VIDGUARD-R1: AI-GENERATED VIDEO DETECTION AND EXPLANATION VIA REASONING MLLMS AND RL

Kyoungjun Park[1], **Yifan Yang**[2], **Juheon Yi**[2], **Shicheng Zheng**[2], **Yifei Shen**[2]
Dongqi Han[2], **Caihua Shan**[2], **Muhammad Muaz**[1], **Lili Qiu**[1,2]
[1]Department of Computer Science, University of Texas at Austin, Austin, TX, USA
[2]Microsoft Research Asia, Shanghai, China

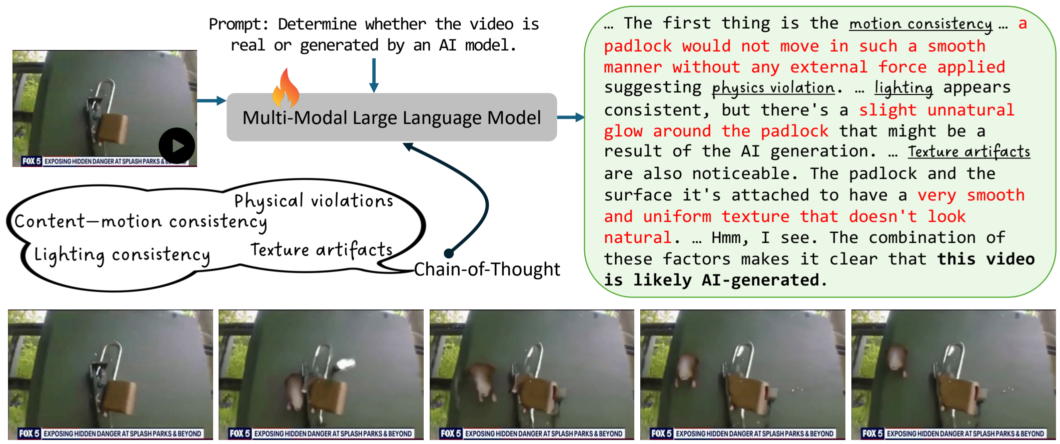


Figure 1: Overall framework of **VidGuard-R1**. We present the first video authenticity detector based on multi-modal large language models (MLLMs), which generates a chain-of-thought reasoning process along with the final answer.

ABSTRACT

With the rapid advancement of AI-generated videos, there is an urgent need for effective detection tools to mitigate societal risks such as misinformation and reputational harm. In addition to accurate classification, it is essential that detection models provide interpretable explanations to ensure transparency for regulators and end users. To address these challenges, we introduce **VidGuard-R1**, the first video authenticity detector that fine-tunes a multi-modal large language model (MLLM) using group relative policy optimization (GRPO). Our model delivers both highly accurate judgments and insightful reasoning. We curate a challenging dataset of 140k real and AI-generated videos produced by state-of-the-art generation models, carefully designing the generation process to maximize discrimination difficulty. We then fine-tune Qwen-VL using GRPO with two specialized reward models that target temporal artifacts and generation complexity. Extensive experiments demonstrate that **VidGuard-R1** achieves state-of-the-art zero-shot performance on existing benchmarks, with additional training pushing accuracy above 95%. Case studies further show that **VidGuard-R1** produces precise and interpretable rationales behind its predictions. The code is publicly available at `https://VidGuard-R1.github.io`.

1 INTRODUCTION

In the past year, we have witnessed unprecedented progress in video generation models, with dramatic improvements in realism and quality. The release of powerful models such as Sora (Brooks et al., 2024), Wan (Wang et al., 2025a), and HunyuanVideo (Kong et al., 2024) has made AI-generated videos more accessible to the public, further blurring the line between synthetic videos and real ones. At the same time, these advancements have led to a series of social risks, including the spread of misinformation, violations of privacy rights, damage to personal reputations, and increased susceptibility to scams and fraud.

Motivated by its practical significance, several pioneering works have been developed to detect AI-generated videos. Early approaches primarily targeted DeepFake-style facial forgeries (Qian et al., 2020b; Tan et al., 2024; Gu et al., 2021), which often assumed single-subject, frontal-face scenarios under constrained settings. These assumptions diverge significantly from open-domain, multi-scene videos produced by modern generative models. More recent detectors leverage spatial-temporal consistency (Ma et al., 2024; Bai et al., 2024b; Liu et al., 2024); however, such methods are limited in capturing higher-level semantic or causal inconsistencies and can be easily bypassed by post-processing techniques. Other methods are trained on curated fake video detection datasets (Chen et al., 2024a; Ni et al., 2025; Kundu et al., 2025), but these benchmarks often lack coverage of newly emerging models and fail to reflect the full diversity of generative capabilities. A recent benchmark (Chen et al., 2024a) shows that even state-of-the-art detectors still struggle to reliably identify videos from advanced models like Sora. Furthermore, these detectors typically offer only binary decisions without accompanying explanations, which raises concerns for transparency, especially when detection outcomes affect content moderation or legal accountability. Users are also more likely to trust detection systems that provide interpretable reasoning.

Recent advances in multi-modal large language models (MLLMs) have significantly enhanced video understanding, enabling detailed explanations of model decisions (Bai et al., 2023; Zhang et al., 2024b). This makes them promising candidates for detecting and explaining AI-generated videos. However, directly applying existing MLLMs, including advanced models like GPT-4o, yields subpar performance on current benchmarks, underscoring the need for supervised fine-tuning (SFT). As an initial step, we applied SFT to the Qwen2.5-VL-7B model (Bai et al., 2025). While the model achieved strong overall performance, it remained limited in its ability to explain why a video is fake, revealing shortcomings in its reasoning capability.

To address this, we adopt reinforcement learning (RL), which has shown promise in enhancing LLM reasoning (Guo et al., 2025). Notably, Video-R1 (Feng et al., 2025) outperforms commercial models on video reasoning tasks. RL enables MLLMs to develop self-improving reasoning via outcome-based rewards. We hypothesize that RL fine-tuning can help models detect subtle temporal and generative artifacts. Key to this is designing effective reward models. Simple binary rewards (e.g., 1 for real, 0 for fake) are insufficient. Instead, we propose two strategies: (1) injecting temporal artifacts into both real and fake videos to encourage temporal reasoning, and (2) assigning higher rewards to videos generated with more diffusion steps, which are harder to detect. Incorporating these into a group relative policy optimization (GRPO) framework leads to over 86% accuracy on our dataset and 95% accuracy on two benchmarks.

- We introduce **VidGuard-R1**, the first video authenticity detector that fine-tunes the MLLM using GRPO. The model leverages the pretrained knowledge of MLLMs for accurate classification and employs reinforcement learning for effective exploration. To further enhance performance, we design two specialized reward models that target temporal artifacts and generation complexity based on diffusion steps.
- We construct a challenging dataset of 140k real/fake video pairs for AI-generated video detection. By employing state-of-the-art generation models and carefully controlling the process, we ensure that distinguishing real from fake is non-trivial.
- **VidGuard-R1** achieves state-of-the-art zero-shot performance on existing benchmarks, with accuracy exceeding 95%. Case studies further highlight its ability to produce accurate and interpretable explanations.

2 Related works

2.1 AI-generated video detection method

In recent years, research on AI-generated video detection has primarily focused on deepfake videos featuring synthetic faces (Pei et al., 2024), with methods based on spatial-temporal consistency, frequency artifacts, and data-driven approaches. However, these approaches often struggle to gen-

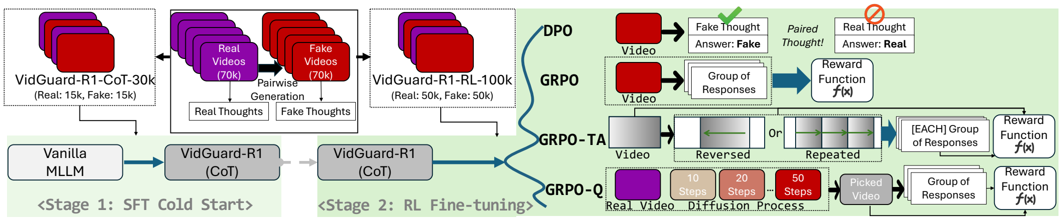


Figure 2: The overall training framework of **VidGuard-R1**, consisting of two stages: (1) supervised fine-tuning (SFT) for chain-of-thought (CoT) initialization, and (2) reinforcement learning-based fine-tuning to enable deeper reasoning.

eralize beyond face-centric content to more diverse, real-world videos. Recently, general video detection methods have emerged: AIGDet (Bai et al., 2024a) captures spatial-temporal anomalies, DeCoF (Ma et al., 2024) exploits frame consistency, and diffusion-based representations have been used to track temporal dynamics (Liu et al., 2024). Other works identify key factors such as appearance, motion, and geometry for classifier training (Chang et al., 2024). MLLMs have also been explored for visual forgery detection: FakeShield (Xu et al., 2024) employs supervised fine-tuning (SFT) for image forgery detection and relies on GPT-4o for explanation generation, while SafeWatch (Chen et al., 2024b) applies both SFT and direct preference optimization (DPO) to train a video guardrail model with transparent reasoning. In contrast, our work is the first to fine-tune a multi-modal large language model (MLLM) using group relative policy optimization (GRPO) for AI-generated video detection, demonstrating strong generalization across a wide range of recent video generative models and benchmark datasets.

2.2 AI-generated video detection dataset

As this is a very recent area of research, only a few benchmarks have been proposed. The generated video dataset (GVD) (Bai et al., 2024a) (11k samples) and GenVideo (Chen et al., 2024a) (with millions of samples) consider settings where both training and test videos are generated by the same series of models. However, these benchmarks lack prompt/image–video pairs, semantic labels, or cross-source settings. GVF (2.8k samples) contains prompts/images–video pairs and semantic labels, but does not provide cross-source settings. GenVidBench (Ni et al., 2025) consists of 100k videos and incorporates cross-source settings, but the video generation models used are less advanced, such as CogVideo and SVD. To address these limitations, we construct a curated dataset of 140,000 real–fake video pairs generated with state-of-the-art video generation models (HunyuanVideo (Kong et al., 2024) and CogVideoX (Yang et al., 2024)), standardizing video properties to mitigate superficial cues and encourage models to focus on intrinsic visual realism.

3 Methodology

Figure 2 illustrates the **VidGuard-R1** framework, which consists of two stages. We first apply supervised fine-tuning (SFT) to the multimodal large language model (MLLM), followed by direct preference optimization (DPO) and group relative policy optimization (GRPO) based on the collected datasets. We further develop two GRPO variants by introducing temporal artifacts and leveraging videos generated with varying diffusion steps.

3.1 Data collection

3.1.1 Data construction for video realism discrimination

High-quality training data is essential for video reasoning in MLLMs.. However, many existing benchmarks for real vs. generated video classification, such as GenVideo (Chen et al., 2024a) and GenVidBench (Ni et al., 2025), exhibit uncontrolled discrepancies in basic metadata—e.g., real videos are often longer than 10 seconds while generated ones are typically under 4 seconds in GenVideo. Moreover, they reveal clear modality-level gaps in motion dynamics and content contrasts between real and generated videos. These differences introduce unintended shortcuts, enabling models to rely on superficial cues like duration or resolution rather than actual visual realism. As a result, **VidGuard-R1** attains over 96% accuracy on both GenVideo and GenVidBench by exploiting such artifacts. To mitigate this reward hacking behavior, we construct a curated dataset with standardized video properties, encouraging models to focus on intrinsic visual content.

We collect real videos from the InternVid (Wang et al., 2023c) and ActivityNet (Caba Heilbron et al., 2015) datasets and generate their corresponding fake counterparts using HunyuanVideo (Kong et al., 2024) and CogVideoX (Yang et al., 2024). We specifically choose these two models because they support conditioning on both the first-frame image and a text description—an essential requirement for generating videos that are contextually aligned with their real counterparts. To achieve such alignment, we provide the generation models with the first frame of each real video along with a textual caption describing its content. For ActivityNet, which lacks native captions, we extract concise descriptions using Qwen2.5-VL 72B. This pairing strategy mitigates content-based biases and forces the model to reason over subtle visual details.

3.1.2 Collecting chain of thought (CoT) annotation

Eliciting deliberate, step-by-step reasoning in MLLMs requires high-quality CoT supervision. To this end, we leverage Qwen-2.5-VL (72B) to extract salient visual cues from each video and guide the model toward a deeper understanding. Specifically, we query the model with critical factors known to distinguish real from generated content—motion consistency, lighting consistency, texture artifacts, and physical plausibility violations. These targeted prompts encourage detailed reasoning grounded in visual evidence.

However, current MLLMs lack the capacity to reliably distinguish real from fake videos on their own. To compensate, we provide ground-truth labels during prompt construction and instruct the model to generate CoT rationales conditioned on the given label. While these rationales do not reflect genuine discrimination ability, they capture rich contextual cues—such as object interactions, background details, and lighting inconsistencies—that are highly informative. These CoT annotations serve as useful clues for subsequent reinforcement learning fine-tuning. For prompt templates used in CoT generation, please refer to our supplementary materials.

3.2 Supervised and RL Fine-tuning

We begin with SFT, where the model is trained to mimic the ground-truth reasoning process. Given a video x and its annotation y from the collected dataset, the model parameters θ are optimized by minimizing the negative log-likelihood $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x)$. To align the model outputs with human preferences, we apply DPO, which updates the model based on pairwise preference data without explicit reward modeling. Given a preferred response y_w and a less-preferred response y_l for the same video x, the DPO loss encourages the model to prefer y_w over y_l compared to a reference model p_{ref}:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\,\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\Big(\beta \log \frac{p_\theta(y_w|x)}{p_{\text{ref}}(y_w|x)} - \beta \log \frac{p_\theta(y_l|x)}{p_{\text{ref}}(y_l|x)}\Big)\right]$$

where $\sigma(\cdot)$ is the sigmoid function and β controls the preference strength. This method allows fine-tuning using preference comparisons without requiring scalar rewards.

Finally, we adopt GRPO from DeepSeek R1 (Guo et al., 2025), which generalizes RLHF to group-level comparisons. Given a query video x and a group of generated outputs $\{o_i\}_{i=1}^{G}$, the model is trained to assign higher probabilities to outputs with higher rewards. The GRPO objective is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}_{(x,o_{1:G})\sim D}\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{p_\theta(o_i|x)}{p_{\text{ref}}(o_i|x)} A_i, \text{clip}\left(\frac{p_\theta(o_i|x)}{p_{\text{ref}}(o_i|x)}, 1-\epsilon, 1+\epsilon\right) A_i\right) - \beta D_{\text{KL}}\left(p_\theta \,\|\, p_{\text{ref}}\right)\right]$$

where ϵ is a clipping threshold and β regularizes the policy to stay close to the reference model. The advantage term A_i normalizes the reward r_i for output o_i within the group, computed as $A_i = \frac{r_i - \mu_x}{\sigma_x}$, where μ_x and σ_x are the mean and standard deviation of $\{r_i\}_{i=1}^{G}$. GRPO thus enables learning from relative ranking among multiple responses, capturing nuanced distinctions in quality across outputs.

3.3 VIDGUARD-R1

3.3.1 OVERVIEW

Figure 2 illustrates the training pipeline of **VidGuard-R1**. Following the data collection procedure, we construct two datasets of different scales: `VidGuard-R1-CoT-30k` and `VidGuard-R1-RL-100k`. We adopt Qwen2.5-VL-7B as the base MLLM and train it using our proposed fine-tuning framework.

The first stage is supervised fine-tuning initialization using the `VidGuard-R1-CoT-30k` dataset, which contains videos paired with chain-of-thought (CoT) annotations. This stage establishes foundational reasoning ability and equips the model with basic cross-modal alignment and visual understanding. The resulting model is referred to as **VidGuard-R1 (CoT)**.

In the second stage, we apply two reinforcement learning methods—DPO and GRPO—to further refine the model on a larger and more diverse dataset, `VidGuard-R1-RL-100k`. DPO aligns the model with high-quality preference signals via pairwise comparisons, requiring the construction of preference pairs. Specifically, since our dataset includes pairwise real and fake videos, each sample is annotated with CoT rationales for both perspectives. For DPO training, we construct preference pairs by swapping these CoTs. For a real video, the CoT supporting its authenticity with the answer "real" serves as the positive annotation, while the CoT from the paired fake video with the answer "fake" is used as the negative annotation. In contrast, GRPO encourages consistent performance across grouped outputs by leveraging structural regularization. As it does not rely on preference annotations, video labels are directly used as reward signals. The resulting models are denoted as **VidGuard-R1 (DPO)** and **VidGuard-R1 (GRPO)**.

We introduce two variants, GRPO-TA and GRPO-Q, to further enhance detection performance. These methods extend the original GRPO framework by adjusting reward values according to the difficulty of detecting fake videos. Detailed descriptions are provided in the following sections.

3.3.2 GRPO WITH TEMPORAL ARTIFACTS (GRPO-TA)

While standard GRPO performs well in video discrimination by leveraging local visual cues—such as pixel distortions and lighting inconsistencies—it often overlooks temporal inconsistencies, which are crucial for detecting generated videos. To address this limitation, we introduce **GRPO with temporal artifacts (GRPO-TA)**, a variant that explicitly promotes temporal reasoning through a contrastive reward adjustment.

We apply two common temporal artifacts: (1) repeating a specific video segment and (2) reversing the frame sequence within a segment. These manipulations are applied probabilistically, with the manipulated region selected based on a Gaussian distribution over the video timeline.

Specifically, for each input query, we generate two sets of model outputs: $\{o_i\}_{i=1}^{G}$ for the original video, and $\{\tilde{o}_i\}_{i=1}^{G'}$ for the corresponding manipulated video with temporal artifacts. These videos should be classified as fake videos. In GRPO-TA, we assign additional rewards when the model correctly classifies temporally manipulated videos as fake. Consider two numbers, $\alpha_1 > \alpha_2$. Detecting temporal artifacts in videos manipulated from real content tends to be more challenging than identifying those derived from fake videos. This is because real videos typically exhibit coherent and natural motion, so temporal manipulations such as frame shuffling or repetition can be subtle and difficult to detect. In contrast, generated videos often contain artifacts like unstable motion or low temporal consistency, which make further manipulations more visually salient. To reflect this asymmetry in difficulty, we assign the model a higher reward α_1 when the original video o_i is real, and a moderate reward α_2 when the original video is fake. This is defined as:

$$w_i = \begin{cases} \alpha_1, & \text{if } \tilde{o}_i = \textit{fake} \text{ and } y_i = \textit{real} \\ \alpha_2, & \text{if } \tilde{o}_i = \textit{fake} \text{ and } y_i = \textit{fake} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where y_i denotes the label of the i-th video. In the experiments, we set the hyperparameters to $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$. This additional reward, w_i, is designed to be applied conditionally. Specifically, for a given sample, we only add w_i to the original GRPO reward if two conditions are met: the model's prediction on the original video (O_i) must be correct, and the overall accuracy on the group of manipulated videos ($\tilde{p}$) must exceed a predefined threshold μ. This ensures that we only reward the model for temporal reasoning when it already has a solid baseline performance. The final reward of GRPO-TA is given by

$$r_i^{\text{GRPO-TA}} = \begin{cases} r_i^{\text{GRPO}} + w_i, & \text{if } o_i \text{ is correct and } \tilde{p} > \mu \\ r_i^{\text{GRPO}}, & \text{otherwise} \end{cases} \tag{2}$$

where r_i denotes the original GRPO reward, set to 1 if the model prediction on the original video is correct, and 0 otherwise. The additional reward w_i is applied only when both the original prediction is correct and the group of responses for the temporally manipulated videos achieves higher accuracy. In the experiments, we set $\mu = 0.8$.

3.3.3 GRPO WITH QUALITY EVOLUTIONARY VIDEOS (GRPO-Q)

Our motivation is to extend the model's capability to detect videos based on quality. Given the subjective nature of quality assessment, we avoid relying on large-scale human annotations. Instead, we leverage diffusion-based video generation by systematically varying the number of reverse diffusion steps to produce videos with distinct quality levels.

As in GRPO-TA, $o_i \in \mathcal{Y}$ and $y_i \in \mathcal{Y}$ denote the model output and ground-truth label, with $\mathcal{Y} = \{\text{real}\} \cup \{\text{fake-}s\}$, where s is the diffusion step. A reward is given for an exact match, and no reward is assigned if the real/fake classification is incorrect. In GRPO-Q, if the model correctly classifies a fake video but selects an incorrect diffusion step, we assign a partial reward based on the distance between the predicted and ground-truth diffusion steps. The GRPO-Q reward is defined as follows:

$$r_i^{\text{GRPO-Q}} = \begin{cases} 0, & \text{if } \big(o_i = \textit{real} \text{ and } y_i \neq \textit{real}\big) \text{or } \big(o_i \neq \textit{real} \text{ and } y_i = \textit{real}\big) \\ \delta, & \text{if } o_i = y_i \\ \left| g\left(o_i, y_i\right) \right|, & \text{if } o_i, y_i \in \mathcal{Y} \setminus \{\textit{real}\}. \end{cases} \tag{3}$$

The first scenario occurs when the model fails to correctly classify the video as real or fake. The second scenario, where $\delta = 1$, represents an exact match in prediction, including the diffusion progression. In the third case, the function $g(\cdot, \cdot)$ maps the step distance to a scalar reward, enabling fine-grained credit assignment based on the similarity in quality. Specifically, we define a progress value $s()$ in the range $[0, 1]$ to indicate the fraction of diffusion steps used, where 0 denotes zero steps, and 1 denotes full completion of the steps. The ground-truth value is $s(y_i)$, and the model will estimate a progress value. We define the reward function as $g(o_i, y_i) = \delta \cdot (1 - |s(o_i) - s(y_i)|)$.

This reward formulation enables the model to move beyond binary discrimination and perform fine-grained analysis of video quality. By learning to associate subtle differences in generation steps with quality variations, the model develops a deeper understanding of the diffusion process and its impact on perceptual realism. As a result, it can not only detect whether a video is fake, but also infer but also estimate the degree of quality degradation in generated videos. This facilitates more interpretable and controllable evaluation of generated content quality.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

4.1.1 DATASET

Our dataset contains 140k videos, balanced between 70k real and 70k generated samples, organized into contextual pairs. The real set comprises 55k videos from InternVid and 15k from ActivityNet, while the generated set includes 50k samples synthesized by HunyuanVideo-I2V (Kong et al., 2024) and 20k by CogVideoX-5B (Yang et al., 2024). We allocate 130k samples for training and 10k for testing, with the latter evenly split between real and generated videos. Within the training data, 30k samples are reserved for chain-of-thought (CoT) learning, denoted as `VidGuard-R1-CoT-30k`, and the remaining 100k are used for reinforcement learning fine-tuning, denoted as `VidGuard-R1-RL-100k`.

Since state-of-the-art generative models still produce relatively short videos (∼129 frames) at modest resolutions, we standardize all real videos to match generated ones by enforcing 49 frames, 8 FPS, 720×480 resolution, and YUV420p format.

Table 1: Comparison of models on our dataset, reported as mean Top-1 accuracy (%). TF denotes transformer.

Method	Type	CogVideoX	HunyuanVideo
SlowFast	CNN	77.87	77.03
I3D	CNN	64.78	62.13
TRN	CNN	68.73	69.87
UniFormer V2	TF	73.95	71.92
TimeSformer	TF	78.53	74.55
VideoSwin	TF	76.81	79.71
MViT V2	TF	58.38	53.91
Qwen2.5-VL-7B	MLLM	50.95	52.83
GPT-4.1 mini	MLLM	54.95	55.31
VidGuard-R1 (CoT)	MLLM	66.18	63.19
VidGuard-R1 (DPO)	MLLM	79.13	80.88
VidGuard-R1 (GRPO)	MLLM	81.30	81.90
VidGuard-R1 (GRPO-TA)	MLLM	82.17	83.72
VidGuard-R1 (GRPO-Q)	MLLM	**84.32**	**86.17**

Table 2: Extended GenVidBench results with **VidGuard-R1** and additional MLLMs, reported as mean Top-1 accuracy (%). TF denotes transformer.

Method	Type	MuseV	SVD	CogVideo	Mora	HD-VG	Mean
SlowFast	CNN	12.25	12.68	38.34	45.93	93.63	41.66
I3D	CNN	8.15	8.29	60.11	59.24	93.99	49.23
TRN	CNN	38.92	26.64	91.34	93.98	93.97	71.26
UniFormer V2	TF	20.05	14.81	45.21	99.21	96.89	57.55
TimeSformer	TF	73.14	20.17	74.80	39.40	92.32	64.28
VideoSwin	TF	62.29	8.01	91.82	45.83	**99.29**	67.27
MViT V2	TF	76.34	**98.29**	47.50	96.62	97.58	79.90
Qwen2.5-VL-7B	MLLM	25.86	27.06	68.51	43.26	71.15	47.30
GPT-4.1 mini	MLLM	26.07	33.78	94.07	57.19	87.64	59.62
VidGuard-R1 (CoT)	MLLM	36.52	16.02	99.35	76.94	99.94	66.09
VidGuard-R1 (GRPO, GenVideo-pretrained, Zero-shot)	MLLM	97.24	96.59	99.88	99.93	88.14	96.37
VidGuard-R1 (GRPO)	MLLM	**97.38**	94.98	**99.90**	**99.99**	95.46	**97.53**

For GRPO-Q fine-tuning, we augment the training set with intermediate generations sampled from diffusion steps 10 to 50. These are labeled with approximate quality levels (20%, 40%, 60%, 80%, and 95%). Specifically, we use 12k real videos, each paired with five generated variants at different diffusion steps, resulting in 72k samples per generation model.

4.1.2 Evaluation protocol

We evaluate three datasets—ours, GenVidBench (Ni et al., 2025), and GenVideo (Chen et al., 2024a)—using the metrics and baselines defined by their respective benchmarks. For ours and GenVidBench, we report **mean Top-1 accuracy**, the average correctness over all predictions. For GenVideo, we follow the original protocol and report **recall** and **F1 score**. All evaluations adhere to the official settings of each benchmark to ensure fair comparison.

4.1.3 Training Setup

We employ Qwen2.5-VL-7B as the base MLLM and conduct all experiments on four NVIDIA A100 GPUs (80GB). Each video is represented by up to 16 frames, where each frame is resized to a 28×28 spatial resolution and mapped to 128 feature channels for encoder input during both training and inference. For GenVideo and GenVidBench, we follow their official evaluation protocols and adopt 8-frame inputs. In GRPO training, we sample 8 responses per input; for GRPO-TA, we additionally sample 4 responses from temporally manipulated variants of the input to enhance robustness against temporal artifacts. Training proceeds in two stages: first, the base model is fine-tuned for one epoch on the CoT dataset, yielding the SFT-CoT MLLM; second, we initialize **VidGuard-R1** with SFT-CoT and perform reinforcement learning for approximately 2,000 steps.

4.2 Main results

4.2.1 Our dataset

We evaluate **VidGuard-R1** on our dataset with several methods, including CNN-based models (SlowFast (Feichtenhofer et al., 2019), I3D (Carreira & Zisserman, 2017), TRN (Zhou et al., 2018)), Transformer-based models (UniFormer V2 Li et al. (2022a), TimeSformer (Bertasius et al., 2021), VideoSwin (Liu et al., 2022), MViT V2 (Li et al., 2022b)), and MLLM-based models (Qwen2.5-VL (Bai et al., 2025) and GPT-4.1 mini (OpenAI, 2025)). For CNN and Transformer models, we use the default training settings provided by the MMAction2 framework (Contributors, 2020).

As shown in Table 1, CNN- and Transformer-based models achieved 53–79% accuracy, with SlowFast and TimeSformer among the top performers. In contrast, Qwen2.5-VL-7B and GPT-4.1 mini exhibited near-random performance, highlighting their limited capability in distinguishing fake videos. *VidGuard-R1 (CoT)*, trained via supervised fine-tuning (SFT) on Qwen2.5-VL-7B, substantially improved accuracy from around 51% to over 66%, yet remained less competitive compared to advanced SOTA methods. This result aligns with the intended role of the SFT stage—as a cold start phase to guide the model toward structured *think* + *answer* responses, emphasizing the extraction of salient visual cues.

Table 3: Extended GenVideo results with **VidGuard-R1** and additional MLLMs, evaluated by F1 score and recall (R)

Method	Detection level	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Mean
NPR (Tan et al., 2024)	Image	R	0.91	0.99	0.99	0.24	0.89	0.57	0.97	0.99	0.94	0.87	0.84
		F1	0.27	0.84	0.91	0.30	0.86	0.59	0.81	0.91	0.81	0.81	0.71
VideoMAE (Tong et al., 2022)	Video	R	0.67	0.96	0.98	0.96	0.77	0.80	0.97	0.96	0.96	0.68	0.87
		F1	0.62	0.95	0.98	**0.96**	0.86	0.87	0.96	0.97	**0.96**	0.79	0.89
MINTIME-CLIP (Coccomini et al., 2024)	Video	R	0.89	1.00	0.98	0.26	0.96	0.98	0.99	1.00	0.84	0.82	0.87
		F1	0.49	0.93	0.96	0.37	0.94	0.92	0.92	0.96	0.84	0.85	0.82
FTCN-CLIP (Zheng et al., 2021)	Video	R	0.87	1.00	0.98	0.17	0.97	0.91	1.00	1.00	0.85	0.82	0.86
		F1	0.78	0.98	0.98	0.29	0.98	0.94	0.98	**0.99**	0.90	0.89	0.87
DeMamba-XCLIP (Chen et al., 2024a)	Video	R	0.98	1.00	0.99	0.65	0.94	0.98	1.00	1.00	0.92	0.89	0.93
		F1	0.64	0.96	0.97	0.75	**0.95**	0.95	0.95	0.97	0.92	**0.91**	0.90
Qwen2.5-VL-7B (Bai et al., 2025)	MLLM	R	0.58	0.56	0.54	0.33	0.43	0.38	0.81	0.63	0.51	0.70	0.54
		F1	0.74	0.72	0.70	0.49	0.60	0.55	0.90	0.77	0.68	0.82	0.70
GPT-4.1 mini (OpenAI, 2025)	MLLM	R	0.43	0.67	0.56	0.54	0.63	0.56	0.92	0.67	0.69	0.69	0.65
		F1	0.60	0.80	0.72	0.70	0.77	0.72	0.96	0.80	0.82	0.82	0.72
VidGuard-R1 (CoT)	MLLM	R	0.92	0.89	0.91	0.90	0.98	0.79	0.99	0.85	0.89	0.87	0.90
		F1	0.90	0.91	0.95	0.89	0.99	0.81	0.95	0.89	0.85	0.88	0.90
VidGuard-R1 (GRPO, GenVidBench-pretrained, Zero-shot)	MLLM	R	0.95	0.98	0.90	0.89	0.97	0.85	0.99	0.93	0.81	0.87	0.92
		F1	0.93	0.93	0.96	0.91	0.99	0.82	0.95	0.89	0.85	0.88	0.91
VidGuard-R1 (GRPO)	MLLM	R	0.95	1.00	0.98	0.94	0.98	0.95	0.97	0.99	0.94	0.91	0.96
		F1	**0.97**	**0.99**	**0.99**	0.91	**0.99**	0.89	**0.99**	**0.99**	0.95	0.90	**0.96**

In the subsequent RL stage, both DPO and GRPO further improved performance by roughly 2% over the best baseline. Our proposed methods—GRPO-TA and GRPO-Q—achieved additional gains of approximately 2% and 5% over GRPO, respectively, demonstrating the effectiveness of temporal artifact supervision and quality-aware reward modeling in enhancing detection accuracy.

4.2.2 GENVIDBENCH BENCHMARK

The GenVidBench dataset comprises approximately 87k training samples and 82k testing samples, with fake videos generated by models such as MuseV (Xia et al., 2024), SVD (Blattmann et al., 2023), CogVideo (Hong et al., 2022), and Mora (Yuan et al., 2024), and real videos sourced from HD-VG (Wang et al., 2023b). We conduct training and evaluation under the cross-source and cross-generator settings as proposed in their benchmark. In addition to the models originally reported in GenVidBench, we evaluate **VidGuard-R1** using the same model families as in our dataset experiments—CNN-based, Transformer-based, and MLLM-based models—including two MLLMs: Qwen2.5-VL and GPT-4.1 mini. `VidGuard-R1 (GRPO, GenVideo-pretrained, Zero-shot)` denotes the zero-shot model pretrained on GenVideo and evaluated on GenVidBench. As shown in Table 2, both the zero-shot model and two fine-tuned variants achieve over 15% higher accuracy compared to prior SOTA methods. Notably, the zero-shot model demonstrates strong generalization, highlighting the effectiveness of pretraining on diverse generative content. Complete detection model results are provided in Appendix B.

4.2.3 GENVIDEO BENCHMARK

The GenVideo dataset comprises approximately 2.2M training samples and 20k testing samples, with generated videos sourced from a diverse set of models, including Sora (OpenAI, 2024), MorphStudio (mor, 2025), Gen2 (Esser et al., 2023b), HotShot (hot, 2025), Lavie (Wang et al., 2025b), Show-1 (Zhang et al., 2024a), MoonValley (moo, 2025), Crafter (Chen et al., 2023), ModelScope (Wang et al., 2023a), and WildScrape (Wei et al., 2024). Following the official evaluation protocol, we benchmark two MLLM baselines and three variants of **VidGuard-R1**. Among these, `VidGuard-R1 (GRPO)` consistently outperforms almost all prior detection methods across videos generated by the various models. As shown in Table 3, it achieves an F1 score improvement of 0.06 compared to DeMamba-XCLIP. Complete detection model results are provided in Appendix B.

4.2.4 PERFORMANCE GAP BETWEEN OUR DATASET AND BENCHMARKS

While **VidGuard-R1** achieves approximately 85% accuracy on our dataset, it obtains significantly higher accuracy—exceeding 95%—on the two benchmark datasets. This discrepancy arises from two key differences. First, the benchmarks exhibit clear discrepancies in video metadata—such as resolution, duration, and frame rate—between real and fake videos, which models can exploit as superficial cues. In contrast, we standardize all videos in our dataset by matching resolution, FPS, and format, thereby forcing models to rely on actual visual content. Second, our dataset ensures strong

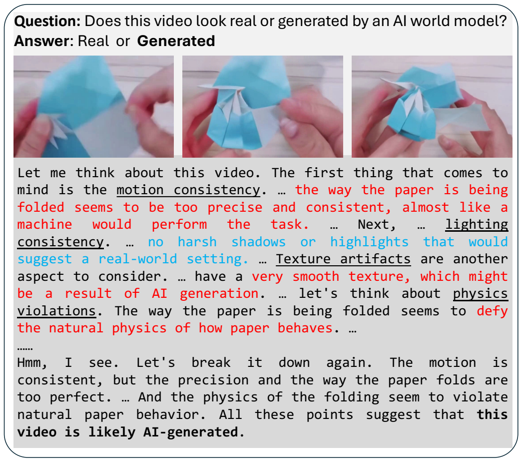


Figure 3: VidGuard-R1 (GRPO): reasoning about an origami folding sequence

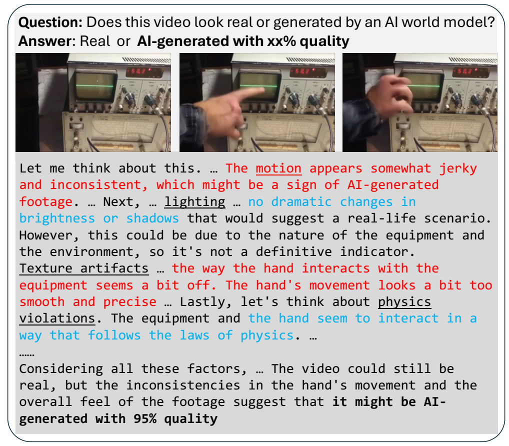


Figure 4: VidGuard-R1 (GRPO-Q): temporal inconsistency observed as unnatural movements

contextual alignment by conditioning generation on the first frame and the corresponding caption of a real video, resulting in more realistic and semantically consistent outputs. In comparison, benchmark datasets often generate fake videos from unrelated prompts and synthetic images, leading to artifacts that make detection easier.

4.2.5 Ablation study

Explanation quality and accuracy comparison Table 4 presents results on the HunyuanVideo (Kong et al., 2024) and CogVideoX (Yang et al., 2024) datasets. We report explanation quality scores, which are rated on a 1–10 scale (with 10 indicating excellent quality and full alie code is publicly availagnment) by GPT-4.1 mini using the LLM-as-a-Judge prompt described in Appendix C. Compared to baseline models such as Qwen2.5-VL-7B and GPT-4.1 mini, our VidGuard-R1 GRPO variants achieve consistent improvements in both classification accuracy and explanation quality.

Table 4: LLM-as-a-Judge explanation scores on our dataset

Method	Expl. Score (HunyuanVideo)	Expl. Score (CogVideoX)
Qwen2.5-VL-7B	5.8	5.6
GPT-4.1 mini	5.8	5.9
VidGuard-R1 (CoT)	6.8	6.9
VidGuard-R1 (DPO)	7.2	8.1
VidGuard-R1 (GRPO)	8.1	8.0
VidGuard-R1 (GRPO-TA)	8.1	8.4
VidGuard-R1 (GRPO-Q)	**8.3**	**8.5**

GRPO-TA reward ablation Table 5 reports an ablation study of GRPO-TA on our dataset by varying the weight parameters α_1 and α_2, which control the relative importance of different temporal artifact types. The highest classification accuracy of 83.57% is achieved with $\alpha_1 = 0.5$ and $\alpha_2 = 0.3$, while the threshold μ is fixed at 0.8 across all experiments.

Table 5: Accuracy (%) for **GRPO-TA** under different reward function parameters α_1 and α_2

α_1	α_2	Accuracy (%)
0.3	0.1	81.31
0.3	0.3	82.59
0.5	**0.3**	**83.57**
0.5	0.5	83.12
0.7	0.5	82.53

GRPO-Q reward ablation Table 6 presents an ablation study on GRPO-Q conducted on our dataset by varying the number of intermediate diffusion steps included per real video during fine-tuning. Using more steps provides richer supervision of video quality progression, improving detection accuracy. The best accuracy of 84.05% is obtained with five steps, which is the setting used in our main experiments.

Table 6: Accuracy (%) for **GRPO-Q** with varying number of intermediate diffusion steps

# of steps (step numbers)	Accuracy (%)
1 (50)	81.63
3 (10, 30, 50)	83.21
5 (10, 20, 30, 40, 50)	**85.80**

4.3 CASE STUDIES ON EXPLANATIONS

Figures 3 and 4 illustrate cases where **VidGuard-R1** correctly identifies videos as generated. The model performs multi-faceted reasoning across motion, lighting, texture, and physical plausibility before arriving at a final decision. Notably, it does not rely on a single cue, but instead accumulates evidence across frames, resembling how humans distinguish fake videos. In each figure, pink highlights denote cues suggesting realism, red indicates artifacts indicative of generation, yellow marks intermediate reasoning steps, and underlines represent several key factors.

For instance, in Figure 3, the smooth hand motion initially suggests realism; however, once the origami folds in a physically implausible manner, the model revises its judgment. In Figure 4, although the lighting and shadows are consistent—typically a cue for authenticity—the model recognizes that this is insufficient in a largely static scene with only a stationary machine and a human hand. In particular, even in its final prediction, the model reflects on earlier realistic cues and acknowledges that *the video could still be real*, underscoring its nuanced, human-like reasoning in assessing video quality. Additional case studies are provided in the Appendix D.

5 CONCLUSION

We propose **VidGuard-R1**, an MLLM-based discriminator that not only detects AI-generated videos with high accuracy but also provides interpretable reasoning. By leveraging reinforcement learning with reward models targeting temporal artifacts and generation quality, **VidGuard-R1** achieves 85% accuracy on our dataset, 97% on GenVidBench, and 96% on GenVideo, substantially surpassing prior state-of-the-art methods. We expect this work to advance MLLM-based video analysis and foster future research on strengthening MLLMs' reasoning.

5.1 LIMITATIONS

Our dataset currently includes fake videos generated exclusively with the HunyuanVideo and CogVideoX. Although it is constructed in a pairwise fashion to ensure contextual similarity between real and fake videos, its generalizability remains limited. Expanding the dataset to incorporate fake videos from a broader range of generative models would result in a more diverse and robust training set, thereby improving its applicability to real-world scenarios.

REFERENCES

Hotshot. Online, 2025. Available:https://huggingface.co/hotshotco/Hotshot-XL/.

Moonvalley. Online, 2025. Available:https://moonvalley.ai/.

Morph studio. Online, 2025. Available:https://www.morphstudio.com/.

Jianfa Bai, Man Lin, and Gang Cao. Ai-generated video detection via spatio-temporal anomaly learning. *arXiv preprint arXiv:2403.16638*, 2024a.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 460–470. Springer, 2024b.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

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

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

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 4724–4733, 07 2017. doi: 10.1109/CVPR.2017.502.

Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv preprint arXiv:2406.19568*, 2024.

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

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

Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*, 2024b.

Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. Mintime: Multi-identity size-invariant video deepfake detection. *IEEE Transactions on Information Forensics and Security*, 19:6084–6096, 2024.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7346–7356, 2023a.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7346–7356, 2023b.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. URL https://arxiv.org/abs/2004.04730.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 6202–6211, 2019.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3473–3481, 2021.

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

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28050–28060, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022b.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2019. URL https://arxiv.org/abs/1811.08383.

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns out i'm not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*, 2024.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Yong Liao, and Haiyang Yu. Decof: Generated video detection via frame consistency, 2024.

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos, 2019. URL https://arxiv.org/abs/1910.12467.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

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 preprint arXiv:2501.11340*, 2025.

OpenAI. Sora. Online, 2024. Available:https://openai.com/index/sora/.

OpenAI. Gpt-4.1. Online, 2025. Available:https://openai.com/index/gpt-4-1/.

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020a. URL https://arxiv.org/abs/2007.09355.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. *AAAI*, 2020.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023b.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023c.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6537–6549, 2024.

Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. Musev: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising. *arxiv*, 2024.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22658–22668, 2023.

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 preprint arXiv:2410.02761*, 2024.

Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework, 2024.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pp. 1–15, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15044–15054, 2021.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. *European Conference on Computer Vision*, 2018.

Table 7: Extended GenVidBench results with **VidGuard-R1** and additional MLLMs, reported as mean Top-1 accuracy (%). TF denotes transformer.

Method	Type	MuseV	SVD	CogVideo	Mora	HD-VG	Mean
SlowFast (Feichtenhofer et al., 2019)	CNN	12.25	12.68	38.34	45.93	93.63	41.66
F3Net (Qian et al., 2020a)	CNN	37.43	37.27	36.46	39.59	52.76	42.52
I3D (Carreira & Zisserman, 2017)	CNN	8.15	8.29	60.11	59.24	93.99	49.23
CFV2 (Nguyen et al., 2019)	CNN	86.26	86.53	10.10	16.90	88.40	60.53
TPN (Yang et al., 2020)	CNN	37.86	8.79	68.25	90.04	97.34	61.52
TIN (Shao et al., 2020)	CNN	33.78	21.47	81.59	79.44	97.88	63.97
TRN (Zhou et al., 2018)	CNN	38.92	26.64	91.34	93.98	93.97	71.26
TSM (Lin et al., 2019)	CNN	70.37	54.70	78.46	70.37	96.76	76.40
X3D (Feichtenhofer, 2020)	CNN	92.39	37.27	65.72	49.60	97.51	77.09
UniFormer V2 (Li et al., 2022a)	TF	20.05	14.81	45.21	99.21	96.89	57.55
TimeSformer (Bertasius et al., 2021)	TF	73.14	20.17	74.80	39.40	92.32	64.28
VideoSwin (Liu et al., 2022)	TF	62.29	8.01	91.82	45.83	**99.29**	67.27
MViT V2 (Li et al., 2022b)	TF	76.34	**98.29**	47.50	96.62	97.58	79.90
Qwen2.5-VL-7B (Bai et al., 2025)	MLLM	25.86	27.06	68.51	43.26	71.15	47.30
GPT-4.1 mini (OpenAI, 2025)	MLLM	26.07	33.78	94.07	57.19	87.64	59.62
VidGuard-R1 (CoT)	MLLM	36.52	16.02	99.35	76.94	99.94	66.09
VidGuard-R1 (GRPO, GenVideo-pretrained, Zero-shot)	MLLM	97.24	96.59	99.88	99.93	88.14	96.37
VidGuard-R1 (GRPO)	MLLM	**97.38**	94.98	**99.90**	**99.99**	95.46	**97.53**

A Additional Setup

To further guide the model during RL training, we incorporate a length-based reward strategy. We promote informative yet concise reasoning by rewarding outputs that are neither too brief nor excessively long. Specifically, if the model predicts the correct answer and the length of the response falls within the range $[l_{\min}, l_{\max}]$, an additional reward ω is assigned. Let l_i be the length of the model's response for the i-th video. The reward is defined as:

$$r_i^{total} = \begin{cases} r_i + \omega, & \text{if } o_i \text{ is correct and } l_{\min} \leq l_i \leq l_{\max} \\ r_i, & \text{otherwise} \end{cases} \tag{4}$$

where we set $\omega = 0.1$, $l_{\min} = 320$, and $l_{\max} = 512$.

B Comprehensive Benchmark Evaluation

In this section, we provide extended benchmark results for **VidGuard-R1** alongside additional MLLMs. Table 7 presents mean Top-1 accuracy on GenVidBench across multiple video datasets, including CNN and Transformer baselines as well as selected MLLM variants. Table 8 reports comprehensive F1 and recall scores on the GenVideo dataset, including all models provided in the official benchmark alongside our MLLM variants. These extended tables offer a complete comparison of performance across all evaluated models.

C Prompt

Figure 5 shows the base prompt used for the real-vs-fake classification task. Annotators are instructed to assess whether a video is real or AI-generated by analyzing key visual and physical properties.

Figures 6 and 7 provide category-specific rationale collection prompts. In particular, Figure 6 presents the prompt for identifying visual cues of realism in real videos, while Figure 7 focuses on spotting artifacts in AI-generated videos. Both prompts guide annotators to evaluate videos across four diagnostic categories: motion consistency, lighting consistency, texture artifacts, and physics violations.

Table 8: Extended GenVideo results with **VidGuard-R1** and additional MLLMs, evaluated by F1 and recall scores

Model	Detection level	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Mean
F3Net (Qian et al., 2020a)	Image	R	0.83	0.99	0.98	0.77	0.57	0.36	0.99	0.99	0.89	0.76	0.81
		F1	0.50	0.94	0.96	0.81	0.69	0.49	0.93	0.96	0.88	0.82	0.80
NPR (Tan et al., 2024)	Image	R	0.91	0.99	0.99	0.24	0.89	0.57	0.97	0.99	0.94	0.87	0.84
		F1	0.27	0.84	0.91	0.30	0.86	0.59	0.81	0.91	0.81	0.81	0.71
STIL (Gu et al., 2021)	Video	R	0.78	0.98	0.98	0.76	0.61	0.53	0.99	0.97	0.94	0.65	0.82
		F1	0.38	0.90	0.94	0.78	0.72	0.62	0.90	0.94	0.88	0.72	0.78
VideoMAE (Tong et al., 2022)	Video	R	0.67	0.96	0.98	0.96	0.77	0.80	0.97	0.96	0.96	0.68	0.87
		F1	0.62	0.95	0.98	**0.96**	0.86	0.87	0.96	0.97	**0.96**	0.79	0.89
MINTIME-CLIP (Coccomini et al., 2024)	Video	R	0.89	1.00	0.98	0.26	0.96	0.98	0.99	1.00	0.84	0.82	0.87
		F1	0.49	0.93	0.96	0.37	0.94	0.92	0.92	0.96	0.84	0.85	0.82
FTCN-CLIP (Zheng et al., 2021)	Video	R	0.87	1.00	0.98	0.17	0.97	0.91	1.00	1.00	0.85	0.82	0.86
		F1	0.78	0.98	0.98	0.29	0.98	0.94	0.98	**0.99**	0.90	0.89	0.87
TALL (Xu et al., 2023)	Video	R	0.91	0.98	0.97	0.83	0.76	0.79	0.99	0.98	0.94	0.66	0.88
		F1	0.26	0.82	0.89	0.74	0.77	0.72	0.81	0.90	0.80	0.67	0.74
CLIP (Radford et al., 2021)	Image	R	0.94	0.99	0.91	0.77	0.88	0.86	0.99	0.99	0.84	0.84	0.90
		F1	0.28	0.84	0.86	0.72	0.85	0.76	0.82	0.91	0.76	0.79	0.76
DeMamba-CLIP (Chen et al., 2024a)	Video	R	0.95	1.00	0.98	0.69	0.92	0.93	1.00	1.00	0.83	0.82	0.91
		F1	0.64	0.96	0.97	0.78	0.94	0.92	0.95	0.98	0.87	0.87	0.89
XCLIP (Ni et al., 2022)	Video	R	0.82	0.99	0.93	0.61	0.79	0.69	0.97	0.99	0.77	0.83	0.84
		F1	0.31	0.88	0.90	0.65	0.82	0.70	0.86	0.93	0.75	0.82	0.76
DeMamba-XCLIP (Chen et al., 2024a)	Video	R	0.98	1.00	0.99	0.65	0.94	0.98	1.00	1.00	0.92	0.89	0.93
		F1	0.64	0.96	0.97	0.75	**0.95**	0.95	0.95	0.97	0.92	**0.91**	0.90
Qwen2.5-VL-7B (Bai et al., 2025)	MLLM	R	0.58	0.56	0.54	0.33	0.43	0.38	0.81	0.63	0.51	0.70	0.54
		F1	0.74	0.72	0.70	0.49	0.60	0.55	0.90	0.77	0.68	0.82	0.70
GPT-4.1 mini (OpenAI, 2025)	MLLM	R	0.43	0.67	0.56	0.54	0.63	0.56	0.92	0.67	0.69	0.69	0.65
		F1	0.60	0.80	0.72	0.70	0.77	0.72	0.96	0.80	0.82	0.82	0.72
VidGuard-R1 (CoT)	MLLM	R	0.92	0.89	0.91	0.90	0.98	0.79	0.99	0.85	0.89	0.87	0.90
		F1	0.90	0.91	0.95	0.89	0.99	0.81	0.95	0.89	0.85	0.88	0.90
VidGuard-R1 (GRPO, GenVidBench-pretrained, Zero-shot)	MLLM	R	0.95	0.98	0.90	0.89	0.97	0.85	0.99	0.93	0.81	0.87	0.92
		F1	0.93	0.93	0.96	0.91	0.99	0.82	0.95	0.89	0.85	0.88	0.91
VidGuard-R1 (GRPO)	MLLM	R	0.95	1.00	0.98	0.94	0.98	0.95	0.97	0.99	0.94	0.91	0.96
		F1	**0.97**	**0.99**	**0.99**	0.91	**0.99**	0.89	**0.99**	**0.99**	0.95	0.90	**0.96**

Figure 8 illustrates the LLM-as-a-Judge prompt used to evaluate rationale quality. In this setting, GPT-4.1 mini rates the quality of model-generated explanations on a 1–10 scale, where a score of 10 corresponds to excellent quality and full alignment with the ground truth rationale.

Prompt for Distinguishing Real from AI-Generated Content

SYSTEM:
A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

USER:
<video> Decide whether a video looks a real one or a generated from the AI world model.

Figure 5: Prompt for identifying realism cues in real videos across four categories

Rationale Collection Prompt for Real Videos

<video> This is a real-world video. Your task is to provide a detailed guide of which specific parts of the video should be examined to identify signs of real across four key categories: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations. For each category, highlight critical areas or elements within the video.

Figure 6: Prompt for identifying realism cues in real videos across four categories

Rationale Collection Prompt for AI-Generated Videos

<video> This video has been generated by an AI model. Your task is to provide a detailed guide on which parts of the video identify signs of generation across four key categories: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations. For each category, highlight critical areas or elements within the video.

Figure 7: Prompt for identifying artifacts in AI-generated videos across four categories

LLM-as-a-Judge Prompt for Rationale Quality Evaluation in Real vs. Generated Video Classification

SYSTEM:
You are an expert judge evaluating the **explanation quality** of a vision-language model (VLM) that decides whether a video is real or AI-generated. The model outputs a binary decision (**real** or **fake**) along with a **rationale** explaining the basis of its decision.
The rationale should focus on four visual diagnostic categories:
Your evaluation should consider the following five criteria:
Accuracy: Does the rationale identify the key generation artifacts or natural signals relevant to the decision?
Alignment with Ground Truth: Does the rationale emphasize the same visual evidence?
Specificity: Is the rationale grounded in the visual content (i.e., free from hallucinations or vague generalities)?
Conciseness: Is the rationale clearly and efficiently worded without unnecessary length or redundancy?
Consistency with Model's Answer: Is the rationale logically coherent with the model's predicted label (real/fake)? For example, if the model said "fake," does the rationale convincingly support this?
Please provide your evaluation in the following format:
Start with "ANALYSIS:" and write a concise paragraph analyzing the rationale with respect to the five criteria.
Then, on the next line, write "SCORE:" followed by a single integer from 1 to 10 (10 = excellent quality and full alignment).

USER:
Let's evaluate the model's rationale!
Ground truth answer (real/fake):
{ground truth answer}
Ground truth rationale:
{ground truth rationale}
Model answer (real/fake):
{model answer}
Model rationale:
{model rationale}

Figure 8: LLM-as-a-Judge Prompt for Rationale Quality Evaluation

D CASE STUDIES ON EXPLANATIONS

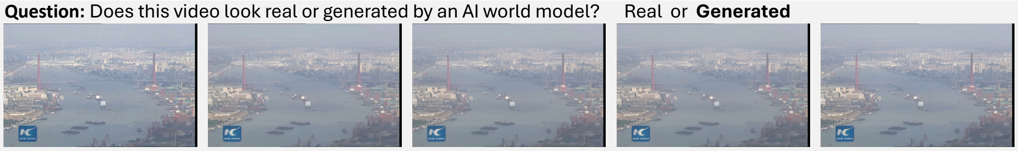
D.1 GENVIDBENCH

Figures 9–12 present inference examples for videos synthesized by four distinct AI models included in the GenVidBench testing dataset: MuseV (Xia et al., 2024), SVD (Blattmann et al., 2023), CogVideo (Hong et al., 2022), and Mora (Yuan et al., 2024).

D.2 GENVIDEO

Figures 13–22 show inference examples for videos generated by ten different AI models included in the GenVideo testing dataset: Sora (OpenAI, 2024), Morph Studio (mor, 2025), Gen2 (Esser et al., 2023a), HotShot (hot, 2025), Lavie (Wang et al., 2025b), Show-1 (Zhang et al., 2024a), Moonvalley (moo, 2025), Crafter (Chen et al., 2023), ModelScope (Wang et al., 2023a), and DreamVideo (Wei et al., 2024).

Question: Does this video look real or generated by an AI world model?   Real or **Generated**



<think>Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the four factors mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: In the video, there are no moving objects that would allow me to assess motion consistency. The scene appears static, with no visible movement of ships, boats, or any other elements. This lack of motion makes it difficult to determine if the video is real or AI-generated based on motion consistency alone.
2. **Lighting Consistency**: The lighting in the video appears to be consistent with what one would expect in a real-world scenario. The shadows and highlights on the buildings and water reflect natural lighting conditions. However, without more context or a comparison to other videos, it's hard to definitively say if the lighting is perfectly consistent with real-world conditions.
3. **Texture Artifacts**: Upon closer inspection, the textures of the buildings, water, and other elements appear somewhat uniform and lack the fine detail and variation one would expect in a real-world photograph. This could be a result of the AI generation process, where textures might not be perfectly realistic.
4. **Physics Violations**: There are no obvious physics violations in this scene. The positioning of the buildings, water, and boats seems plausible and follows the expected physics of a harbor or cityscape. However, the lack of motion and the uniformity of textures could still be indicative of AI generation.
Considering these points, the video does not exhibit clear motion, and the textures and lighting are somewhat uniform, which could suggest AI generation. However, the absence of motion and the lack of fine detail make it challenging to definitively conclude.
In conclusion, while the video does not exhibit clear motion or physics violations, the uniformity of textures and the lack of fine detail suggest that it could be AI-generated. However, more context or a comparison to other videos would be needed for a more definitive assessment.</think>
<answer> Generated </answer>

Figure 9: VidGuard-R1 (GRPO) inference on a MuseV-generated video

Question: Does this video look real or generated by an AI world model?   Real or **Generated**

<think>Let me think about this video carefully. The scene appears to be a chaotic event, possibly a protest or a police intervention, with people in motion. As I analyze the video, I'll consider the factors mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: The individuals in the video seem to be moving in a way that is somewhat realistic. However, the speed and direction of the movement appear slightly inconsistent at times. This could be due to the AI's attempt to simulate human movement, but the slight discrepancies might indicate a lack of perfect synchronization.
2. **Lighting Consistency**: The lighting in the video appears to be consistent with natural daylight. There are no sudden changes in brightness or shadows that would suggest artificial lighting. The shadows and highlights seem to follow the expected patterns for an outdoor scene.
3. **Texture Artifacts**: Upon closer inspection, there are some subtle artifacts in the textures of the clothing and the environment. The edges of objects and people seem a bit jagged, which is a common issue with AI-generated images. This could be due to the AI's attempt to render textures in a realistic manner, but the artifacts are noticeable.
4. **Physics Violations**: There are no obvious physics violations in the video. The people and objects are moving in a way that is plausible for a real-world scenario. The interactions between the individuals and the environment also seem natural.
Considering these observations, the video does exhibit some inconsistencies that are typical of AI-generated content. The motion and texture artifacts are subtle but present, and the overall scene is consistent with what one might expect from a real-world event. However, the slight discrepancies in motion consistency and the presence of texture artifacts suggest that this video is likely generated by an AI model rather than a real recording.
In conclusion, while the video appears to be a realistic depiction of a chaotic event, the subtle inconsistencies in motion and texture artifacts point towards it being an AI-generated video.</think>
<answer> Generated </answer>

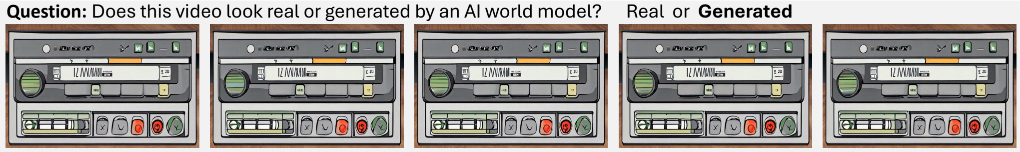
Figure 10: VidGuard-R1 (GRPO) inference on a SVD-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the factors of motion consistency, lighting consistency, texture artifacts, and physics violations.
1. **Motion Consistency**: The video shows a plush toy dinosaur inside a box. The movement of the toy appears to be smooth and natural, suggesting that it might be a real video. However, the toy is not moving in a way that would be expected in a real-world scenario, as it seems to be floating or moving in a way that defies gravity. This could be a physics violation.
2. **Lighting Consistency**: The lighting in the video appears consistent and natural. There are no harsh shadows or unusual highlights that would suggest the use of artificial lighting or post-processing. This is a good sign that the video could be real.
3. **Texture Artifacts**: The texture of the plush toy and the box looks realistic. There are no visible artifacts or distortions that would indicate the use of AI generation. The fabric of the toy and the cardboard of the box have a natural appearance.
4. **Physics Violations**: The most striking aspect is the movement of the toy. It appears to be floating or moving in a way that defies the laws of physics. This is a significant red flag for AI-generated content, as it is highly unlikely for a real video to capture such an anomaly.
Considering all these points, the video seems to have a physics violation that is not typical in real-world footage. While the motion and lighting appear consistent, the unnatural movement of the toy suggests that this video might be AI-generated. The combination of these factors makes it highly improbable that this is a real video.</think>
<answer> Generated </answer>

Figure 11: VidGuard-R1 (GRPO) inference on a CogVideo-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think>Let me think about this. The image provided is a still frame of what appears to be a vintage radio or cassette player. It's a detailed and realistic-looking illustration, but I need to consider the factors mentioned to determine if it's real or AI-generated.
1. **Motion Consistency**: The image is a still frame, so there's no motion to analyze. However, if this were a video, I would look for any inconsistencies in the movement of the objects, such as the tape or the buttons, to see if they move smoothly or if there are any jarring motions.
2. **Lighting Consistency**: The lighting in the image appears consistent across the entire frame. There are no shadows or highlights that seem out of place. If this were a real video, I would expect the lighting to change slightly as the camera moves or as the scene changes.
3. **Texture Artifacts**: The texture of the radio and the cassette tape looks very detailed and realistic. There are no visible pixelation or distortion that would suggest it's AI-generated. The lines and edges are sharp and well-defined.
4. **Physics Violations**: There are no physics violations in this image. Everything looks as it should, with the cassette tape in the correct position and the buttons in a logical arrangement.
Given these observations, the image appears to be a high-quality illustration rather than a real video. The lack of motion, the consistent lighting, the detailed textures, and the absence of physics violations all point towards this being an AI-generated image rather than a real video.
Hmm, let's break it down. The still nature of the image, the detailed and realistic textures, and the absence of motion and physics violations all support the conclusion that this is an AI-generated image. It's a convincing illustration, but it doesn't have the dynamic qualities of a real video.</think>
<answer> Generated </answer>

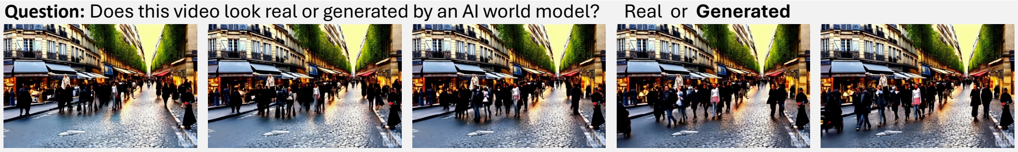
Figure 12: VidGuard-R1 (GRPO) inference on a Mora-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this. When evaluating whether a video is AI-generated, I consider several factors: motion consistency, lighting consistency, texture artifacts, and physics violations. Let's break it down.
1. **Motion Consistency**: In the video, the wolves are shown in a natural, fluid motion. Their movements appear smooth and consistent with how animals would move in the wild. There are no sudden jerks or unnatural movements that would typically indicate AI generation. The wolves' postures and the way they interact with each other seem realistic.
2. **Lighting Consistency**: The lighting in the video is consistent with what one would expect in a night scene. The moonlight illuminates the wolves and the surrounding environment in a way that is natural and not overly artificial. The shadows and highlights are well-balanced, which adds to the realism of the scene.
3. **Texture Artifacts**: The texture of the wolves and the environment appears smooth and natural. There are no visible pixelation or unusual textures that are often seen in AI-generated images. The fur of the wolves looks realistic, and the environment, including the trees and the ground, has a natural texture that is consistent with what one would expect in a forest at night.
4. **Physics Violations**: The wolves are shown howling and moving in a way that is physically plausible. There are no obvious physics violations, such as the wolves floating or moving in an unrealistic manner. The howling and the movement of the wolves are in sync with their natural behavior.
Considering all these factors, the video does not exhibit any significant motion inconsistencies, lighting inconsistencies, texture artifacts, or physics violations. The scene appears to be a real video of wolves in a natural setting, captured in a way that is consistent with real-world photography or video.
Therefore, based on the analysis, it seems that this video is likely a real video rather than an AI-generated one.</think>
<answer> Generated </answer>

Figure 13: VidGuard-R1 (GRPO) inference on a Sora-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think> Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the four factors you mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: In the video, people are walking, and their movements appear somewhat natural. However, the speed and fluidity of their movements might not be entirely consistent with real-world human motion. AI-generated videos often have a slightly artificial feel to the movement, which can be subtle but noticeable.
2. **Lighting Consistency**: The lighting in the video seems to be consistent with what you'd expect in a real-world setting. The shadows and highlights on the people and buildings appear natural, suggesting that the lighting was carefully simulated to match real-world conditions.
3. **Texture Artifacts**: Upon closer inspection, the textures of the people's clothing and the environment (like the cobblestones and building facades) appear somewhat uniform and lack the fine detail and variation you'd typically see in a real photograph or video. This could be a sign that the textures are generated by an AI model.
4. **Physics Violations**: There are no obvious physics violations in this video. The people are walking in a way that is consistent with how they would move in a real-world scenario. The shadows and reflections also align with what you'd expect in a real environment.
Considering these points, the video seems to have some artificial qualities, particularly in the texture and motion consistency. While it's not definitively clear if this is an AI-generated video, the subtle inconsistencies in texture and motion suggest that it might be. However, without more advanced analysis tools, it's challenging to be certain. The video could still be a high-quality real-world recording that has been edited or enhanced in some way.</think>
<answer> Generated </answer>

Figure 14: VidGuard-R1 (GRPO) inference on a Morph Studio-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this. The image provided is a still frame from what appears to be an animated video. When considering the factors of Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations, I'll analyze the image step by step.
1. **Motion Consistency**: The character in the image is in a dynamic pose, suggesting movement. However, since this is a still frame, we cannot determine motion consistency. The character's pose looks smooth and well-animated, which is typical for high-quality animation.
2. **Lighting Consistency**: The lighting in the image is bright and evenly distributed, which is consistent with a well-lit animated scene. There are no harsh shadows or inconsistencies that would suggest a real-world environment.
3. **Texture Artifacts**: The texture of the character and the background appears smooth and consistent with digital animation. There are no visible signs of pixelation or irregularities that would indicate a real-world video.
4. **Physics Violations**: The character is depicted in a way that defies real-world physics. For example, the character is floating in the air, which is not possible in real life. This is a clear indication that the image is from an animated video.
Considering these points, it seems that the image is indeed an AI-generated video. The character's pose, lighting, and the physics involved all point towards animation rather than a real-world video. The absence of motion in the still frame does not affect this conclusion, as the analysis is based on the visual elements present.
In summary, the image exhibits characteristics typical of AI-generated animation, such as smooth textures, consistent lighting, and physics violations, which are not present in real-world videos.</think>
<answer> Generated </answer>

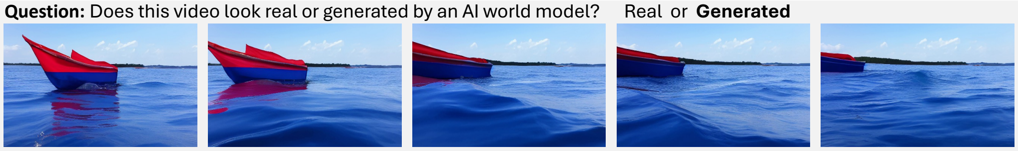
Figure 15: VidGuard-R1 (GRPO) inference on a Gen2-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this. When evaluating whether a video is real or AI-generated, I consider several factors: motion consistency, lighting consistency, texture artifacts, and physics violations. Let's break it down.
1. **Motion Consistency**: In the video, the Ferris wheel appears to be in motion, but the movement seems somewhat jerky and inconsistent. Real-world objects, especially large structures like Ferris wheels, typically exhibit smoother motion. The way the gondolas are moving might suggest a lack of natural fluidity, which is a common sign of AI-generated content.
2. **Lighting Consistency**: The lighting in the video appears to be natural, with shadows and highlights that align with what one would expect from sunlight. There are no obvious artifacts or anomalies in the lighting that would suggest the use of artificial lighting techniques typically employed in AI-generated videos.
3. **Texture Artifacts**: The texture of the Ferris wheel and the gondolas looks quite uniform and lacks the fine details that would be expected in a real-world photograph. AI-generated images often have a slightly artificial look, with less variation in texture and detail.
4. **Physics Violations**: The Ferris wheel is rotating, and the gondolas are moving in a way that seems plausible. However, the motion of the gondolas appears a bit too perfect and lacks the slight imperfections one would expect in real-world footage. The gondolas seem to move in a way that doesn't quite match the natural physics of a Ferris wheel.
Considering these observations, the video exhibits some inconsistencies in motion and texture that are characteristic of AI-generated content. The motion of the Ferris wheel and gondolas is not entirely natural, and the texture of the objects is somewhat artificial. While the lighting is consistent with natural sunlight, the overall impression is that this video is likely AI-generated.</think>
<answer> Generated </answer>

Figure 16: VidGuard-R1 (GRPO) inference on a HotShot-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think>Let me think about this video and the factors you've mentioned. The video shows a red and blue boat floating on calm blue waters under a clear sky. As I analyze it, I'll consider the motion consistency, lighting consistency, texture artifacts, and physics violations.
1. **Motion Consistency**: The boat appears to be moving smoothly across the water. There are no abrupt changes in its position or orientation, which suggests a consistent motion. However, the lack of visible ripples or waves around the boat might indicate a slight inconsistency in the water's movement, which could be a result of the AI generation process.
2. **Lighting Consistency**: The lighting in the video seems natural, with the sun casting shadows that align with the direction of the light source. The colors of the boat and the water appear vibrant and consistent with what one would expect in a sunny day. There are no noticeable artifacts or anomalies in the lighting that would suggest the video is AI-generated.
3. **Texture Artifacts**: The water's surface looks smooth and lacks the fine details that would be expected in a real photograph or video. The boat's surface also appears somewhat uniform, which might be a result of the AI's attempt to simplify textures for aesthetic purposes.
4. **Physics Violations**: The boat is floating on water without any visible support or propulsion, which is a plausible scenario. However, the lack of any movement or interaction with the water, such as ripples or waves, might be a physics violation. In a real scenario, the boat would create some disturbance in the water, even if it's moving slowly.
Considering these observations, the video seems to have some inconsistencies in motion and texture, which are common in AI-generated images. The lighting and physics violations are subtle but noticeable. Therefore, it is likely that this video has been generated by an AI model, but it is still quite plausible that it could be a real video with some post-processing or simplification.</think>
<answer> Generated </answer>

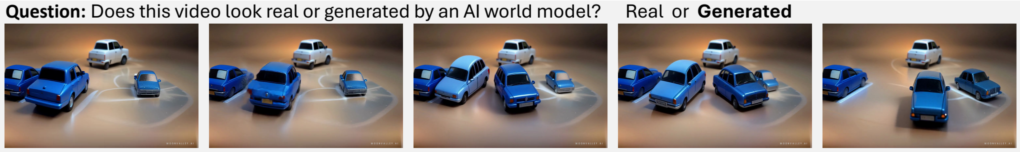
Figure 17: VidGuard-R1 (GRPO) inference on a Lavie-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this video carefully. The scene depicts a group of dancers performing on a stage with vibrant lighting. As I analyze the video, I'll consider the factors mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: The dancers appear to be moving in a coordinated manner, which is typical for a choreographed performance. However, upon closer inspection, the movements seem slightly unnatural. The fluidity of the dancers' movements is not entirely consistent with what I would expect from real human dancers. There are some abrupt transitions that might suggest the video is not entirely natural.
2. **Lighting Consistency**: The lighting on the stage is dynamic and colorful, which is a common feature in live performances. The lighting changes in a way that is visually appealing and consistent with what one would expect from a professional stage setup. There are no obvious inconsistencies in the lighting that would suggest the video is AI-generated.
3. **Texture Artifacts**: The dancers' costumes and the stage itself appear to have a realistic texture. There are no visible artifacts or distortions that would indicate the video is AI-generated. The textures of the costumes and the stage are smooth and consistent with what one would expect in a real video.
4. **Physics Violations**: The dancers' movements are generally consistent with what one would expect from human dancers. There are no obvious physics violations, such as dancers floating or moving in ways that defy gravity. The movements are smooth and coordinated, which is typical for a well-rehearsed performance.
After considering all these factors, I'm inclined to think that this video is likely AI-generated. The slight inconsistencies in the dancers' movements and the lack of a completely natural flow suggest that the video might be a generated image or a heavily edited video. The lighting and texture appear consistent with real footage, but the motion inconsistencies are a strong indicator that this is not a real video.
In conclusion, while the video has some realistic elements, the motion inconsistencies are significant enough to suggest that this video is AI-generated.</think>
<answer> Generated </answer>

Figure 18: VidGuard-R1 (GRPO) inference on a Show-1-generated video

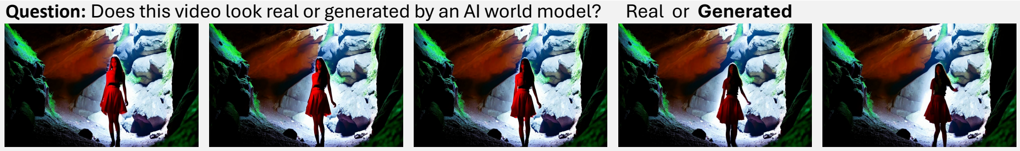
Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think>Let me think about this video carefully. The image shows a scene with toy cars, and I need to determine if it's real or AI-generated. Let's break it down by the factors mentioned:
1. **Motion Consistency**: The cars appear to be moving in a way that looks somewhat realistic, but the motion seems a bit stiff and not entirely natural. The cars are positioned in a way that suggests they are moving forward, but the way the wheels are turning and the cars' positions don't quite align with real-world physics. This could be a sign of AI-generated content.
2. **Lighting Consistency**: The lighting in the scene is quite uniform and doesn't show any significant shadows or highlights that would be expected in a real-world scenario. This could be a result of the AI model trying to create a consistent lighting environment, but it might not match the natural lighting conditions.
3. **Texture Artifacts**: The cars and the background have a smooth, almost plastic-like texture. This is not typical of real-world objects, which usually have more varied and complex textures. The lack of detail in the textures could be another indicator of AI generation.
4. **Physics Violations**: The cars are positioned in a way that suggests they are moving, but the physics of their movement don't seem to follow real-world rules. For example, the cars appear to be moving without any visible force or interaction with the ground, which is not how real cars would behave.
Considering all these points, it seems that the video has some inconsistencies that are typical of AI-generated content. The motion, lighting, and texture artifacts, along with the physics violations, suggest that this is likely an AI-generated video rather than a real-world recording. However, it's important to note that AI models can sometimes produce content that closely mimics real-world scenarios, so the distinction might not always be clear-cut.</think>
<answer> Generated </answer>

Figure 19: VidGuard-R1 (GRPO) inference on a Moonvalley-generated video

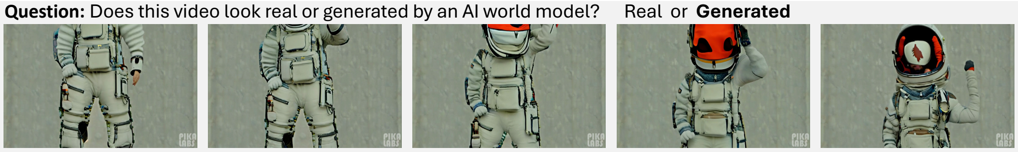
Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think>Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the four factors you mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: In the video, the woman appears to be walking through a cave. Her movement looks smooth and natural, which is a good sign. However, the motion is somewhat exaggerated, which might be a result of the AI generation process. Real-life videos typically show more subtle and less exaggerated movements.
2. **Lighting Consistency**: The lighting in the video seems to be consistent with what one would expect in a cave. The light source appears to be natural, possibly sunlight filtering through an opening. The shadows and highlights are realistic, which supports the idea that this could be a real video.
3. **Texture Artifacts**: The textures of the cave walls and the woman's dress appear quite detailed and realistic. There are no obvious signs of pixelation or distortion that one would typically associate with AI-generated images. The colors and details in the environment look natural and well-rendered.
4. **Physics Violations**: The woman's movement through the cave seems to follow the laws of physics. There are no sudden or unrealistic movements that would suggest a physics violation. The way she walks and the way the light interacts with her and the environment are consistent with real-world physics.
Considering all these points, the video seems to be quite realistic. The motion is smooth and natural, the lighting is consistent with a real cave, the textures are detailed and realistic, and there are no physics violations. However, the slight exaggeration in the woman's movement might be a subtle indicator that this is an AI-generated video.
In conclusion, while the video appears to be real, the slight exaggeration in motion could suggest that it is generated by an AI model.</think>
<answer> Generated </answer>

Figure 20: VidGuard-R1 (GRPO) inference on a Crafter-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**



<think>Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the four factors you've mentioned: Motion Consistency, Lighting Consistency, Texture Artifacts, and Physics Violations.
1. **Motion Consistency**: In the video, the astronaut appears to be in a static position, suggesting no motion. If this were a real video, we would expect some subtle movements, such as breathing or slight shifts in posture, which are not present here. This lack of motion could indicate that the video is AI-generated.
2. **Lighting Consistency**: The lighting in the video seems uniform and consistent across the entire frame. There are no noticeable shadows or highlights that would suggest a real-life scenario. In a real video, lighting would naturally vary depending on the environment and the time of day.
3. **Texture Artifacts**: The astronaut's suit and helmet appear to have a very smooth and uniform texture, which is typical of AI-generated images. Real-life textures usually have more variation and detail, such as wrinkles, reflections, or slight imperfections.
4. **Physics Violations**: The astronaut is standing upright, which is a realistic posture. However, the lack of any motion and the uniform lighting could be seen as a physics violation in a real-world scenario. Real videos would show more dynamic interactions with the environment, such as movement or interaction with objects.
Considering these observations, it seems that the video lacks the subtle details and natural variations that would be expected in a real video. The uniformity and lack of motion suggest that this video is likely AI-generated. The absence of realistic lighting and texture artifacts further supports this conclusion.
Hmm, I see. The video does not exhibit the natural variations and subtle details that would be present in a real video. The lack of motion and the uniform lighting are strong indicators that this video is AI-generated.</think>
<answer> Generated </answer>

Figure 21: VidGuard-R1 (GRPO) inference on a ModelScope-generated video

Question: Does this video look real or generated by an AI world model? Real or **Generated**

<think>Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the factors of motion consistency, lighting consistency, texture artifacts, and physics violations.
1. **Motion Consistency**: The astronaut riding the horse appears to be in motion, but the movement looks somewhat unnatural. The horse's legs and the astronaut's body seem to move in a way that doesn't quite align with real-life physics. The motion looks slightly jerky and inconsistent, which is a common sign of AI-generated content.
2. **Lighting Consistency**: The lighting in the video seems to be consistent with what one might expect in a real outdoor setting. However, the lighting on the astronaut and the horse is somewhat flat and lacks the depth and variation that would be present in a real scene. This could be due to the AI model's limitations in generating realistic lighting.
3. **Texture Artifacts**: The textures of the astronaut's suit and the horse's coat appear somewhat artificial. There are no visible details that would suggest a real-life material, which is another indicator that this might be AI-generated.
4. **Physics Violations**: The most striking aspect is the physics violation. An astronaut in a spacesuit would not be able to ride a horse in a real-world scenario. The combination of these elements suggests that this is not a real video but rather an AI-generated one.
Considering all these points, it seems that this video likely falls into the category of AI-generated content. The inconsistencies in motion, the flat lighting, the artificial textures, and the physics violations all point towards this conclusion. However, it's important to note that AI models can sometimes produce content that closely mimics real-life scenarios, so it's always a good idea to cross-reference with other sources or experts if possible.</think>
<answer> Generated </answer>

Figure 22: VidGuard-R1 (GRPO) inference on a DreamVideo-generated video