arXiv:2507.14632v2 [cs.CV] 31 Jul 2025

BusterX++: Towards Unified Cross-Modal AI-Generated Content Detection and Explanation with MLLM

Haiquan Wen[1] Tianxiao Li[1] Zhenglin Huang[1] Yiwei He[1] Guangliang Cheng[1]†
[1]University of Liverpool, UK
† Corresponding author. E-mail: guangliang.cheng@liverpool.ac.uk
Project Page: https://github.com/l8cv/BusterX

Abstract

Recent advances in generative AI have dramatically improved image and video synthesis capabilities, significantly increasing the risk of misinformation through sophisticated fake content. In response, detection methods have evolved from traditional approaches to multimodal large language models (MLLMs), offering enhanced transparency and interpretability in identifying synthetic media. However, current detection systems remain fundamentally limited by their single-modality design. These approaches analyze images or videos separately, making them ineffective against synthetic content that combines multiple media formats. To address these challenges, we introduce ***BusterX++****, a novel framework designed specifically for cross-modal detection and explanation of synthetic media. Our approach incorporates an advanced reinforcement learning (RL) post-training strategy that eliminates cold-start. Through Multi-stage Training, Thinking Reward, and Hybrid Reasoning, BusterX++ achieves stable and substantial performance improvements. To enable comprehensive evaluation, we also present* ***GenBuster++****, a cross-modal benchmark leveraging state-of-the-art image and video generation techniques. This benchmark comprises 4,000 images and video clips, meticulously curated by human experts using a novel filtering methodology to ensure high quality, diversity, and real-world applicability. Extensive experiments demonstrate the effectiveness and generalizability of our approach.*

1. Introduction

Recent advancements in generative AI, epitomized by models like Seedance 1.0 [17] and GPT-4o [36], have unlocked tremendous potential across sectors such as advertising, education, and entertainment. These technologies have revolutionized content creation by enabling the production of highly realistic images and videos. Nevertheless, this power has a darker side as well: The increasing prevalence of AI-generated content on social media platforms is blurring the lines between authentic and synthetic media. This phenomenon not only challenges our ability to verify content credibility but also raises profound concerns about the integrity of information in the digital age. To address these challenges, researchers have prioritized AI-Generated content (AIGC) detection as a critical area of study. Consequently, the development of robust detection frameworks has become increasingly urgent.

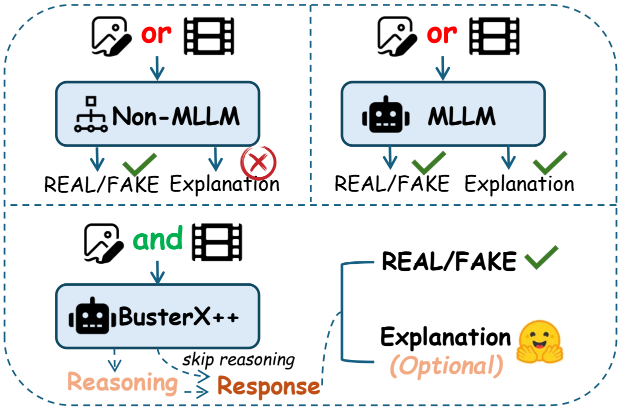


Figure 1. The framework comparisons. Existing non-MLLM methods can only classify without providing explanations. MLLM-based methods enhance explainability but are limited to single-image/video modality inputs. In contrast, BusterX++ is a reasoning MLLM framework that supports both image and video modalities and offers optional explanations.

Initially, non-MLLM (Multimodal Large Language Model) methods [1, 6, 8, 15, 35, 40, 45, 46, 54, 68] dominated the field of AI-generated content detection, primarily focusing on single-modality detection of images or videos. While these methods achieved acceptable accuracy within their specific domains, they exhibited significant limitations in generalizing to unseen generation techniques and cross-modal scenarios. In addition, the lack of interpretability in their decision-making process raised concerns about transparency and trustworthiness in practical applications.

Recently, the focus has shifted to MLLM-based approaches [7, 9, 23, 24, 26, 30, 38, 59, 60, 64, 69], which have considerably enhanced the transparency and interpretability of detection outcomes in both AI-generated image and video detection fields. Notable works in this area include FakeShield [64], SIDA [23], X²-DFD [9] for image detection, and MM-Det [44] and BusterX [59] for video detection.

However, these methods have the following limitations: **1)** They are restricted to single-modality image or video detection and have not sufficiently explored the cross-modal capabilities of MLLM. **2)** Most MLLM + RL (Reinforcement Learning) methods [24, 59] rely on a resource-intensive cold-start before RL stage. But the performance improvement in the RL stage is often restricted.

To overcome these limitations, we introduce **BusterX++**, a novel framework designed for unified cross-modal AI-generated content detection and interpretation. As shown in Figure 1, it can detect and explain image and video in a unified framework. Leveraging the cross-modal capabilities of MLLM, we employ a unified training strategy incorporating both images and videos during the post-training phase. Our experiments reveal that joint training across modalities leads to mutually beneficial performance gains, with each modality enhancing the other's detection accuracy. Additionally, we propose a novel **cold-start-free** post-training strategy, integrating Multi-stage Training, Thinking Reward, and Hybrid Reasoning to achieve stable and substantial performance improvements. By directly starting with RL, our model demonstrates superior generalization and adaptability compared to cold-start-dependent methods. Furthermore, our model offers an optional non-thinking mode, which allows it to directly output classification results without engaging in detailed reasoning. This flexibility not only reduces computational overhead, but also enhances the model's practicality in scenarios where rapid classification is prioritized, while still maintaining its advanced reasoning capabilities for cases requiring detailed explanations.

While BusterX++ advances AI-generated content detection technology, the lack of a suitable benchmark to comprehensively evaluate its cross-modal capabilities posed a substantial challenge. Although there are some existing datasets and benchmarks like So-Fake-Set [24] and GenBuster-200K [59] in the field, they still have the following issues: **1)** They focus on single-modalities, such as image or video. **2)** They lack fine-grained human curation and contain some not realistic enough content.

Recognizing this critical gap, we created **GenBuster++**, a cross-modal benchmark designed to meet the demands of modern AI-generated content detection. It leverages state-of-the-art image and video generation techniques and consists of 4,000 images and video clips. Each fake sample undergoes a rigorous two-stage filtering process. Experts first identify samples perceived as real from a mixed set of both real and fake content, then isolate those that are actually synthetic, to ensure high-quality, diversity, and real-world relevance.

Our contributions can be summarized as follows.

1. We introduce **BusterX++**, a unified cross-modal detection and explanation framework designed to identify AI-generated content across both images and videos. In addition to its multimodal capabilities, BusterX++ achieves state-of-the-art performance on single-modality benchmarks.
2. We adopt a novel RL post-training strategy **without cold-start**, which effectively improves the model's final performance compared to the cold-start-dependent strategy.
3. We introduce **GenBuster++**, a cross-modal high-quality benchmark consisting of 4,000 images and video clips. It provides a reliable standard for cross-modal AI-generated content detection and strongly supports evaluation and application expansion.

2. Related Work

2.1. AI-Generated Content Detection Benchmark

The landscape of AI-generated content detection benchmarks has evolved to reflect the growing complexity of synthetic media. The early benchmarks [12, 20, 27, 28, 40, 62, 72] focused mainly on GAN-generated facial forgeries, consistent with the initial focus of deepfake technology on the modification of human identities. With the development of more sophisticated generative models, the focus gradually shifted toward generating diverse and realistic content. In the image domain, benchmarks such as GenImage [71], DMImage [11], TrueFake [14], MMTD-SET [64], SID-Set [23], and So-Fake-Set [24] have significantly expanded the scope of research. Similarly, the video domain has witnessed substantial progress with benchmarks like GVF [32], GenVideo [8], GenVidBench [34], GenBuster-200K [59]. These benchmarks leverage state-of-the-art video generators to produce high-quality synthetic content that closely mimics real-world scenarios. They have paved the way for more robust and versatile detection frameworks capable of addressing the challenges posed by modern synthetic content.

Despite significant advancements, existing benchmarks have notable limitations. Most are confined to single-modality data, restricting their utility in assessing cross-modal model capabilities. Additionally, most of them lack fine-grained human curation, resulting in inconsistent quality. To address these gaps, GenBuster++ addresses these challenges by integrating both image and video modalities, leveraging two-stage filtering to ensure that each sample is high-quality, diversity, and real-world relevance.

2.2. AI-Generated Content Detection Method

Traditional AI-generated content detection methods are primarily non-MLLM [1, 6, 8, 15, 35, 40, 45, 46, 54, 68], focusing on binary classification tasks. These methods achieved

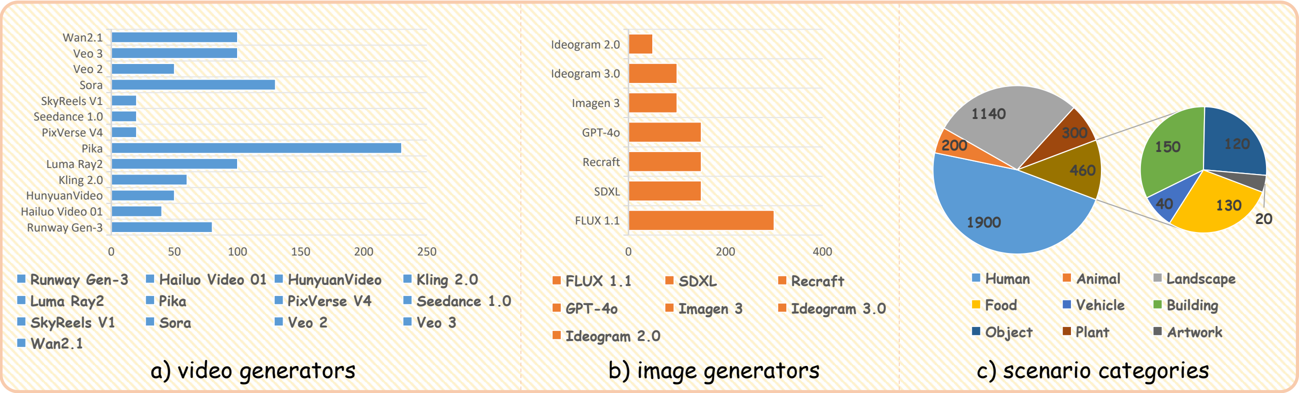


Figure 2. Overview of GenBuster++. a) GenBuster++ incorporates video samples generated by 13 state-of-the-art video generators. b) It also includes image samples from 7 state-of-the-art image generators. c) GenBuster++ covers diverse real-world scenarios, including 9 different categories and each sample undergoes a rigorous filtering process.

reasonable accuracy on in-domain data but exhibited limited generalization to unseen generative techniques and lacked interpretability. Recent advancements introduced MLLM-based methods in both AI-generated image and video detection fields, such as AntifakePrompt [7], FakeShield [64], SIDA [23], X²-DFD [9], FakeVLM [60], LEGION [26], FakeScope [30], VLF-FFD [38], AIGI-Holme [69], So-Fake-R1 [24], MM-Det [44], BusterX [59], which enhanced detection transparency.

However, these approaches remained confined to single-modality inputs, failing to fully leverage the cross-modal capabilities of MLLMs. To address these limitations, we propose BusterX++, which demonstrates superior generalization and adaptability in cross-modal detection and explanation.

2.3. Multimodal Large Language Model

Recent advancements in MLLM have emphasized enhancing cross-modal reasoning capabilities. Commercial models like GPT-4o [36], Gemini 2 [47], Claude 4 [2], and Kimi-k1.5 [48] have demonstrated remarkable capabilities. On the open-source side, models such as Qwen-VL [4] and InternVL [70] stand out. These models allocate additional computational resources to complex reasoning tasks, pushing the limits of existing benchmarks. The community has even introduced extremely challenging benchmarks like Humanity's Last Exam [39] to benchmark the models' limitation. BusterX++ further explores the reasoning capabilities of MLLM in the domain of AI-generated content detection.

2.4. Post-Training in Large Language Model

The post-training phase of LLM has advanced rapidly, with supervised fine-tuning (SFT) and reinforcement learning (RL) being widely used. Models such as DeepSeek-R1 [13] and Kimi-k1.5 [48] have demonstrated remarkable reasoning abilities thanks to RL. In particular, DeepSeek-R1 introduced Group Relative Policy Optimization (GRPO) [13], a variant of PPO [41], which effectively enhances the model's reasoning capacity. The community has also made numerous attempts to improve GRPO, such as REINFORCE++ [22], DAPO [65] and GVPO [67]. Building on this foundation, models like Qwen3 [51] and Seed1.6-Thinking [42] introduced hybrid thinking and multi-stage RL post-training. Reward Models [31, 37, 55] still play a crucial role in guiding and shaping the behavior of reasoning. SophiaVL-R1 [16] proposed a thinking reward model that supervises the overall quality of the reasoning content.

Some studies [10, 61] indicate that "SFT Memorizes, RL Generalizes": SFT tends to memorize training data, whereas RL enhances the model's visual reasoning abilities. While the cold-start + RL paradigm has been widely adopted in modern LLM post-training, some research suggests that it may not be essential [61]. We therefore explore a novel post-training strategy **without cold-start** and achieved encouraging results.

3. Benchmark

3.1. Motivation

In the field of deepfake detection, research has primarily focused on facial data due to its significant societal impact. With the rapid advancement of generative models, attention has increasingly shifted toward wider AI-generated video content, which poses broader challenges and risks. Although several datasets and benchmarks [8, 11, 14, 23, 24, 32, 34, 59, 64, 71] have been developed for AI-generated content detection, they still have the following limitations: **1) Single-Modality Focus**: Existing work predominantly concentrates on images or videos. **2) Limited Human Curation**: Most datasets lack fine-grained human curation, and contain some not realistic enough content. It will affect the models' evaluation around real-world scenarios.

To address these limitations, we introduce **GenBuster++**, a cross-modal benchmark for MLLM evaluation as shown in

Figure 2. The benchmark contains samples from 7 state-of-the-art image generators (FLUX [29], GPT-4o [36], etc.) and 13 video generators (Seedance 1.0 [17], SkyReels V1 [43], etc.). We design a rigorous filtering process to guarantee each sample is high-quality and real-world relevance.

3.2. Benchmark Construction

Our benchmark leverages state-of-the-art image and video generation techniques and consists of 4,000 images and video clips. It has two parts: real images/videos from real-world scenarios, and synthetic images/videos that simulate real-world conditions. The details are as follows:

Data Sources: We first employ the following methods to construct a data pool for subsequent filtering.

Real Images/Videos. We source a large number of real images and videos from OpenVid-1M HD [33], a dataset that covers a diverse range of real-world scenarios. These samples are carefully pre-filtered to ensure they come from a wide variety of scenes.

Fake Images/Videos. We source a large number of fake images and videos from multiple sources. This includes samples from MagicArena (high-rated sample). In addition, we construct the following pipeline for generation: we use Reddit's official API to obtain social media images that cover a wide range of real-world scenarios. Then, we employ Qwen-2.5-VL [5] to generate detailed captions that describe the content of these source images. These captions then serve as prompts for various diffusion models to generate synthetic content. Finally, there are models such as FLUX [29] and GPT-4o [36] for image generation and Seedance 1.0 [17], SkyReels V1 [43] for video generation in the pool.

Data Filtering: At this stage, we introduce a rigorous filtering strategy to ensure that the final samples are of high-quality, diverse in content, and closely aligned with real-world scenarios.

Real Images/Videos. We first filter out a large number of low-quality samples based on resolution, video frame rate, and video bitrate. Next, we eliminate duplicate content from the same origin clip. Finally, we manually check each sample one by one, removing those with extensive watermarks, anime content, or an obviously synthetic background. This process results in a final set of 1,000 real images and 1,000 fake videos.

Fake Images/Videos. To ensure the high quality and realism of fake samples, we implement a novel two-stage filtering pipeline. First, we create a mixed pool that contains both real and fake samples from the previous data pool. Experts carefully review this pool to identify samples that appeared real. In the second stage, these identified samples are re-examined to separate those that were actually synthetic. This process results in a final set of 1,000 fake images and 1,000 fake videos.

Post-processing. We employ the same way as BusterX to post-process the samples. For image, we standardize the resolution to 1024×1024. For video clip, we standardize the resolution to 1920×1080, the duration to 5 seconds, and the frame rate to 24 FPS. Figure 3 shows some samples. This unified post-processing approach offers several benefits: **1) Elimination of Encoding Biases**: By using HEVC encoding with x265, we eliminate potential biases from underlying different encoding preferences. **2) Consistency Across Sources**: Standardization ensures consistency across videos or images generated by different models, which may have varying original resolutions and frame rates.

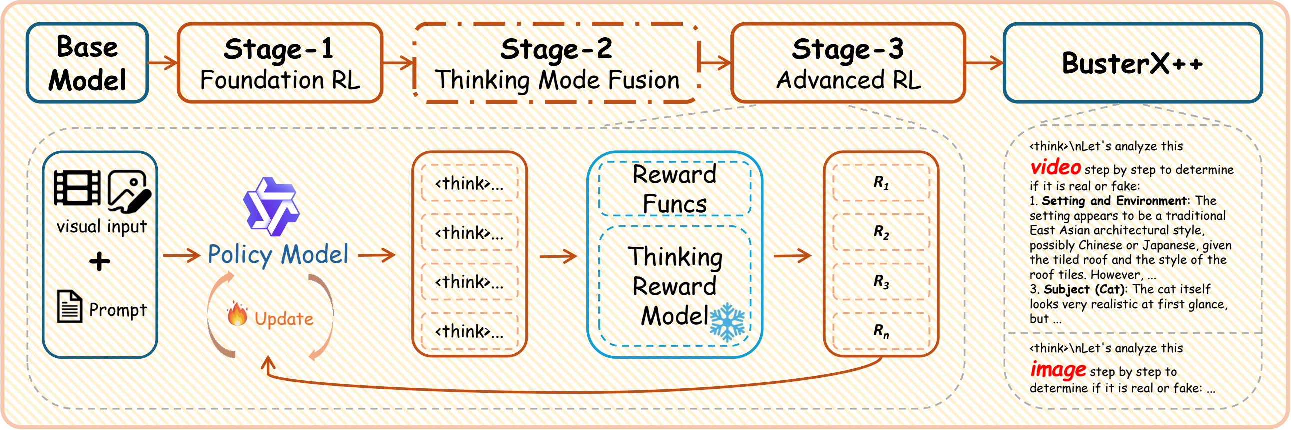
Figure 3. Visual Examples from GenBuster++.

4. Method

In this section, we introduce BusterX++, a novel framework for detecting AI-generated images and videos with detailed explanations. As shown in Figure 4, we train a reasoning MLLM with Chain-of-Thought (CoT) [58] and reinforcement learning (RL). We expand the training data modality by mixing images and videos during post-training and refine the post-training strategy to boost the model's final performance.

4.1. Challenges of Cold-Start

Existing MLLM + RL methods [24, 59] mostly depend on a SFT cold-start stage before RL. However, despite accelerating training, the performance improvement in the RL stage is often restricted. We conjecture that this limitation may result from the reasoning quality of CoT data used in the cold-start phase, which could undermine the model's reasoning ability. Human judgments about the "fakeness" of images or videos are often based on subtle, intuitive, and multi-dimensional cues such as unnatural reflections, inconsistent lighting, slight motion artifacts, or uncanny valley effects. Given the non-linear and non-linguistic nature of cross-modal preferences in our task, it is extremely challenging to precisely elaborate a linear thinking chain for "why



<think>\nLet's analyze this *video* step by step to determine if it is real or fake:
1. **Setting and Environment**: The setting appears to be a traditional East Asian architectural style, possibly Chinese or Japanese, given the tiled roof and the style of the roof tiles. However, ...
3. **Subject (Cat)**: The cat itself looks very realistic at first glance, but ...

<think>\nLet's analyze this *image* step by step to determine if it is real or fake: ...

Figure 4. The framework of Buster++. Our method uses MLLM to analyze image/video and determine if it is AI-generated, providing reasoning for its decision. We adopt a multi-stage training strategy, progressively refining the model's capabilities from foundational classification to advanced reasoning. Full response details are in the appendix.

it is fake." [25] Thus, generating CoT data from MLLMs with prompt engineering is not only difficult to guarantee quality but also potentially degrade the model's reasoning performance. Consequently, we abandon the cold start and achieve promising results.

4.2. Multi-Stage Training

We employ multi-stage training strategy, setting different learning objectives for each stage to stabilize the training process. As BusterX, we adopt Dynamic sAmpling Policy Optimization (DAPO) [65] as our reinforcement learning strategy. DAPO samples a group of outputs $\{o_i\}_{i=1}^{G}$ for each question q from the old policy $\pi_{\theta_{old}}$. The reward model is then used to score the outputs, yielding $\{r_i\}_{i=1}^{G}$ correspondingly. Then DAPO optimizes the policy model π_θ by maximizing the following objective:

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

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

$$r_{i,t}(\theta)=\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})} \tag{2}$$

$$\hat{A}_{i,t}=\frac{r_i-\text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})} \tag{3}$$

Stage-1 Foundation RL: In this stage, we focus on building the fundamental capabilities of the model. We use RL to train the model on basic classification tasks. The model learns to distinguish between real and AI-generated content by receiving rewards based on the accuracy. This helps the model quickly grasp the basic characteristics and features of different types of data, laying a solid foundation for subsequent training.

Table 1. Thinking Mode Fusion: Chat Template.

Thinking Mode	Non-Thinking Mode
<|im_start|>user {query} /think<|im_end|> <|im_start|>assistant <think> {thinking_content} </think> <answer> {response} </answer><|im_end|>	<|im_start|>user {query} /no_think<|im_end|> <|im_start|>assistant <think> </think> <answer> {response} </answer><|im_end|>

Stage-2 Thinking Mode Fusion: This stage involves SFT to switch between reasoning and performance. We introduce a Thinking Mode Fusion mechanism inspired by Qwen3 [51]. Our SFT data collection of several hundred samples from Stage-1 model. As the template shown in Table 1, the model is trained to generate detailed CoT explanations and direct answers. The model thinks when there is no specific instruction or a /think prompt, and directly outputs answers when a /no_think prompt is encountered. Compared to thinking mode which often uses hundreds of tokens, this flexibility allows it to adapt to different scenarios, saving resources while maintaining performance. Ablation study shows this stage has minimal impact on performance.

Stage-3 Advanced RL: In the final stage, we further enhance the model's response quality using advanced RL techniques. The model continues to receive rewards for ac-

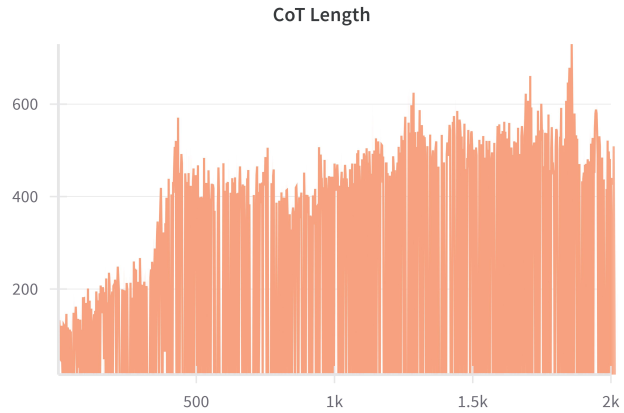


Figure 5. Stage-3: CoT Length (token) increases naturally with training steps.

curate classifications and well-formatted responses. We employ a mix of /think and /no_think prompts to train, thereby enhancing the model's ability to switch in two modes. Additionally, we introduce a Thinking Reward provided by an external model. This reward evaluates the quality of the model's thinking process, replacing the Length Reward used in BusterX, and the CoT Length also increases naturally with this approach (Figure 5). It encourages the model to produce more reasonable and comprehensive interpretations. However, we find that applying the Thinking Reward too early in Stage-1 can destabilize training, which underscores the importance of our multi-stage approach.

4.3. Reward Functions

We design the following reward functions and employ different ones in each RL stage:

- **Format**: If the model response adheres to the specified format (i.e., <think>...</think>and <answer>...</answer>), it receives a reward of $r_{\text{fmt}} = 0$; otherwise, it receives a penalty of $r_{\text{fmt}} = -1$.
- **Soft Overlong**: When the response length exceeds the maximum value, we define a penalty interval.

$$r_{\text{overlong}} = \begin{cases} 0, & L_{\text{gen}} \le L_{\max} - L_{\text{cache}} \\ \frac{(L_{\max} - L_{\text{cache}}) - L_{\text{gen}}}{L_{\text{cache}}}, & L_{\max} - L_{\text{cache}} < L_{\text{gen}} \le L_{\max} \\ -1, & L_{\max} < L_{\text{gen}} \end{cases} \tag{4}$$

- **Accuracy**: If the model classifies correctly, it receives a reward of $r_{\text{acc}} = 1$; otherwise, $r_{\text{acc}} = 0$.
- **Hybrid Thinking**: If the response adheres to the Template in Table 1, it receives $r_{\text{hybrid}} = 0$; otherwise, if the model skips thinking in Thinking Mode or thinks in Non-Thinking Mode, it receives $r_{\text{hybrid}} = -1$.
- **Thinking Reward**: This reward evaluates the quality of the response y_{res} by an external model $\mathcal{M}$. A higher score indicates a more reasonable and comprehensive thinking process.

$$r_{\text{think}} = \begin{cases} 0, & \text{/no\_think Mode} \\ \min(r_{\text{acc}}, \mathcal{M}(y_{\text{res}})), & \text{otherwise} \end{cases} \tag{5}$$

Where $0 \le r_{\text{think}} \le 1$.

The total reward for each RL stage is computed as follows:

$$\begin{aligned} R_{\text{stage-1}} &= r_{\text{fmt}} + r_{\text{overlong}} + r_{\text{acc}} \\ R_{\text{stage-3}} &= R_{\text{stage-1}} + r_{\text{hybrid}} + r_{\text{think}} \end{aligned} \tag{6}$$

5. Experiments

We adpot Qwen2.5-VL-7B-Instruct [4] as our base model, and SophiaVL-R1-Thinking-Reward-Model-3B [16] as our external model for Thinking Reward. We sample 16 frames at a rate of 4 FPS for video-level detection. For model training, we employ LoRA [21] Parameter-Efficient Fine-Tuning with a rank of 16 and alpha of 32. The model is trained using a learning rate of 1×10^{-5} with bfloat16 to optimize computational efficiency. We evaluate the accuracy (ACC) for each subcategory in GenBuster++.

Table 2. Performance comparison on So-Fake-Set.

Methods	Acc	F1
CnnSpott [54]	91.2	90.8
DiffForensics [66]	91.7	91.4
DIRE [57]	91.9	91.7
HIFI-Net [18]	88.4	83.9
SIDA [23]	91.9	91.5
So-Fake-R1 [24]	93.2	92.9
BusterX++	**93.9**	**93.7**
BusterX++(**/no_think**)	92.3	92.1

Table 3. Comparison on GenBuster-200K and its OOD Benchmark.

Methods	Test Set		OOD Benchmark	
	ACC	F1	ACC	F1
3D ResNet [19]	70.6	73.5	65.6	70.6
3D ResNeXt [19]	72.6	75.5	65.1	71.0
Vivit [3]	78.5	81.2	76.2	79.4
VideoMAE [53]	79.1	81.7	76.9	80.3
DeMamba [8]	82.0	83.9	79.3	82.0
BusterX [59]	85.5	85.5	84.8	85.1
BusterX++	**88.3**	**88.3**	**92.4**	**92.3**
BusterX++(**/no_think**)	87.5	87.4	91.5	91.5

5.1. Performance on Single-Modality Benchmarks

We begin by validating the effectiveness of BusterX++ on two of the latest high-quality single-modality benchmarks. **So-Fake-Set** [24] is the largest social media image forgery benchmark, and **GenBuster-200K** [59] is the most recent high-resolution video forgery dataset, containing video clips generated by both open-source and commercial models. As shown in Tables 2 and 3, BusterX++ sets new state-of-the-art results on both datasets. Notably, our approach surpasses the MLLM baseline BusterX [59] and So-Fake-R1 [24], which employ a standard cold-start + RL method. BusterX++ starts directly with reinforcement learning, achieving 0.6% improvement (ACC) on So-Fake-Set and 2.8% on GenBuster-200K Test Set, while retaining stronger zero-shot robustness (7.6% improvement) on GenBuster-200K out-of-domain benchmark.

Table 4. Comparison with existing MLLMs on GenBuster++.

Methods	Image		Video		Overall
	Real	Fake	Real	Fake	
MiMo-VL-7B-RL [52]	91.9	12.8	86.1	42.7	58.4
InternVL3-8B [70]	82.1	18.4	80.1	41.3	55.5
Keye-VL-8B [49]	98.3	1.8	95.7	11.8	51.9
MiniCPM-o 2.6 [50]	27.1	78.8	78.4	29.8	53.3
Qwen2.5-Omni-7B [63]	78.7	28.6	81.4	31.4	55.0
Qwen2.5-VL-7B [4]	92.4	8.9	92.6	27.7	55.4
BusterX [59]	79.2	54.3	86.4	53.1	68.3
BusterX++	80.4	76.2	95.3	57.9	**77.5**
BusterX++**(/no_think)**	80.5	74.4	96.4	55.9	76.8

5.2. Performance on GenBuster++

To assess cross-modal capabilities, we re-train the model on a mixture of image and video data from the aforementioned datasets. As shown in Table 4, BusterX++ achieves notable overall accuracy. Compared to BusterX, it shows enhanced adaptability in handling diverse data modalities. This indicates our model's strong cross-modal generalization ability. Moreover, in the Non-Thinking Mode, BusterX++ retains virtually the same performance as the Thinking Mode with only approximately a 0.7% drop, underscoring its flexibility and practical utility.

Table 5. Cold Start vs. Non-Cold Start.

Strategy			Image		Video		Overall
Cold-Start	Stage-1	Stage-3	Real	Fake	Real	Fake	
✓	-	-	72.4	64.7	80.5	51.9	67.4
✓	✓	-	77.3	67.6	88.1	53.7	71.7
✓	✓	✓	81.0	65.9	91.4	53.2	72.9
-	✓	-	78.6	63.4	86.8	48.6	69.4
-	✓	✓	81.2	76.7	94.1	57.5	**77.4**

5.3. Cold Start vs. Non-Cold Start

We compare the traditional cold-start training strategy with our proposed non-cold start approach on GenBuster++. To establish the cold-start baseline, we collect about 1k Chain-of-Thought samples. We skip Stage-2 as it does not affect performance. As shown in Table 5, although the non-cold start strategy initially underperforms (69.4% vs. 71.7%) compared to the cold-start method in Stage-1, it demonstrates significantly superior performance (77.4% vs. 72.9%) after Stage-3 training. By directly starting with RL and incorporating multi-stage training, our model achieves better generalization and adaptability. The results highlight the effectiveness of our training strategy in enhancing the model's final performance.

5.4. Ablation Study

Data Modality. We compare the performance of models trained on single-modality data with cross-modal data. As shown in Table 6, single-modality trained models still underperform the cross-modal approach within their own modality, exhibit certain limitations.

Table 6. Ablation Study for Data Modality.

Data Modality		Image		Video		Overall
Image	Video	Real	Fake	Real	Fake	
✓	-	79.0	72.3	81.2	52.9	71.4
-	✓	78.7	65.4	92.4	55.0	72.9
✓	✓	80.4	76.2	95.3	57.9	**77.5**

Training Strategy. We conduct an ablation study on our 3-stage training strategy. As shown in Table 7, each training stage demonstrates stable performance improvement, with the full pipeline ultimately achieving superior results. Stage-2 has minimal impact on the final performance. This highlights the stability and effectiveness of our incremental training approach.

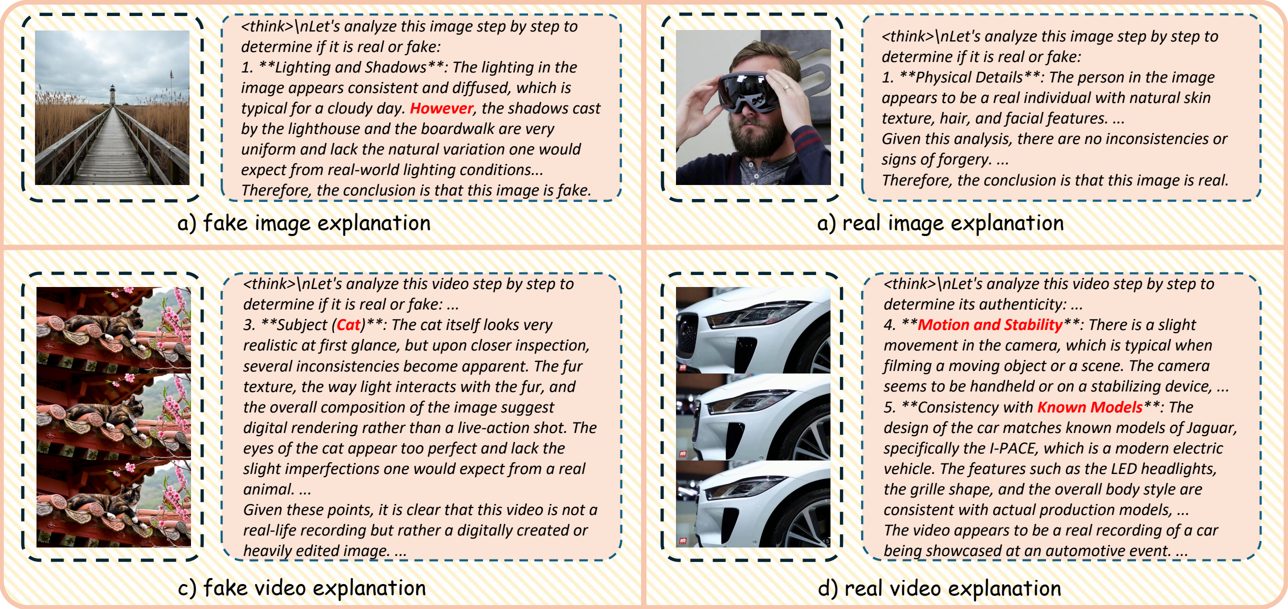
Table 7. Ablation Study for Training Strategy.

Strategy			Image		Video		Overall
Stage-1	Stage-2	Stage-3	Real	Fake	Real	Fake	
✓	-	-	78.6	63.4	86.8	48.6	69.4
✓	✓	-	78.5	63.0	87.2	48.4	69.3
✓	-	✓	81.2	76.7	94.1	57.5	77.4
✓	✓	✓	80.4	76.2	95.3	57.9	**77.5**

These results validate the effectiveness of our comprehensive training approach in enhancing the model's cross-modal capabilities and overall performance.

5.5. Robustness Study

We further assessed the robustness of BusterX++ against common visual perturbations to simulate real-world scenarios, including JPEG Compression (quality=70), Gaussian Noise (σ=5), Gaussian Blur, and their Degradation Cascade inspired by Real-ESRGAN [56]. As shown in Table 8, BusterX++ shows remarkable stability against these low-level distortions, even without specific training on degraded


<think>\nLet's analyze this image step by step to determine if it is real or fake:
1. **Lighting and Shadows**: The lighting in the image appears consistent and diffused, which is typical for a cloudy day. **However**, the shadows cast by the lighthouse and the boardwalk are very uniform and lack the natural variation one would expect from real-world lighting conditions...
Therefore, the conclusion is that this image is fake.

a) fake image explanation

<think>\nLet's analyze this image step by step to determine if it is real or fake:
1. **Physical Details**: The person in the image appears to be a real individual with natural skin texture, hair, and facial features. ...
Given this analysis, there are no inconsistencies or signs of forgery. ...
Therefore, the conclusion is that this image is real.

a) real image explanation

<think>\nLet's analyze this video step by step to determine if it is real or fake: ...
3. **Subject (**Cat**)**: The cat itself looks very realistic at first glance, but upon closer inspection, several inconsistencies become apparent. The fur texture, the way light interacts with the fur, and the overall composition of the image suggest digital rendering rather than a live-action shot. The eyes of the cat appear too perfect and lack the slight imperfections one would expect from a real animal. ...
Given these points, it is clear that this video is not a real-life recording but rather a digitally created or heavily edited image. ...

c) fake video explanation

<think>\nLet's analyze this video step by step to determine its authenticity: ...
4. ****Motion and Stability****: There is a slight movement in the camera, which is typical when filming a moving object or a scene. The camera seems to be handheld or on a stabilizing device, ...
5. **Consistency with **Known Models****: The design of the car matches known models of Jaguar, specifically the I-PACE, which is a modern electric vehicle. The features such as the LED headlights, the grille shape, and the overall body style are consistent with actual production models, ...
The video appears to be a real recording of a car being showcased at an automotive event. ...

d) real video explanation

Figure 6. Case Study. Full responses are provided in appendix.

Table 8. Robustness evaluation of BusterX++.

Perturbation			Image		Video		Overall
JPEG	Noise	Blur	Real	Fake	Real	Fake	
-	-	-	80.4	76.2	95.3	57.9	77.5
✓	-	-	82.1	67.2	94.5	55.6	74.9
-	✓	-	76.4	66.7	95.1	49.2	71.9
-	-	✓	91.6	66.4	93.9	57.6	77.4
✓	✓	✓	90.8	53.5	97.0	40.8	70.5

data. While Gaussian Blur appears to have only a marginal effect on overall accuracy, it produces a mild decline within sub-categories. BusterX++ still delivers solid performance under the harshest degradation cascade, highlights its robustness and practical value in diverse and unpredictable real-world settings.

5.6. Case Study

We provide a detailed case study in Figure 6 to showcase the model's performance in distinguishing between real and fake. Through our analysis, we have some interesting findings from BusterX++'s explanations.

1. **Stable Reasoning**: The model demonstrates a systematic approach to analyze visual content. This step-by-step reasoning allows for a thorough understanding of the content's authenticity.
2. **Attention to low-level details**: The model pays close attention to details that are easily overlooked. It can spot subtle anomalies like unnatural motion blur or irregularities in lighting and shadows, which are crucial for determining authenticity.
3. **Advanced Reasoning**: The model begins to demonstrate advanced reasoning by activating the pre-training knowledge. This enables it to use external knowledge to help determine whether the object is inconsistent with the real world.

6. Conclusion and Limitations

In this paper, we present BusterX++, a novel framework for unified cross-modal AI-generated content detection and explanation. To support evaluation and application expansion, we introduce GenBuster++, a cross-modal benchmark that leverages state-of-the-art image and video generation techniques. By leveraging Multi-stage Training, Thinking Reward, and Hybrid Reasoning, BusterX++ sets new state-of-the-art results in GenBuster++ and other latest benchmarks. Despite promising progress, we recognize certain limitations and aim to address them through future optimization.

1) Generative Technology Adaptation. While our framework demonstrates effectiveness in detecting fake samples from GenBuster-200K, we observe that samples generated with the latest technologies in GenBuster++ present a more significant challenge. This indicates that the rapid evolution of generative methods necessitates continuous adaptation and enhancement of our detection framework to maintain its efficacy.

2) Potential Bottleneck in Post-training. Post-training phase may be approaching a performance bottleneck in our task. To further advance the capabilities of MLLM, future work should explore other training stages.

BusterX++: Towards Unified Cross-Modal AI-Generated Content Detection and Explanation with MLLM

Supplementary Material

Contents of the Appendices:

Section A. Design of Prompts.

Section B. Non-Thinking Mode Response.

Section C. Complete Responses.

Section D. Additional Examples.

A. Design of Prompts

Based on some experiments, we have carefully designed the following prompts to guide the answer.

System Prompt

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 Prompt: IMAGE

Please analyze whether there are any inconsistencies or obvious signs of forgery in the image, and finally come to a conclusion: Is this image real or fake?
Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process.
Then, just answer this MCQ with a single letter:
Q: Is this image real or fake?
Options:
A) real
B) fake

User Prompt: VIDEO

Please analyze whether there are any inconsistencies or obvious signs of forgery in the video, and finally come to a conclusion: Is this video real or fake?
Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process.
Then, just answer this MCQ with a single letter:
Q: Is this video real or fake?
Options:
A) real
B) fake

B. Non-Thinking Mode Response

Response in Non-Thinking Mode (/no_think).

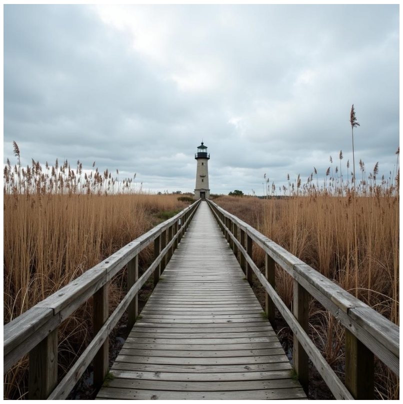
Assistant Prompt

<think>

</think>

<answer>
A
</answer>

C. Complete Responses



FAKE IMAGE

<think>

Let's analyze this image step by step to determine if it is real or fake:

*1. **Lighting and Shadows**: The lighting in the image appears consistent and diffused, which is typical for a cloudy day. However, the shadows cast by the lighthouse and the boardwalk are very uniform and lack the natural variation one would expect from real-world lighting conditions. This suggests that the lighting has been digitally manipulated, which is a common technique in digital forgery.*

*2. **Perspective and Depth**: The perspective of the boardwalk leading towards the lighthouse is perfectly straight and converges at a single point, which is a strong indicator of digital manipulation. In reality, such a straight and precise perspective would be highly unlikely due to the natural curvature of the earth and the imperfections in real-world photography.*

*3. **Vegetation and Texture**: The reeds on either side of the boardwalk have a uniform texture and color that lack the natural variations and imperfections one would expect from real vegetation. The reeds appear too uniform and lack the subtle differences in color and texture that are characteristic of real plants.*

*4. **Lighthouse Structure**: The lighthouse itself looks like it could be real, but upon closer inspection, there are several telltale signs of digital alteration. The edges and corners of the lighthouse appear too sharp and clean, which is not typical of a real structure. Additionally, the lighting on the lighthouse is perfectly even, which is not how light typically behaves on a real lighthouse, especially given the overcast sky.*

*5. **Boardwalk and Railings**: The boardwalk and railings have a very uniform appearance and texture, which is inconsistent with the natural wear and tear one would expect from real wood. The grain and knots in the wood are missing, and the overall finish looks too perfect, which is a hallmark of digital creation rather than a real photograph.*

*6. **Sky and Clouds**: The sky and clouds in the background are rendered with a high level of detail and consistency that is often achieved through digital rendering. The clouds have a uniform density and lack the natural variability and texture that one would expect from real clouds captured in a photograph.*

*7. **Overall Composition**: The composition of the image is highly symmetrical and balanced, which is a common characteristic of digitally created images. Real photographs often have more organic and less symmetrical compositions due to the natural randomness of real-world scenes.*

Given these points, it is clear that this image exhibits numerous signs of digital manipulation and enhancement. There are no elements that suggest this is a real photograph taken with a camera. Instead, all indicators point to this being a digitally created or heavily edited image.

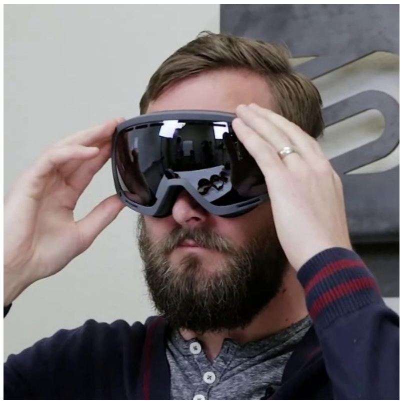
Therefore, the conclusion is that this image is fake. The answer is: B) fake.

</think>

<answer>

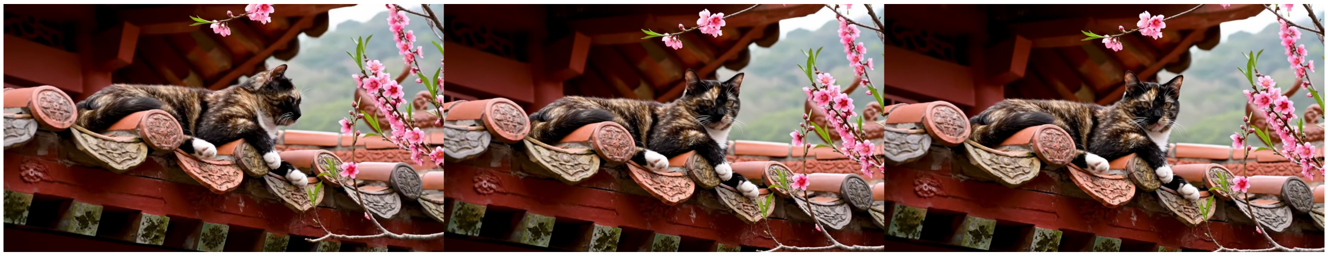
B

</answer>



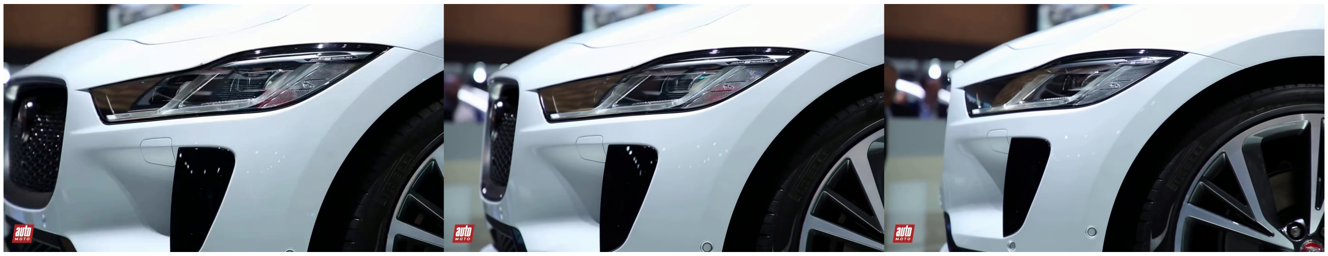
REAL IMAGE

<think>
Let's analyze this image step by step to determine if it is real or fake:
*1. **Physical Details**: The person in the image appears to be a real individual with natural skin texture, hair, and facial features. There are no visible signs of digital manipulation that would indicate this is a synthetic creation. The lighting and shadows are consistent with natural light, which further supports the realism of the image.*
*2. **Clothing and Accessories**: The person is wearing a real-looking sweater with a buttoned shirt underneath. The fabric texture and the way the clothing fits the body are consistent with real-life garments. The ring on the finger is also a real accessory that could be worn by a person.*
*3. **Background and Environment**: The background shows a wall and what appears to be part of a logo or sign, which looks like it could be in a real-world setting. There are no elements that suggest this is a staged or digitally created environment.*
*4. **Human Features**: The person's beard, mustache, and overall facial structure appear natural and unaltered. There are no signs of digital enhancement or distortion that would be typical of a fake image. The hands are positioned in a way that suggests they are holding something, which is a common action in real-life scenarios.*
*5. **Technical Quality**: The image quality is high, with clear details and no visible artifacts that one might expect from a low-quality or AI-generated image. The colors and contrasts are natural, and there are no indications of being a render or a highly stylized artwork.*
Given this analysis, there are no inconsistencies or signs of forgery. All elements of the image point towards it being a real photograph of a person in a real-life setting.
Therefore, the conclusion is that this image is real.
The answer is: A) real.
</think>
<answer>
A
</answer>



FAKE VIDEO

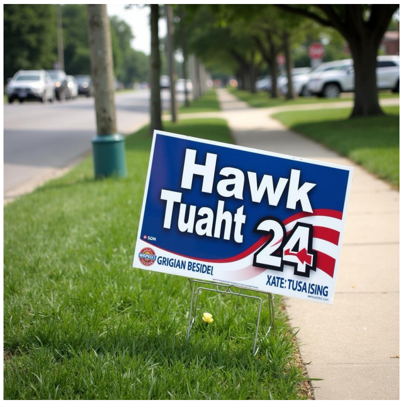
<think>
Let’s analyze this video step by step to determine if it is real or fake:
1. **Setting and Environment**: The setting appears to be a traditional East Asian architectural style, possibly Chinese or Japanese, given the tiled roof and the style of the roof tiles. However, the presence of cherry blossoms (which are typically associated with spring in East Asia) and the overall aesthetic suggest a highly stylized or digitally enhanced environment. Real-life scenes usually have more natural variations in lighting, shadows, and textures that are not present here.
2. **Lighting and Shadows**: The lighting is uniform and lacks the natural variations one would expect from sunlight filtering through clouds or trees. The shadows are consistent and lack the depth and complexity that would be seen in a real-world setting. This suggests that the lighting has been controlled and applied post-production, which is a telltale sign of digital manipulation.
3. **Subject (Cat)**: The cat itself looks very realistic at first glance, but upon closer inspection, several inconsistencies become apparent. The fur texture, the way light interacts with the fur, and the overall composition of the image suggest digital rendering rather than a live-action shot. The eyes of the cat appear too perfect and lack the slight imperfections one would expect from a real animal. Additionally, the positioning of the cat on the roof tiles seems unnatural for a real cat, as they would not lie in such a manner without support.
4. **Background Elements**: The background includes elements like the cherry blossoms and the distant landscape. These elements are also highly stylized and do not match the typical appearance of a real-world setting. The colors are vibrant and saturated in a way that is not characteristic of natural photography. The absence of any real-world imperfections or anomalies further supports the idea that this is a digitally created scene.
5. **Overall Composition**: The composition of the video is highly curated and aesthetically pleasing, which is common in staged photos or videos intended for artistic or promotional purposes. Real-life videos typically have a more organic feel, with less control over every element in the frame. The presence of the cat, the precise placement of the flowers, and the overall symmetry suggest a high level of digital manipulation.
Given these points, it is clear that this video is not a real-life recording but rather a digitally created or heavily edited image. Therefore, the conclusion is that this video is fake.
The answer is: B) fake
</think>
<answer>
B
</answer>



REAL VIDEO

<think>
Let’s analyze this video step by step to determine its authenticity:
1. **Visual Quality and Clarity**: The video shows a close-up of a white car, likely a Jaguar, based on the design elements visible. The image quality is high, with sharp details on the car’s surface, including the headlights, grille, and wheel. There are no pixelation issues or artifacts that would suggest digital manipulation or forgery.
2. **Lighting and Shadows**: The lighting appears consistent and natural, which is typical for a real-life setting. There are shadows cast by the car, which are realistic and align with the direction of light sources that one would expect in an indoor exhibition or showroom environment. These shadows are not uniform or overly artificial, which would be indicative of a staged or digitally created scene.
3. **Background Elements**: In the background, there are blurred elements that resemble other cars and possibly exhibition stands. This suggests that the video was taken at a car show or similar event where multiple vehicles are displayed. The presence of these background elements adds to the realism and indicates that this is not a controlled studio environment but rather a live setting.
4. **Motion and Stability**: There is a slight movement in the camera, which is typical when filming a moving object or a scene. The camera seems to be handheld or on a stabilizing device, capturing the car from different angles as it might be showcased during a presentation. This movement is smooth and natural, further supporting the idea that this is a real recording rather than a CGI creation.
5. **Consistency with Known Models**: The design of the car matches known models of Jaguar, specifically the I-PACE, which is a modern electric vehicle. The features such as the LED headlights, the grille shape, and the overall body style are consistent with actual production models, reinforcing the authenticity of the video.
Given these points, there are no inconsistencies or signs of forgery. The video appears to be a real recording of a car being showcased at an automotive event.
Therefore, the conclusion is that this video is real.
The answer is: A) real
</think>
<answer>
A
</answer>

D. Additional Examples



FAKE IMAGE

<think>

Let’s analyze this image step by step to determine if it is real or fake:

1. **Language and Spelling**: The sign contains words that appear to be a mix of English and what seems to be a constructed language or a fictional dialect. Words like "Hawk," "Tuaht," "24," "GRIGIAN BESIDEL," and "XATE: TUSA ISING" do not correspond to any known language or proper English spelling. This suggests that the image is not a representation of a real political campaign but rather a creation for artistic or illustrative purposes.

2. **Design Elements**: The design includes elements that are typical of political campaign signs, such as the use of bold colors (blue, red, white), the inclusion of a candidate’s name, and the year "24" which could represent 2024. However, the combination of these elements with the non-English words creates a surreal and artificial look.

3. **Context and Setting**: The sign is placed on a grassy area next to a sidewalk, which is a common location for political campaign signs. The background shows a street with parked cars and trees, which appears to be a typical suburban setting. However, the overall composition and the way the sign is presented give away that this is not a photograph of a real-world scene. The lighting and shadows are consistent with a digital rendering rather than a natural outdoor setting.

4. **Technical Indicators**: Upon closer inspection, there are several technical indicators that suggest this is not a real photograph:

- **Blurriness and Focus**: The background is intentionally blurred, which is often done in digital images to draw attention to the foreground subject. This technique is commonly used in AI-generated images to create a sense of depth and focus.
- **Lighting and Shadows**: The lighting on the sign and the surrounding area is perfectly even and lacks the natural variations one would expect from sunlight filtering through trees. The shadows are uniform and lack the complexity that would be present in a real-world scenario.
- **Edge and Pixelation**: There are subtle artifacts and pixelation around the edges of the sign and the text, which are characteristic of AI-generated images. These artifacts are not typically seen in high-resolution photographs taken with real cameras.

5. **Purpose and Intent**: The purpose of this image appears to be artistic or illustrative, possibly for a political satire or a creative project. The combination of real-world elements with fictional text and design elements is a clear indicator that this is not a real photograph but rather a digitally created image.

Given these points, it is evident that this image is not a real photograph but a digitally created or manipulated image. Therefore, the conclusion is that this image is fake.

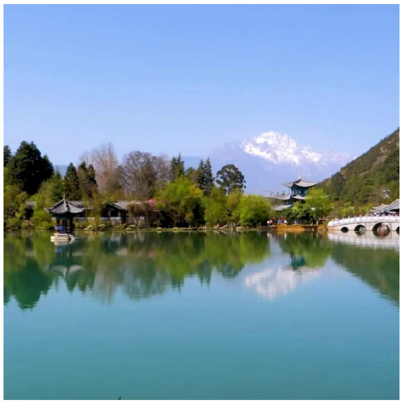
The answer is: B) fake

</think>

<answer>

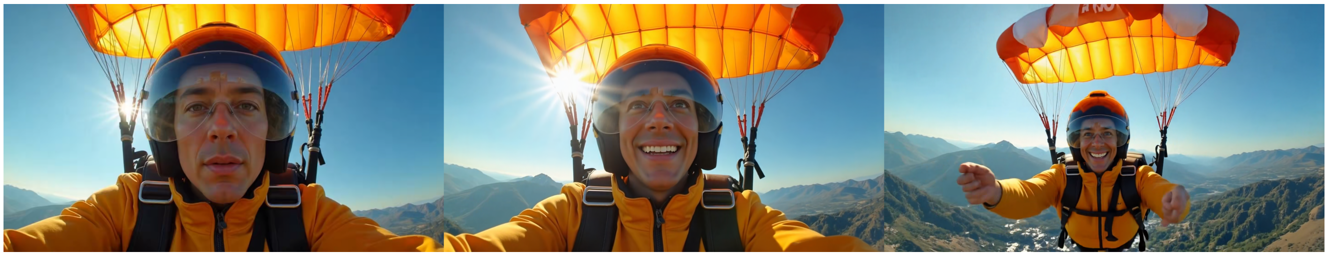
B

</answer>



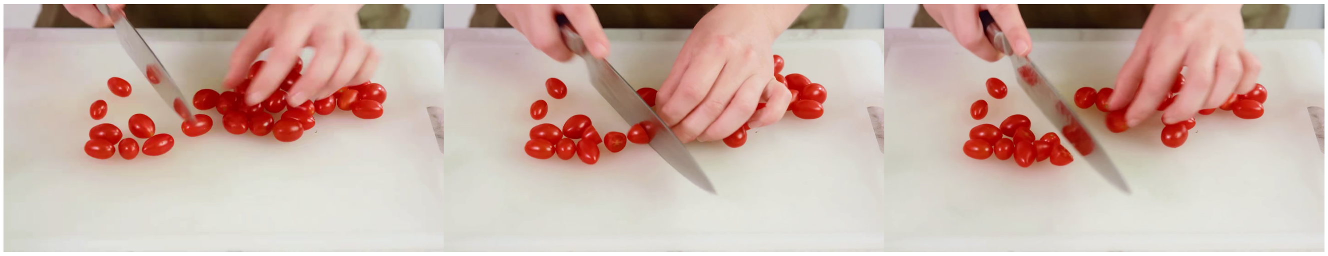
REAL IMAGE

<think>
Let’s analyze this image step by step to determine if it is real or fake:
1. **Scenery and Architecture**: The image depicts a serene landscape with traditional Chinese-style architecture, including pavilions and a bridge, set against a backdrop of mountains and a clear sky. The structures appear to be well-preserved and integrated into the natural environment, which is characteristic of many real-world locations in China known for their historical and cultural significance.
2. **Lighting and Shadows**: The lighting in the image is natural and consistent with what one would expect from a sunny day. The shadows cast by the trees and buildings are realistic and align with the direction of the sunlight, suggesting that the image was taken during daylight hours under clear weather conditions.
3. **Reflections**: The water surface reflects the surrounding scenery with remarkable clarity and detail. The reflections are sharp and symmetrical, indicating that the water is calm and undisturbed. This level of reflection quality is typical of real-world photographs taken in such settings, where the water’s stillness allows for precise mirroring of the scene above.
4. **Vegetation**: The trees and plants are lush and vibrant, with varying shades of green that suggest healthy foliage. The presence of different types of trees and shrubs adds to the authenticity of the scene, as it is common to find such diverse vegetation in well-maintained natural parks or scenic areas.
5. **Background Elements**: The distant mountain range with snow-capped peaks is a realistic feature often found in high-altitude regions of China. The snow appears to be genuine and not artificial, which supports the idea that this is a real photograph rather than a digital creation.
6. **Overall Composition**: The composition of the image, with its balanced elements—water, land, sky, and architecture—is typical of professional photography aimed at capturing the beauty of natural and cultural landscapes. There are no visible signs of digital manipulation or artistic alterations that would typically be present in a fake image.
7. **Technological Verification**: Upon closer inspection, there are no indications of digital artifacts, blurring issues, or other common signs of image manipulation that are often used to identify fake images. The image appears to be a high-resolution capture of a real location, likely taken with a professional camera given the sharpness and detail.
Given this analysis, there are no inconsistencies or signs of forgery. The image presents a realistic depiction of a natural and cultural landscape, likely from a real-world location in China.
Therefore, the conclusion is that this image is real.
The answer is: A) real
</think>
<answer>
A
</answer>



FAKE VIDEO

<think>
Let's analyze this video step by step to determine if it is real or fake:
1. **Clothing and Equipment**: The individual is wearing a helmet with a visor, a yellow jacket, and what appears to be a harness system for paragliding. These items are typical for paragliding activities. However, upon closer inspection, there are several elements that suggest this is not a real, live-action video but rather a digitally created or heavily edited image.
2. **Lighting and Shadows**: The lighting in the video is extremely bright and uniform, which is unusual for a real-life scenario. The sun appears to be directly overhead, casting no significant shadows on the person's face or equipment. In reality, paragliders would have more dynamic lighting and shadows due to the angle of the sun relative to their position.
3. **Background and Environment**: The background features a clear blue sky and distant mountains. While these elements could be real, the way they are presented here lacks the depth and texture one would expect from a real photograph. The mountains appear too flat and lack the natural variations in color and shading that are typical of real landscapes.
4. **Parachute and Ropes**: The parachute is orange and appears to be inflated, which is correct for paragliding. However, the ropes connecting the parachute to the harness look overly smooth and lack the natural wear and tear one would expect from real paragliding gear. Additionally, the perspective and angle of the ropes do not align with how they would typically appear in a real paragliding setup.
5. **Posture and Expression**: The person's posture and expression seem exaggerated for a real paraglider. Real paragliders often have a more relaxed and focused demeanor, especially when in motion. The smile and the way the person is holding the camera also suggest a staged or animated scene rather than a spontaneous moment captured during a real paragliding session.
6. **Technical Quality**: There are no visible graininess, noise, or other artifacts that one would expect from a real-life video recorded with a camera. The image is too clean and has a high level of detail that is characteristic of digital rendering or heavy post-processing.
7. **Overall Consistency**: When considering all these factors together, there are numerous inconsistencies that point towards this being a digitally created or heavily edited image. The level of detail and the way the elements are combined suggest the use of 3D modeling and rendering software rather than a real-life capture.
Given this analysis, it is clear that this video is not a real-life recording but rather a fake creation, likely intended for artistic or promotional purposes.
Therefore, the conclusion is that this video is fake.
The answer is: B) fake
</think>
<answer>
B
</answer>



REAL VIDEO

<think>

Let’s analyze this video step by step to determine if it is real or fake:

*1. **Physical Evidence**: The video shows hands actively cutting cherry tomatoes on a cutting board. The hands appear to be those of a person, and the movement is consistent with human hand movements. There are no signs of digital manipulation or animation that would suggest this is not a real-life scenario.*

*2. **Tools and Objects**: The tools used, such as the knife and the cutting board, are typical kitchen items. They are positioned and used in a manner that is consistent with real-world cooking activities. There are no indications of these being props or models.*

*3. **Lighting and Shadows**: The lighting in the video is natural and consistent with indoor lighting conditions. Shadows cast by the hands and objects are realistic and do not exhibit any anomalies that would suggest digital alteration.*

*4. **Motion and Texture**: The motion of the knife slicing through the tomatoes is smooth and fluid, which is characteristic of real-life actions. The texture of the tomatoes and the cutting board also appears natural and not overly polished, which is typical for real food and kitchen surfaces.*

*5. **Background and Environment**: The background is simple and unobtrusive, focusing attention on the task at hand. There are no elements that suggest a staged or artificial environment. The simplicity and clarity of the scene are indicative of a real-life recording rather than a highly produced or animated sequence.*

*6. **Consistency**: Throughout the video, there is a consistent flow of action without any abrupt changes or glitches that one might expect from a digitally created video. The continuity of the action and the lack of any visual artifacts further support the conclusion that this is a real video.*

Given this analysis, there are no inconsistencies or signs of forgery. All elements point towards this being a real video capturing a real-life cooking activity.

Therefore, the conclusion is that this video is real.

The answer is: A) real

</think>

<answer>

A

</answer>

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In WIFS, 2018. 1, 2

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com`, 2024. 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In CVPR, 2021. 6

[4] 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, 2025. 3, 6, 7

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. Arxiv, 2025. 4

[6] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting generated images by real images only. Arxiv, 2023. 1, 2

[7] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. Arxiv, 2024. 2, 3

[8] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. Arxiv, 2024. 1, 2, 3, 6

[9] Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. Arxiv, 2025. 2, 3

[10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. Arxiv, 2025. 3

[11] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. Arxiv, 2022. 2, 3

[12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. In CVPR, 2020. 2

[13] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Arxiv, 2025. 3

[14] Stefano Dell'Anna, Andrea Montibeller, and Giulia Boato. Truefake: A real world case dataset of last generation fake images also shared on social networks. Arxiv, 2025. 2, 3

[15] Junxian Duan, Yuang Ai, Jipeng Liu, Shenyuan Huang, Huaibo Huang, Jie Cao, and Ran He. Test-time forgery detection with spatial-frequency prompt learning. ICCV, 2024. 1, 2

[16] Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. Arxiv, 2025. 3, 6

[17] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, Xiaonan Nie, Zhiwu Qing, Yuxi Ren, Li Sun, Zhi Tian, Rui Wang, Sen Wang, Guoqiang Wei, Guohong Wu, Jie Wu, Ruiqi Xia, Fei Xiao, Xuefeng Xiao, Jiangqiao Yan, Ceyuan Yang, Jianchao Yang, Runkai Yang, Tao Yang, Yihang Yang, Zilyu Ye, Xuejiao Zeng, Yan Zeng, Heng Zhang, Yang Zhao, Xiaozheng Zheng, Peihao Zhu, Jiaxin Zou, and Feilong Zuo. Seedance 1.0: Exploring the boundaries of video generation models. Arxiv, 2025. 1, 4

[18] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. Arxiv, 2024. 6

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In CVPR, 2018. 6

[20] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In CVPR, 2021. 2

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In ICLR, 2022. 6

[22] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. Arxiv, 2025. 3

[23] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In CVPR, 2025. 2, 3, 6

[24] Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, and Guangliang Cheng. So-fake: Benchmarking and explaining social media image forgery detection. Arxiv, 2025. 2, 3, 4, 6, 7

[25] Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In ICCV, 2025. 5

[26] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. Legion: Learning to ground and explain for synthetic image detection. Arxiv, 2025. 2, 3

[27] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. Arxiv, 2022. 2

[28] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In ICCV, 2021. 2

[29] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. Arxiv, 2025. 4

[30] Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics. Arxiv, 2025. 2, 3

[31] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In ICLR, 2023. 3

[32] Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Qinglang Guo, Haiyang Yu, Yong Liao, and Pengyuan Zhou. Decof: Generated video detection via frame consistency: The first benchmark dataset. Arxiv, 2024. 2, 3

[33] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. Arxiv, 2024. 4

[34] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video. Arxiv, 2025. 2, 3

[35] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In CVPR, 2023. 1, 2

[36] OpenAI. Gpt-4o system card. Arxiv, 2024. 1, 3, 4

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. NeurIPS, 35:27730–27744, 2022. 3

[38] Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. Mllm-enhanced face forgery detection: A vision-language fusion solution. Arxiv, 2025. 2, 3

[39] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and Arnav Chopra et al. Humanity's last exam. Arxiv, 2025. 3

[40] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In ICCV, pages 1–11, 2019. 1, 2

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. Arxiv, 2017. 3

[42] ByteDance Seed. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. Arxiv, 2025. 3

[43] SkyReels-AI. Skyreels v1: Human-centric video foundation model. `https://github.com/SkyworkAI/SkyReels-V1`, 2025. 4

[44] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In NeurIPS, 2024. 2, 3

[45] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In CVPR, 2023. 1, 2

[46] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. Arxiv, 2020. 1, 2

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. Arxiv, 2023. 3

[48] Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. Arxiv, 2025. 3

[49] Kwai Keye Team. Kwai keye-vl technical report. Arxiv, 2025. 7

[50] OpenBMB Team. Minicpm-o 2.6: Openbmb's lightweight multimodal model. `https://github.com/OpenBMB/MiniCPM-o`, 2024. 7

[51] Qwen Team. Qwen3 technical report. Arxiv, 2025. 3, 5

[52] Xiaomi LLM-Core Team. Mimo-vl technical report. Arxiv, 2025. 7

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS, 2022. 6

[54] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020. 1, 2, 6

[55] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. Arxiv, 2025. 3

[56] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In ICCVW, 2021. 7

[57] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In ICCV, pages 22388–22398. IEEE, 2023. 6

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In NeurIPS, 2022. 4

[59] Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. Arxiv, 2025. 2, 3, 4, 6, 7

[60] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. Arxiv, 2025. 2, 3

[61] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. Arxiv, 2025. 3

[62] Junhao Xu, Jingjing Chen, Xue Song, Feng Han, Haijun Shan, and Yu-Gang Jiang. Identity-driven multimedia forgery detection via reference assistance. In ACMMM, 2024. 2

[63] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. Arxiv, 2025. 7

[64] 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. In ICLR, 2025. 2, 3

[65] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. Arxiv, 2025. 3, 5

[66] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In CVPR, pages 12765–12774, 2024. 6

[67] Kaichen Zhang, Yuzhong Hong, Junwei Bao, Hongfei Jiang, Yang Song, Dingqian Hong, and Hui Xiong. Gvpo: Group variance policy optimization for large language model post-training. Arxiv, 2025. 3

[68] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In CVPR, 2021. 1, 2

[69] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. Arxiv, 2025. 2, 3

[70] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. Arxiv, 2025. 3, 7

[71] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. Arxiv, 2023. 2, 3

[72] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In ACMMM, 2020. 2