.3	50.0
	T@5%	9.8	8.5	4.0
7-detector mean	AUC	**33.4**	**50.9**	**44.6**
	T@5%	**2.2**	**6.3**	**4.0**
Skyra-SFT				
Official timestamp	FakeR	63.2	33.8	16.4
Frame index	FakeR	94.9	65.1	45.8
Skyra-RL				
Official timestamp	FakeR	70.4	38.2	20.7
Frame index	FakeR	97.2	70.5	54.9
BusterX++				
Released pipeline	FakeR	27.0	6.0	4.0
Mean across fine-tuned MLLM settings	**FakeR**	**70.5**	**42.7**	**28.3**

From first-frame to first+last-frame I2V, FakeR decreases by 15.6–19.3 points across the four Skyra settings, and the mean across all five fine-tuned MLLM settings falls from 42.7% to 28.3%. BusterX++ FakeR decreases from 6.0% to 4.0%; its smaller change reflects the already low FakeR rather than stable detection. Both Skyra prompt formats follow the same direction, showing that the decline is not specific to the temporal-label format.

Traditional detectors do not show the same monotonic pattern. Moving from first-frame to first+last-frame I2V increases AUC for CNNSpot, UnivFD, and DeCoF, but decreases it for NPR, ForgeLens, D3, and ReStraV. Their mean AUC falls from 50.9% to 44.6%, and mean T@5% falls from 6.3% to 4.0%, while detector-specific AUC changes range from −25.6 to +6.0 points. For Skyra under the official-timestamp prompts and for BusterX++, the FakeR decrease from first-frame to first+last-frame I2V persists across all three seeds (Appendix C.2). Generation settings therefore do not impose a shared difficulty ordering across detector families: more real-image conditioning is associated with progressively lower FakeR for the fine-tuned MLLMs, whereas traditional-detector AUC changes in detector-specific directions.

Takeaway. Across T2V, first-frame I2V, and first+last-frame I2V, the mean FakeR of the evaluated MLLMs fine-tuned for AI-generated video detection falls from 70.5% to 42.7% and 28.3%, while traditional-detector AUCs move in both directions.

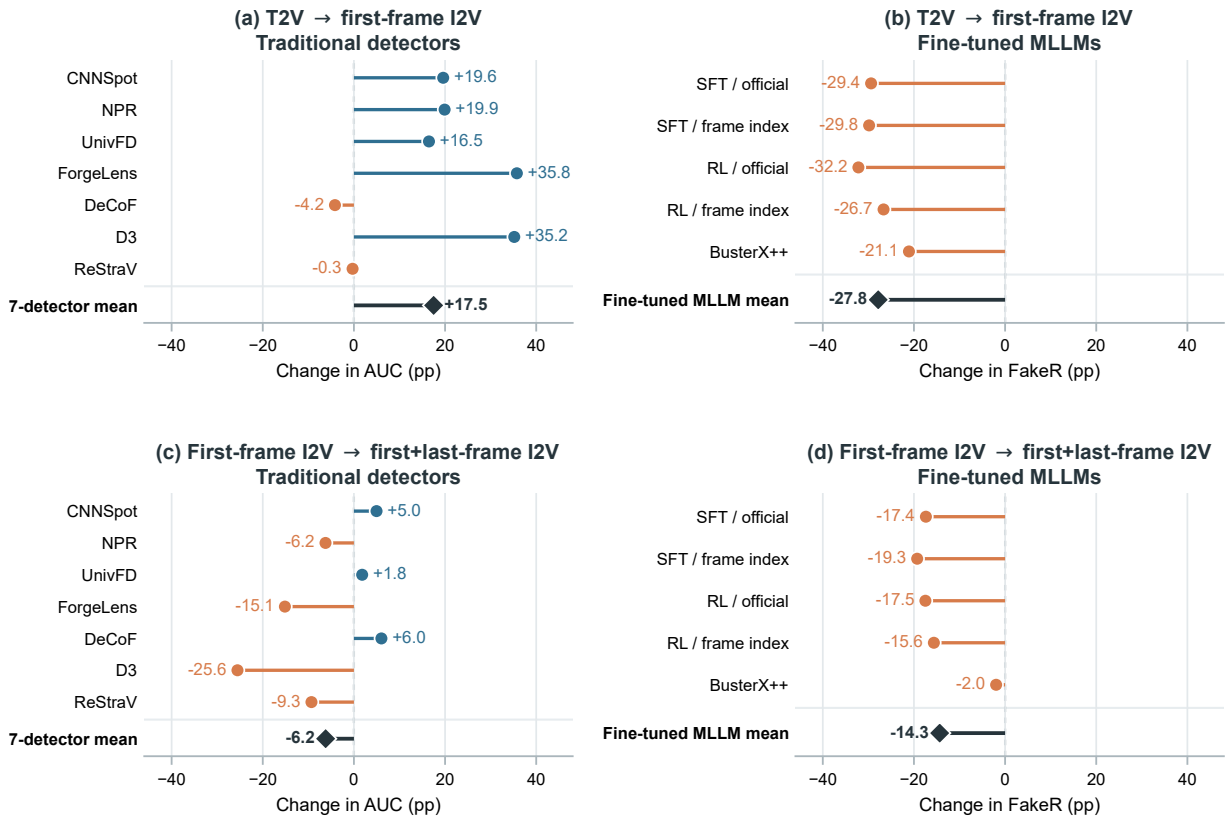


Figure 10 | Changes in detectability across Wan2.2 generation settings. Each panel reports the second setting minus the first setting named in its title. Panels (a) and (c) show paired AUC for the seven traditional detectors, while panels (b) and (d) show FakeR for the five fine-tuned MLLM settings. Positive values indicate higher AUC or FakeR in the second setting; negative values indicate lower values. Diamonds denote detector-family means.

4.2.3. Detection Results Vary Little Across Sampling Seeds

Sampling can change a video's appearance and motion even when its prompt and conditioning image are fixed. All primary results for the four open-source RA-Bench generators use seed 0. To determine whether the detector–source patterns depend on this choice, we generate two additional realizations, with seeds 42 and 123, from the same 1,830 anchors per generator. We hold the prompts, conditioning images, and all other generation settings fixed. Closed-source providers are omitted because their APIs do not expose a controllable seed, and the fixed-duration Wan2.2 control is omitted because this analysis concerns the four open-source RA-Bench generators. Table 6 summarizes the comparison.

Table 6 | Detection results remain stable across generation seeds. Results for seeds 0, 42, and 123 on the same 1,830 anchors per generator: (a) source-level mean AUC over seven traditional detectors; (b) source-wise FakeR and BAcc ranges for the fine-tuned MLLMs. Skyra-SFT and Skyra-RL use the released official-timestamp prompt, and BusterX++ uses its released evaluation pipeline.

(a) Traditional-detector AUC

Seed	Wan2.2 dynamic	Wan2.2 Lightning	LTX	OmniWeaving
0	50.85	54.24	57.26	56.12
42	50.39	54.49	57.77	56.59
123	50.72	54.47	56.72	56.12
Δ	0.46	0.25	**1.05**	0.47

Δ is the maximum minus the minimum across seeds; all AUC values are in %. Bold marks the largest Δ in each panel. Full results are in Appendix C.3.

Traditional detectors vary little at the source level. The largest range among the source-level seven-detector AUC means is 1.05 points, for LTX. Seed 0 differs from the three-seed averages by at most 0.20 AUC points. Although individual detector–source cells vary more, the 28-cell AUC pattern remains highly consistent (pairwise Spearman 0.978–0.989). Only three cells cross 50% AUC, and all remain close to random ranking under every seed (Appendix C.3). Sampling therefore changes some local estimates without altering the broader detector–source pattern.

(b) Fine-tuned MLLM FakeR

	Skyra-SFT					Skyra-RL					BusterX++				
Source	**0**	**42**	**123**	**ΔF**	**ΔB**	**0**	**42**	**123**	**ΔF**	**ΔB**	**0**	**42**	**123**	**ΔF**	**ΔB**
Wan2.2 dynamic	33.77	32.84	31.42	2.35	1.17	38.20	39.45	37.65	1.80	0.90	5.96	6.61	6.34	0.66	0.33
Wan2.2-Lightning	32.46	30.78	30.11	2.35	1.17	37.32	36.50	34.86	**2.46**	**1.23**	9.07	9.18	8.58	0.60	0.30
LTX	61.91	62.90	63.01	1.09	0.55	69.84	70.44	70.60	0.77	0.38	4.10	3.88	3.61	0.49	0.25
OmniWeaving	16.45	17.32	16.07	1.26	0.63	20.22	20.66	20.49	0.44	0.22	6.12	6.83	6.39	0.71	0.36

Δ is the maximum minus the minimum across seeds; F and B denote FakeR and BAcc, respectively; all metrics are in %. The largest Δ within each panel is bold. Detector-specific results and protocol details are in Appendix C.3.

The fine-tuned MLLMs show similarly limited variation. The largest FakeR/BAcc ranges are 2.35/1.17 points for Skyra-SFT, 2.46/1.23 for Skyra-RL, and 0.71/0.36 for BusterX++. Because the real-video control is shared, BAcc changes are driven entirely by FakeR. Skyra's source differences persist under every seed, while BusterX++ remains below 10% FakeR for every source–seed combination. Neither pattern is therefore specific to the seed-0 realizations used in the primary benchmark.

Takeaway. Across the three tested seeds, the source-level seven-detector mean AUC and fine-tuned MLLM FakeR vary only slightly.

4.2.4. Summary of Generation Properties

Generation quality and conditioning affect detector families differently. Within sources, stronger Condition Fidelity and temporal consistency are associated with weaker evidence for the generated class, whereas dynamic content strengthens Gemini's fake evidence but leaves the traditional-detector mean nearly unchanged. Across T2V, first-frame I2V, and first+last-frame I2V, the mean FakeR of the evaluated MLLMs fine-tuned for AI-generated video detection falls from 70.5% to 42.7% and 28.3%, while traditional-detector AUCs move in both directions. Across the three tested seeds, the source-level seven-detector mean AUC and fine-tuned MLLM FakeR vary only slightly.

4.3. Human Perception and Social Dissemination

4.3.1. Human Recognition Varies Sharply Across Generation Sources

Whether generated videos can mislead viewers is central to their societal risk, but detector performance alone does not measure human recognition. We therefore conduct a source-unaware human evaluation on RA-Bench. Twenty reviewers inspect videos in reviewer-specific randomized orders and choose among *Real*, *Uncertain*, and *Generated*, with real and generated videos from different sources

interleaved. Each video in the primary analysis receives three independent judgments. Reviewers label 60.3% of generated-video judgments as *Generated*. Real videos are recognized more often, with 71.9% of judgments labeled *Real*; however, 22.8% are labeled *Generated* and 5.3% *Uncertain*. Real videos depicting crisis events can therefore also be mistaken for generated content.

Table 7 | Human recognition varies sharply by generation source. Stage 1 response shares are in %. For generated sources, *Generated* is human FakeR. The *Generated total* row pools judgments, whereas Open/Closed averages weight generation sources equally.

Video source	Real	Unc.	Generated
Real videos	**71.9**	5.3	22.8
Generated total	33.9	5.8	**60.3**
Open-source generators			
Wan2.2 dynamic	27.0	5.5	67.5
Wan2.2 Lightning	28.5	4.9	66.6
LTX	30.3	5.1	64.6
OmniWeaving	19.5	4.7	75.8
Open avg.	**26.3**	**5.0**	**68.6**
Closed-source generators			
HappyHorse	37.2	6.9	55.9
Runway	34.8	5.6	59.7
Kling	47.7	7.3	45.1
Seedance2.0	51.9	7.4	40.7
Hailuo	31.8	5.3	63.0
Closed avg.	**40.6**	**6.5**	**52.9**

The pooled result masks substantial differences across generation sources. The source-equal FakeR is 68.6% for the four open-source generators but 52.9% for the five closed-source generators. Seedance2.0 and Kling are the most difficult to recognize as generated, with FakeR values of 40.7% and 45.1%, respectively, whereas OmniWeaving reaches 75.8%. This pattern is not driven by a small subset of reviewers: all 20 reviewers obtain higher source-equal FakeR on open-source than on closed-source generators, and leaving out any one reviewer preserves the complete nine-source ranking. Detailed protocol, source-level counts, and reviewer-level robustness checks are provided in Appendix D.1; generated videos repeatedly judged *Real* form the candidate pool studied next.

Takeaway. Reviewers identify 68.6% of open-source videos as generated but only 52.9% of closed-source videos, with the rates falling to 40.7% for Seedance2.0 and 45.1% for Kling.

4.3.2. RA-Bench-HumanProof Remains Difficult Across Detector Families

We construct **RA-Bench-HumanProof** through two stages. Of the 16,038 generated videos in the standard review stream, Stage 1 retains 1,080 that all three assigned reviewers label *Real*. Two additional reviewers independently reassess these candidates using the same source-unaware protocol, and a video is retained only when both again label it *Real*. The resulting 633 generated videos have therefore been labeled *Real* by all five reviewers. RA-Bench-HumanProof contains 119 open-source and 514 closed-source videos, including 160 from Kling and 159 from Seedance2.0.

For detector evaluation, each generated video in RA-Bench-HumanProof is paired with its matched real anchor. Because the source composition is determined by human selection rather than a predefined quota, we compare these results with a source-matched RA-Bench reference that uses the same source proportions. Table 8 reports representative detectors from all three families. For discrete outputs, we report balanced accuracy (BAcc) and fake recall (FakeR); for continuous scores, we report paired AUC and T@5%. The final column reports the same metric pair for the source-matched RA-Bench reference: BAcc/FakeR for discrete outputs and AUC/T@5% for continuous scores. Full construction details, response pairs, source counts, coverage audits, and comparison rules are provided in Appendix D.2.

The seven traditional detectors average 47.5% AUC and 4.4% T@5% on RA-Bench-HumanProof, compared with 49.5% and 4.6% under source-matched RA-Bench weighting. This limited change does not imply reliable detection: both evaluations remain close to random ranking and provide little generated-video recall at a 5% false-positive rate. Gemini-3.1-Pro-Preview exhibits a clearer association with human difficulty. Binary and Diagnostic BAcc decrease from 61.2% and 61.0% to 54.7% and 54.5%, while their FakeR decreases from 49.9% and 47.3% to 34.3% and 30.0%, respectively. Rating AUC decreases from 61.5% to 54.9%, and its T@5% decreases from 5.6% to 4.3%.

Table 8 | Human-deceptive videos remain difficult for current detectors. RA-Bench-HumanProof contains 633 generated videos, each paired with its matched real anchor. Discrete-output methods report BAcc and FakeR, while continuous-score methods report paired AUC and T@5%. The final column reports BAcc/FakeR for discrete outputs and AUC/T@5% for continuous scores on full RA-Bench after weighting its source-specific results by the RA-Bench-HumanProof source proportions. All values are percentages. Skyra official-timestamp results follow the released protocol; frame index is the timestamp-free control.

Configuration	RA-Bench-HumanProof				Source-matched RA-Bench BAcc/FakeR or AUC/T@5%
	Discrete output		Continuous score		
	BAcc	FakeR	AUC	T@5%	
Traditional detectors					
CNNSpot	–	–	40.5	3.6	43.5 / 3.8
NPR	–	–	50.3	3.8	50.4 / 4.0
UnivFD	–	–	39.4	0.6	39.8 / 1.0
ForgeLens	–	–	54.9	8.8	58.4 / 10.4
DeCoF	–	–	59.0	1.9	59.2 / 1.7
D3	–	–	33.1	6.8	39.5 / 6.2
ReStraV	–	–	55.2	5.2	55.6 / 4.9
7-detector mean	–	–	**47.5**	**4.4**	**49.5 / 4.6**
Zero-shot multimodal models					
Gemini-3.1-Pro-Preview Binary	54.7	34.3	–	–	61.2 / 49.9
Gemini-3.1-Pro-Preview Diagnostic	54.5	30.0	–	–	61.0 / 47.3
Gemini-3.1-Pro-Preview Rating	–	–	54.9	4.3	61.5 / 5.6
Fine-tuned MLLM detectors					
Skyra-SFT, official timestamp	72.0	55.1	–	–	73.9 / 60.4
Skyra-SFT, frame index	53.7	51.3	–	–	55.3 / 55.7
Skyra-RL, official timestamp	74.5	63.0	–	–	75.1 / 66.3
Skyra-RL, frame index	53.7	58.8	–	–	56.1 / 63.3
BusterX++, released pipeline	49.4	3.9	–	–	49.9 / 6.2

The fine-tuned MLLMs change by at most 2.4 BAcc points relative to their source-matched references, while FakeR decreases by 2.3–5.3 points across all five configurations. This limited variation does not establish content-based robustness. The official-timestamp Skyra configurations retain the prior identified in Section 4.1.3; replacing timestamps with frame indices reduces both checkpoints to 53.7% BAcc. BusterX++ reaches 49.4% BAcc, with 3.9% FakeR and 94.9% RealR, indicating a strong tendency to predict *Real*. Neither behavior provides reliable detection of human-deceptive videos.

RA-Bench-HumanProof exposes different failure patterns across detector families: Gemini loses much of its fake-side evidence, traditional detectors remain weak before and after human selection, and the stronger official Skyra results retain the timestamp prior. Human and detector failures therefore overlap, but they are not equivalent. Together, these results show that videos that mislead reviewers are also difficult for current detectors. This overlap is particularly consequential for generated videos depicting crisis events, which repeatedly appear real to reviewers while receiving weak or unreliable fake evidence from the evaluated detector families.

Takeaway. RA-Bench-HumanProof remains difficult across detector families: Gemini Binary and Diagnostic reach 54.7% and 54.5% BAcc, the seven traditional detectors average 47.5% AUC, both timestamp-free Skyra checkpoints reach 53.7% BAcc, and BusterX++ reaches only 3.9% FakeR.

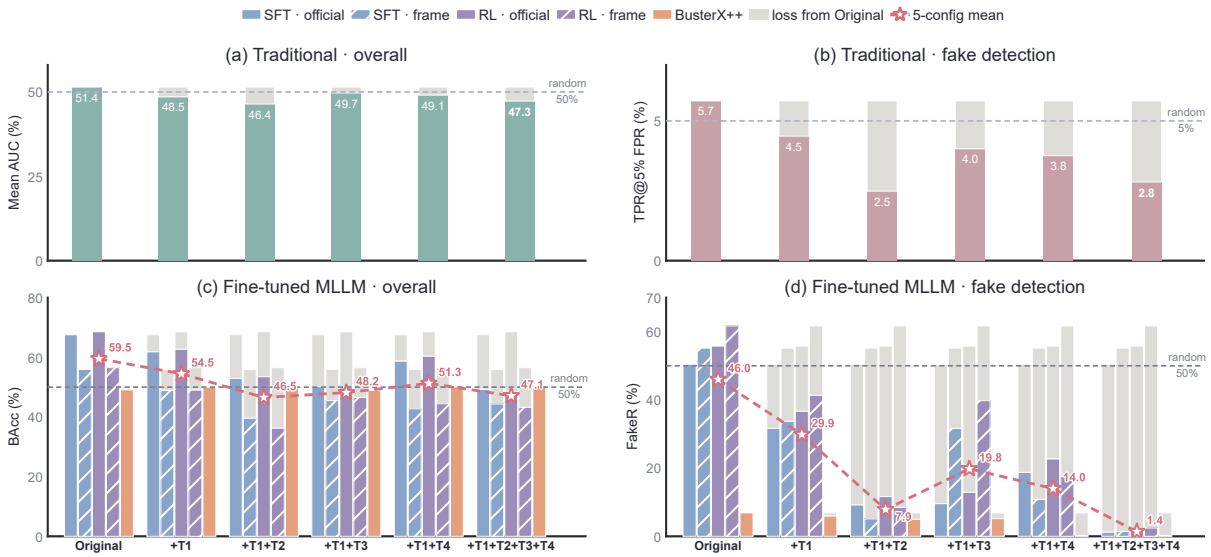


Figure 11 | Detection under the RA-Bench-LastMile social dissemination simulation. Original retains the standardized clips; T1 denotes VP9-to-H.264 transcoding, T2 denotes 0.5× spatial downsampling, T3 denotes conversion to 8 fps, and T4 denotes a synthetic news badge. T1+T2, T1+T3, and T1+T4 isolate each added operation, while Full combines T1–T4. (a–b) Mean AUC and T@5% over the seven traditional detectors. (c–d) BAcc and FakeR for the fine-tuned MLLM detectors. Gray extensions show the loss from each configuration's Original result, hatched bars denote frame-index prompts, and the red dashed line gives the mean over the five fine-tuned configurations. Results are equal-source means over the nine RA-Bench generators.

4.3.3. Social Dissemination Simulation Further Weakens Detection

Videos rarely reach viewers or moderation systems in their standardized form. We therefore introduce **RA-Bench-LastMile**, a controlled social dissemination simulation for evaluating detector reliability. It contains 150 real-event anchors spanning 41 of the 44 L2 categories in RA-Bench, together with their matched videos from all nine generation sources. Original retains the standardized clips; T1 applies a common VP9-to-H.264 transcode; T1+T2, T1+T3, and T1+T4 add spatial downsampling, frame-rate reduction, or a synthetic news badge to T1; and Full combines all four operations. Every condition is applied identically to each generated video and its matched real anchor. Appendix D.3 provides the complete protocol and detector-level results.

Traditional detectors are already weak on Original and respond inconsistently across the social dissemination simulation. As shown in Figure 11(a–b), their mean AUC decreases from 51.4% on Original to 48.5% after T1 and 47.3% under Full, while mean T@5% falls from 5.7% to 2.8%. Detector-specific responses differ sharply: ForgeLens decreases from 61.6% to 35.6%, whereas DeCoF increases from 59.9% to 62.3% and D3 from 49.3% to 54.1%. The Spearman correlation between the Original and Full AUC rankings is consequently only 0.07. The social dissemination simulation can therefore change which detector appears strongest without yielding reliable separation.

Fine-tuned MLLMs instead show a systematic shift toward *Real* as the operations in the social dissemination simulation accumulate. Averaged over the five configurations, BAcc decreases from 59.5% on Original to 54.5% after T1 and 47.1% under Full; FakeR falls from 46.0% to 29.9% and then to 1.4%. Under Full, the four Skyra configurations retain only 1.2–2.4% FakeR despite 84.0–97.3% RealR, while BusterX++ reaches 0.2% FakeR and 100% RealR. The near-50% BAcc therefore reflects

a collapse toward predicting *Real*, not preserved generated-video detection.

The isolated additions to T1 clarify which operations drive this shift. Spatial downsampling lowers mean FakeR from 29.9% to 7.9%. Conversion to 8 fps reduces BAcc by 11.7–12.9 points for the official-timestamp Skyra configurations, but by only 2.4–3.4 points for their frame-index controls, consistent with the temporal-grid dependence in Section 4.1.3. The news badge lowers mean FakeR from 29.9% to 14.0% while increasing mean RealR from 79.1% to 88.5%. Because the underlying scene is unchanged, this presentation cue alone shifts predictions toward *Real*. These results show that both signal degradation and presentation changes can suppress evidence for the generated class, and Full combines their effects.

Takeaway. On RA-Bench-LastMile, Full reduces mean FakeR across the five fine-tuned configurations from 46.0% to 1.4% as their predictions shift toward *Real*.

4.3.4. Summary of Human Perception and Social Dissemination

Human recognition varies substantially across generation sources. Reviewers identify 68.6% of open-source videos as generated, but only 52.9% of closed-source videos; the rates fall to 40.7% for Seedance2.0 and 45.1% for Kling. The 633 generated videos labeled *Real* by all five reviewers form RA-Bench-HumanProof. On RA-Bench-HumanProof, Gemini Binary and Diagnostic reach only 54.7% and 54.5% BAcc, while the seven traditional detectors average 47.5% AUC and 4.4% T@5%. Both timestamp-free Skyra checkpoints reach 53.7% BAcc, and BusterX++ identifies only 3.9% of the videos as fake. On RA-Bench-LastMile, Full reduces mean FakeR across the five fine-tuned configurations from 46.0% to 1.4% as their predictions shift toward *Real*. Generated videos depicting crisis events can therefore be difficult for both viewers and detectors, and the social dissemination simulation can make them still harder to flag.

5. Discussion

Detector reliability is conditional. RA-Bench shows that detector reliability cannot be summarized by a single benchmark score. Traditional detectors fall sharply from their public-reference performance and change rank across generation sources, so performance on one generator does not establish transfer to another. Zero-shot multimodal models vary with prompt format and generation source, while MLLMs fine-tuned for AI-generated video detection depend on temporal-label format or strongly favor the *Real* class. No evaluated detector family therefore provides a decision rule that transfers reliably across sources and evaluation protocols. As generators evolve, detectors tuned to fixed sources or protocols may lose reliability. Evaluation should therefore separate source generalization, prompt sensitivity, protocol dependence, and class bias. These results also motivate systems that combine evidence from multiple detectors and forensic tools through modular or agent-based workflows rather than relying on a single classifier.

Generation properties affect detector families differently. Within each generation source, stronger Condition Fidelity and temporal consistency are associated with weaker fake evidence, but the most relevant dimensions differ across detector families. Gemini is most sensitive to Condition Fidelity and assigns more fake evidence to dynamic clips, whereas the traditional-detector mean changes more with the VBench-I2V Quality Score, Subject Consistency, and Imaging Quality. Across T2V, first-frame I2V, and first+last-frame I2V, the mean FakeR of the evaluated MLLMs fine-tuned for AI-generated video detection falls from 70.5% to 42.7% and 28.3%, while traditional-detector AUCs move in both directions. Detection difficulty therefore cannot be summarized by a single quality score or a common ordering of generation conditions. The small variation across three sampling seeds further indicates that these patterns are not specific to one sampled realization.

Human deception identifies socially consequential failures. Human review reveals a failure mode that detector scores alone cannot capture. Reviewers label 22.8% of real-video judgments as *Generated*, while 633 generated videos are labeled *Real* by all five assigned reviewers. These 633 videos form **RA-Bench-HumanProof**, on which Gemini performs near chance under all three prompt formats. Traditional detectors show only a small additional decline, but their source-matched RA-Bench performance is already close to random. RA-Bench-HumanProof therefore isolates generated videos that repeatedly mislead viewers and remain difficult across detector families. When these videos depict crisis events, they may circulate as authentic evidence and distort public understanding before verification.

Social dissemination shifts detector predictions toward *Real*. RA-Bench-LastMile uses a social dissemination simulation to measure how detector behavior changes without altering the underlying event content. Under Full, the mean AUC of the seven traditional detectors falls from 51.4% to 47.3%, and their ranking changes substantially. The evaluated MLLMs fine-tuned for AI-generated video detection show a sharper failure on generated videos: their mean FakeR falls from 46.0% to 1.4% as predictions shift toward *Real*. Downsampling, frame-rate reduction, and a news-style badge each contribute to this shift. Generated videos depicting crisis events may therefore pass automated screening and reach viewers without an authenticity warning, which may give misinformation more time to spread before verification.

Limitations. RA-Bench provides a controlled evaluation of AI-generated video detection in real-world crisis settings rather than a complete account of how such content may be misused. First, the current release is necessarily a snapshot of a rapidly changing field. It covers nine I2V generation sources and visual-only detection, but new generators and detectors may quickly change the difficulty of the benchmark. RA-Bench should therefore be maintained through versioned updates that add new generation sources, detection methods, and social dissemination settings. Second, the benchmark studies I2V generation and social dissemination as separate, controlled stages, with the latter represented through a social dissemination simulation. In practice, malicious actors may use multi-stage forgery pipelines that combine generation with selective editing, audio synthesis, contextual captions, and repeated platform processing. Future versions should evaluate these end-to-end workflows while retaining traceable controls over each stage. Finally, human judgments are collected in a controlled interface without the surrounding social context that can influence credibility online.

6. Future Directions

Based on our proposed benchmark and comprehensive analysis, we present several promising directions to inspire future research:

- **Developing robust detectors for increasingly realistic AI-generated videos.** Although significant progress has been achieved in AI-generated video detection, our analysis demonstrates a substantial performance gap when existing detection methods are applied to realistic videos produced by state-of-the-art generators (van den Oord and Roman, 2024; Team Seedance et al., 2026). More advanced detectors should be designed to leverage multiple cues, such as temporal consistency, lighting variations, and object motion, to ensure robust performance across diverse video scenarios and post-processing operations.
- **Establishing evaluation benchmarks and approaches for detector interpretability.** MLLM-based detection methods can provide detailed evidence beyond the simple logits produced by traditional approaches, yet the reliability of such reasoning remains largely unverified. A promising future direction is to construct benchmarks that comprehensively evaluate the correctness and faithfulness of generated evidence. Building upon this, future work could

explore MLLM-based agents (Yao et al., 2025; Xie et al., 2024) and post-training approaches (Guo et al., 2025; Agarwal et al., 2024) to decompose AI-generated video detection into structured reasoning steps, thereby further enhancing transparency.

- **Incorporating active watermarking for reliable AI-generated video detection.** To improve the reliability of AI-generated video detection, active watermarking techniques in video generation models (Fernandez et al., 2024; Hu et al., 2025; Su et al., 2025) represent another important research direction. Such methods can embed identifiable signals during generation, making AI-generated videos easier to detect. Our benchmark can be further extended to systematically evaluate the robustness of active watermarking methods against post-processing operations during social dissemination. Additionally, future work should move beyond solely relying on passive detectors and explore their co-design with active watermarking to keep pace with increasingly realistic AI-generated videos.
- **Extending AI-generated video detection to audio-visual settings.** Since existing methods mainly rely on visual evidence to detect AI-generated videos, our benchmark focuses on measuring this visual detection ability. However, as video generation models such as Seedance 2.0 (Team Seedance et al., 2026) have started to natively generate audio-video aligned content rather than only silent video frames, AI-generated video detection should also move beyond visual cues alone. Future work could jointly explore audio and visual evidence for more reliable and comprehensive AI-generated video detection, and construct corresponding benchmarks to evaluate such multimodal capabilities.
- **Enhancing the realism of video generation models.** Although our benchmark is originally designed for detecting AI-generated videos, it can also be leveraged in the opposite direction to promote more realistic video generation. Our analysis reveals that higher-quality AI-generated videos are generally harder for both humans and detectors to identify, suggesting that detection results can serve as a useful signal for measuring generation realism. Therefore, the best-performing existing detectors, as well as future detectors developed under this benchmark, can be used as robust reward models during the post-training process (Xu et al., 2026; Liu et al., 2025) of video generation models.

7. Conclusion

We introduce **RA-Bench**, a benchmark for AI-generated video detection that uses **R**eal videos as **A**nchors. RA-Bench contains 17,886 videos, comprising 1,830 real-video anchors across 10 social-risk categories and 16,056 generated clips from four open-source and five closed-source generators. We evaluate detector generalization, generation properties, and human perception and social dissemination. None of the three detector families generalizes consistently across generation sources: traditional-detector performance falls sharply from public-reference results, zero-shot multimodal models remain sensitive to prompts and sources, and fine-tuned MLLMs depend on timestamp cues or exhibit strong class bias. Generation properties also affect detector families differently. Condition Fidelity, temporal consistency, dynamic content, and real-image conditioning show different relationships with traditional-detector and MLLM outputs, while source-level detection patterns remain stable across three sampling seeds. Human review identifies 633 generated videos judged *Real* by all five assigned reviewers, forming **RA-Bench-HumanProof**, on which Gemini performs near chance under all three prompt formats. **RA-Bench-LastMile** further shows that the social dissemination simulation reduces mean FakeR across the evaluated MLLMs fine-tuned for AI-generated video detection from 46.0% to 1.4%. Together, these results show that current methods still cannot reliably detect realistic AI-generated videos in real-world crisis settings.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824. PMLR, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

Weiliang Chen, Wenzhao Zheng, Yu Zheng, Lei Chen, Jie Zhou, Jiwen Lu, and Yueqi Duan. Genworld: Towards detecting ai-generated real-world simulation videos. *arXiv preprint arXiv:2506.10975*, 2025a.

Yingjian Chen, Lei Zhang, and Yakun Niu. Forgelens: Data-efficient forgery focus for generalizable forgery image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Cloudflare. Happyhorse 1.0 i2v. https://developers.cloudflare.com/ai/models/alibaba/hh1-i2v/, 2026. Accessed: 2026-06-30.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

Pierre Fernandez, Hady Elsahar, I Zeki Yalniz, and Alexandre Mourachko. Video seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024.

Google. Gemini 3.1 pro and gemini api model documentation. https://ai.google.dev/gemini-api/docs/models, 2026. Accessed: 2026-06-30.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Yoav HaCohen et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Peisong He, Leyao Zhu, Jiaxing Li, Shiqi Wang, and Haoliang Li. Exposing ai-generated videos: A benchmark dataset and a local-and-global temporal defect based detection method. *arXiv preprint arXiv:2405.04133*, 2024.

Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Videoshield: Regulating diffusion-based video generation models via watermarking. *arXiv preprint arXiv:2501.14195*, 2025.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems*, pages 20672–20705. Curran Associates, Inc., 2025.

Xuan Ju et al. Miradata: A large-scale video dataset with long durations and structured captions. In *Advances in Neural Information Processing Systems*, 2024.

Kling AI. KlingAI open platform: Image-to-video api documentation. https://kling.ai/document-api/apiReference/model/imageToVideo, 2026. Accessed: 2026-06-30.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning. *arXiv preprint arXiv:2512.15693*, 2025.

Xingming Liao, Meiyu Zeng, Canyu Chen, Nankai Lin, Zhuowei Wang, and Aimin Yang. Chameleon: Benchmarking detection and backtracking on commercial-grade ai-generated videos. In *Proceedings of the International Conference on Multimedia Retrieval*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, pages 82155–82192. Curran Associates, Inc., 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

Zhengxiong Luo et al. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379*, 2025.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency. In *2025 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2025.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for ai-generated video detection. *arXiv preprint arXiv:2601.11035*, 2026.

MiniMax. MiniMax Hailuo 2.3: A New Level of Complex Video Performance & Media Agent. https://www.minimax.io/news/minimax-hailuo-23, 2025. Published: October 28, 2025; accessed: July 4, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A 6-million benchmark for ai-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15582–15590, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

Kaihang Pan et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Runway. Runway api: Available models. https://docs.dev.runwayml.com/guides/models/, 2026. Accessed: 2026-06-30.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Zihan Su, Xuerui Qiu, Hongbin Xu, Tangyu Jiang, Junhao Zhuang, Chun Yuan, Ming Li, Shengfeng He, and Fei Richard Yu. Safe-sora: Safe text-to-video generation via graphical watermarking. In *Advances in Neural Information Processing Systems*, pages 156697–156720. Curran Associates, Inc., 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, pages 10078–10093, 2022.

Aäron van den Oord and Elias Roman. State-of-the-art video and image generation with veo 2 and imagen 3. https://blog.google/innovation-and-ai/models-and-research/google-labs/video-image-generation-update-december-2024/, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*, 2025a.

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*, 2025b.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11269–11277, 2026.

Huanjin Yao, Ruifei Zhang, Jiaxing Huang, Jingyi Zhang, Yibo Wang, Bo Fang, Ruolin Zhu, Yongcheng Jing, Shunyu Liu, Guanbin Li, et al. A survey on agentic multimodal large language models. *arXiv preprint arXiv:2510.10991*, 2025.

Appendix

Table of Contents

A RA-Bench Construction and Data Documentation **32**
- A.1 L2 Taxonomy and Source Distribution . . . 32
- A.2 Source Inventory and Rights Basis . . . 33
- A.3 Automated Preprocessing Details . . . 33
- A.4 Manual Review Details . . . 34
- A.5 Postprocessing Details . . . 34
- A.6 Generation Details. . . . 35
- A.7 Qualitative Real-versus-Generated Examples. . . . 40

B Detector Evaluation Protocols and Extended Results **41**
- B.1 Detector Evaluation Protocol and Model Inputs . . . 41
- B.2 Traditional-Detector Reference Transfer and Operating Points . . . 43
- B.3 Zero-Shot Multimodal Model Protocol and Extended Analysis . . . 45
- B.4 Fine-Tuned MLLM Protocol Sensitivity and Class-Conditional Behavior . . . 49

C Generation Factors and Robustness Analysis **51**
- C.1 Generation Quality and Detectability . . . 51
- C.2 Generation Settings and Detectability. . . . 53
- C.3 Stability Across Generation Seeds . . . 55

D Human Evaluation and Social Dissemination **57**
- D.1 Human Evaluation Protocol and Source-Level Recognition . . . 57
- D.2 RA-Bench-HumanProof Construction and Detector Evaluation . . . 58
- D.3 RA-Bench-LastMile Protocol and Complete Results . . . 61

A. RA-Bench Construction and Data Documentation

A.1. L2 Taxonomy and Source Distribution

Section 3.1 organizes the 675 collected source videos using a two-level taxonomy. The 10 L1 domains represent broad areas of social risk, while the 44 L2 categories specify the event scope within each domain. Table 9 provides the complete category definitions and source counts.

Taxonomy assignments are made before scene segmentation and manual review. The counts therefore refer to source videos rather than review clips or released anchors. We retain the collected source distribution without imposing uniform category quotas or rebalancing at this stage.

Table 9 | L2 taxonomy and source distribution of RA-Bench. Counts refer to the 675 collected source videos before scene segmentation and manual review. L1 subtotals are shown in parentheses.

L2	Event scope	#
L1-01 Weather and natural disasters (75)		
L2-01a	Windstorms, hurricanes, and tornadoes	17
L2-01b	Wildfire spread	15
L2-01c	Earthquakes, building damage, and post-disaster street scenes	23
L2-01d	Flooding	20
L1-02 War and armed conflict (60)		
L2-02a	Battlefield overviews and urban war damage	15
L2-02b	Airstrikes, missile launches, and distant explosions	15
L2-02c	Military aircraft, naval vessels, and armored vehicle formations	15
L2-02d	Official military exercises, parades, and public training	15
L1-03 Politics and governance (60)		
L2-03a	Official press briefings and public announcements	15
L2-03b	Parliamentary and government-hall meetings	15
L2-03c	Public speeches by political leaders	15
L2-03d	Diplomatic meetings, signing ceremonies, and summit photo calls	15
L1-04 Public safety (69)		
L2-04a	Urban lockdowns, police lines, and security deployment	15
L2-04b	Urban unrest and arson scenes in distant, CCTV, or aerial views	24
L2-04c	Counter-terrorism drills and large-scale emergency evacuation	15
L2-04d	Search-and-rescue operations in mountains, at sea, or urban areas	15
L1-05 Accidents and infrastructure failures (93)		
L2-05a	Aviation incidents, including emergency landings, runway excursions, and airport emergency response	22
L2-05b	Rail transit incidents, including train derailments and subway evacuation	11
L2-05c	Highway multi-vehicle crashes and large bus or truck accidents in distant or aerial views	16
L2-05d	Industrial disasters, including factory explosions, fires, and chemical leaks	17
L2-05e	Major infrastructure damage, including bridge or high-rise collapse and repair operations	27
L1-06 Economic and social panic (60)		
L2-06a	Cash-withdrawal queues or bank-run scenes outside bank branches	12
L2-06b	Trading floors or market screens showing abrupt stock-market drops	12
L2-06c	Empty supermarket shelves and panic buying	12
L2-06d	Energy shortages, including gas-station queues and city-scale power restrictions	12
L2-06e	Official or regulatory briefings and hearings on financial crises	12
L1-07 Public health (60)		
L2-07a	Official briefings on epidemics and public-health emergencies	15
L2-07b	Hospital exteriors, emergency entrances, and ambulance dispatch without patient close-ups	15
L2-07c	Mass vaccination sites and medical queues in distant group views	15
L2-07d	City-scale public-health measures, including disinfection and health checkpoints	15
L1-08 Technology (60)		
L2-08a	Technology or AI product launches with stage demonstrations	13
L2-08b	Robots and autonomous vehicles in public spaces, such as inspection or delivery	18
L2-08c	Network outages and data-center failure scenes, including machine-room and operations views	16
L2-08d	Technology and developer conference demos	13
L1-09 Space, exploration, and anomalies (63)		
L2-09a	Rocket launches, booster recovery, and launch-site views	12
L2-09b	Space-station exterior views and public-license in-orbit satellite video	13
L2-09c	Rare astronomical phenomena, including eclipses and meteor showers	14
L2-09d	Polar and deep-sea scientific exploration video	12
L2-09e	Aerial anomalies	12
L1-10 Large public events (75)		
L2-10a	Major sports broadcasts, including wide views and scoreboards	15
L2-10b	Post-game celebrations and fan gatherings in public squares	15
L2-10c	Large concerts and music festivals with stage and crowd views	15
L2-10d	Award ceremonies, premieres, and red-carpet events centered on public figures	15
L2-10e	City-scale festivals, lantern shows, New Year fireworks, and public celebrations	15
Total		**675**

A.2. Source Inventory and Rights Basis

Table 10 summarizes the 675 source videos by provenance group and major contributor. Government and public-institution sources contribute 348 videos, and open-license repositories contribute 191; together they account for 539 of the 675 source videos (79.9%). These counts are measured before scene segmentation and do not depend on how many clips are later retained from each source video.

The *Rights basis* column condenses the source-level provenance and terms recorded during collection; it does not by itself establish permission to redistribute a clip. An aggregated row may contain more than one rights basis, while the final redistribution decision is made separately for each source during the rights review described in Appendix A.5.

Table 10 | Source inventory and recorded rights basis. Counts refer to source videos before scene segmentation. Combined entries (e.g., "public domain / official reuse") indicate that the aggregated row contains sources with different recorded rights bases.

Source	#	Rights basis
Government & public institutions (348)		
DVIDS	240	US public domain
The White House	22	Open / free license
NTSB	16	Platform terms
EU institutions (Council, Parliament, Commission)	38	Official reuse terms
NASA (incl. SVS)	14	Public domain / official reuse
NOAA	10	US public domain
Other agencies (USGS, NSF, Defense.gov)	8	Public domain / official reuse
Open-license repositories (191)		
Wikimedia Commons	131	Open / free license, public domain
Pexels	60	Pexels license
News & broadcast platforms (105)		
YouTube (general uploads)	75	Platform terms
Reuters (Video, YouTube)	16	Platform / editorial terms
Associated Press (AP, AP Archive)	13	Editorial terms
Other broadcast	1	Editorial terms
Other official channels (31)		
Corporate channels (Microsoft, Google, AWS, NVIDIA, OpenAI, Meta, Apple, Samsung, and others)	19	Official reuse / case-by-case
Transportation & organizational channels (state DOTs, airports, others)	12	Platform terms
Total	**675**	

A.3. Automated Preprocessing Details

Scene segmentation. We apply PySceneDetect's `ContentDetector` with a threshold of 27.0 and a minimum scene length of 15 frames. Detected scenes shorter than 5.0 s are merged with an adjacent segment to avoid fragmenting a coherent event into very short clips. Detection uses the PyAV decoding backend. We then split the source videos with FFmpeg using re-encoding rather than stream copy, which avoids restricting the cut points to existing keyframes. Applied to the 675 source videos, this procedure produces the 5,774 clips reviewed in Section 3.3.

Near-duplicate prefilter. We uniformly sample 8 frames from each clip and encode them with a ResNet-18 pretrained on ImageNet-1K. Comparisons are restricted to the next 4 clips from the same source video, since the main source of local redundancy is repeated content across adjacent scenes. Two sampled frames are treated as a match when their cosine similarity is at least 0.90. A clip pair is then flagged only when its bidirectional matched-frame ratio is at least 0.50, its mean cosine similarity over matched frames is at least 0.90, and its clip-level cosine similarity is at least 0.92. The prefilter produces 1,269 pairwise duplicate warnings grouped into 350 clusters. These warnings are advisory:

no clip is removed automatically, and the corresponding pairs are presented to the Round 1 reviewers for confirmation.

A.4. Manual Review Details

Round 1 review. Seven volunteers review all 5,774 clips. Each clip is assessed independently by two reviewers, yielding 11,548 clip–reviewer decisions. Assignments are balanced by semantic subcategory and automatic prefilter status so that no reviewer receives a disproportionate share of any event type or of the suspected near-duplicates.

Review criteria and actions. Reviewers assess visual quality, semantic fit to the assigned subcategory, duration suitability, duplicate content, and source-related risks, such as privacy-sensitive content or intrusive platform overlays. Each reviewer assigns one of three actions: *retain*, *reject*, or *uncertain*. The near-duplicate warnings from Appendix A.3 serve only as review aids; the reviewers make the inclusion decision from the video content and the predefined criteria.

Round 2 adjudication. A clip is routed to Round 2 when the two Round 1 reviewers disagree or when either reviewer selects *uncertain*. Four additional volunteers serve as adjudicators, with each routed clip assigned to one adjudicator for a final decision. After adjudication, 2,426 clips are retained as the standard sample pool and 3,348 are rejected. Thus, all 5,774 clips receive a resolved binary outcome before the retained clips enter postprocessing (Section 3.4; Appendix A.5).

A.5. Postprocessing Details

This subsection reports the exact count-changing operations, encoding settings, and source-level rights review used in Section 3.4.

Duration bounding. Each clip is bounded to a 3–15 s window: clips shorter than 3 s are discarded, while clips longer than 15 s are truncated. This step removes one clip and truncates 444, leaving 2,425 length-bounded clips.

Encoding standardization. We re-encode the video stream of every retained clip with H.264 using `libx264`, preset `slow`, and CRF 17. The same release encoding is later applied to generated clips so that codec choice and encoder configuration do not serve as label cues.

Homogeneity pruning. We apply the mutually exclusive per-source schedule to reduce overrepresented uploads without imposing equal category quotas. Sources with at most 10 retained clips are unchanged. For sources with 11–20, 21–30, 31–40, and 41–50 clips, we remove one clip in every 5, 4, 3, and 2, respectively; for sources with more than 50 clips, we remove approximately 70% at random. Density pruning removes 519 clips. Removing seven additional groups of manually identified repetitive clips discards another 76, for 595 removals in total.

Rights review. Rights are determined per source, with one decision covering all clips from that source. Two of the seven reviewers handled a manual queue of 247 sources whose licenses were not directly determinable; the remaining sources had determinable licenses from their origin. Each anchor carries its documented rights basis and release mode as metadata. The public release distributes 1,319 anchors with `public_cleared` or `public_conditional` status as media; the remaining 511 are represented by metadata and source URLs only.

Count reconciliation. Starting from the 2,426 clips retained after manual review, duration bounding removes one clip and homogeneity pruning removes 595. The resulting real set contains 1,830 clips from 338 source videos, totaling 18,448 s with a mean duration of 10.08 s.

A.6. Generation Details

Two-stage caption and prompt construction. Captioning and prompt construction use Gemini 3.1 Pro Preview (Google, 2026), queried through its native video endpoint under the model identifier `gemini-3.1-pro-preview`. Stage A converts the audio-stripped real anchor and anonymized clip metadata into the structured visual caption defined in Table 11. Stage B converts this caption and the target duration into a generator-agnostic English prompt, a negative prompt, a Chinese translation retained only for bookkeeping, and a duration hint. The English prompt and negative prompt are shared across generation sources. In the reproduced instructions, "text-to-video prompt" refers to this textual condition; every benchmark generation call also receives the real anchor's first frame and is executed as I2V.

Structured output and interface. The Stage A user message supplies only anonymized identifiers, L1/L2 labels and definitions, duration, frame rate, and resolution. Its output follows Table 11. Stage B receives this JSON together with `target_duration_sec`. Numeric duration is excluded from the generation prompt itself and appears only in `duration_hint`, formatted as `"{t}s"`.

Table 11 | Stage A structured caption schema. The six core fields describe visible content; the auxiliary fields support prompt construction, uncertainty tracking, and downstream claim-level analysis.

Field	Content
Core visual fields	
`short_caption`	one-sentence summary
`main_object_caption`	foreground subject(s)
`background_caption`	scene / setting
`camera_caption`	shot type and camera motion
`style_caption`	capture / visual style
`action_caption`	event and visible motion
Auxiliary fields	
`dense_caption`	integrated 120–200 word paragraph
`text_in_video`	on-screen text as {text, position, role}, PII redacted
`salient_entities`	common-noun entity list
`salient_actions`	visible-action list
`claim_candidates`	{claim, evidence_type, confidence} list
`uncertainty_notes`	free-text unknowns
`category_consistency`	{matches_declared_category, notes}

Exact prompt text. The two boxes below reproduce the complete Stage A and Stage B messages. The bracketed headers separate the system instruction from the user template for presentation and were not included in either message. Braced expressions denote fields populated separately for each clip.

Stage A: Structured visual captioning

```
[SYSTEM INSTRUCTION]
You are an expert video annotator constructing a research benchmark for AI-generated
video detection. Your task is to produce a structured, visually grounded caption from
a real-world video clip. The caption will serve as benchmark metadata and will be used to derive
the text
prompt for video generation models (Wan, Hunyuan, LTX, Veo, Kling, Sora).

Adopt the following principles strictly:
1. VISUAL-ONLY. The input video has no audio track. Describe only what is visible.
2. RE-GENERATION ORIENTED. The caption must let a text-to-video model reproduce a
   visually equivalent clip. Prioritize subject appearance, action, location, lighting,
   camera behavior, and visible motion.
3. GROUNDED. Describe only what is directly observable. Do not infer news context,
   identities, causal explanations, or event significance from visual style alone.
   When an attribute is not visually supported, mark it "unknown" in uncertainty_notes.
4. NEUTRAL TONE. Avoid evaluative adjectives (tragic, shocking, heroic). Stay descriptive.
5. NO NAMED IDENTITIES. Do not name real people, organizations, locations, countries,
   brands, or events, even if recognizable. Use generic descriptors.
6. PRIVACY-SAFE ON-SCREEN TEXT. Record visible burned-in text in text_in_video, but
   redact identifying strings and PII with bracketed generic tokens.
7. SHOT STRUCTURE AWARENESS. State whether the clip is one continuous shot or contains
   hard cuts / multiple distinct shots.
8. L2 CATEGORY CONSISTENCY. Judge category_consistency against the declared L2 label,
   not only the broad L1 category.
9. TEMPORAL COVERAGE. Describe the clip as a whole, including visible temporal progression.
10. ENGLISH OUTPUT. All caption fields are in English.

Output a single JSON object conforming to the schema. No preamble, no code fences.

[USER TEMPLATE]
<task>
Annotate the attached real-world video clip into a structured caption following the
MiraData / Any2Caption schema. The caption is benchmark metadata and the source for
downstream text-to-video prompt construction.
</task>

<context>
clip_id: {clip_id}
parent_video_id: {parent_video_id}
category_l1: {l1_code} ({l1_label})
category_l2: {l2_code} ({l2_label})
category_l2_definition: {l2_label}
clip_duration_sec: {duration_sec}
fps: {fps}
resolution: {width}x{height}
notes: sensitive-impact taxonomy; apply identity-free anonymization.
</context>

Return a single JSON object with the 13 schema fields (six core caption fields plus
seven auxiliary fields).
```

Stage B: Generator-agnostic prompt construction

```
[SYSTEM INSTRUCTION]
You convert a structured video caption into a single generator-agnostic text prompt for
the text-to-video generation stage of an AI-generated video detection benchmark.

Single objective: produce one English prompt, plus one Chinese translation for metadata
bookkeeping, from the same scene description and target duration for use with
text-to-video models (Wan, Hunyuan, LTX, Veo, Kling, Sora). The same prompt is sent to
every generator so prompt wording is fixed across the comparison; do not tailor wording
to any single generator.

Strict requirements for the English prompt:
- GROUNDED. Use only content present in the source structured caption.
- COMPLETE. Integrate subject, action, setting, camera, style, and lighting/quality.
- CONCISE. Target 100-180 words, ideally 120-160.
- VISUAL. Concrete nouns, verbs, spatial / lighting / camera terms.
- IDENTITY-FREE. No real names, brands, places, countries, organizations, events,
  phone numbers, addresses, license plates, or bracket tokens.
- TEMPORALLY SPECIFIC BUT NOT NUMERIC. Use natural pace words, no numeric durations.
- SINGLE-SHOT. Assume one continuous take.
- VISUAL-ONLY. Do not describe sounds, dialogue, music, or audio cues.
- GENERATOR-NEUTRAL. Plain descriptive English that common T2V encoders can parse.

Output a single JSON object: English prompt, Chinese translation, negative prompt,
duration hint.

[USER TEMPLATE]
<source_structured_caption>
{stage_a_json}
</source_structured_caption>

<generation_target>
target_duration_sec: {target_duration_sec}
</generation_target>

<instructions>
Compose the English generation prompt as a single coherent paragraph ordered as:
1. [Subject + Action]   2. [Setting + Environment]   3. [Camera]   4. [Style]

Length 100-180 words (ideally 120-160); never below 80. Do not pad.

SINGLE-SHOT FRAMING: if the source caption mentions cuts, describe only the first shot.

Then produce:
- generation_prompt_zh: a faithful Chinese translation of generation_prompt_en.
- negative_prompt: a generator-neutral English negative prompt, <=40 words.
- duration_hint: format "{target_duration_sec:g}s".

Hard rules:
- Do NOT introduce content absent from the caption, or copy from claim_candidates.
- Do NOT use named identities, places, brands, organizations, events, or bracket tokens.
- Do NOT mention audio, dialogue, music, or numeric seconds in the prompts.
- Do NOT tailor the prompt to any specific generator. Output JSON only, no fences.
</instructions>
```

End-to-end worked example. The following example shows how a Stage A visual description is converted into the shared Stage B prompt. Only the short-caption field is shown for Stage A; the full record follows Table 11.

Worked example: structured caption to generation prompt

Clip ID
`L1-01_L2-01a_0003__scene_003`

Duration hint
`6.602s`

Stage A output: `short_caption`
An aerial view from an aircraft window showing a spinning turboprop engine against a backdrop of dense clouds.

Stage B output: `generation_prompt_en`
On the right side of the frame, a turboprop engine with a light-colored cylindrical housing and dark propeller blades operates continuously. Throughout the clip, the dark blades spin rapidly, creating a distinct motion blur as the aircraft maintains a steady forward flight path. The scene is an aerial view set high above a vast, textured layer of whitish-grey clouds below. In the upper sky, a dense layer of dark grey clouds looms, while a faint band of orange light glows at the distant horizon under natural, muted lighting. The shot is captured from an interior passenger perspective, looking out through an aircraft window. The camera remains mostly static, featuring only minor, steady vibrations typical of in-flight recording. The footage is presented in an amateur smartphone video style, exhibiting slight low-light noise in the darker cloud areas alongside the realistic motion blur of the active propeller.

Stage B output: `negative_prompt`
low quality, blurry, distorted face, warped hands, extra limbs, flicker, oversaturation, text watermark, logo, jittery camera, cartoon, anime, painting, illustration.

Generator execution. All generation sources receive the same Stage B prompt and the real anchor's first frame. Table 12 summarizes the generator, request settings, realized output format, and retained counts. Open-source settings follow Wan2.2 (Team Wan, 2025), LTX-2 (HaCohen et al., 2026), and OmniWeaving (Pan et al., 2026); provider calls follow the corresponding HappyHorse (Cloudflare, 2026), Runway (Runway, 2026), Kling (Kling AI, 2026), Seedance (Team Seedance et al., 2026), and Hailuo (MiniMax, 2025) interfaces. Internal inference parameters are available only for the locally executed sources and are reported separately in Table 13.

Duration mapping. For an anchor of length $t_{\text{real}} \in [3, 15]$ s, the shared target is $t_{\text{gen}} = \text{clamp}(t_{\text{real}} \cdot 8/15,\ 2,\ 8)$ s. This target preserves duration variation across anchors while remaining compatible with the operating ranges of the evaluated generators.

Each source then maps t_{gen} to its supported temporal grid. The Wan2.2 and OmniWeaving sources use a $4n{+}1$ grid at 16 fps, whereas LTX uses an $8n{+}1$ grid at 24 fps; both policies impose a floor of 33 frames. The separate Wan2.2 fixed-duration control uses 81 frames (5.06 s) and is excluded from the benchmark count. Seedance accepts integer durations from 4 to 15 s and therefore realizes 4–8 s under our target policy.

The Hailuo 768P I2V endpoint exposes only 6- and 10-second outputs.[3] We therefore use 7.5 s as the boundary: target durations at or above 7.5 s are generated as 10-second clips, whereas shorter targets are generated as 6-second clips. This provider-specific discretization explains why Hailuo does not follow the continuous 2–8 s target range reported for the shared policy.

[3] MiniMax Image-to-Video API documentation: https://platform.minimax.io/docs/api-reference/video-generation-i2v.

Table 12 | Generation sources and realized outputs. Execution strings and request settings are taken from the recorded generation manifests. n counts clips in the primary paired benchmark set; additional seed repeats and the Wan2.2 fixed-duration control are excluded from the 16,056 benchmark clips.

Source	Execution string / request setting	Output res.	fps	Temporal setting	n
Open-source benchmark generators					
Wan2.2 dynamic	`Wan2.2-I2V-A14B`	832×480	16	$4n$+1 frames (≥33)	1,830
Wan2.2-Lightning	`Wan2.2-Lightning-I2V-A14B-NFE4-V1`	1280×720	16	$4n$+1 frames (≥33)	1,830
LTX	`ltx-2.3-22b-dev`	1536×1024	24	$8n$+1 frames (≥33)	1,830
OmniWeaving	`HY-OmniWeaving`	848×480	16	$4n$+1 frames (≥33)	1,830
Closed-source benchmark generators					
HappyHorse	`happyhorse-1.0-i2v` 720P, 16:9	1264×730	24	3–8 s	1,787
Runway	`gen4.5` 1280:720	1280×720	24	2–8 s	1,805
Kling	`kling-v3` mode `std`	1264×728	24	3–8 s	1,790
Seedance2.0	`doubao-seedance-2-0-260128`; ratio-conditioned	mixed[†]	24	4–8 s	1,524
Hailuo	`MiniMax-Hailuo-2.3-Fast`; 768P	1330×768	24	6 or 10 s	1,830
Auxiliary duration control					
Wan2.2 fixed[*]	`Wan2.2-I2V-A14B`	832×480	16	81 frames (5.06 s)	1,830

[†]Seedance outputs comprise 1,452 clips at 1280×720, 41 at 960×960, and 31 at 720×1280. [*]The fixed-duration clips are used only as an auxiliary control. Each provider was queried on all 1,830 anchors; smaller retained counts reflect provider-specific content-safety rejections rather than deliberate subsampling.

Table 13 | Internal inference settings for locally executed sources. Provider APIs do not expose corresponding sampler-level controls. The dynamic and fixed Wan2.2 settings differ only in their temporal policy.

Source	Conditioning size	Steps	Sampler / inference setting
Wan2.2 dynamic/fixed	output resolution	40	UniPC; shift 5.0; guidance (3.5, 3.5)
Wan2.2-Lightning	output resolution	4	Euler + 4-step LoRA; guidance (1.0, 1.0)
LTX	768×512	30	distilled LoRA 0.8; cfg/STG/rescale 3.0/1.0/0.7
OmniWeaving	output resolution	30	480p, 16:9

LTX generates at its base resolution and applies `ltx-2.3-spatial-upscaler-x2-1.1` to reach 1536×1024; its text model is `gemma-3-12b-it-qat-q4_0-unquantized`. Wan2.2-Lightning applies `Wan2.2-I2V-A14B-4steps-lora-rank64-Seko-V1` to the `Wan2.2-I2V-A14B` base.

A.7. Qualitative Real-versus-Generated Examples

Figures 12 and 13 extend the paired examples in Figure 4 along category coverage and generator variation. Figure 12 covers all ten L1 domains using one matched pair per domain, sampled from several benchmark generation sources. Figure 13 instead holds the anchor, prompt, and first-frame reference fixed while varying the open-source generation configuration. Each filmstrip contains six uniformly sampled frames. The dashed outline marks the real first frame supplied to the generator; columns indicate relative positions within each clip and are not timestamp-aligned across rows.

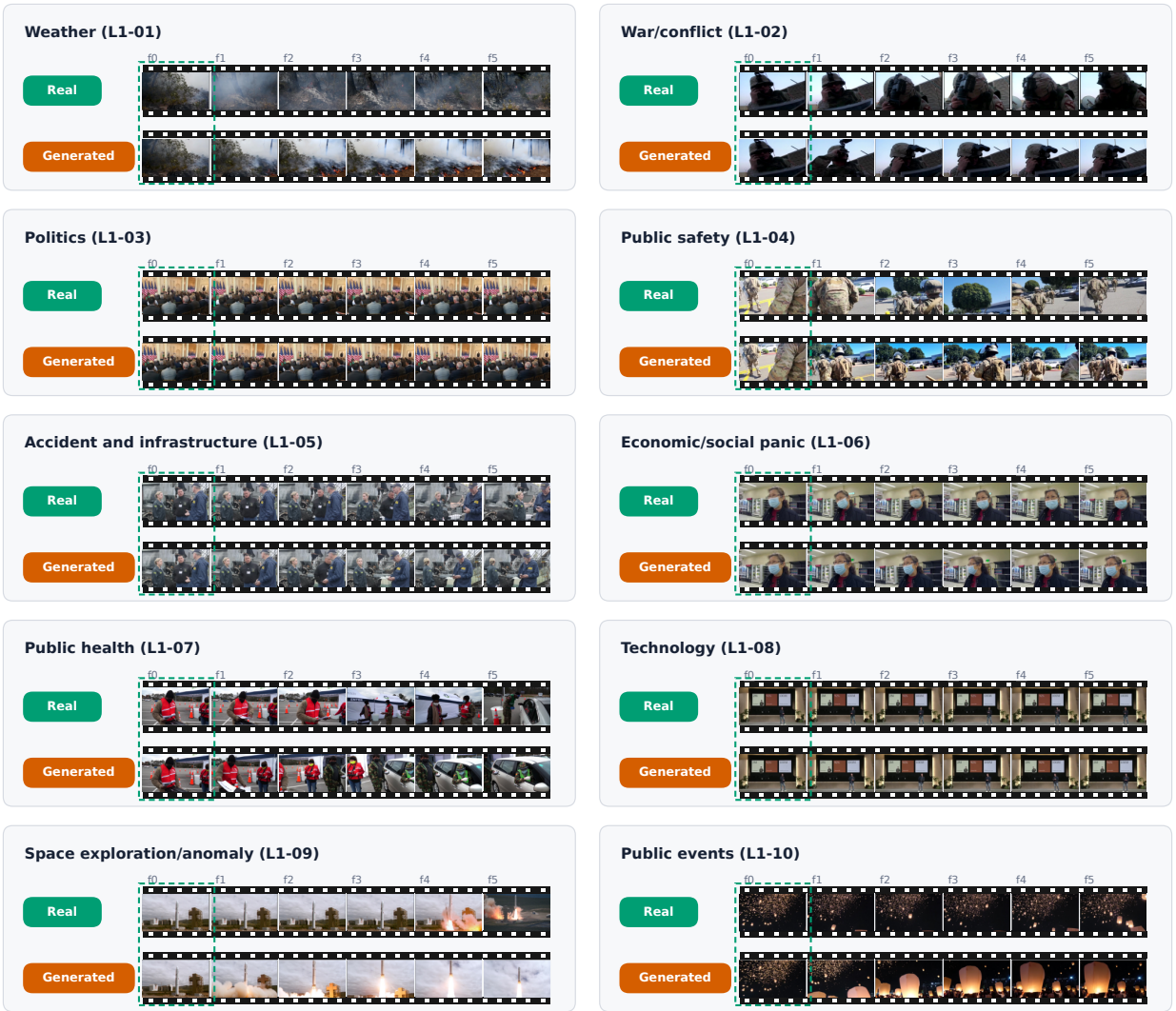


Figure 12 | Qualitative coverage across all ten social-risk domains. Each panel pairs a real anchor (top) with one generated counterpart (bottom); each continuation starts from the highlighted conditioning frame, retains the event-specific scene appearance, and introduces new motion and content.

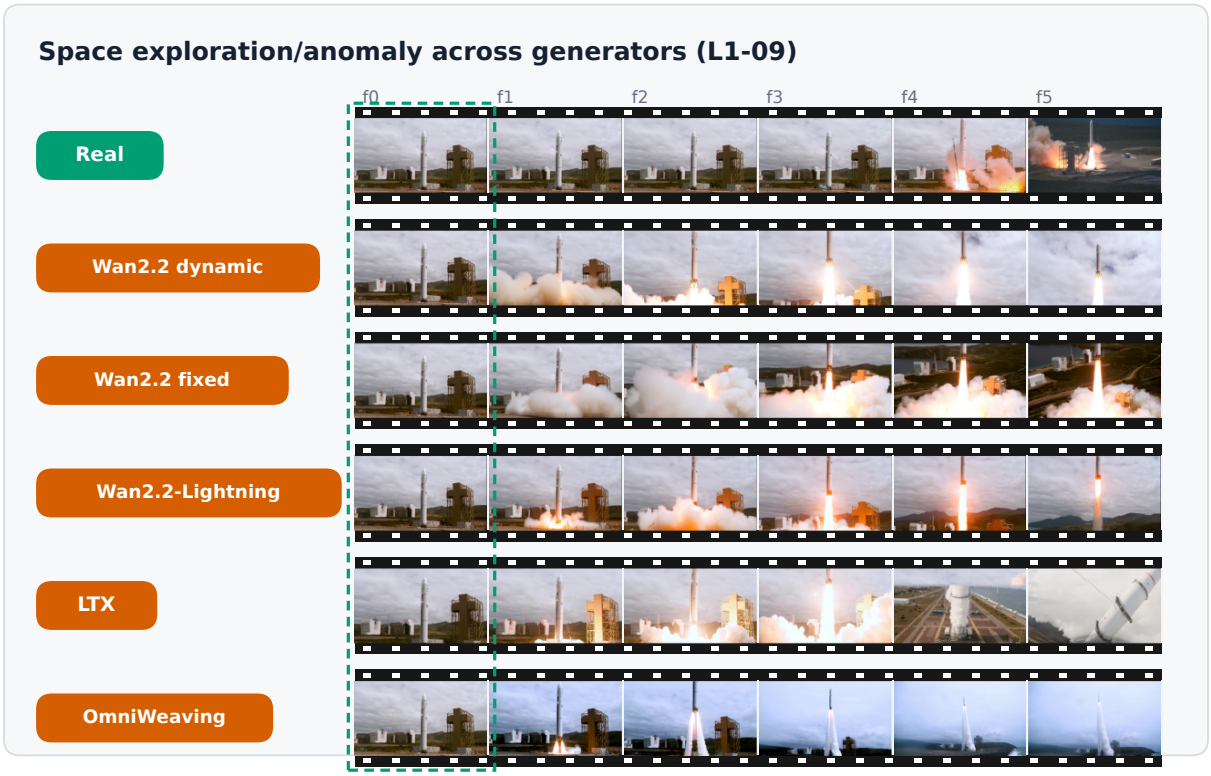


Figure 13 | Cross-generator comparison under matched conditioning. The top row is a real space-launch anchor; subsequent rows show Wan2.2 dynamic, its auxiliary fixed-duration control, Wan2.2-Lightning, LTX, and OmniWeaving under the same prompt and first-frame reference. The outputs preserve the initial launch-pad layout but differ in subsequent motion and visual drift.

B. Detector Evaluation Protocols and Extended Results

B.1. Detector Evaluation Protocol and Model Inputs

For each generation source, we evaluate every generated video together with its matched real anchor. Each open-source generator contributes 1,830 such pairs; for each closed-source generator, evaluation is restricted to the anchors for which the provider returned a generated video. Benchmark-level results give equal weight to the nine RA-Bench generation sources. The fixed-duration Wan2.2 control is marked with an asterisk wherever it is reported and is excluded from these averages.

Unless an ablation states otherwise, we preserve each method's released checkpoint or API version, visual preprocessing, temporal sampling policy, and score-extraction rule. Continuous-score methods are evaluated with paired AUC and TPR at 5% FPR (T@5%); discrete outputs are evaluated with BAcc, macro-F1, and FakeR. For continuous outputs, the released score direction is fixed across all sources; scores are not inverted post hoc when AUC falls below 50%. We do not apply source-specific inversion because it would require knowing the generator identity and would not measure cross-source transfer. We additionally report real recall (RealR) when analyzing class preference. Zero-shot multimodal models receive sampled video frames only; no filename, generator identity, seed, dataset label, or other auxiliary metadata is provided. Table 14 records the visual input and output used for each method family.

Zero-shot model identifiers and decoding. The local Qwen3.5 checkpoints are `Qwen3.5-0.8B`, `Qwen3.5-2B`, `Qwen3.5-4B`, `Qwen3.5-9B`, `Qwen3.5-27B`, `Qwen3.5-35B-A3B`, and `Qwen3.5-122B-A10B`. Their inference disables thinking and uses greedy decoding. API evaluations use the invoked identifiers `qwen3.7-plus` (thinking mode), `gemini-3.1-pro-preview`, and `gpt-5.5`; Gemini uses temperature 0. These API evaluations were conducted in June–July 2026. Because the providers expose model aliases rather than immutable public snapshots, we report the exact invoked identifiers and access period rather than infer unrecorded version numbers.

Table 14 | Detector inputs and evaluated outputs. Unless a control is explicitly noted, visual preprocessing and temporal sampling follow the released implementations. The table reports the visual evidence presented to each method and the output used to compute the common metrics; prompt-specific fields and interventions are detailed in the corresponding zero-shot and fine-tuned appendices.

Method	Visual input / temporal sampling	Evaluation output
Traditional detectors		
CNNSpot, NPR, UnivFD, ForgeLens	Eight uniformly spaced full-video frames; official ImageNet or CLIP normalization	Per-frame fake probabilities averaged into one clip score
DeCoF	Eight frames within the first 32 decoded frames; center crop and 224×224 resize	Softmax fake probability from the eight-frame stack
D3	One 3 s window sampled at 8 fps; 16 frames when available; 10% center crop and 224×224 resize	Continuous `dis_2nd_std` score
ReStraV	Center 2 s window; 24 frames resized to 224×224; DINOv2 ViT-S/14 features	Temporal-geometry MLP score, `prob_fake` = 1 − `prob_real`
Zero-shot multimodal models		
Qwen3.5 family Qwen3.7-Plus Gemini-3.1-Pro-Preview GPT-5.5	Sixteen temporally ordered frames, short side 256; no auxiliary metadata	Binary verdict, 0–100 Rating with verdict, or five-aspect Diagnostic with verdict
MLLMs fine-tuned for AI-generated video detection		
Skyra-SFT / Skyra-RL	Sixteen frames at round$(i(N-1)/15)$, short side 256; official timestamps or frame-index control	Real/fake verdict parsed from `<answer>`
BusterX++ (Wen et al., 2025b)	Released image-mode pipeline sampled at 2 fps	A/B multiple-choice answer parsed from `\boxed{}`

The methods do not observe identical temporal evidence. Frame-level models sample isolated frames across the clip, DeCoF emphasizes the opening frames, D3 and ReStraV evaluate local windows, and the multimodal models receive broader multi-frame summaries. These differences are part of the released inference protocols, so the main comparison treats each detector as an end-to-end system.

Temporal-sampling control. The controlled experiment in Section 4.1.1 changes only the temporal allocation of a fixed eight-frame input. Uniform-8 selects eight frames at uniformly spaced positions. For Global–Local-8, each clip is first scanned at 4 fps with a 160-pixel short side. We construct a camera-compensated eventness curve from residual motion, structural change, and their temporal variation, then select four consecutive frames centered on the strongest response and four uniformly spaced frames. The coarse scan is used only for routing and is not passed to the detector; detector weights, detector-side preprocessing, and score aggregation remain unchanged. The experiment covers eight fixed-length RA-Bench sources and the auxiliary fixed-duration Wan2.2 control, yielding

nine evaluated settings; dynamic-duration Wan2.2 is excluded. We report macro-average AUC over these settings and estimate the five-detector mean confidence interval with a source-block bootstrap. This ablation average is separate from the benchmark-level RA-Bench averages defined above.

B.2. Traditional-Detector Reference Transfer and Operating Points

Reference alignment and rank transfer. The public-reference column in Table 2 provides high-performance context rather than a matched-domain baseline. Six detectors share the LTX-I2V evaluation reported by AIGVDBench (Ma et al., 2026); ReStraV instead uses its reported VidProM AUROC (Internò et al., 2025). The source paper calls this metric AUROC; because it is equivalent to ROC AUC, we denote it as AUC throughout. We therefore restrict the rank-transfer analysis to the six detectors evaluated on the same public reference. ReStraV remains in the absolute-performance comparison but does not enter the correlation.

Table 15 | Public detector rankings do not transfer to RA-Bench. Public AUC and rank use the shared AIGVDBench LTX-I2V reference. RA-Bench AUC is the source-equal mean over the nine benchmark generation sources; the fixed-duration Wan2.2 control is excluded. ΔRank is public rank minus RA-Bench rank, so a positive value denotes an improved relative rank on RA-Bench. AUC values are in %.

Detector	Public AUC	Public rank	RA-Bench AUC	RA-Bench rank	ΔRank
ForgeLens	92.9	1	61.1	1	0
UnivFD	89.9	2	40.7	6	−4
DeCoF	81.5	3	60.9	2	+1
CNNSpot	81.4	4	44.4	4	0
D3	77.7	5	43.5	5	0
NPR	67.6	6	52.2	3	+3

The public and RA-Bench mean rankings have a Spearman correlation of only 0.26. UnivFD moves from second to sixth, whereas NPR moves from sixth to third. This is not a uniform performance loss: such a loss would reduce AUC while largely preserving detector order. The per-source correlations in Table 2 range from −0.37 to 0.54, with a median of 0.31; Seedance2.0 reverses the public ordering most strongly. Together with the source-specific leaders reported in the main text, these changes show that the preferred detector varies with the generation source.

Practical operating points. Table 16 complements the AUC and T@5% results in Table 2 with stricter endpoints from the same ROC sweep. T@1% and T@5% denote TPR at 1% and 5% FPR, respectively, whereas F@95% denotes FPR at 95% TPR. These values do not use a detector's default threshold. For random ranking, the corresponding values are 1%, 5%, and 95%. The auxiliary Wan2.2 fixed-duration control is excluded from all Open averages, consistent with the main evaluation.

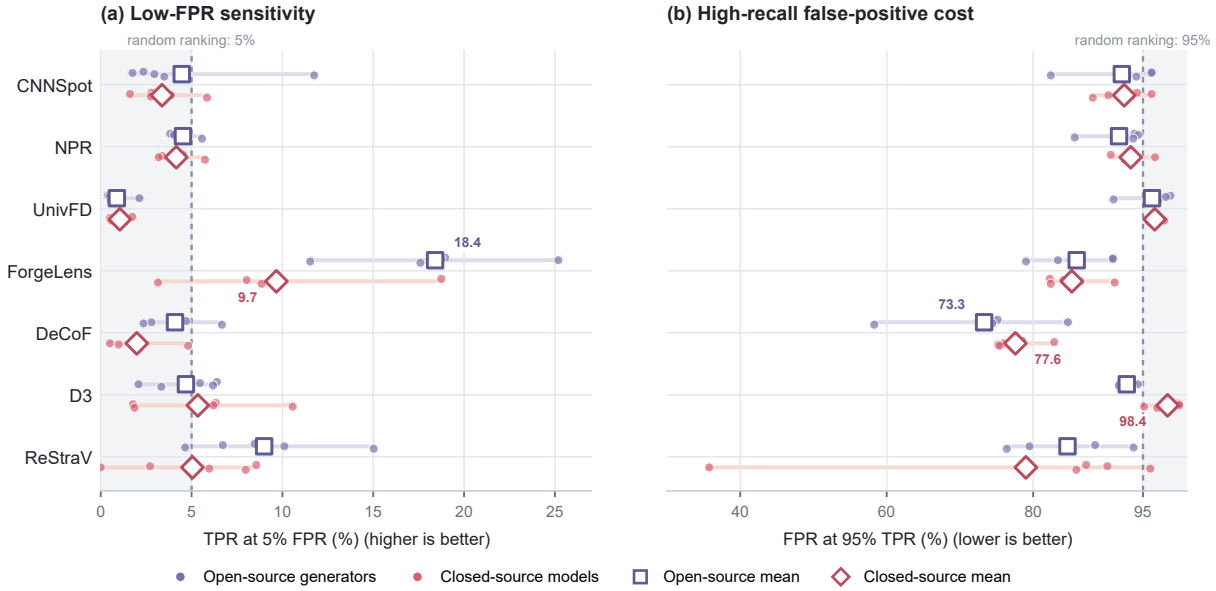


Figure 14 | Operating-point variability across generation sources. Each filled point is one detector–source pair; horizontal segments span the four open-source or five closed-source generators in RA-Bench, and hollow markers denote source-equal means. The Wan2.2 fixed-duration control is excluded. (a) TPR at 5% FPR. (b) FPR at 95% TPR. Dashed lines mark random-ranking values, and shaded regions indicate worse-than-random operating points.

Table 16 | Traditional detectors remain weak at both low-FPR and high-recall operating points. Cells are source-equal means over the four open-source and five closed-source generators in RA-Bench, in %. Higher is better for T@1% and T@5%; lower is better for F@95%. The Wan2.2 fixed-duration control is excluded. Bold marks the best value within each source group.

	Open-source generators			Closed-source generators		
Detector	**T@1%↑**	**T@5%↑**	**F@95%↓**	**T@1%↑**	**T@5%↑**	**F@95%↓**
Image / frame-level detectors						
CNNSpot	0.5	5.1	91.1	0.2	3.4	92.4
NPR	0.7	4.6	91.1	0.6	4.2	93.3
UnivFD	0.2	1.0	95.8	0.2	1.1	96.6
ForgeLens	**5.3**	**18.3**	84.7	**2.5**	**9.7**	85.3
Video / temporal-level detectors						
DeCoF	0.6	3.9	**73.2**	0.4	2.0	**77.6**
D3	0.6	4.5	93.0	1.2	5.3	98.4
ReStraV	1.8	9.6	83.7	1.0	5.0	79.0

The stricter endpoints show that the low AUCs in Table 2 translate into weak practical operating regions. Across the nine RA-Bench sources, the seven-detector mean ranges from 0.4% to 1.7% at 1% FPR and from 2.8% to 7.5% at 5% FPR. Conversely, reaching 95% TPR requires a mean FPR of 83.1–91.8%. At the detector–source level, 54 of 63 pairs remain below 10% T@5%, and 52 require at least 80% FPR to reach 95% TPR. Even LTX and OmniWeaving, the two sources with the highest seven-detector mean AUC, reach only 6.0% and 7.5% T@5%, while their F@95% remains 85.4% and 85.5%. These results are obtained from complete ROC curves; they therefore cannot be attributed to an unfavorable default threshold.

Similar AUC can also conceal different operating trade-offs. ForgeLens and DeCoF are nearly tied in mean AUC on both the open-source generators (64.2% versus 63.4%) and closed-source generators (58.7% versus 58.9%). Across the open/closed groups, ForgeLens provides higher T@5% (18.3%/9.7%) but requires F@95% of 84.7%/85.3%. DeCoF provides lower T@5% (3.9%/2.0%) but reduces F@95% to 73.2%/77.6%. Figure 14 further shows that group means can hide source-specific instability: ReStraV's F@95% ranges from 35.8% on Kling to 96.0% on Seedance2.0. No evaluated detector therefore maintains both low false-positive rates and high recall across generation sources.

B.3. Zero-Shot Multimodal Model Protocol and Extended Analysis

Unified evaluation protocol. All zero-shot multimodal models receive the same 16 temporally ordered frames at a short-side resolution of 256 pixels. No filename, source, seed, dataset, watermark, or other metadata is provided. We evaluate three prompt formats: Binary requests a categorical verdict, Diagnostic requests five aspect scores and a verdict, and Rating requests a continuous AI-generation score and a verdict. The visual evidence and real-versus-generated decision task remain fixed across the three formats.

All aggregate results give equal weight to generation sources rather than individual clips. Open covers the four open-source RA-Bench generators, Closed covers the five closed-source generators, and All covers all nine sources. The fixed-duration Wan2.2 setting is an auxiliary temporal control and is excluded from these averages.

Table 17 | Output fields and metrics for the three zero-shot prompts. Binary and Diagnostic use the explicit *Overall Verdict* for their primary classification metrics. Rating uses the continuous score for paired AUC; its verdict and the Diagnostic aspect scores are analyzed separately.

Prompt	Requested output	Primary metric	Additional analysis
Binary	Overall Verdict	BAcc / macro-F1	FakeR and RealR
Rating	0–100 rating and verdict	Paired rating AUC	Verdict macro-F1, FakeR, and RealR
Diagnostic	Five 0–100 aspect scores and verdict	Verdict BAcc / macro-F1	Per-aspect paired AUC and exact-five collapse

Canonical prompt templates. The following templates define the three reported settings. The shared prefix fixes the visual input, prohibits metadata, and asks the model to judge the clip as a whole. Rating and Diagnostic scores measure the apparent degree of AI generation rather than confidence.

Prompt 1: Binary verdict

```
You are evaluating whether a video clip is AI-generated.
You will receive 16 frames sampled from the video clip in temporal order.
Base your judgment only on the visible content of these frames.
Do not use or assume any metadata, filename, source, seed, dataset, or watermark information.
Judge the clip as a whole.

Output exactly the following line and nothing else:
Overall Verdict: real/fake
```

Prompt 2: AI-generation rating

```
You are evaluating whether a video clip is AI-generated.
You will receive 16 frames sampled from the video clip in temporal order.
Base your judgment only on the visible content of these frames.
Do not use or assume any metadata, filename, source, seed, dataset, or watermark information.
Judge the clip as a whole.

AI-Generation Rating: an integer from 0 to 100, divided into five levels based on the degree to
which the video appears to be AI-generated.
0-19: No visible signs of AI generation.
20-39: Low degree of AI generation, with only weak or isolated signs.
40-59: Moderate degree of AI generation, with mixed or inconclusive signs.
60-79: High degree of AI generation, with clear signs from one or more visual cues.
80-100: Very high degree of AI generation, with strong and consistent signs.
The rating reflects how AI-generated the content appears, not how confident you are.

Output exactly the following two lines and nothing else:
AI-Generation Rating: <integer 0-100>
Overall Verdict: real/fake
```

Prompt 3: Five-aspect diagnostic

```
You are evaluating whether a video clip is AI-generated.
You will receive 16 frames sampled from the video clip in temporal order.
Base your judgment only on the visible content of these frames.
Do not use or assume any metadata, filename, source, seed, dataset, or watermark information.
Judge the clip as a whole.

Rate the clip on the five aspects below. For each aspect, give an integer from 0 to 100 for how
strong the signs of AI generation are on that aspect, in five levels:
0-19: No visible signs of AI generation.
20-39: Only weak or isolated signs of AI generation.
40-59: Mixed or inconclusive signs of AI generation.
60-79: Clear signs of AI generation from one or more visual cues.
80-100: Strong and consistent signs of AI generation.
The rating reflects how AI-generated the content appears, not how confident you are.

The five aspects:
1. Texture & Material: surface textures and materials, such as oversmoothing, repetition, melting
, or implausible details.
2. Structure & Local: object, body, face, hand, and text structure, such as warping, incorrect
proportions, extra or missing parts, or splicing.
3. Lighting, Color & Optical: lighting, shadows, colors, reflections, and other optical effects.
4. Temporal: temporal appearance across the ordered frames, such as flicker, popping, duplication
, frozen regions, or identity inconsistency.
5. Motion & Physical: motion, interactions, trajectories, and physical plausibility across the
ordered frames.

Output exactly the following lines and nothing else:
Texture & Material: <integer 0-100>
Structure & Local: <integer 0-100>
Lighting, Color & Optical: <integer 0-100>
Temporal: <integer 0-100>
Motion & Physical: <integer 0-100>
Overall Verdict: real/fake
```

Aggregate performance. Table 18 extends Table 3 to every evaluated zero-shot model. Aggregate performance remains close to chance for many model–prompt pairs, and no prompt improves consistently with model scale. Qwen3.5-122B-A10B improves from 53.0/45.7 BAcc/macro-F1 under Binary to 54.8/50.8 under Diagnostic, whereas Qwen3.5-27B declines from 53.2/46.7 to 51.3/37.2. The two Qwen3.5 mixture-of-experts models perform better under Rating, while the smaller dense models remain near chance. Qwen3.7-Plus and Gemini-3.1-Pro-Preview also vary between source groups: their Rating AUC decreases from 62.6 to 55.4 and from 66.6 to 61.2, respectively, between

open- and closed-source generators.

Table 18 | Source-equal zero-shot results. Binary and Diagnostic report BAcc/macro-F1; Rating reports paired AUC/verdict macro-F1 (top/bottom). The Open and Closed columns average four and five sources, respectively, and All averages all nine. The fixed-duration Wan2.2 control does not enter these averages. Values are in %.

Model	Prompt	Open	Closed	All
Qwen3.5 dense models				
Qwen3.5-0.8B	Binary	55.2 49.5	53.6 46.9	54.3 48.1
	Diagnostic	49.8 34.7	49.9 34.8	49.8 34.8
	Rating	50.4 33.3	52.3 33.3	51.5 33.3
Qwen3.5-2B	Binary	50.1 33.6	50.3 33.9	50.2 33.8
	Diagnostic	51.4 43.8	50.3 42.5	50.8 43.0
	Rating	52.8 48.8	51.9 48.5	52.3 48.7
Qwen3.5-4B	Binary	53.7 47.5	51.0 45.4	52.2 46.3
	Diagnostic	50.8 38.8	50.0 37.1	50.3 37.8
	Rating	52.4 47.5	50.6 45.9	51.4 46.6
Qwen3.5-9B	Binary	54.1 50.2	52.4 48.8	53.2 49.4
	Diagnostic	50.2 33.9	50.0 33.6	50.1 33.7
	Rating	49.9 34.8	48.9 34.5	49.4 34.6
Qwen3.5-27B	Binary	55.3 49.8	51.6 44.2	53.2 46.7
	Diagnostic	51.9 38.6	50.7 36.1	51.3 37.2
	Rating	52.3 35.5	50.3 34.6	51.2 35.0
Qwen3.5 mixture-of-experts models				
Qwen3.5-35B-A3B	Binary	55.8 48.0	52.4 42.4	53.9 44.9
	Diagnostic	52.8 40.8	50.7 36.6	51.6 38.5
	Rating	60.3 59.1	53.4 52.6	56.5 55.5
Qwen3.5-122B-A10B	Binary	54.2 47.5	52.0 44.2	53.0 45.7
	Diagnostic	58.9 56.3	51.5 46.4	54.8 50.8
	Rating	59.9 59.2	52.5 52.7	55.8 55.6
Frontier multimodal models				
Qwen3.7-Plus (thinking)	Binary	57.2 51.0	53.3 45.2	55.0 47.8
	Diagnostic	57.9 53.0	53.3 46.2	55.3 49.2
	Rating	62.6 53.4	55.4 47.3	58.6 50.0
Gemini-3.1-Pro-Preview	Binary	66.5 66.3	60.9 60.2	63.4 62.9
	Diagnostic	66.3 65.9	60.6 59.7	63.1 62.5
	Rating	66.6 66.7	61.2 60.0	63.6 63.0
GPT-5.5	Binary	52.5 39.1	51.2 36.6	51.8 37.7
	Diagnostic	51.8 37.4	50.8 35.4	51.2 36.3
	Rating	64.3 38.0	60.3 36.0	62.1 36.9

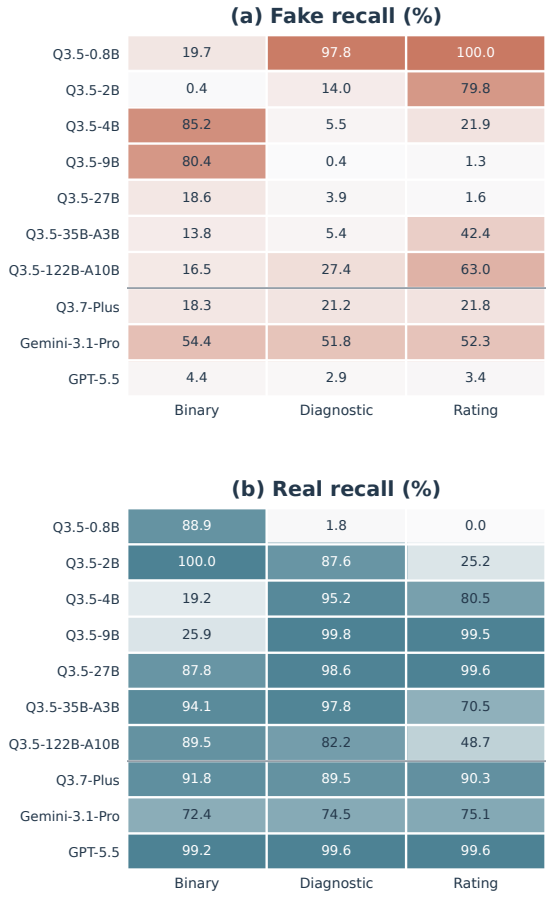


Figure 15 | Prompt-dependent class recall. Each heatmap reports a source-equal mean over the nine RA-Bench generators; Rating uses its explicit Overall Verdict. Several Qwen3.5 models reverse their class preference across prompts. Gemini remains comparatively balanced, whereas GPT-5.5 consistently predicts *Real*.

Prompt-dependent operating points. Figure 15 shows why BAcc alone is insufficient for comparing the three prompts. Qwen3.5-0.8B changes from a Real-favoring Binary rule to predicting nearly every video as fake under Diagnostic and Rating. Qwen3.5-4B and Qwen3.5-9B move in the opposite direction: Binary favors fake, whereas the structured prompts favor real. These models can all remain near 50% BAcc because errors on one class offset correct decisions on the other, despite representing qualitatively different operating points.

Gemini-3.1-Pro-Preview changes little across prompt formats: its FakeR remains between 51.8% and 54.4%, while RealR remains between 72.4% and 75.1%. Qwen3.7-Plus is also stable but remains tilted toward real. GPT-5.5 shows why prompt stability alone is insufficient: its FakeR remains below 4.5% under every prompt while RealR exceeds 99%. Reporting FakeR and RealR is therefore necessary

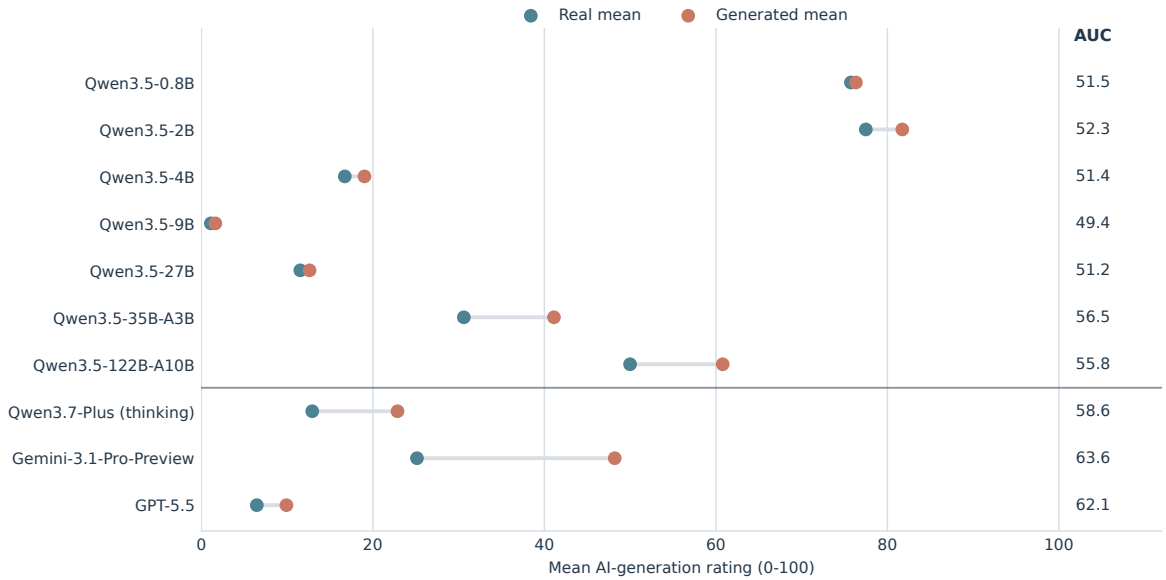


Figure 16 | Model-specific use of the 0–100 Rating scale. Gray-blue and coral markers show source-equal means for matched real and generated videos; the right column reports paired AUC. Large differences in score location across models do not imply corresponding differences in separability.

to distinguish prompt-invariant detection behavior from a persistent class preference.

Absolute Rating scores are not directly comparable across models. Figure 16 shows that models use the 0–100 scale very differently. Qwen3.5-2B assigns mean scores of 81.7 to generated videos and 77.5 to real videos, whereas Qwen3.5-9B assigns 1.6 and 1.1. Both remain close to chance AUC. A high absolute rating therefore does not imply stronger separation, and scores from different models should not be compared as generation probabilities.

Continuous ranking and categorical decisions can also diverge within one model. GPT-5.5 obtains 62.1% paired AUC despite assigning low scores to both classes, but its explicit verdict yields only 36.9% macro-F1 because it predicts almost every video as real. Gemini-3.1-Pro-Preview is the only evaluated model with both paired AUC and verdict macro-F1 above 60%. We therefore report paired AUC for the continuous score and evaluate the explicit verdict separately rather than deriving an additional prediction with a post hoc threshold.

The Diagnostic prompt often collapses to one global score. Figure 17 evaluates each Diagnostic dimension as a continuous signal. For Qwen3.5-0.8B, 4B, 9B, and 27B, aspect AUC remains near chance while exact-five collapse exceeds 94%. These models usually repeat one global score across all requested dimensions rather than provide aspect-specific judgments.

Qwen3.7-Plus reaches 55.6–57.5% aspect AUC, and Qwen3.5-122B-A10B reaches 54.5–54.8%, but both still collapse a majority of complete outputs. Gemini-3.1-Pro-Preview reaches 62.6–63.8% aspect AUC, yet assigns the same value to all five dimensions in 55.2% of complete outputs. GPT-5.5 rarely collapses the five values, but its Temporal and Motion scores reach only 44.9% and 48.6% AUC. A useful diagnostic therefore requires both distinct outputs and dimension-specific separation.

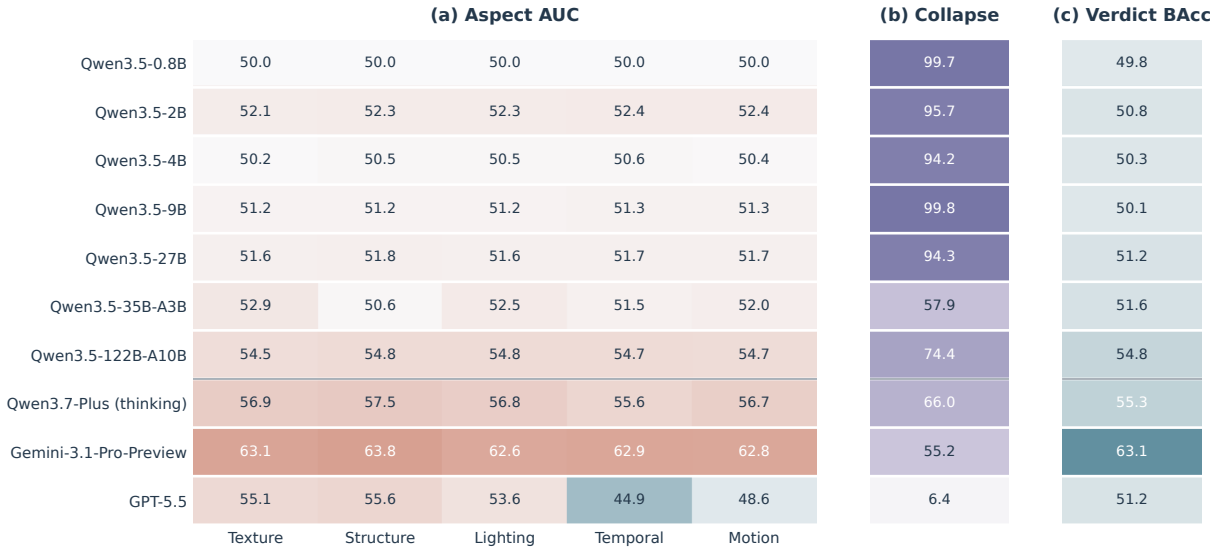


Figure 17 | Separability and collapse of the five Diagnostic scores. Panel (a) reports source-equal paired AUC for each requested dimension. Panel (b) reports the fraction of complete outputs that assign exactly the same value to all five dimensions. Panel (c) reports BAcc from the explicit Overall Verdict.

These analyses separate three behaviors that aggregate BAcc cannot distinguish. FakeR and RealR reveal prompt-induced class preference, Rating AUC measures continuous ranking, and Diagnostic AUC together with collapse tests whether structured responses contain aspect-specific information. We therefore report the class-conditional and structured-output analyses alongside Table 3 rather than treating one summary score as sufficient evidence of zero-shot detection ability.

B.4. Fine-Tuned MLLM Protocol Sensitivity and Class-Conditional Behavior

We evaluate Skyra under two temporal-label settings. The official prompt prefixes each of the 16 sampled frames with its absolute timestamp. The frame-index prompt replaces these prefixes with `[Frame=01]` through `[Frame=16]`, while preserving the images and their order. Both settings cover the nine RA-Bench generation sources and the auxiliary Wan2.2 fixed-duration control. Because the visual input is unchanged, the comparison measures sensitivity to the temporal-label representation.

Table 19 shows that replacing timestamps with frame indices changes the class preference rather than reducing recall for both classes. Skyra-SFT gains 4.0 points in source-equal FakeR but loses 32.0 points in full-set RealR; Skyra-RL gains 5.5 points in FakeR but loses 34.5 points in full-set RealR. Both BAcc values consequently fall to about 55%.

Our audit of the released ViF-CoT-4K and ViF-Bench metadata (Li et al., 2025) finds a label-correlated temporal grid. In ViF-CoT-4K, the final timestamp is exactly 5.00 seconds for 81 of 2,017 real samples (4.0%) and 1,074 of 2,017 generated samples (53.2%). In ViF-Bench, the corresponding proportions are 97 of 165 real samples (58.8%) and 2,994 of 2,997 generated samples (99.9%). Because these timestamps are exposed directly in the prompt, the final temporal tag can act as a label prior.

Table 20 and Figure 18 show the corresponding behavior on RA-Bench. Under the official-timestamp prompt, every eligible source has higher fake recall for exact-5-second clips than for other clips. The source-equal gap ranges from 29.0 to 82.8 points for Skyra-SFT and from 23.7 to 80.3 points for Skyra-RL. Replacing timestamps with frame indices reduces the mean gap from 48.6 to 1.5 points

Table 19 | Temporal-label controls for Skyra. FakeR and BAcc are source-equal results over the nine RA-Bench generation sources, whereas Full RealR is measured on all 1,830 real anchors; Wan2.2 fixed* is excluded from the benchmark averages. Δ_5 is the source-equal fake-recall difference between exact-5-second and other clips over the eight RA-Bench generation sources containing both strata. Values are in %, and Δ_5 is in percentage points.

Model	Temporal label	Full RealR	FakeR	BAcc	Δ_5
Skyra-SFT	official timestamp	87.4	49.6	68.5	+48.6
	frame index	55.4	53.6	54.4	+1.5
Skyra-RL	official timestamp	84.0	55.1	69.5	+43.4
	frame index	49.5	60.6	54.9	+1.4

Table 20 | Fake recall by final temporal tag. Results cover the nine benchmark sources and the auxiliary Wan2.2 fixed-duration control. *5* denotes clips whose final official timestamp token is exactly [T=5.00 s]; *other* denotes all remaining clips. Wan2.2 fixed* contains only the 5-second stratum, whereas Hailuo contains no exact-5-second clip. Replacing timestamps with frame indices removes the large 5-versus-other gap on all eight RA-Bench generation sources containing both strata.

	Count		Skyra-SFT				Skyra-RL			
			Timestamp		Frame index		Timestamp		Frame index	
Source	5	other	5	other	5	other	5	other	5	other
Open-source generators										
Wan2.2 dynamic	74	1,756	97.3	31.1	63.5	65.1	98.6	35.6	70.3	70.6
Wan2.2 fixed*	1,830	0	96.4	–	62.8	–	97.7	–	68.4	–
Wan2.2-Lightning	74	1,756	94.6	29.8	66.2	61.8	97.3	34.8	70.3	67.6
LTX	96	1,734	100.0	59.8	3.1	6.9	99.0	68.2	5.2	10.1
OmniWeaving	74	1,756	95.9	13.1	47.3	46.3	97.3	17.0	58.1	56.3
Closed-source generators										
HappyHorse	302	1,485	97.4	68.4	66.6	63.3	98.3	74.7	72.5	70.0
Runway	300	1,505	97.7	68.3	69.3	66.1	98.7	74.2	75.3	72.8
Kling	298	1,492	98.7	67.7	64.1	62.2	99.0	74.9	74.2	70.9
Seedance2.0	269	1,255	97.0	51.6	55.4	51.7	97.8	59.0	64.7	60.9
Hailuo	0	1,830	–	22.5	–	56.6	–	28.3	–	64.6

and from 43.4 to 1.4 points, respectively.

The two single-stratum sources provide complementary evidence. Wan2.2 fixed contains only exact-5-second clips; frame indices reduce FakeR from 96.4% to 62.8% for Skyra-SFT and from 97.7% to 68.4% for Skyra-RL. Hailuo has no exact-5-second clips, and FakeR instead rises from 22.5% to 56.6% and from 28.3% to 64.6%. These opposite shifts are consistent with a temporal prior that favors a fake verdict at 5 seconds and a real verdict otherwise.

Switching to frame indices does not eliminate source dependence. Under frame indices, source-wise BAcc still ranges from 31.1% to 61.0% for Skyra-SFT and from 29.7% to 61.3% for Skyra-RL. The source patterns of SFT and RL remain strongly aligned under both timestamps (Spearman $\rho = 0.93$) and frame indices ($\rho = 0.95$). In particular, both variants retain very low frame-index FakeR on LTX. RL therefore shifts the operating point but does not remove the shared source-specific failure pattern.

BusterX++ exhibits a different class-conditional failure. Across the nine benchmark sources and Wan2.2 fixed*, its released pipeline attains only 4.1–9.1% FakeR while RealR remains at 93.0–93.8%. It predicts *Real* for approximately 93.6% of the paired inputs, largely independent of generation source. This behavior explains why its macro-F1 remains low despite high RealR in Table 4. The Skyra and BusterX++ results therefore show why balanced and class-conditional metrics are necessary

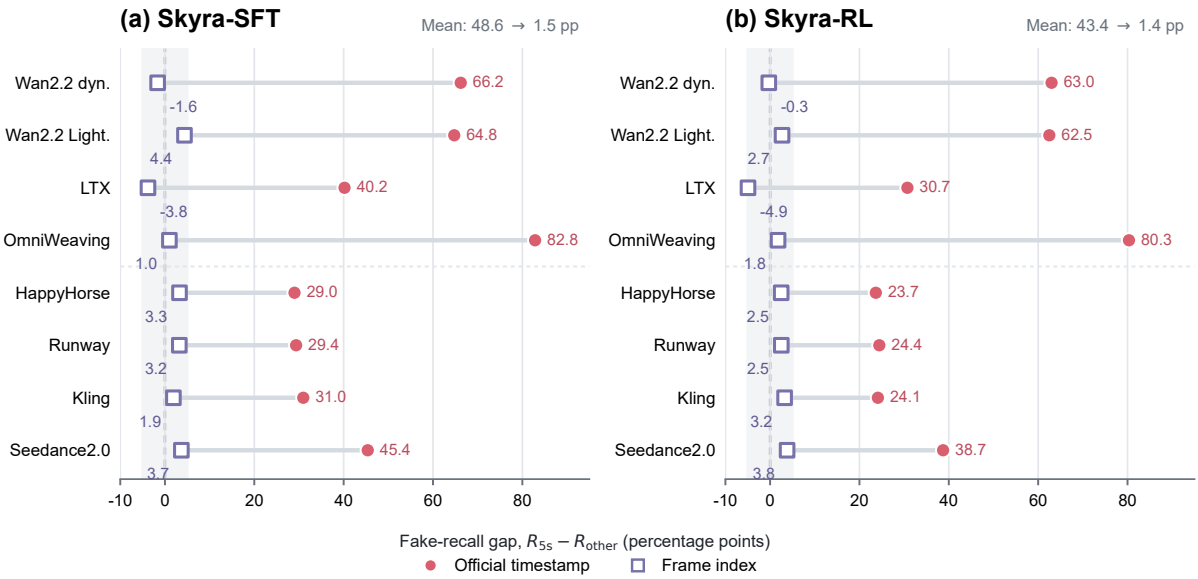


Figure 18 | Exact-5-second fake-recall gap under the Skyra prompt ablation. For each of the eight RA-Bench generation sources containing both strata, points show $\Delta_5 = R_{5s} - R_{other}$. Replacing timestamps with frame indices reduces the source-equal mean gap from 48.6 to 1.5 points for Skyra-SFT and from 43.4 to 1.4 points for Skyra-RL. Both prompts receive the same sampled frames. The auxiliary Wan2.2 fixed* control and Hailuo are excluded because each contains only one stratum.

when evaluating fine-tuned MLLMs.

C. Generation Factors and Robustness Analysis

C.1. Generation Quality and Detectability

VBench++ I2V protocol. We run the released VBench++ I2V evaluators on all 16,056 generated clips. The evaluation covers Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Imaging Quality, together with Video–Image Subject Consistency and Video–Image Background Consistency. The official VBench-I2V Quality Score combines the first six dimensions after normalization with the released ranges, using weight 0.5 for Dynamic Degree and weight 1 for each remaining dimension. We define Condition Fidelity as the mean of the two normalized Video–Image consistency scores and Combined Quality as the equal-weight mean of the VBench-I2V Quality Score and Condition Fidelity. Camera Motion is excluded because RA-Bench prompts do not provide the controlled camera-motion labels required by that evaluator. Condition Fidelity and Combined Quality are therefore benchmark-specific aggregates rather than the official VBench-I2V I2V Score and Total Score.

All eight dimension scores are produced by the released evaluators on RA-Bench inputs. Because RA-Bench uses real-event conditioning images and prompts rather than the VBench++ Image Suite, these results characterize the present benchmark and are not intended for direct comparison with the official leaderboard. Source-level Subject and Background Consistency retain the released frame-level aggregation. The association analysis instead uses one score vector per clip so that video duration does not determine statistical weight. All eight dimensions are complete for all 16,056 clips.

Association analysis. For each continuous score, we compute percentile ranks within every generation source and Dynamic Degree group. We fit a model with fixed effects for each source–group

combination, give every source equal total weight, and report the estimated change associated with an interquartile increase. This specification retains the continuous scores rather than dividing clips into discrete quality groups. Gemini Binary and Diagnostic outputs are represented by fake indicators, while Rating is scaled to $[0, 1]$. Traditional-detector fake scores are converted to empirical percentiles within each detector and source before averaging across the seven detectors, preventing detector-specific score scales from dominating the mean. The valid sample counts are 15,347 for Binary, 15,890 for Diagnostic, 15,451 for Rating, and 16,052 for the traditional-detector mean.

Dynamic Degree is binary in the released evaluator and is analyzed separately through a dynamic-minus-static contrast with generation-source and real-anchor fixed effects. Confidence intervals use cluster-robust standard errors at the real-anchor level to account for multiple generated clips derived from the same anchor. All estimates are interpreted as adjusted associations within RA-Bench rather than causal effects of quality on detector behavior.

Table 21 | Source-level VBench++ profile on RA-Bench. (a) Aggregate scores and conditioning consistency. The VBench-I2V Quality Score follows the released six-dimension normalization and weighting. Condition Fidelity averages normalized Video–Image Subject and Background Consistency, and Combined Quality gives the two aggregates equal weight. I2V-S and I2V-B are the corresponding raw evaluator outputs. (b) Raw means for the six dimensions of the VBench-I2V Quality Score. Source-level Subject and Background Consistency follow the released frame-level aggregation. All values except *N* are percentages. These are custom-input results on RA-Bench rather than VBench++ Image Suite leaderboard scores.

(a) Aggregate scores and conditioning consistency

Generation source	*N*	Quality Score	Condition Fidelity	Combined Quality	I2V-S	I2V-B
Open-source models						
Wan2.2 dynamic	1,830	76.2	94.0	85.1	94.4	95.9
Wan2.2-Lightning	1,830	76.4	96.4	86.4	96.8	97.4
LTX	1,830	74.9	92.5	83.7	93.5	94.5
OmniWeaving	1,830	72.9	95.3	84.1	95.8	96.6
Closed-source models						
HappyHorse	1,787	76.9	92.6	84.8	93.6	94.6
Runway	1,805	78.0	95.0	86.5	95.5	96.5
Kling	1,790	76.7	93.1	84.9	93.9	95.0
Seedance2.0	1,524	76.4	91.1	83.7	92.3	93.5
Hailuo	1,830	77.0	96.3	86.7	96.7	97.4

(b) Six VBench-I2V Quality Score dimensions

Generation source	Subject	Background	Motion	Dynamic	Aesthetic	Imaging
Open-source models						
Wan2.2 dynamic	88.1	92.0	98.0	69.4	50.6	66.1
Wan2.2-Lightning	89.9	92.1	98.2	63.7	50.0	67.4
LTX	86.0	89.9	98.7	74.4	47.4	62.8
OmniWeaving	87.5	91.0	99.0	52.8	47.9	58.1
Closed-source models						
HappyHorse	88.4	91.8	98.7	66.7	50.2	69.4
Runway	88.9	91.3	98.8	76.8	50.5	69.5
Kling	90.3	93.4	99.0	63.4	49.3	65.1
Seedance2.0	88.5	92.2	98.6	63.6	49.6	67.9
Hailuo	89.6	93.0	99.0	63.3	50.0	66.3

The source-level profile is descriptive rather than the basis of the main analysis. The VBench-I2V Quality Score spans 72.9–78.0, Condition Fidelity spans 91.1–96.4, and Combined Quality spans 83.7–86.7. These narrow ranges do not imply similar detection difficulty. LTX and Seedance2.0, for

example, both obtain a Combined Quality score of 83.7, while their Gemini Diagnostic fake recalls are 74.5% and 32.3%, respectively. We therefore rely on within-source clip-level associations rather than source-level quality rankings.

Table 22 | Adjusted associations between VBench++ scores and fake-side detector outputs. Except for Dynamic Degree, each cell reports the estimated change associated with an interquartile increase within each source and Dynamic Degree group, followed by its 95% confidence interval. Negative values indicate weaker fake evidence at higher scores. Dynamic Degree reports the adjusted dynamic-minus-static contrast. Gemini Binary and Diagnostic values are percentage points of fake recall, Gemini Rating is expressed on a 0–100 scale, and traditional values are percentile points of the mean normalized fake score across seven detectors.

Quality score or dimension	Gemini Binary	Gemini Diagnostic	Gemini Rating	7-detector mean
Aggregate scores				
VBench-I2V Quality Score	−6.1 [−8.2, −3.9]	−9.6 [−11.7, −7.5]	−6.4 [−8.1, −4.6]	−11.5 [−12.3, −10.8]
Condition Fidelity	−10.6 [−12.5, −8.6]	−14.4 [−16.3, −12.5]	−10.2 [−11.8, −8.5]	−9.5 [−10.1, −8.8]
Combined Quality	−8.2 [−10.3, −6.1]	−12.4 [−14.4, −10.4]	−8.3 [−10.0, −6.6]	−10.9 [−11.6, −10.2]
Condition fidelity and temporal consistency				
Video–Image subject consistency	−9.6 [−11.6, −7.7]	−13.2 [−15.1, −11.3]	−9.4 [−11.0, −7.7]	−9.3 [−10.0, −8.7]
Video–Image background consistency	−10.9 [−12.8, −8.9]	−14.9 [−16.7, −13.0]	−10.3 [−11.9, −8.7]	−9.0 [−9.7, −8.3]
Subject consistency	−11.4 [−13.4, −9.3]	−14.8 [−16.7, −12.9]	−10.7 [−12.4, −9.0]	−10.1 [−10.8, −9.4]
Background consistency	−9.4 [−11.4, −7.4]	−13.6 [−15.5, −11.7]	−9.3 [−10.9, −7.6]	−7.6 [−8.3, −6.9]
Motion smoothness	−7.8 [−9.7, −5.9]	−7.9 [−9.7, −6.0]	−7.0 [−8.6, −5.5]	−2.1 [−2.9, −1.4]
Frame-level quality				
Aesthetic quality	1.4 [−0.7, 3.6]	−0.4 [−2.6, 1.7]	−0.8 [−2.6, 1.0]	−4.1 [−4.9, −3.2]
Imaging quality	1.7 [−0.4, 3.9]	0.6 [−1.5, 2.7]	1.8 [0.1, 3.6]	−9.4 [−10.2, −8.7]
Motion amount				
Dynamic Degree	6.5 [4.0, 9.0]	8.4 [5.8, 10.9]	5.5 [3.3, 7.7]	0.1 [−0.4, 0.6]

Table 22 shows three distinct patterns. Condition Fidelity has the largest aggregate association with all three Gemini outputs, whereas the VBench-I2V Quality Score has the largest association with the traditional-detector mean. The four subject and background consistency measures are negative across every detector output, while Aesthetic Quality and Imaging Quality show no clear association with Gemini Diagnostic. Dynamic Degree changes in the opposite direction for Gemini but is nearly null for the traditional-detector mean. A single aggregate quality score therefore combines dimensions with different detector-specific associations.

C.2. Generation Settings and Detectability

Generation protocol. We evaluate the same 1,830 anchor-derived prompts under T2V, first-frame I2V, and first+last-frame I2V generation. T2V uses no real-image condition, while the two I2V settings use the matched first frame or the matched first and last frames, respectively; the first-frame setting is the protocol used by RA-Bench. We use the same anchor-derived prompts and seeds across all three settings.

We generate all three settings for seeds 0, 42, and 123. The main comparison uses seed 0 and evaluates the seven traditional detectors, both official-timestamp and frame-index variants of Skyra-SFT and Skyra-RL, and the released BusterX++ pipeline. Traditional-detector AUC and T@5% use the same frozen score vectors for the 1,830 paired real anchors. For the fine-tuned MLLMs, the real-video control is fixed within each configuration, and we report BAcc, FakeR, and macro-F1. Unparseable BusterX++ responses are retained as abstentions and counted as incorrect. Table 23 gives the complete seed-0 classification results.

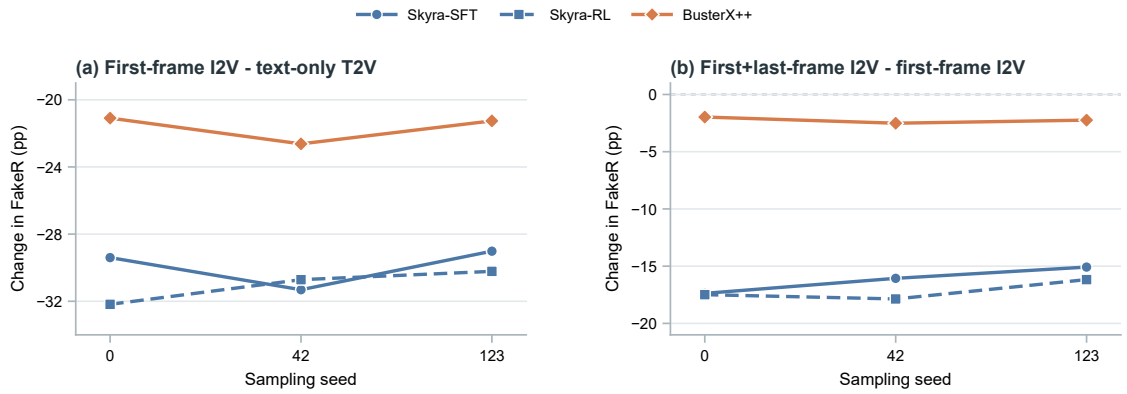


Figure 19 | Cross-seed stability of the generation-setting effects. Panel (a) reports first-frame-I2V-minus-T2V changes in FakeR, and panel (b) reports first+last-frame-I2V-minus-first-frame-I2V changes. Skyra-SFT and Skyra-RL share the same color and are distinguished by solid and dashed lines, while BusterX++ is shown in orange. Negative values indicate lower FakeR in the second setting. The panels use different vertical ranges to show cross-seed variation at the scale of each contrast.

Table 23 | Complete seed-0 fine-tuned MLLM results across generation settings. The real-video control is fixed within each configuration. RealR is listed in official-timestamp/frame-index order for Skyra. All values are percentages; Δ is first+last minus first frame. BusterX++ abstentions are counted as incorrect.

Configuration	Metric	T2V	First frame (I2V)	First+last frames (I2V)	End-frame effect Δ
Skyra-SFT (RealR: 87.4 / 55.4)					
Official timestamp	BAcc	75.3	60.6	51.9	−8.7
	FakeR	63.2	33.8	16.4	−17.4
	macro-F1	74.9	57.5	45.0	−12.6
Frame index	BAcc	75.2	60.2	50.6	−9.6
	FakeR	94.9	65.1	45.8	−19.3
	macro-F1	74.1	60.1	50.5	−9.7
Skyra-RL (RealR: 84.0 / 49.5)					
Official timestamp	BAcc	77.2	61.1	52.3	−8.7
	FakeR	70.4	38.2	20.7	−17.5
	macro-F1	77.1	58.9	47.0	−11.9
Frame index	BAcc	73.3	60.0	52.2	−7.8
	FakeR	97.2	70.5	54.9	−15.6
	macro-F1	71.7	59.6	52.2	−7.4
BusterX++ (RealR: 93.7)					
Released pipeline	BAcc	60.4	49.8	48.9	−1.0
	FakeR	27.0	6.0	4.0	−2.0
	macro-F1	55.4	37.9	36.0	−1.9

Because the real-video control is fixed within each configuration, differences in BAcc across generation settings are driven by FakeR. Both Skyra prompt variants follow the same direction when the matched last frame is added: FakeR decreases by 17.4 and 19.3 points for Skyra-SFT and by 17.5 and 15.6 points for Skyra-RL. BusterX++ decreases by only 2.0 points because its FakeR is already 6.0% under first-frame I2V.

For the cross-seed analysis, Table 24 and Figure 19 report the released official-timestamp prompts for Skyra-SFT and Skyra-RL together with BusterX++. This matches the protocol used for the seed analysis in Section 4.2.3 while retaining one released setting for each fine-tuned MLLM detector.

Table 24 | Fine-tuned MLLM FakeR across generation settings and seeds. Each setting contains the same 1,830 anchors for seeds 0, 42, and 123. Skyra uses the official-timestamp prompt, matching the paper's cross-seed protocol. End-frame effect is first+last minus first frame; Seed range is the maximum minus the minimum across the three seeds. All values are percentages. BusterX++ abstentions are counted as incorrect.

Configuration	Generation setting	Seed 0	Seed 42	Seed 123	Seed range
Skyra-SFT					
Official timestamp	T2V	63.2	64.2	60.4	3.7
	First frame	33.8	32.8	31.4	2.4
	First+last frames	16.4	16.8	16.3	0.4
	End-frame effect	−17.4	−16.1	−15.1	2.3
Skyra-RL					
Official timestamp	T2V	70.4	70.2	67.9	2.5
	First frame	38.2	39.5	37.7	1.8
	First+last frames	20.7	21.6	21.5	0.9
	End-frame effect	−17.5	−17.9	−16.2	1.7
BusterX++					
Released pipeline	T2V	27.0	29.2	27.6	2.2
	First frame	6.0	6.6	6.3	0.7
	First+last frames	4.0	4.1	4.1	0.1
	End-frame effect	−2.0	−2.5	−2.2	0.5

The direction of each change is unchanged across the three seeds. For first+last-frame versus first-frame I2V, the FakeR decrease ranges from 15.1 to 17.4 points for Skyra-SFT, from 16.2 to 17.9 points for Skyra-RL, and from 2.0 to 2.5 points for BusterX++. The small cross-seed ranges show that the seed-0 results in the main paper do not arise from a single sampled realization.

C.3. Stability Across Generation Seeds

Controlled comparison. The primary RA-Bench clips from the four open-source generators use seed 0. We generate two additional sets with seeds 42 and 123, each covering the same 1,830 real-video anchors per generator. For every source–anchor pair, the prompt, conditioning image, and all generation settings other than the seed are unchanged. The paired real-video set is also shared across the three evaluations. Closed-source providers are omitted because their APIs do not expose a controllable seed, and the fixed-duration Wan2.2 control is omitted because this analysis concerns the four open-source RA-Bench generators.

Traditional detectors. The source-level seven-detector AUC means in Table 6(a) vary by at most 1.05 points across seeds. The seed-0 source means differ from their three-seed averages by at most 0.20 AUC points. Table 25 reports all 28 detector–source comparisons and shows where the remaining variation occurs.

The largest detector-specific range is 5.74 AUC points for ReStraV on LTX. The 28-cell detector–source pattern nevertheless remains highly similar across seeds. Pairwise Spearman correlations over all 28 AUC values are 0.981, 0.978, and 0.989 for seed pairs 0/42, 0/123, and 42/123, respectively. Only NPR on Wan2.2 dynamic, D3 on Wan2.2-Lightning, and UnivFD on LTX cross 50% AUC, and each remains close to random ranking under all three seeds.

Fine-tuned MLLMs. The seed changes only the generated-video side of each paired evaluation, so RealR is constant for a given model. The resulting BAcc range is therefore half of the FakeR range. Table 26 reports the complete source-wise comparison.

The largest FakeR ranges are 2.35 points for Skyra-SFT, 2.46 points for Skyra-RL, and 0.71 points

for BusterX++; the corresponding BAcc ranges are 1.17, 1.23, and 0.36 points. Skyra retains the same pronounced source differences under all three seeds. BusterX++ remains below 10% FakeR for every source–seed combination. The source-level conclusions are therefore not driven by the seed-0 generations used in the primary benchmark.

Table 25 | Detector-specific traditional results across generation seeds. Each cell reports AUC. Δ is the maximum minus the minimum across seeds. All values are percentages; bold marks the largest range in the table.

Detector	Seed 0	Seed 42	Seed 123	Δ
Wan2.2 dynamic				
CNNSpot	35.50	34.78	35.38	0.72
NPR	50.64	49.79	50.38	0.85
UnivFD	32.72	33.23	33.84	1.11
ForgeLens	59.99	60.23	59.89	0.34
DeCoF	62.47	63.18	63.29	0.82
D3	55.36	53.70	54.93	1.67
ReStraV	59.28	57.84	57.32	1.96
Wan2.2-Lightning				
CNNSpot	43.78	44.36	43.54	0.82
NPR	55.17	54.89	55.03	0.29
UnivFD	42.96	44.28	43.76	1.32
ForgeLens	69.61	70.04	69.85	0.43
DeCoF	56.70	56.95	57.05	0.36
D3	49.09	50.05	49.65	0.97
ReStraV	62.41	60.86	62.40	1.55
LTX				
CNNSpot	63.99	62.02	60.29	3.70
NPR	58.65	59.74	58.62	1.11
UnivFD	51.16	50.03	49.54	1.62
ForgeLens	64.11	62.07	61.87	2.24
DeCoF	62.47	63.42	61.07	2.35
D3	58.01	58.90	57.77	1.13
ReStraV	42.45	48.19	47.87	**5.74**
OmniWeaving				
CNNSpot	45.76	45.99	45.03	0.96
NPR	50.93	51.94	51.69	1.01
UnivFD	39.88	39.94	39.94	0.06
ForgeLens	62.95	62.83	63.34	0.51
DeCoF	72.11	72.26	71.48	0.77
D3	52.85	52.91	52.70	0.22
ReStraV	68.34	70.23	68.64	1.89

Table 26 | Source-wise seed sensitivity of the fine-tuned MLLMs. Skyra-SFT and Skyra-RL use the official-timestamp prompt, and BusterX++ uses its released evaluation pipeline. FakeR and BAcc are reported in %; Δ denotes the maximum minus the minimum across seeds. Bold marks the largest range for each metric.

Generation source	FakeR			Δ FakeR	Δ BAcc
	Seed 0	Seed 42	Seed 123		
Skyra-SFT					
Wan2.2 dynamic	33.77	32.84	31.42	2.35	1.17
Wan2.2-Lightning	32.46	30.78	30.11	2.35	1.17
LTX	61.91	62.90	63.01	1.09	0.55
OmniWeaving	16.45	17.32	16.07	1.26	0.63
Skyra-RL					
Wan2.2 dynamic	38.20	39.45	37.65	1.80	0.90
Wan2.2-Lightning	37.32	36.50	34.86	**2.46**	**1.23**
LTX	69.84	70.44	70.60	0.77	0.38
OmniWeaving	20.22	20.66	20.49	0.44	0.22
BusterX++					
Wan2.2 dynamic	5.96	6.61	6.34	0.66	0.33
Wan2.2-Lightning	9.07	9.18	8.58	0.60	0.30
LTX	4.10	3.88	3.61	0.49	0.25
OmniWeaving	6.12	6.83	6.39	0.71	0.36

D. Human Evaluation and Social Dissemination

D.1. Human Evaluation Protocol and Source-Level Recognition

Analysis set. The human study is drawn from the 17,886-video RA-Bench pool: 1,830 real anchors and 16,056 generated clips from four open-source generators and five closed-source providers. The reported Stage 1 results use the standard three-review stream, which contains 17,850 videos and 53,550 judgments. This analysis set comprises 1,812 real videos and 16,038 generated videos, with each video assigned to three different reviewers. The remaining 36 videos were reserved for internal assignment-level quality control and are excluded from the reported recognition statistics and RA-Bench-HumanProof construction.

Reviewers, training, and assignment. The 20 Stage 1 reviewers and two additional Stage 2 reviewers are undergraduate and graduate students recruited from universities in China and abroad. Before annotation, all reviewers complete the same standardized training on the review interface and label definitions. In Stage 1, the interface presents one video at a time together with an optional 16-frame contact sheet and asks the reviewer to choose *Real*, *Uncertain*, or *Generated*. No source or label information is shown. Real and generated videos from different sources are interleaved in a reviewer-specific randomized order, and assignments are balanced across reviewers. Reviewers may replay a video and revise their immediately preceding response. The interface does not enforce a minimum viewing time, but reviewers are instructed to inspect each video before submitting.

Source-level recognition. A judgment is counted as correct when a real video is labeled *Real* or a generated video is labeled *Generated*; *Uncertain* is reported separately and is not counted as correct. Table 27 reports both the number of reviewed items and the response distribution for every source.

Table 27 | Stage-1 human recognition by source. The standard analysis stream contains three judgments per video. Response shares are in %; for generated sources, *Judged Generated* is human FakeR. Open and Closed averages give equal weight to each generation source, whereas Generated total pools all generated-video judgments.

Source-level response distribution					
Video source	**Videos**	**Judgments**	**Judged Real**	**Uncertain**	**Judged Generated**
Real videos	1,812	5,436	**71.9**	5.3	22.8
Open-source generators					
Wan2.2 dynamic	1,828	5,484	27.0	5.5	67.5
Wan2.2-Lightning	1,828	5,484	28.5	4.9	66.6
LTX	1,828	5,484	30.3	5.1	64.6
OmniWeaving	1,828	5,484	19.5	4.7	75.8
Open avg.	–	–	**26.3**	**5.0**	**68.6**
Closed-source generators					
HappyHorse	1,785	5,355	37.2	6.9	55.9
Runway	1,803	5,409	34.8	5.6	59.7
Kling	1,788	5,364	47.7	7.3	45.1
Seedance2.0	1,522	4,566	51.9	7.4	40.7
Hailuo	1,828	5,484	31.8	5.3	63.0
Closed avg.	–	–	**40.6**	**6.5**	**52.9**
Generated total	**16,038**	**48,114**	**33.9**	**5.8**	**60.3**

Overall, 60.3% of judgments on generated videos and 71.9% of judgments on real videos are correct. Generated-video recognition varies substantially by source. The source-equal human FakeR is 68.6% for open-source generators and 52.9% for closed-source providers. Seedance2.0 and Kling are the most difficult to recognize as generated, with FakeR values of 40.7% and 45.1%, whereas OmniWeaving reaches 75.8%. The pooled human FakeR of 60.3% therefore masks these source-level differences.

Reviewer-level variation and source-level stability. Reviewers use different response thresholds: individual FakeR ranges from 18.8% to 87.5%, RealR from 13.7% to 93.3%, and the share of *Uncertain* responses from 0.0% to 30.4%. The source-level result is nevertheless stable. Every reviewer obtains higher source-equal FakeR on open-source than on closed-source generators; the reviewer-level difference has a median of 15.7 points and ranges from 0.6 to 38.2 points. Leaving out any one reviewer preserves the complete nine-source ranking, and the largest change in a source-level response rate is 2.9 points.

D.2. RA-Bench-HumanProof Construction and Detector Evaluation

Two-stage selection. Stage 1 retains a generated video as a candidate only when all three assigned reviewers label it *Real*. This criterion selects 1,080 of the 16,038 generated videos in the primary analysis. Two additional reviewers then independently reassess all candidates using the same interface and response options, without source or class labels. A candidate enters RA-Bench-HumanProof only when both additional reviewers also label it *Real*; any *Uncertain* or *Generated* response excludes the video. The second stage retains 633 videos, or 58.6% of the Stage-1 candidates. Each retained video therefore receives five *Real* judgments.

Table 28 reports the source composition before and after Stage 2 together with the complete response-pair counts.

Table 28 | Two-stage construction and source composition of RA-Bench-HumanProof. (a) Source-wise retention from generated videos labeled *Real* by all three Stage 1 reviewers to those also labeled *Real* by both Stage 2 reviewers. Retention rates are in %. (b) Complete Stage 2 response pairs; only the Real/Real cell enters RA-Bench-HumanProof.

(a) Source-wise retention

Source	Stage 1 3/3 Real	Final 5/5 Real	Retained
Open-source generators			
Wan2.2 dynamic	49	21	42.9
Wan2.2-Lightning	77	40	51.9
LTX	97	49	50.5
OmniWeaving	23	9	39.1
Open total	**246**	**119**	**48.4**
Closed-source generators			
HappyHorse	134	74	55.2
Runway	122	74	60.7
Kling	235	160	68.1
Seedance2.0	249	159	63.9
Hailuo	94	47	50.0
Closed total	**834**	**514**	**61.6**
All sources	**1,080**	**633**	**58.6**

(b) Stage-2 response pairs

Reviewer 1	Reviewer 2	Count	Kept
Real	Real	633	Yes
Real	Generated	276	No
Generated	Real	86	No
Generated	Uncertain	1	No
Generated	Generated	84	No

RA-Bench-HumanProof contains 119 open-source and 514 closed-source videos. Closed-source candidates have a higher retention rate than open-source candidates (61.6% versus 48.4%). Kling and Seedance2.0 contribute 160 and 159 videos, respectively, and together account for 50.4% of RA-Bench-HumanProof. This composition follows directly from the five-reviewer selection criterion rather than a predefined source quota.

Paired evaluation and source-matched reference. Every generated video in RA-Bench-HumanProof retains the identifier of its real anchor. Detector evaluation therefore uses 633 generated–real pairs, corresponding to 511 unique real videos because several generated videos may share an anchor. For discrete outputs, we report BAcc together with FakeR and RealR. For continuous scores, we report paired AUC and the fake-video true-positive rate at a 5% real-video false-positive rate (T@5%).

RA-Bench-HumanProof contains a larger share of Seedance2.0 and Kling videos than full RA-Bench. An unweighted comparison would therefore combine human-selection effects with a change in source composition. For each metric, we compute the RA-Bench reference as $\sum_s n_s M_s / \sum_s n_s$, where M_s is the full-RA-Bench result for source s and n_s is its RA-Bench-HumanProof count. This reference matches the RA-Bench-HumanProof source proportions while retaining all videos within each RA-Bench source. It controls the source mixture, but not the conditional selection induced by the five human judgments.

Table 29 | Complete detector results on RA-Bench-HumanProof. (a) Continuous-score configurations report paired AUC and T@5%. (b) Discrete-output configurations report BAcc, FakeR, and RealR. *Matched ref.* reports the corresponding AUC/T@5% or BAcc/FakeR pair on source-matched RA-Bench, and Δ is the RA-Bench-HumanProof AUC or BAcc minus its matched reference. All values are in %. Skyra official-timestamp results follow the released protocol; frame index removes the numerical timestamp values.

(a) Continuous-score configurations

Configuration	RA-Bench-HumanProof AUC	T@5%	Matched ref. AUC	Matched ref. T@5%	Δ
Traditional detectors					
CNNSpot	40.5	3.6	43.5	3.8	-3.0
NPR	50.3	3.8	50.4	4.0	-0.1
UnivFD	39.4	0.6	39.8	1.0	-0.4
ForgeLens	54.9	8.8	58.4	10.4	-3.5
DeCoF	59.0	1.9	59.2	1.7	-0.2
D3	33.1	6.8	39.5	6.2	-6.4
ReStraV	55.2	5.2	55.6	4.9	-0.4
7-detector mean	**47.5**	**4.4**	**49.5**	**4.6**	**-2.0**
Zero-shot multimodal models					
Gemini-3.1-Pro-Preview Rating	54.9	4.3	61.5	5.6	-6.6

(b) Discrete-output configurations

Configuration	BAcc	FakeR	RealR	Matched ref. BAcc	Matched ref. FakeR	Δ
Zero-shot multimodal models						
Gemini-3.1-Pro-Preview Binary	54.7	34.3	75.2	61.2	49.9	-6.5
Gemini-3.1-Pro-Preview Diagnostic	54.5	30.0	79.0	61.0	47.3	-6.5
Fine-tuned MLLM detectors						
Skyra-SFT, official timestamp	72.0	55.1	88.8	73.9	60.4	-1.9
Skyra-SFT, frame index	53.7	51.3	56.0	55.3	55.7	-1.6
Skyra-RL, official timestamp	74.5	63.0	85.9	75.1	66.3	-0.6
Skyra-RL, frame index	53.7	58.8	48.7	56.1	63.3	-2.4
BusterX++, released pipeline	49.4	3.9	94.9	49.9	6.2	-0.5

All configurations in Table 29 contain predictions for all 633 RA-Bench-HumanProof videos. The traditional detectors, Gemini, Skyra-RL, and BusterX++ also contain 633 matched-real predictions. Skyra-SFT contains 632 valid matched-real predictions after excluding one invalid record; its BAcc and RealR use the available predictions.

Detector-family results. The seven traditional detectors show a modest mean AUC decrease of 2.0 points, from 49.5% under source-matched RA-Bench weighting to 47.5% on RA-Bench-HumanProof. Their individual AUCs span only 33.1%–59.0%, and their mean T@5% changes from 4.6% to 4.4%. These small or heterogeneous changes do not indicate robustness: the source-matched baseline is already close to random ranking and provides little generated-video recall at a 5% false-positive rate.

Gemini shows the clearest alignment with human difficulty. Binary and Diagnostic BAcc each decrease by 6.5 points, while their FakeR decreases by 15.6 and 17.3 points, respectively. Rating AUC decreases by 6.6 points and T@5% by 1.3 points. The loss is concentrated on the generated side: Binary and Diagnostic FakeR fall to 34.3% and 30.0%, while RealR remains 75.2% and 79.0%. On RA-Bench-HumanProof, Gemini therefore provides substantially weaker fake evidence without a comparable loss on the matched real videos.

The fine-tuned MLLMs require a different interpretation. Their BAcc changes by at most 2.4 points, and FakeR decreases by 2.3–5.3 points across all five configurations. Skyra-SFT and Skyra-RL retain

72.0% and 74.5% BAcc with official timestamps, but these configurations preserve the timestamp prior identified in Section 4.1.3. Replacing timestamps with frame indices reduces both checkpoints to 53.7% BAcc. BusterX++ reaches 49.4% BAcc by labeling only 3.9% of generated videos as fake while retaining 94.9% of real videos. Their limited changes from the source-matched reference therefore reflect an already weak visual decision rule or a strong class preference, not reliable detection of human-deceptive content.

Interpretation and scope. RA-Bench-HumanProof separates three failure patterns: Gemini loses much of its fake-side evidence on videos that mislead reviewers; traditional detectors remain weak before and after human selection; and the higher official-timestamp Skyra results retain a protocol prior. Human and detector failures therefore overlap, but they are not equivalent. The most consequential cases lie at their intersection: videos that repeatedly appear real to reviewers also receive weak or unreliable fake evidence from the evaluated detector families.

RA-Bench-HumanProof is constructed conditionally and does not replace full RA-Bench. Stage 2 responses determine membership and therefore cannot provide an independent estimate of human accuracy on the retained videos. Its source distribution is skewed toward Seedance2.0 and Kling, which motivates the source-matched reference. The 633 generated–real pairs also include repeated real anchors; interval estimates should therefore resample the real-anchor identifier rather than treat all pairs as independent.

D.3. RA-Bench-LastMile Protocol and Complete Results

Evaluation subset. RA-Bench-LastMile uses anchors for which a real video and generated videos from all nine RA-Bench sources are available. To preserve coverage without allowing large L2 subcategories to dominate, let N_l denote the number of common anchors in subcategory l. We retain $q_l = \lfloor 0.1N_l + 0.5 \rfloor$ anchors when $N_l > 10$, one anchor when $1 \leq N_l \leq 10$, and none when $N_l = 0$. A fixed SHA-256 ordering of normalized clip identifiers determines the retained anchors within each subcategory. This procedure selects 150 real-event anchors spanning 41 of the 44 L2 subcategories. Each condition contains these 150 real videos and 1,350 matched generated videos, for 1,500 videos per condition and 9,000 video instances across the six conditions. The Wan2.2 fixed-duration variant serves as an auxiliary control and is not included as a separate generation source.

Social dissemination simulation. All operations are applied identically to a generated video and its matched real anchor. Table 30 summarizes the six conditions. T1 encodes each video with VP9 (CRF 36, `-b:v 0`) and then H.264 (CRF 28, `medium` preset). The spatial, temporal, and presentation operations are each evaluated as an addition to this common transcode, and Full applies all four operations in sequence.

Table 30 | Social dissemination simulation conditions. Each condition contains 150 real videos and 1,350 matched generated videos. T2–T4 are evaluated as additions to the common T1 transcode.

Condition	Operations, in order
Original	Standardized input clip; no additional encoding.
T1	VP9 encoding followed by H.264 transcoding.
T1+T2	0.5× spatial downsampling, followed by T1.
T1+T3	Conversion to 8 fps, followed by T1.
T1+T4	Synthetic news badge, followed by T1.
Full	Spatial downsampling, 8-fps conversion, news badge, and T1.

Figure 20 | Illustration of the news-badge transformation. Matched real and Seedance2 videos are shown before and after adding the synthetic news badge. The same overlay rule is applied to both classes while preserving the remaining scene content. We show a wildfire aftermath scene (top) and an aviation accident investigation (bottom).

As illustrated in Figure 20, the synthetic badge is placed at the upper-right corner of the active picture. Its width is 10.04% of the active-picture width, with a 1.67% right margin. The overlay is an experimental presentation cue rather than source attribution and does not imply affiliation with or endorsement by any news organization. If a clip already contains an upper-right news badge, the existing badge is retained and no second badge is added. Audio tracks and container metadata are removed from every transformed condition; all transformed outputs use H.264, yuv420p, and `faststart`.

Evaluation and uncertainty. For each traditional detector, we compute metrics separately on each generation source and its matched real anchors, then average the nine source-level values with equal weight. Classification metrics for the fine-tuned MLLM detectors follow the same protocol. Confidence intervals use 2,000 bootstrap replicates, stratified by L1 category and clustered by real-event anchor, with random seed 20260718. All reported configurations contain complete predictions for the nine sources under all six conditions. We repeat the analysis after excluding manually identified clips with a pre-existing upper-right badge. This exclusion does not materially change the results: across method–condition cells, the largest absolute changes are 2.8 percentage points for AUC and 4.3 points for FakeR.

Traditional-detector sensitivity. The seven-detector mean AUC decreases by 4.2 points under Full, but this average conceals sharply different responses. ForgeLens loses 26.0 points, whereas DeCoF and D3 gain 2.4 and 4.8 points. The detector ordering is consequently unstable: its Spearman correlation with Original is 0.36 after T1, 0.07 after T1+T2, 0.46 after T1+T3, 0.29 after T1+T4, and 0.07 under Full. These gains for individual detectors do not indicate reliable performance under the social dissemination simulation because T@5% remains low throughout, reaching only 2.8% for the seven-detector mean under Full.

Fine-tuned MLLM sensitivity. The isolated additions to T1 reveal different failure patterns. Spatial downsampling produces the largest FakeR loss, reducing the five-configuration mean from 29.9% to 7.9%. Conversion to 8 fps lowers BAcc by 11.7 and 12.9 points for the two official-timestamp Skyra configurations, compared with 3.4 and 2.4 points for their frame-index controls. This gap is consistent with sensitivity to the temporal representation rather than only to visual degradation. The news badge lowers mean FakeR from 29.9% to 14.0% while increasing RealR from 79.1% to 88.5%, shifting predictions toward *Real* even though the scene content is unchanged. Under Full, every fine-tuned configuration has at most 2.4% FakeR.

Table 31 | Complete detector results on RA-Bench-LastMile. Each cell reports the primary metric on top and the secondary metric below: AUC/T@5% for traditional detectors and BAcc/FakeR for fine-tuned MLLM detectors. All values are percentages and equal-source means over the nine RA-Bench generation sources.

Detector / configuration	Original	T1	T1+T2	T1+T3	T1+T4	Full
Traditional detectors			**AUC** / T@5%			
CNNSpot	44.5 2.4	48.2 2.4	48.0 2.4	50.1 2.1	47.5 1.9	46.5 2.2
NPR	52.2 3.8	45.6 3.4	44.9 3.0	45.9 2.7	45.5 3.5	43.5 3.0
UnivFD	37.1 2.1	36.6 1.8	36.0 0.3	38.1 1.0	37.2 1.9	37.4 0.2
ForgeLens	61.6 17.3	44.7 6.1	32.6 2.2	46.3 6.5	44.9 4.8	35.6 3.0
DeCoF	59.9 3.2	63.0 3.9	60.1 0.4	65.1 5.0	63.0 3.3	62.3 1.2
D3	49.3 4.4	49.5 5.5	52.2 4.7	50.4 5.9	53.3 4.4	54.1 5.4
ReStraV	55.4 6.8	52.2 8.2	51.1 4.5	52.0 4.8	52.0 6.5	51.6 4.7
7-detector mean	**51.4** **5.7**	**48.5** **4.5**	**46.4** **2.5**	**49.7** **4.0**	**49.1** **3.8**	**47.3** **2.8**
Fine-tuned MLLM detectors			**BAcc** / FakeR			
Skyra-SFT, official timestamp	67.6 50.4	61.8 31.6	52.9 9.2	50.1 9.6	58.7 18.7	49.3 1.2
Skyra-SFT, frame index	55.9 55.2	48.9 33.7	39.6 5.1	45.5 31.6	42.7 10.8	44.4 1.4
Skyra-RL, official timestamp	68.6 55.8	62.7 36.7	53.5 11.6	49.8 12.9	60.3 22.7	48.5 1.7
Skyra-RL, frame index	56.5 61.7	49.0 41.3	36.2 8.4	46.6 39.8	44.4 17.6	43.2 2.4
BusterX++	49.1 6.9	50.0 5.9	50.5 5.0	48.9 5.2	50.1 0.2	50.1 0.2

Table 32 | Uncertainty of Full-condition changes relative to Original. Values are percentage-point changes with 95% anchor-cluster bootstrap confidence intervals.

Method	Metric	Change [95% CI]
Seven-detector mean	AUC	−4.2 [−5.6, −2.7]
Skyra-SFT, timestamp	BAcc	−18.3 [−21.1, −15.4]
Skyra-SFT, frame index	BAcc	−11.5 [−15.0, −8.3]
Skyra-RL, timestamp	BAcc	−20.0 [−23.1, −16.9]
Skyra-RL, frame index	BAcc	−13.3 [−16.7, −10.1]

Interpretation. All reported intervals remain below zero under L1-stratified, anchor-cluster resampling. BusterX++ is omitted from this change table because its BAcc remains near 50% in both conditions; its FakeR nevertheless falls from 6.9% to 0.2%, showing that a stable BAcc can conceal a stronger shift toward *Real*.