BusterX: MLLM-Powered AI-Generated Video Forgery Detection and Explanation

Haiquan Wen[1*] **Yiwei He**[1*] **Zhenglin Huang**[1*] **Tianxiao Li**[1] **Zihan Yu**[1]
Xingru Huang[5] **Lu Qi**[4] **Baoyuan Wu**[3] **Xiangtai Li**[2] **Guangliang Cheng**[1†]

[1]University of Liverpool [2]Nanyang Technological University
[3]The Chinese University of Hong Kong, Shenzhen
[4]Wuhan University [5]Hangzhou Dianzi University
[*] Equal Contribution. [†] Corresponding author: guangliang.cheng@liverpool.ac.uk
Project Page: `https://github.com/l8cv/BusterX`

Abstract

As generative video models become increasingly realistic, detecting AI-generated videos requires systems that offer both accuracy and interpretability. However, applying Multimodal Large Language Models (MLLMs) to video forensics is currently limited by outdated datasets, simplistic evaluation protocols, and a reliance on black-box classification. To address these issues, we introduce a comprehensive dataset, benchmark, and baseline model for video forgery detection. First, we present **GenBuster-200K**, a fair dataset of over 200,000 high-quality videos sourced from state-of-the-art generators, featuring diverse real-world scenarios. Second, we propose **GenBuster-Bench**, a diagnostic benchmark spanning three progressive tracks (In-Domain, Out-of-Domain, and In-the-Wild) to evaluate models across *domain shifts* and *generational shifts*. Finally, we develop **BusterX**, a simple MLLM baseline with Reinforcement Learning (RL) training. Instead of direct binary classification, BusterX formulates detection as a visual reasoning task, where the generated reasoning chain serves as detector itself. Experimental results demonstrate that BusterX outperforms several leading MLLMs in both detection accuracy and rationale quality.

1 Introduction

Recent models, such as Seedance 2.0 [7], enable users to synthesize highly realistic videos with minimal effort. While beneficial for creative applications, this accessibility also facilitates malicious uses, such as political misinformation and financial fraud. Consequently, there is a growing need for reliable and interpretable video forgery detection systems. To provide human-readable insights, recent research has started exploring Multimodal Large Language Models (MLLMs) as an alternative to traditional binary classifiers.

Despite their potential, applying MLLMs to video forensics is currently limited by the lack of high-quality data, simplistic evaluation protocols, and a reliance on direct classification. Specifically, we identify these critical limitations in the current research landscape: **1) Outdated and Biased Data:** Existing video datasets [5, 30, 9, 32, 18, 19] frequently rely on early-stage generators, producing easily recognizable artifacts. Furthermore, they often overlook demographic fairness, failing to account for gender and racial diversity. **2) Aggregated and Black-Box Evaluation:** Current benchmarks predominantly treat video forgery detection as a simple binary classification task on an aggregated test set. This approach obscures a model's specific weaknesses—whether it fails to generalize to out-of-domain models or breaks down under real-world social media degradations. Crucially, they

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison with existing datasets. GenBuster outperforms prior works in scale, visual quality and fairness (balanced gender, age, and ethnicity distributions). In-the-wild refers to whether the dataset includes videos sourced from social media.

Dataset	Latest Model	Scale	Generators	High Quality	Efficient Encoding	Resolution	FPS	Fairness	Focus on Real-World	In-the-wild Data
GVD [5]	2024/01	11k	11	✗	✗	Variable	-	-	-	✓
GVF [30]	2024/03	2.8k	6	✗	✗	Variable	-	-	-	✗
GenVideo [9]	2024/04	2271k	20	✗	✗	Variable	8-24	-	-	✓
DVF [44]	2024/10	7k	8	✓	✗	Variable	8-30	-	✓	✗
GenBuster	**2026/02**	203k	23	✓	✓	1024x1024	24	✓	✓	✓

also ignore the *explainability* of detection models. In forensic scenarios, providing a human-readable rationale is just as important as the binary verdict itself. **3) Lack of Forensic Reasoning:** Recent methods that adapt MLLMs for detection either use them as auxiliary to assist binary classification, or merely to generate post-hoc explanations. Neither approach treats step-by-step physical and spatiotemporal reasoning as the detection mechanism itself.

To address these limitations, we introduce three contributions:

1) GenBuster-200K: A High-Quality and Fair Dataset. We introduce a curated dataset comprising over 200,000 high-resolution real and AI-generated video clips. Diverging from prior works, GenBuster-200K leverages the latest generators, enforces rigorous fairness constraints (e.g., balanced gender, ethnicity), and focuses heavily on high-fidelity, real-world scenarios. As shown in Table 1, it successfully overcomes the authenticity and bias limitations of previous collections.

2) GenBuster-Bench: A Progressive Diagnostic Benchmark. To replace aggregated evaluations, we propose *the first* benchmark that decomposes the video forgery detection challenge along two critical axes: *domain shifts* and *generational shifts*. This dual-axis design partitions the evaluation into three progressive tracks: *In-Domain* (Era: 2024), *Out-of-Domain* (Era: 2025), and *In-the-Wild* (Era: 2026). Furthermore, we establish a standardized *MLLM-as-a-Judge* protocol to exclusively evaluate the forensic grounding and reasoning depth of the MLLMs.

3) BusterX: A Simple MLLM Baseline with RL Training. We propose BusterX, which formulates forgery detection as a visual reasoning task where the reasoning chain itself serves as the detector. To the best of our knowledge, it is the *first* MLLM-based video forgery detection framework with Reinforcement Learning (RL) training, empowering the model to uncover physical inconsistencies without expensive text annotations. Experimental results demonstrate that BusterX outperforms several leading MLLMs (e.g., Qwen3.5 [36], Claude-Sonnet-4.6 [3]) in both accuracy and interpretability.

2 GenBuster-200K: A High-Quality and Fair Dataset

Motivation. While recent pioneering datasets [44, 9] have catalyzed research in general video forgery detection, they are increasingly bottlenecked by outdated generative techniques, limited semantic diversity, and a heavy reliance on obviously fictional or low-realism content. To bridge the gap between current academic datasets and imminent real-world threats, we introduce GenBuster-200K. Designed to address these limitations, our dataset is constructed based on three principles: 1) **High Visual Quality**, as we employ state-of-the-art video generation models to ensure the synthetic videos exhibit high realism; 2) **Large Scale**, comprising over 200,000 high-resolution video clips; 3) **Real-World Focus and Fairness**. We deliberately filter out cartoonish prompts and enforce strict semantic constraints to ensure balanced demographic representations (e.g., gender, age, ethnicity), while actively mitigating harmful occupational and cultural stereotypes in practical, real-world scenarios.

2.1 Data Construction

Data Sources. GenBuster-200K consists of two parts: real videos and synthetic videos that simulate real-world conditions. Figure 1 illustrates our data construction pipeline.

Real Videos. More than 100K real videos sourced from OpenVid-1M [31], covering a highly diverse spectrum of real-world, in-the-wild scenarios to ensure comprehensive coverage. These videos were carefully filtered to ensure high quality and eliminate any duplicate clips.

Synthetic Videos. More than 100K videos were generated with high realism and diversity. We adopt text-to-video generation, as it affords high semantic controllability. We developed a Synthetic Prompt

Table 2: Data distribution for GenBuster.

Split	Real	Fake	ID				OOD								Wild											
			CogVideoX	EasyAnimate	HunyuanVideo	LTX-Video	Gen3	Jimeng	Kling	Luma	Pika	Sora	Vidu	Wanx	Gen4.5	Hailuo2.3	Kling2.5 Turbo	Kling2.6	Pixverse v5	Seedance1.5 Pro	Seedance2.0	Sora2	Veo3.1	Wan2.5	Wan2.6
GenBuster-200K	100k	100k	30k	40k	10k	20k	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GenBuster-Bench: ID	1000	1000	300	400	100	200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GenBuster-Bench: OOD	-	1000	-	-	-	-	100	100	100	100	100	200	150	150	-	-	-	-	-	-	-	-	-	-	-
GenBuster-Bench: Wild	-	150	-	-	-	-	-	-	-	-	-	-	-	-	10	20	15	20	10	20	15	10	10	10	10

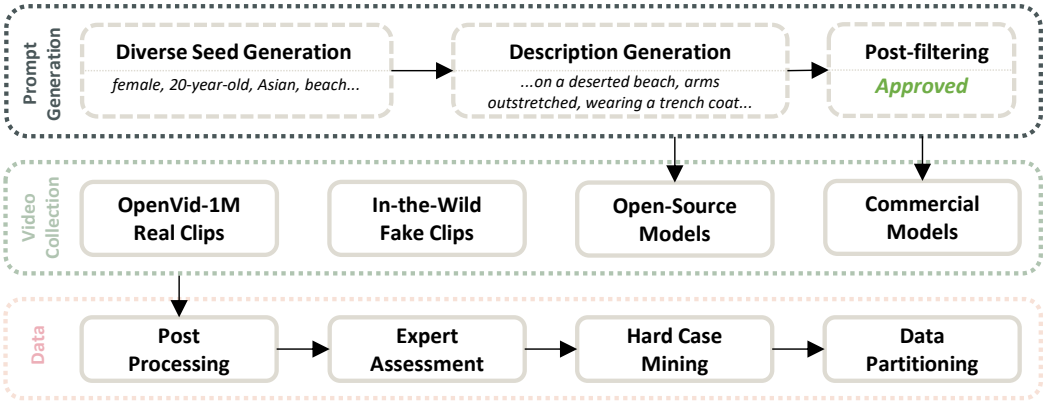


Figure 1: Data construction pipeline of GenBuster.

Generation Pipeline (with full details provided in the appendix) and experimented with several generative models such as HunyuanVideo [20] and LTX-Video [15]. For each model, we generated video frames with different hyper-parameters to further ensure visual diversity.

Post-processing. We perform unified post-processing **on both fake and real videos**. This process standardizes the resolution to 1024×1024, the duration to 5 seconds, and the frame rate to 24 FPS. This ensures consistency across videos generated by different models, which have varying original resolutions and frame rates. Additionally, we employ HEVC encoding using x265 with default parameters to encode the videos in the yuv420p10le format. This achieves a better balance between file size and video quality, allowing higher video quality at same bitrate. The unified post-processing eliminates potential biases from underlying encoding preferences (e.g., commonly used x264), which is crucial for ensuring the generalizability of forgery detection models across diverse sources. All videos are stored without audio after post-processing.

Expert Assessment. To ensure the fundamental quality and ethical alignment of our dataset, we engaged 5 well-trained annotators to perform rigorous quality control on the generated pool. Beyond technical filtering—such as removing corrupted video files, rendering failures, and clips with severe visual artifacts—this screening process included a strict **fairness audit**. Annotators were explicitly instructed to identify and discard videos exhibiting harmful demographic stereotypes, cultural misrepresentations, or unsafe content. Through this rigorous curation, we arrive at the final data distribution presented in Table 2.

2.2 Fairness-Aware Construction

Adhering to the principle of **fairness**, our pipeline applies explicit semantic constraints during seed generation to promote demographic diversity in human-centric contexts. These constraints aim to balance gender representation, sample ages from a normal distribution, and include ethnic groups from diverse geographic regions. To further mitigate intersectional bias, we avoid disproportionately associating particular demographic groups with specific occupations, thereby reducing stereotypical demographic–occupational associations in the generated prompts.



Figure 2: Visual Examples from GenBuster-Bench.

3 GenBuster-Bench: A Progressive Diagnostic Benchmark

3.1 Progressive Design

Existing video forgery detection benchmarks [30] typically evaluate models on a monolithic test set. This aggregated approach is inherently flawed, as it obscures whether a model's failure stems from a lack of foundational forensic knowledge, poor cross-generator adaptability, or vulnerability to real-world degradations. Furthermore, generative techniques are **evolving** at an unprecedented pace, rendering static benchmarks quickly obsolete. To provide a comprehensive, diagnostic, and future-proof evaluation, we introduce GenBuster-Bench. It strategically decomposes the challenge along two critical axes—*domain shifts* and *generational shifts*—into three *progressive* tracks:

Level 1: In-Domain (ID) Track – Foundational Capability **(Era: 2024).** This track is directly constructed using the identical real video data sources and established open-source generators from GenBuster-200K. It serves as a standard test set to evaluate a detector's fundamental baseline ability to perceive known spatial-temporal artifacts on seen distributions and generators.

Level 2: Out-of-Domain (OOD) Track – Generalization **(Era: 2025).** Attackers consistently leverage the newest, undisclosed generative models. To simulate this temporal arms race, the OOD track exclusively comprises videos from unseen, state-of-the-art commercial models representing the rapid capability leap of 2025. It strictly assesses whether a detector learns universal forensic laws rather than overfitting to outdated generator noise.

Level 3: In-the-Wild (Wild) Track – Real-World Robustness **(Era: 2026).** Representing the most imminent and complex threat landscape, this is the ultimate stress test. We meticulously curated high-fidelity "hard negatives" from cutting-edge models and subjected them to dynamic real-world social media degradations. Furthermore, this track exclusively evaluates the *explainability* of MLLMs strictly on correctly predicted samples, ensuring models can provide professional forensic rationales even when confronting the most deceptive future forgeries.

Through this dual-axis progressive paradigm, GenBuster-Bench transcends a simple leaderboard, serving as a robust diagnostic compass to evaluate the temporal lifespan and real-world reliability of future AI-generated video detectors.

3.2 Track Construction and Curation

Unified Pipeline. To ensure a rigorous and standardized evaluation, all tracks in GenBuster-Bench strictly adopt the identical generation pipeline and post-processing protocols (both real and fake) as described for the GenBuster-200K. Figure 2 presents a selection of video examples from GenBuster, showcasing its high-quality and diversity.

In-Domain (ID). This track evaluates models on seen generators from GenBuster-200K. To prevent performance saturation and rigorously stress-test the detectors, the synthetic videos in this test set are purposefully generated at **higher specifications** (e.g., increased sampling steps) than training set.

Out-of-Domain (OOD). This track is designed to assess cross-generator generalization. It exclusively features videos from the unseen commercial models. Crucially, because many of these proprietary state-of-the-art models lack accessible APIs, our experts manually generated and collected these videos by directly interacting with their official web interfaces.

In-the-Wild (Wild) Track. Representing the most imminent threat landscape, the Wild track is specifically engineered to evaluate model robustness against real-world distributions. To construct this track, we collected high-fidelity forgeries actively circulating on prominent social platforms. Notably, this track incorporates content from the most advanced generative engines to date, such as Seedance 2.0. To maximize the forensic challenge, we implemented a highly stringent **expert-driven curation** process: from an initial pool of approximately 750 candidate videos generated by 2026-era models, our forensic experts manually selected only the top 150 samples. This selection process strictly eliminated videos with any perceptible artifacts, retaining only those that are visually indistinguishable from authentic footage to ensure a rigorous stress test.

3.3 Evaluation Protocols and Metrics

Standardized Evaluation Protocol. To ensure a fair and reproducible assessment across diverse architectures, we establish a standardized spatiotemporal sampling protocol. For all evaluated models, we sample frames at 2.0 FPS from each video. We generally resize video frames to a target resolution of approximately 256×256. However, recognizing the specific patch size constraints inherent to various Vision Transformers (ViTs), we allow the actual resized dimensions to vary slightly to strictly align with the closest multiple of each model's required patch size (e.g., 252×252 for Qwen2.5-VL).

Evaluation Metrics. To comprehensively evaluate the performance of our model, we use overall accuracy (ACC) to measure the model's detection ability. Additionally, for every **unseen generator** in OOD track, we report its per-generator accuracy. Note that since our ID track is balanced between authentic and forged videos, we report the ACC on the `real` and `fake` subsets separately to expose potential model biases (e.g., over-forgery or conservative bias).

4 BusterX: A Simple MLLM Baseline with RL Training

In this section, we introduce BusterX, an MLLM baseline for detecting AI-generated videos and providing detailed explanations using label-only, outcome-supervised RL. The training is general-purpose: it requires only video clips and binary real/fake labels, making it directly applicable to any forgery detection dataset without additional annotation.

RL Training. Benefiting from the powerful visual reasoning capabilities of modern MLLMs, we bypass the traditional supervised fine-tuning phase and **directly** proceed with RL training. We adopt DAPO [56] over GRPO [13] primarily because DAPO effectively prevents *entropy collapse* inherent to standard GRPO. Maintaining a healthy policy entropy is crucial, as it encourages the model to continuously explore the policy space and generate complex, multi-step reasoning chains instead of prematurely converging on superficial solutions.

Reward Functions. The total reward R is a weighted sum of three components, designed to enforce format and encourage correct and detailed reasoning: $R = r_{\text{format}} + r_{\text{accuracy}} + r_{\text{soft_overlong}}$

- **Format**: If the model adheres to the specified format (i.e., `<think>...</think>\n\n...`), it receives a reward of $r_{\text{format}} = 0$; otherwise, it receives a penalty of $r_{\text{format}} = -1$.
- **Accuracy**: If formatted and correct, it receives $r_{\text{accuracy}} = 1$; otherwise, $r_{\text{accuracy}} = 0$.
- **Soft Overlong Penalty**: This component penalizes excessively long responses to prevent computational waste and mitigate repetitive token generation. The penalty is defined as:

$$r_{\text{soft_overlong}} = \begin{cases} 0, & L_{\text{gen}} \leq L_{\max} - L_{\text{cache}} \\ \frac{(L_{\max} - L_{\text{cache}}) - L_{\text{gen}}}{L_{\text{cache}}}, & L_{\max} - L_{\text{cache}} < L_{\text{gen}} \leq L_{\max} \\ -1, & L_{\max} < L_{\text{gen}} \end{cases} \tag{1}$$

Table 3: Comparison of state-of-the-art methods on GenBuster-Bench. The benchmark comprises three progressive tracks: ID evaluates performance on seen generators; OOD assesses cross-generator generalization on unseen commercial models; and Wild stress-tests real-world robustness.

Methods	ID (2024)			OOD (2025)										Wild (2026)	
	Real	Fake	ACC	Sora	Pika	Gen3	Luma	WanX	Kling	Jimeng	Vidu	Fake	ACC	Fake	ACC
Classical Detectors															
3D ResNeXt [16]	84.4	60.8	72.6	31.0	75.0	39.0	42.0	48.7	22.0	65.0	46.0	44.7	64.6	39.3	61.9
Vivit [4]	92.9	64.1	78.5	74.0	52.0	69.0	53.0	62.0	55.0	33.0	68.7	60.6	76.8	46.7	69.8
VideoMAE [49]	93.3	64.9	79.1	48.5	79.0	65.0	52.0	76.7	24.0	55.0	72.0	59.5	76.4	50.7	72.0
DeMamba [9]	**93.8**	70.2	82.0	65.0	75.0	73.0	56.0	46.0	68.0	71.0	84.7	66.9	80.4	48.0	70.9
Closed-Source MLLMs															
GPT-5.2 (High) [34]	93.3	61.8	77.6	43.0	77.0	78.0	91.0	66.7	36.0	27.0	33.3	54.5	73.9	50.7	72.0
Grok-4.1-fast [53]	76.1	77.5	76.8	78.0	81.0	60.0	90.0	80.0	87.0	**92.0**	73.3	79.6	77.9	35.3	55.7
Claude-Haiku-4.5 [2]	84.7	58.7	71.7	53.0	40.0	72.0	80.0	60.0	31.0	46.0	46.7	53.5	69.1	19.3	52.0
Claude-Sonnet-4.6 [3]	61.6	83.2	72.4	84.0	**91.0**	**93.0**	94.0	93.3	72.0	61.0	80.0	83.9	72.8	63.3	62.5
Open-Source MLLMs															
Qwen3.5-4B [36]	83.1	48.9	66.0	30.0	69.0	68.0	70.0	52.7	44.0	43.0	60.7	52.4	67.8	29.3	56.2
Qwen3.5-27B [36]	76.3	54.8	65.6	43.0	70.0	72.0	89.0	61.3	47.0	43.0	54.0	58.0	67.2	38.0	57.2
Qwen3.5-122B-A10B [36]	79.4	54.7	67.1	57.0	87.0	73.0	93.0	66.0	49.0	40.0	68.0	65.7	72.6	42.0	60.7
domain-specific MLLMs															
BusterX	83.0	**90.4**	**86.7**	**91.0**	89.0	92.0	**99.0**	**94.0**	**93.0**	87.0	**94.7**	**92.5**	**87.8**	**72.0**	**77.5**

† *Note:* The ACC for the OOD and Wild tracks is calculated as the macro-average of the track's fake accuracy and the ID Real accuracy.

5 Evaluation

Setup. We adopt Qwen3.5-4B [36] as our base model and train BusterX on GenBuster-200K. We employ the AdamW [28] with a learning rate of 1×10^{-6}, which aligns with established community best practices for stable RL training. During training, we use DeepSpeed ZeRO, `bfloat16` precision. More details are shown in Appendix D.

5.1 Main Results

We evaluate SOTA baselines and BusterX across the three progressive tracks of GenBuster-Bench to assess detection accuracy and generalization. The detailed results are presented in Table 3. Before delving into specific track performances, we highlight our primary observation:

> ***Finding:*** While traditional detectors and general MLLMs experience severe performance degradation on real-world manipulations (the *Wild* track), BusterX achieves higher accuracy on this track than the evaluated baselines, supporting the effectiveness of its RL training approach.

In-Domain (ID) Performance and Bias Mitigation. On the ID track, BusterX achieves the highest overall accuracy among all open-source MLLMs and classical detectors. More importantly, breaking down the ID performance into `Real` and `Fake` accuracy exposes severe inherent biases in existing models. For instance, Qwen3.5-122B exhibits a conservative bias (79.4% on Real, but 54.7% on Fake). Conversely, Claude-Sonnet-4.6 suffers from an over-forgery bias, which ruins its performance on real videos (61.6% Real ACC). In contrast, BusterX maintains a balanced performance, demonstrating that our training paradigm successfully aligns the model's forensic perception without compromising its judgment on authentic videos.

Out-of-Domain (OOD) Generalization. The OOD track measures cross-generator generalization exclusively on unseen commercial models. BusterX achieves a remarkable 87.8% accuracy, significantly outperforming all classical detectors and MLLMs.

In-the-Wild (Wild) Robustness. When confronting real-world social media in the Wild track, the performance of most classical detectors and generic MLLMs severely collapses (e.g., Claude-Haiku-4.5 drops to 19.3% Fake). BusterX achieves 72.0% Fake accuracy and 77.5% ACC, outperforming the evaluated baselines on this track. These results support the effectiveness of BusterX under the challenging conditions represented by the Wild track. Notably, we also observe a clear scaling law within the Qwen3.5 family (4B, 27B, 122B): the detection accuracy on in-the-wild videos exhibits a steady upward trend as the model scales. This indicates that larger parameter sizes bring consistent benefits in handling highly diverse, real-world forgeries.

5.2 Explanation Evaluation

MLLM-as-a-Judge. We evaluate explanation quality using an MLLM-as-a-Judge experiment on the Wild track, employing Gemini 3 Flash Preview [48] as an impartial automated judge. To maintain strict visual consistency, the judge evaluates the rationales using the exact same frame sequences provided to the models during the detection phase. Crucially, evaluating the rationale of an incorrect prediction is inherently flawed, as it forces the model to hallucinate justifications for a false premise. To prevent this, our protocol is *strictly accuracy-decoupled*—we exclusively prompt the judge to evaluate explanations for samples from the Wild track where the evaluated model successfully predicted the ground-truth label. The judge scores the generated rationales on a discrete 1-5 scale across three critical dimensions: 1) **Forensic Visual Grounding**, which assesses the identification of systemic rendering artifacts rather than superficial entity descriptions; 2) **Reasoning Depth & Logic Structure**, which evaluates cognitive depth and causal physical reasoning; and 3) **Factual Accuracy & Pseudo-Precision Penalty**, which penalizes hallucinations and unverifiable claims (e.g., falsely citing specific frame numbers). The final explainability score is derived by averaging these dimensions and normalizing the result to the $[0, 1]$ interval. The full judge prompt is provided in Appendix F. Table 4 compares explanation scores for four representative MLLMs. BusterX achieves the highest score of 84.7 among these MLLMs.

Table 4: Explanation scores on correctly predicted Wild samples.

Model	Score
GPT-5.2 (High) [34]	48.9
Grok-4.1-fast [53]	81.0
Claude-Sonnet-4.6 [3]	78.3
BusterX	**84.7**

False-Positive Analysis. Across all false-positive rationales, 89.6% explicitly contain terms such as "render," "3D," or "simulation," suggesting a tendency to attribute visual effects and stylized content to AI generation. We manually reviewed 20 false-positive cases by comparing the cited evidence with the video frames: 16 contained observable cues incorrectly attributed to AI generation, while 4 contained hallucinated evidence absent from the video. Representative errors include treating a floating visual-effect icon or a 2D explosion effect as evidence of AI generation, and misinterpreting hands and motion textures in a natural scene as generation artifacts. These cases distinguish incorrect attribution of visible evidence from hallucination of the evidence itself.

5.3 Case Study

As illustrated in Figure 3, we evaluate BusterX against advanced proprietary MLLMs (e.g., GPT-5.2 and Claude-Sonnet-4.6) when faced with highly realistic synthetic videos. While commercial models often rely on superficial visual coherence and fail to detect deep fakes, BusterX demonstrates superior frame-level spatiotemporal localization and deep reasoning grounded in world knowledge. Our qualitative analysis reveals three key capabilities:

1) Uncovering Physical Inconsistencies via Spatiotemporal Grounding: In the coastal wave video (Figure 3), baseline MLLMs incorrectly assess the video as real, fooled by seemingly coherent rock textures and wave motion. Conversely, BusterX isolates fatal spatiotemporal anomalies. It explicitly tracks the timeline, noting that the water transitions from calm to a massive crashing wave abruptly between 0.0s and 2.0s. Furthermore, BusterX identifies unnatural shifts in lighting and texture.

2) Assessing Logical Consistency: BusterX's ability to track long-horizon logic is highly effective in complex urban scenes (Appendix H). When analyzing a synthetic street video featuring a tram, BusterX catches severe object permanence violations—specifically, a crowd of pedestrians spontaneously appearing on a previously empty platform at the 5.0s mark.

3) Validating Realism through World Knowledge: When analyzing real-world footage (Appendix H), BusterX does not just look for low-level artifacts; it actively interprets the specific event content by grounding its visual observations in world knowledge. For instance, in an outdoor cooking scenario, the model identifies specific cultural and environmental contexts, such as the subject's traditional attire (kurta) and the rural setting framed by a dirt ground and rice paddies. Furthermore, it applies culinary common sense to evaluate the scene's logic, and even questioning whether the observed cooking techniques align with real-world practices for making fritters (vadas). By corroborating these semantic details with practical physical indicators, BusterX comprehensively verifies both the event logic and physical realism to confidently authenticate genuine videos.



Figure 3: Case Study. Faced with highly realistic synthetic video, MLLMs like GPT-5.2 and Claude-Sonnet-4.6 fail, whereas BusterX successfully performs frame-level spatiotemporal localization to uncover underlying physical inconsistencies. Full responses are provided in Appendix G.

Table 5: Robustness evaluation under various degradations.

Degradation	ID	OOD	Wild
None	**86.7**	**87.8**	**77.5**
50% Frame	84.8	86.2	76.0
25% Frame	83.7	85.5	76.2
JPEG 80	86.0	87.8	76.8
JPEG 70	85.4	87.3	77.1

Table 6: Ablation on w/o Thinking and RL loss type.

Variant	ID	OOD	Wild
BusterX	**86.7**	**87.8**	**77.5**
w/o Thinking	69.4	79.9	70.0
w/ GRPO loss	85.9	85.1	70.9

Figure 4: Ablation on reward funcs.



5.4 Ablation and Robustness Studies

Reasoning chain matters. As detailed in Table 6, skipping the reasoning chain (w/o Thinking) causes a severe performance collapse across all tracks. This confirms our core premise that formulating detection as a step-by-step visual reasoning task is essential.

DAPO vs. GRPO. Table 6 also highlights the impact of the RL objective. Replacing DAPO with standard GRPO loss noticeably degrades accuracy. We observe that this drop is primarily driven by severe *entropy collapse* during GRPO training. This premature collapse drastically restricts the policy space, severely hindering the model's ability to explore and formulate the in-depth physical reasoning chains crucial for forgery detection. By effectively mitigating this collapse, DAPO ensures robust exploration and optimization of the reasoning process. More details are provided in Appendix C.

Ablation on Reward functions. As shown in Figure 4, we evaluate the training stability across different reward configurations. Without the soft overlong penalty (w/o soft_overlong) or the format constraint (w/o format), the training exhibits significant fragility; it fails to converge and often collapses into repetitive patterns, demonstrating the necessity of our reward design.

Robustness. Table 5 demonstrates BusterX maintains consistent performance against video perturbations without specific degradation training, ensuring real-world practicality.

6 Related Work

6.1 AI-Generated Video Detection Datasets

Early datasets primarily focused on face manipulation and generation, typically employing GANs, VAEs, or various swapping techniques to produce synthetic videos. Datasets such as FaceForensics++ [41], FakeAVCeleb [19], WildDeepfake [61], and KoDF [22] are categorized as deepfake video detection dataset [32]. In recent years, identity-driven forgery datasets have also emerged, with IDForge [54] being a representative example. Recent advances in generative technology have significantly improved the quality of video generation, expanding synthetic content beyond facial manipulations to encompass full-scene natural video synthesis. This shift has broadened the research focus from facial forgeries to holistic video generation, driving the development of diverse datasets for AI-generated video detection. For instance, GenVideo [9] and GenVidBench [32] are large-scale benchmarks built on multiple AI generation models, covering a diverse range of semantic content. However, they are often limited by issues such as relatively low visual quality and lack of fairness. In contrast, GenBuster addresses these challenges by offering higher-quality data, a larger volume, emphasis on fairness, and focus on real-world applicability.

6.2 AI-Generated Video Detection Methods

Previous methods [51, 33, 52, 14, 47, 46] primarily formulated AI-generated content detection as a binary classification problem, focusing mainly on image-based deepfake detection. These methods utilized various architectures, such as CNNs and Transformers. To enhance the interpretability of the models, Huang et al. [17] recently propose a method for localizing altered regions in images and generating textual explanations, with a specific emphasis on detecting image editing manipulations. As AI-generated video quality advances, the demand for robust detection methods grows increasingly urgent. Early works primarily focused on face-centric deepfake detection [60, 11, 35, 10]. For instance, Zheng et al. [60, 11] develop methods that identify spatial inconsistencies to distinguish real from synthetic content. Meanwhile, Pellicer et al. [35] propose a model that learns person-specific prototypes by analyzing patterns in real videos. However, these methods struggle to generalize to in-the-wild AI-generated video detection. To address these limitations, recent works have explored alternative strategies. DeMamba [9] detects synthetic videos by analyzing small spatiotemporal zones, capturing subtle pixel changes. UNITE [21] enhances synthetic video detection by capturing full-frame manipulations using a transformer-based architecture with domain-agnostic features. Reiss et al. [37] propose the concept of fact-checking for detection and introduce FACTOR, a practical recipe that effectively counters critical attacks such as face swapping and audio-visual synthesis. To address these challenges, BusterX not only detects video authenticity but also provides detailed explanations by treating AI-generated video detection as a visual reasoning task, thereby enhancing the detection analysis.

6.3 Multimodal Large Language Models

MLLMs have evolved significantly in recent years. For instance, BLIP-2 [24] and LLaVA [26] integrated vision encoders with large autoregressive language models, enabling image-conditioned text generation and multimodal reasoning. Qwen2.5-VL [6] further advances these developments by improving vision-language alignment through a more efficient vision encoder and enhanced cross-modal fusion mechanisms. Concurrently, there is growing interest in leveraging MLLMs for fine-grained pixel-level understanding tasks [58, 57, 38–40]. Recently, the integration of MLLMs has introduced a new paradigm in AIGC detection [62]. For image forgery detection, models such as AntifakePrompt [8], ForgeryGPT [27], FakeShield [55], ForgeryTalker [25], ForgerySleuth [45], and SIDA [17] utilize the capabilities of MLLMs to enhance detection accuracy and provide robust support for localization and explanation tasks. In the video domain, MM-Det [44] leverages the strong visual-linguistic representations of MLLMs by generating a Multi-Modal Forgery Representation, which enhances its ability to detect unseen manipulated content. Reinforcement Learning (RL) has demonstrated its effectiveness in enhancing LLM reasoning capabilities [1, 48] following the Supervised Fine-Tuning (SFT) stage [29]. Notably, DeepSeek-R1 [13] introduced Group Relative Policy Optimization (GRPO), a variant of the Proximal Policy Optimization (PPO) algorithm [42], to significantly improve model reasoning [43]. The community has made numerous recent efforts to further refine GRPO, with DAPO [56] and GSPO [59] being representative examples. To enable the

detection model to not only classify videos as real or fake but also provide explanations that align with human understanding, we apply RL fine-tuning to the video forgery detection task, representing the first attempt of this paradigm in the research domain.

7 Conclusion

In this paper, we address critical vulnerabilities in AI-generated video detection through three contributions: the GenBuster-200K dataset, the GenBuster-Bench benchmark, and the BusterX detection framework. GenBuster-200K addresses the authenticity and distributional limitations found in prior datasets. GenBuster-Bench provides structured diagnostic evaluation along temporal and domain axes, and introduces an MLLM-as-a-judge to assess forensic explainability independently of classification performance. BusterX leverages reinforcement learning to train MLLMs toward transparent, step-by-step physical reasoning, moving beyond opaque classification outputs.

Limitations. Generated explanations may hallucinate details or misattribute observable cues to AI generation. Verifying these explanations against video frames assesses visual grounding rather than causal faithfulness. Our framework is also limited to the visual modality, leaving audio manipulation and audiovisual semantic inconsistencies unaddressed — both of which are common in real-world deepfake content. Future work will extend toward audio-visual multimodal detection.

Acknowledgments

This work was funded by the Engineering and Physical Sciences Research Council [grant number EP/X525741/1] Impact Acceleration Accounts Grant.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. Arxiv, 2023.

[2] Anthropic. Introducing claude haiku 4.5, 2026. URL https://www.anthropic.com/news/claude-haiku-4-5/.

[3] Anthropic. Introducing claude sonnet 4.6, 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6/.

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In CVPR, 2021.

[5] Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In PRCV, 2024.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. Arxiv, 2025.

[7] ByteDance. Seedance 2.0, 2026. URL https://seed.bytedance.com/en/seedance2_0.

[8] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. Arxiv, 2023.

[9] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. Arxiv, 2024.

[10] Yize Chen, Zhiyuan Yan, Siwei Lyu, and Baoyuan Wu. X^2-dfd: A framework for eXplainable and eXtendable deepfake detection. Arxiv, 2024.

[11] Sara Concas, Simone Maurizio La Cava, Roberto Casula, Giulia Orru, Giovanni Puglisi, and Gian Luca Marcialis. Quality-based artifact modeling for facial deepfake detection in videos. In CVPR, 2024.

[12] DeepSeek-AI. Deepseek-v3 technical report. Arxiv, 2024.

[13] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Arxiv, 2025.

[14] Junxian Duan, Yuang Ai, Jipeng Liu, Shenyuan Huang, Huaibo Huang, Jie Cao, and Ran He. Test-time forgery detection with spatial-frequency prompt learning. ICCV, 2024.

[15] 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, 2024.

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In CVPR, 2018.

[17] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In CVPR, 2025.

[18] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. GLFF: global and local feature fusion for ai-synthesized image detection. IEEE Trans. Multim., 2024.

[19] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. Arxiv, 2022.

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Junkun Yuan, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yanxin Long, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. Arxiv, 2024.

[21] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. Arxiv, 2024.

[22] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In ICCV, 2021.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In ACM SIGOPS, 2023.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML, 2023.

[25] Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. A large-scale interpretable multi-modality benchmark for facial image forgery localization. Arxiv, 2024.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023.

[27] Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. Arxiv, 2024.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019.

[29] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. Arxiv, 2023.

[30] Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency. Arxiv, 2024.

[31] 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. Arxiv, 2024.

[32] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video. Arxiv, 2025.

[33] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In CVPR, 2023.

[34] OpenAI. Introducing gpt-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

[35] Alvaro Lopez Pellicer, Yi Li, and Plamen Angelov. Pudd: towards robust multi-modal prototype-based deepfake detection. In CVPR, 2024.

[36] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

[37] Tal Reiss, Bar Cavia, and Yedid Hoshen. Detecting deepfakes without seeing any. Arxiv, 2023.

[38] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In CVPR, 2024.

[39] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In CVPR, 2025.

[40] Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld 2: Learning transferable knowledge from real-world videos. In CVPR, 2026.

[41] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In ICCV, pages 1–11, 2019.

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. Arxiv, 2017.

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. Arxiv, 2024.

[44] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In NeurIPS, 2024.

[45] Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. Forgerysleuth: Empowering multimodal large language models for image manipulation detection. Arxiv, 2024.

[46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In CVPR, 2023.

[47] 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 CVPR, 2024.

[48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. Arxiv, 2023.

[49] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In NeurIPS, 2022.

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. Arxiv, 2023.

[51] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In CVPR, 2020.

[52] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In CVPR, 2023.

[53] X.AI. Grok 4.1 fast and agent tools api, 2025. URL https://x.ai/news/grok-4-1-fast/.

[54] Junhao Xu, Jingjing Chen, Xue Song, Feng Han, Haijun Shan, and Yu-Gang Jiang. Identity-driven multimedia forgery detection via reference assistance. In ACMMM, 2024.

[55] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. Arxiv, 2024.

[56] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. Arxiv, 2025.

[57] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. Arxiv, 2025.

[58] Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. Arxiv, 2025.

[59] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. Arxiv, 2025.

[60] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In CVPR, 2021.

[61] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In ACMMM, 2020.

[62] Yueying Zou, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. Survey on ai-generated media detection: From non-mllm to mllm. Arxiv, 2025.

A Societal Impact

The proliferation of AI-generated video poses tangible risks to public discourse, including the spread of misinformation, non-consensual synthetic media, and manipulation of political or legal evidence. This work contributes tools intended to help counter these harms. At the same time, we acknowledge the dual-use nature of this research. Detailed forensic analysis of generator artifacts may inadvertently inform efforts to produce harder-to-detect content. We release our benchmark and model with the expectation that they will be used to strengthen detection capabilities, and encourage the community to consider access controls when deploying forensic tools in adversarial settings.

B Prompt Generation Pipeline in GenBuster-200K

The main process comprises three stages:

Stage-1: Diverse Seed Generation. This stage is explicitly engineered for **Fairness**: keyword collection begins with human-centric contexts where we systematically enforce balanced sampling across gender, age, ethnicity and intersecting attributes, then extends to animal, scenery and other non-human domains to ensure the resulting semantic space is globally inclusive and bias-resistant. This methodical approach ensures our initial semantic space is both inclusive and comprehensive, spanning a wide range of perspectives.

Stage-2: Description Generation. Building on the prior output from stage-1, the forthcoming synthetic video descriptions will be generated by an LLM using a predefined prompt. Specifically, the process includes: **1)** Prompt Constraints Guidance: The prompt introduces physical and aesthetic constraints, with manual adjustments to cover multiple scenarios. These constraints guide the LLM to generate descriptions that are both realistic and contextually appropriate. **2)** Multi-model Overfitting Resistance: To combat model-specific bias, we randomly sample from a diverse pool of over 20 open-source LLMs, spanning varied architectures and parameter scales (e.g., DeepSeek-V3 [12], LLaMA [50], etc.). We employ high-temperature sampling to maximize output diversity during generation.

Stage-3: Post-filtering. After generating the prompt, we employ a multi-step post-filtering mechanism to ensure the quality and reliability of the generated prompts. Specifically, the post-filtering process includes the following key steps: **1)** Length Control: We first check the length of each generated

prompt to ensure it falls within a predefined range. Prompts that are too short may lack sufficient context, while those that are excessively long can introduce unnecessary complexity or redundancy. **2)** Special Character Check: We then scan for any special characters or formatting issues that may have been inadvertently introduced during the generation process. **3)** Readability Scoring: The LLM Judge assigns a readability score based on linguistic coherence, grammatical correctness, and overall clarity. This step is crucial because low-quality prompts (those with poor readability) tend to produce low-quality generated videos that are easily detectable, which would undermine the dataset's utility for training robust detectors. More details are provided in the appendix. By integrating these steps, we can ensure that the generated prompts are of high quality.

C DAPO

We utilize DAPO [56] as our reinforcement learning strategy. DAPO samples a group of outputs $\{o_i\}_{i=1}^{G}$ for each question q from the old policy $\pi_{\theta_{old}}$. The reward model is then used to score the outputs, yielding $\{r_i\}_{i=1}^{G}$ correspondingly. Then DAPO optimizes the policy model π_θ by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = \quad & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\right. \\ & \left.\text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right] \\ \text{s.t.} \quad & 0 < \Big|\{o_i \mid \texttt{is_equivalent}(a, o_i)\}\Big| < G,\end{aligned} \tag{2}$$

where ε_{low} and $\varepsilon_{\text{high}}$ are hyper-parameters, and $\hat{A}_{i,t}$ is the advantage:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}. \tag{3}$$

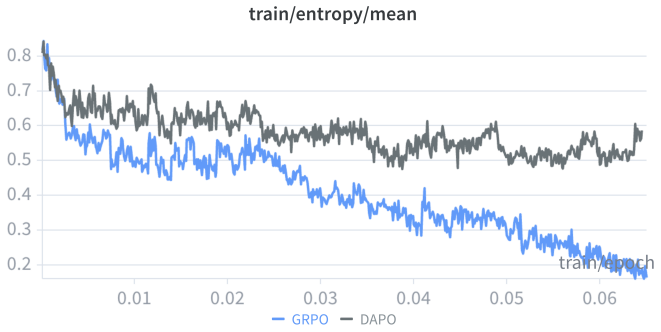


Figure 5: Training entropy: DAPO vs. GRPO.

As shown in Figure 5, DAPO effectively prevents entropy collapse inherent to standard GRPO.

D Experimental Details

In this section, we provide comprehensive details regarding the training configurations, evaluation parameters, and hyperparameter selections utilized in our study to ensure full reproducibility.

To minimize generation overhead, we integrate `vLLM` [23] for both training rollouts and evaluation. Its high-throughput capabilities enable us to maintain a large training rollout batch size. For each query, we generate $G = 8$ trajectories at a temperature of $T = 1.0$. The maximum completion length is set to $4,096$ tokens. We apply a clipping bound of $[0.8, 1.28]$ for the policy ratio.

E Prompt Design

Based on extensive experiments, we have carefully designed the following prompt to guide the answer.

User Prompt

Please analyze whether there are any inconsistencies or obvious signs of forgery in the video, and finally come to a conclusion: Is this video real or fake?

Then, just answer this MCQ with a single letter:
Q: Is this video real or fake?
Options:
A) real
B) fake

Put the final answer in $\backslash\backslash boxed\{...\}$

F MLLM-as-a-judge

To rigorously evaluate the rationale and forensic reasoning capabilities of the models, we employ **Gemini 3 Flash Preview** as Judge. The full prompt template is detailed below:

User Prompt

Please carefully watch the provided video.
Here is the rationale generated by the evaluated model:

"{explanation}"

Evaluate the rationale strictly using the following 1-5 discrete rubric for three dimensions.

—
Dimension 1: Forensic Visual Grounding (Systemic vs. Superficial)
Evaluates whether the model merely describes objects (image captioning) or genuinely grounds its analysis in systemic forensic artifacts (lighting, physics, rendering).
** 1 (Vague/Captioning): Mentions nothing specific, or just describes the scene's content (e.g., "there is a man and a car") without identifying actual forensic flaws.*
** 2 (Coarse Entity): Mentions specific entities being "fake" or "unnatural" without explaining the physical manifestation.*
** 3 (Local Feature): Locks onto isolated local features (e.g., "blurry face").*
** 4 (Holistic/Systemic Grounding): Perfectly identifies systemic rendering artifacts rather than just objects. This includes artificial lighting uniformity, lack of real-world analog camera noise/entropy, overly perfect textures, or static motion rendering.*
** 5 (Advanced Spatiotemporal): Seamlessly combines spatial artifacts with temporal dynamics, or provides a master-level grounding of cinematic, environmental, and textural inconsistencies across the sequence.*

Dimension 2: Reasoning Depth & Logic Structure
Evaluates the cognitive depth, use of deductive reasoning, and structural rigor of the analysis.
** 1 (Basic/Tautological): "It is fake because it is AI-generated" or zero logical flow.*
** 2 (Surface Flaw): Points out a flaw but provides no causal or physical reasoning.*
** 3 (Mechanism-Level): Explains the flaw using basic generative defect mechanisms.*
** 4 (Deductive Physical Analysis): Contrasts the observed digital rendering against expected real-world physics (e.g., "Real-world videos would have natural shadows... but this shows uniform artificial lighting").*
** 5 (Structured Forensic Chain-of-Thought): Employs a highly structured, step-by-step reasoning process (e.g., breaking down lighting, posture, background, and camera artifacts systematically). It reads like a methodical, professional digital forensic investigation report.*

Dimension 3: Factual Accuracy & Pseudo-Precision Penalty
Evaluates if the model fabricated details or used deceptive "pseudo-precision" to appear accurate. Compare strictly against the ACTUAL video frames.
** 1 (Severe Hallucination / Pseudo-Precision): The model fabricates key evidence. **CRITICAL PENALTY:** If the model acts like a "frame-counter" and cites highly specific frame numbers (e.g., "In frame 5...", "In frame 11...") or instantaneous color swaps that you CANNOT strictly verify in your limited visual sample, it is hallucinating precision and MUST receive a 1.*
** 2 (Partial Over-interpretation): Half factual, half forced correlation.*
** 3 (Slight Hallucination): Mostly correct, but slightly mischaracterizes a minor visual detail.*
** 4 (Mostly Accurate): Perfectly aligns with the video content without inventing unverifiable data points.*
** 5 (Absolutely Rigorous & Objective): Fully grounded in the true pixels, lighting, and overarching physics of the video. Zero hallucination, avoiding unverifiable pseudo-precision, maintaining an objective forensic tone.*

—
Output JSON Format
You must return the evaluation strictly in the following JSON format. Provide a brief analysis (1-2 sentences) before assigning the integer score (1-5) for each dimension.

```
{
 "dim1": {
 "analysis": "<your step-by-step reasoning>",
 "score": <1-5>
 },
 "dim2": {
 "analysis": "<your step-by-step reasoning>",
 "score": <1-5>
 },
 "dim3": {
 "analysis": "<your step-by-step reasoning>",
 "score": <1-5>
 }
 }
```

G Complete Responses

The model's full responses in Case Study. This is a fake video, the answer should be B.



GPT-5.2

The frames show coherent, physically plausible wave motion (cresting, impact against the rock face, spray/foam expansion, then foam dispersal and flowing around the tide pool rocks). The rock textures, lighting, and cloud patterns remain consistent across frames without the typical AI/video-forgery artifacts (warping edges, temporal "melting," inconsistent reflections, or unstable fine details). No obvious compositing seams or repeated/looped foam patterns stand out.

Conclusion: this looks like a real video.

$\backslash\backslash boxed\{A\}$

Claude-Sonnet-4.6

I'll analyze the frames of this video showing a coastal rocky scene with waves.

Analysis

Scene consistency: - The location shows a rocky coastal area with a natural rock arch, tidal pools, and ocean waves - The rock formations, arch structure, and general geography remain consistent throughout all frames

Motion and progression: - The video shows a natural wave sequence: calm water → approaching wave → wave crashing against rocks → water spreading/receding - The wave movement follows physically plausible fluid dynamics - The timing and progression of the wave appears natural

Lighting and color: - There is a gradual, subtle shift in lighting/color temperature across frames (slightly cooler to slightly warmer tones) - This could indicate time-lapse or color grading, but appears consistent with natural lighting changes at dusk/dawn

Potential anomalies: - The color grading shifts somewhat dramatically between early and later frames - However, this appears to be natural lighting variation during the golden hour/dusk period - Water movement, foam patterns, and spray all appear physically consistent - No obvious digital artifacts, cloning, or unnatural elements detected

Overall assessment: The wave dynamics are natural, the rock formations remain consistent, and the motion follows expected physical behavior. The color shifts appear to be natural lighting changes during twilight conditions.

Conclusion This video appears to be **real** - a genuine timelapse/video of ocean waves crashing against a rocky coastal formation with a natural arch.

$\backslash\backslash boxed\{A\}$

BusterX

<think>
The user wants me to determine if the video is real or fake based on a sequence of frames showing a coastal scene.

1. **Analyze the video content:**
* **Subject:** A rocky coastline with an arch, tide pools, and waves crashing.
* **Action:** The video progresses from a calm scene with small ripples to a large wave crashing against the rocks.
* **Details:**
* 0.0s: Calm water, rocks, tide pools. The sky has clouds.
* 0.5s: Similar scene, maybe a slight movement in the water.
* 1.0s: Still relatively calm, but the water looks a bit more turbulent.
* 1.5s: A wave is starting to form in the distance.
* 2.0s: A wave is breaking near the shore.
* 2.5s: The wave is crashing, creating white foam.
* 3.0s: The wave is fully crashing against the rocks.
* 3.5s: The wave is receding slightly but still turbulent.
* 4.0s: More white foam, the wave has hit the rocks.
* 4.5s: The aftermath of the wave, water flowing over the rocks.
* 5.0s: The scene is now quite active with white water.

2. **Evaluate for real vs. fake indicators:**
* **Physics/Physics-like motion:** The wave progression looks consistent but slightly unnatural in some ways. In a real video, a wave would build up and crash. Here, it seems to jump from calm to a huge crash almost instantly between 0.0 and 2.0s. In reality, a wave takes time to build. The transition from calm water (0.0) to a crashing wave (2.5) happens very quickly, suggesting a generated video or a very specific edit.
* **Texture/Details:** Look at the rocks and the water.
* The rocks look consistent but have a slightly smooth, almost painted or rendered texture in some frames (especially 0.0-1.0). In later frames, they look more like realistic rocks. Wait, looking closely at the rocks, they seem consistent but the lighting changes drastically.
* The water texture: In 0.0-1.0, it's calm and reflective. In 2.0+, it's churning white foam. This is a common AI generation technique (motion generation).
* **The most telling part:** The morph or transition. The video starts with a still, almost like a photo, and then suddenly a wave crashes. In a real video, you'd see the wave building up. Here, it's almost like a video generation prompt where the prompt is a rocky coast with a wave crashing. The lighting changes significantly between 0.0 and 5.0. In 0.0, it's a soft, maybe slightly overcast or sunset light. In 5.0, the lighting seems similar but the motion blur and foam suggest a different pass.
* **Specific AI artifacts:**
* Look at the texturing of the rocks. In 0.0, the rock on the far left is distinct. In 5.0, it's in the same spot but the texture looks slightly different or just different lighting.
* Look at the tide pools or shallow water. In 0.0, you see reflections. In 5.0, the water is churning.
* **The most obvious fake sign:** The motion of the wave. In 0.0-1.0, the water is almost glassy and calm. In 2.0, a white blob appears. By 2.5, it's a full crash. This is a classic motion or video generation artifact where the model tries to simulate physics but often gets the timing wrong. A real wave wouldn't just pop into existence like that. It would build up.
* **The sky:** The clouds move, but the lighting on the rocks changes significantly between the calm start and the crash. In 0.0, the sun seems low (sunset/sunrise). In 5.0, it's still low but the lighting on the wet rocks changes drastically, suggesting a different render or a very inconsistent motion simulation.

3. **Conclusion:** The video exhibits classic AI generation characteristics: inconsistent physics (wave appearing out of nowhere), inconsistent lighting/texture shifts, and a motion that feels like a simulation rather than a camera recording. It looks like a video generation model (like SVD or similar) taking a prompt or an image and generating motion.

4. **Final check:** The prompt asks for a single letter MCQ answer.
* Q: Is this video real or fake?
* Options: A) real B) fake
* Answer: B
</think>

The video shows a consistent but physically inconsistent progression. The wave appears out of nowhere between 0.0s and 2.0s, and the lighting/texture on the rocks shifts significantly between the calm start and the crashing wave, typical of AI video generation (like SVD or similar tools).

$\\boxed{B}$

H Additional Examples of BusterX



FAKE

<think>
The user wants me to determine if the video is real or fake based on a sequence of frames showing a street scene with buses and a tram.

1. **Analyze the video content:**
* **Subject:** A red Coca-Cola tram is moving along tracks. There are other buses (yellow, blue) on the road next to it. People are walking on the sidewalk and waiting at a platform.
* **Motion:** The tram is moving forward. The buses are moving or stopped. People are walking.
* **Details:**
* 0.0s: Tram is moving forward. A yellow bus is behind it. A blue bus is further back. People are on the sidewalk.
* 0.5s: Similar scene. The tram is moving. The yellow bus is moving.
* 1.0s: The tram is moving. The yellow bus is moving.
* 1.5s: The tram is moving. The yellow bus is moving.
* 2.0s: The tram is moving. The yellow bus is moving.
* 2.5s: The tram is moving. The yellow bus is moving.
* 3.0s: The tram is moving. The yellow bus is moving. People are on the platform.
* 3.5s: The tram is moving. The yellow bus is moving. People are on the platform.
* 4.0s: The tram is moving. The yellow bus is moving. People are on the platform.
* 4.5s: The tram is moving. The yellow bus is moving. People are on the platform.
* 5.0s: The tram is moving. The yellow bus is moving. People are on the platform.

2. **Evaluate for real vs. fake indicators:**
* **Physics/Motion:** The motion looks consistent but slightly unnatural in some frames. For example, in 0.0s, the tram is moving forward. In 5.0s, it's still moving forward. The cars and buses are moving in a way that suggests a simulation. Look at the wheels and reflections.
* **Object details:**
* The tram has Coca-Cola branding clearly visible. It's a specific model (looks like a double-decker tram).
* The buses are also distinct. One is yellow and blue, another is orange and blue.
* The environment looks like a city street (Hong Kong maybe? The architecture and the tram suggest it).
* **Key AI/Generative artifacts:**
* **Textures:** Look at the road surface and the tram tracks. In 0.0s, the tracks are clear. In 5.0s, the tracks are clear. But look at the reflections on the tram in 0.0 vs 5.0. In 0.0, it's reflecting the surroundings. In 5.0, it's reflecting the surroundings.
* **Motion blur/Physics:** Look at the people. In 0.0, people are walking. In 5.0, people are standing on the platform. Wait, looking closely at the people.
* 0.0: People walking on the far left sidewalk.
* 5.0: People standing on the platform on the right. This suggests a consistent scene but maybe inconsistent object persistence? No, the people seem to be consistent but their actions change slightly between frames, which is common in video generation.
* **The most telling part is the motion and physics of the objects.**
* Look at the yellow bus. In 0.0, it's behind the tram. In 5.0, it's still behind the tram but seems to have moved forward relative to the camera? Actually, looking at the position, the bus is consistently on the left side.
* Look at the tram's position. It moves from the center towards the right foreground.
* **The most obvious fake sign is the text on the objects.** Look at the Coca-Cola logo on the tram. It's consistent but the text rendering changes slightly between frames (0.0 vs 5.0). In 0.0 it's clear. In 5.0 it's clear. But look at the text on the bus in 0.0. It's blurry/indistinct. In 5.0, it's also blurry.
* **Look at the reflections and lighting.** The lighting changes significantly between 0.0 and 5.0. In 0.0, it's bright sunlight, strong shadows. In 5.0, still bright sunlight, but the shadows seem slightly different or the sun position shifts? Actually, looking at the shadows, the sun seems consistent but the lighting on the tram changes drastically between 0.0 and 5.0. In 0.0, the tram is fully lit. In 5.0, it's fully lit. But look at the reflection on the tram's side in 0.0 vs 5.0. In 0.0, you see the street. In 5.0, you see the street.
* **Look at the glint on the tram.** In 0.0, there's a reflection on the front window. In 5.0, same reflection.
* **Look at the people.** In 0.0, people are walking on the far left. In 5.0, people are standing on the platform on the right. This suggests the video generation might be hallucinating or changing the scene context significantly between frames. Wait, looking closer at the people.
* 0.0: People walking on the far left sidewalk.
* 5.0: People standing on the platform on the right. This is a huge inconsistency. In a real video, people would stay in their locations. In a generated video, objects often drift or change context.
* Also, look at the people on the platform in 0.0. There are none visible clearly. In 5.0, there are people standing on the platform. This suggests a video generation where the prompt is a video of a tram but the consistency is low.
* Look at the cars in the background. In 0.0, there's a car driving away in the distance. In 5.0, the background is different? In 0.0, you see a street curving away. In 5.0, you see the same street but the perspective shifts.
* **Look at the tram's movement.** In 0.0, it's moving forward. In 5.0, it's moving forward. But look at the wheels. In 0.0, the wheels are on the tracks. In 5.0, the wheels are on the tracks.
* **Look at the text on the tram.** 108 is visible in 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0. Consistent.
* **Look at the Coca-Cola logo.** Consistent.
* **Look at the bus.** The yellow bus is consistently there.
* **Conclusion:** The inconsistencies in object persistence (people moving from left sidewalk to right platform) and lighting/texture shifts strongly suggest this is an AI-generated video or a digital simulation. The motion blur and physics feel slightly off but consistent enough for a simulation. The key is the people's location change. In 0.0, people are on the far left sidewalk. In 5.0, people are on the platform on the right. That's a hallucination/inconsistency.

3. **Final check:**
* Real video: People would stay in place.
* Fake video: People can move or change location.
* Therefore, the video is fake.
</think>

The video shows inconsistencies in object persistence and physics. For example, the people on the far left sidewalk in the first few frames are suddenly standing on the platform on the right in the later frames. Also, the lighting and reflections shift in ways typical for video generation (like the glare on the tram).

$\\boxed{B}$



REAL

<think>
The user wants me to determine if the video is real or fake based on a sequence of frames showing a man cooking outdoors.

1. **Analyze the video content:**
* **Subject:** A man in traditional white clothing (kurta and cap) is sitting outdoors, likely in a field or garden (rice paddies are visible in the background).
* **Action:** He is preparing food. Specifically, he's making something like vada or fritters.
* **Objects:**
* A large metal pot filled with a yellow batter (likely gram flour or similar).
* A smaller metal bowl with red tomatoes.
* A black wok/pan with hot oil.
* A metal tray with already fried, golden-brown fritters (vadas).
* **Details:**
* In the first few frames, he's holding a tomato and dipping it into the batter? No, wait, looking closer at 0.0-0.5, he's holding a tomato in one hand and maybe a utensil or just gesturing. Actually, at 0.0 he's holding a tomato. At 0.5 he's holding a tomato and maybe a small bowl? No, just the tomato. Wait, at 0.0 he's holding a tomato in his left hand and a small metal bowl in his right? No, looking at 0.0, he's holding a tomato in his left hand and a small metal bowl with more tomatoes in his right hand. Wait, looking closer at 0.0, he's holding a tomato in his left hand and a small metal bowl with tomatoes in his right hand.
* At 0.5, he's holding the tomato and the bowl.
* At 1.0, he's dipping something into the oil? No, he's holding the tomato and the bowl. Wait, at 1.0 he's holding the tomato and the bowl.
* At 2.0, he's holding the tomato and the bowl.
* At 2.5, he's holding the tomato and the bowl.
* At 3.0, he's holding the tomato and the bowl.
* At 4.0, he's holding the tomato and the bowl.
* At 4.5, he's dipping a tomato into the yellow batter? No, looking closely, he's holding a tomato in his left hand and a small metal bowl in his right. Wait, at 4.5 he's dipping something into the batter. It looks like he's dipping a tomato? That's unusual. Or maybe he's just holding the bowl. Let's re-examine.
* Actually, looking at the sequence: He has a large pot of yellow batter. He has a tray of fried vadas. He has a pan of oil. He has a bowl of tomatoes.
* At 0.0, he is holding a tomato in his left hand and a small metal bowl with tomatoes in his right hand.
* At 4.0, he is holding the tomato and the bowl.
* At 4.5, he is dipping the tomato into the yellow batter? That seems like a weird way to make vadas. Usually, you dip dough balls. Maybe he's making a tomato-based dish? Or maybe he's just holding the tomato and the bowl is separate. Wait, looking at 4.5, his hand is in the batter. Is he dipping the tomato? That would be very strange. Maybe he's dipping a small piece of dough? Or maybe he's just holding the bowl and the tomato is separate. Let's look at the hand motion. At 4.5, his right hand is in the batter. His left hand is holding the tomato. This suggests he might be dipping the tomato? That's odd. Or maybe he's just mixing something?
* Regardless of the specific cooking technique (which might be a specific regional dish), the key is the setting and the objects.
* **Setting:** Outdoors, natural light, dirt ground, rice fields in the background. This strongly suggests a real, likely rural or semi-rural setting.
* **Camera:** Handheld, slightly shaky, close-up but wide enough to see the context. This is typical for social media content (like TikTok or Instagram Reels) showing street food or cooking.
* **Conclusion:** This is clearly a real video recording of a real event (cooking outdoors). It has the characteristics of a street food or cooking process video often found on social media. It's not AI generated or a digital forgery. The lighting, the textures of the food, the natural background all scream real video.

2. **Formulate the final answer:**
* Identify the key indicators: Real-world setting (outdoors, fields), real objects (metal pots, food), natural lighting, realistic physics (sitting on the ground, holding objects).
* Select the correct MCQ option: A) real.
</think>

The video shows a real-world cooking scene outdoors with natural lighting and realistic objects (metal pots, food, fields).

$\\boxed{A}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper acknowledges limitations about the dataset.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: The paper does not focus on theoretical results but rather on experimental and methodological contributions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper states that the code, model, and dataset will be released.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides comprehensive details about the training and test settings.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details about the training and test settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper includes appropriate statistical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the computational resources about data generation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the positive impacts of improved deepfake detection and the potential negative impacts of misuse of the technology.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators of existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The data is well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The qualitative checks were internal technical evaluations by author-team members with deepfake detection experience. No external participants or crowd workers were recruited, and no additional compensation was provided.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Author-team members evaluated model outputs as part of the technical analysis; the evaluators were not studied as research subjects, and no external participants were recruited.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used for writing and editing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.