
IVY-FAKE: A Unified Explainable Framework and Benchmark for Image and Video AIGC Detection

Wayne Zhang*
 π^3 AI Lab

Changjiang Jiang*
Wuhan University

Zhonghao Zhang
 π^3 AI Lab

Chenyang Si
Nanjing University

Fengchang Yu
Wuhan University

Wei Peng
Stanford University

Abstract

The rapid advancement of Artificial Intelligence Generated Content (AIGC) in visual domains has resulted in highly realistic synthetic images and videos, driven by sophisticated generative frameworks such as diffusion-based architectures. While these breakthroughs open substantial opportunities, they simultaneously raise critical concerns about content authenticity and integrity. Many current AIGC detection methods operate as black-box binary classifiers, which offer limited interpretability, and no approach supports detecting both images and videos in a unified framework. This dual limitation compromises model transparency, reduces trustworthiness, and hinders practical deployment. To address these challenges, we introduce **IVY-FAKE**, a novel, unified, and large-scale dataset specifically designed for explainable multimodal AIGC detection. Unlike prior benchmarks, which suffer from fragmented modality coverage and sparse annotations, IVY-FAKE contains over 150,000 richly annotated training samples (images and videos) and 18,700 evaluation examples, each accompanied by detailed natural-language reasoning beyond simple binary labels. Building on this, we propose **Ivy Explainable Detector (IVY-XDETECTOR)**, a unified AIGC detection and explainable architecture that jointly performs explainable detection for both image and video content. Our unified vision-language model achieves state-of-the-art performance across multiple image and video detection benchmarks, highlighting the significant advancements enabled by our dataset and modeling framework. Our data is publicly available at <https://huggingface.co/datasets/AI-Safeguard/Ivy-Fake>.

1 Introduction

Artificial Intelligence Generated Content (AIGC) has advanced rapidly, particularly in visual modalities such as images and videos. These developments have been driven by powerful generative models, notably diffusion-based architectures like DALL-E (Betker et al., 2023), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022a), which have redefined state-of-the-art performance in text-to-image synthesis. More recently, models such as OpenAI’s SORA (Brooks et al., 2024) have extended AIGC into video generation, producing visually coherent, minute-long videos conditioned on text prompts with remarkable fidelity and semantic consistency.

While these breakthroughs unlock transformative applications across digital media, advertising, and entertainment, they simultaneously exacerbate challenges related to content authenticity. The growing realism of synthetic content increasingly blurs the boundary between genuine and fabricated media, raising critical concerns regarding misinformation, content provenance, and public trust.

*Equal contribution

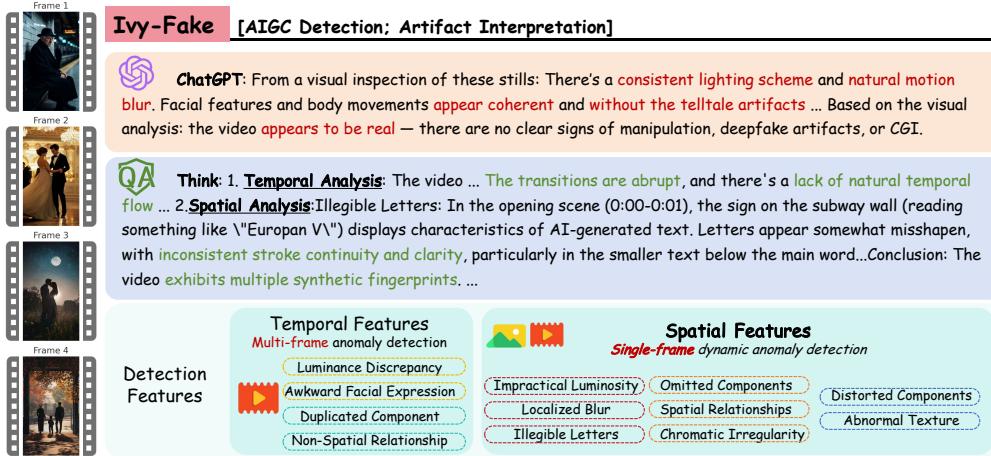


Figure 1: Overview of the IVY-FAKE framework: By conducting in-depth analysis of temporal and spatial artifacts, the framework enables explainable detection of AI-generated content.

To date, most AIGC detection methods have framed the problem as a binary classification task: determining whether content is authentic or AI-generated. However, these approaches typically offer limited interpretability, which undermines transparency and user confidence. Many rely on low-level statistical cues or implicit features without providing insight into which regions of an image or video contribute to the detection outcome. Furthermore, current detection datasets exhibit constraints in terms of generator diversity, modality coverage, and annotation quality, hindering rigorous evaluation of detection models under varied and complex real-world scenarios.

In response to these limitations, Multimodal Large Language Models (MLLMs) have emerged as promising candidates for AIGC detection. By integrating visual and linguistic reasoning, MLLMs offer unique advantages, enabling not only authenticity assessment but also natural language explanations of detected anomalies. This capacity for explainability enhances the interpretability and credibility of detection results, positioning MLLMs as valuable tools for mitigating risks associated with synthetic media proliferation.

Nevertheless, existing benchmarks remain inadequate for evaluating explainable AIGC detection. Datasets such as AIGCDetectionBenchmark (Zhong et al., 2023) and GenVideo (Chen et al., 2024a) provide only binary labels, while more recent resources like LOKI (Ye et al., 2025) attempt to incorporate fine-grained anomaly annotations across modalities but remain limited in scale and diversity. Similarly, FakeBench (Li et al., 2024c) emphasizes explainable fake image detection, yet omits video content, and FakeClue (Wen et al., 2025) offers extensive image-level annotations but lacks integrated video data. This fragmentation results in significant gaps in benchmarking and hinders unified progress.

To address these challenges, we propose IVY-FAKE, a comprehensive benchmark designed to evaluate explainable multimodal AIGC detection. IVY-FAKE offers: 1) Diverse Multimodal Data, a large-scale dataset comprising 94,781 annotated images and 54,967 annotated videos for training, along with 18700 samples per modality for evaluation. 2) Explainable Annotations, rich annotations that extend beyond binary labels to include detailed reasoning, enabling nuanced evaluation of models' interpretability and explanatory capabilities. Based on our proprietary dataset, we introduce IVY-XDETECTOR, a model for detecting AI-generated images and videos and explaining the associated artifacts. As demonstrated by Figure 1, unlike other MLLMs, IVY-XDETECTOR excels at spotting generative artifacts, spatial in images and both spatial and temporal in videos. By integrating multiple spatial and temporal feature extractors, it detects image-level artifacts and video-level temporal inconsistencies with superior accuracy. Our main contributions are summarized as follows:

- **Unified Vision-Language Detector:** We introduce the first unified vision–language detector that provides detailed, explainable AIGC detection for both images and videos, achieving state-of-the-art performance across multiple detection and explainability benchmarks.
- **Unified Multimodal Detection & Explainability Dataset:** We present the first large-scale benchmark for explainable AIGC detection across images and videos, featuring 5.4 million

labeled samples plus 150 000 instances enriched with fine-grained visual annotations and textual justifications to enable transparent, multimodal evaluation and future research.

2 Related Work

2.1 Synthetic Content Detection

Due to growing concerns about the misuse of synthetic data (Deng et al., 2024), research on AI-generated content (AIGC) detection has expanded rapidly in recent years. Most existing models for AI-generated images and videos formulate the task as binary classification, simply predicting whether the content is "real" or "fake." Representative examples include CNN-based AIGVDet (Bai et al., 2024), CNNSpot (Wang et al., 2020) and Transformer-based models such as DIRE (Wang et al., 2023) and AIDE (Yan et al., 2025). Meanwhile, several works have explored the application of multimodal large language models (MLLMs) (Zheng et al., 2024; Guo et al., 2025b; Zheng et al., 2023a) to AIGC detection, including FakeBench (Li et al., 2024c), LoKI (Ye et al., 2025), Synartifact (Cao et al., 2024) and Bi-LORA (Keita et al., 2025). However, these approaches largely overlook the importance of interpretability in AIGC detection.

Some efforts attempt to introduce interpretability by leveraging spatial annotations (Dong et al., 2022) or frequency-domain artifact analysis (Zhang et al., 2023). Nevertheless, the resulting explanations are often difficult for humans to comprehend, as they lack clarity in natural language. This limitation is particularly evident in the video domain, where AI-generated content frequently exhibits obvious flaws, e.g., incoherent frame transitions and object inconsistency, that are easily noticed and reasoned about by humans (Deng et al., 2024). FakeClue (Wen et al., 2025) introduces the use of vision-language models (VLMs) to provide interpretability for image-level detection, but it does not offer a unified framework that integrates both images and videos.

2.2 Datasets for Synthetic Content Detection.

Early datasets for synthetic content detection, such as CNNSpot (Wang et al., 2020), primarily collected fake images generated by GAN-based models (Goodfellow et al., 2014; Zhu et al., 2017; Brock et al., 2018). However, with the advent of more advanced generative architectures like diffusion models (Ho et al., 2020; Dhariwal, Nichol, 2021; Rombach et al., 2022b; Hertz et al., 2022; Nichol et al., 2021) and their variants, the authenticity of generated content has significantly increased, making it more challenging for detection models to discern. This has spurred the development of newer datasets, including ArtiFact (Cao et al., 2024), GenImage (Zhu et al., 2023b), and WildFake (Hong et al., 2025). GenImage (Zhu et al., 2023b) comprises images from the 1000 ImageNet (Russakovsky et al., 2015) categories, generated by eight state-of-the-art generators such as Stable Diffusion (Rombach et al., 2022a) and Midjourney. Nevertheless, these datasets predominantly focus on image-based content. More recently, datasets emphasizing interpretability have also been introduced. FakeClue (Wen et al., 2025) contains a large amount of image data with explainability annotations but lacks video data. LOKI (Ye et al., 2025) offers data across 26 different categories and includes 18,000 distinct questions; however, its volume of multimodal data is relatively small and primarily suited for evaluation rather than comprehensive model training. Therefore, a critical gap exists for a unified benchmark encompassing both image and video modalities to rigorously evaluate the performance of contemporary AIGC detectors.

3 Dataset

The IVY-FAKE dataset is a comprehensive dataset specifically designed for explainable multimodal AIGC detection. It contains a substantial scale of data, comprising 94,781 images and 54,967 videos in the training set, alongside 8,731 images and 9,956 videos allocated for testing. To ensure extensive coverage of potential AIGC scenarios, the dataset features significant content diversity, encompassing numerous categories such as animals, objects, human portraits, scenes, documents, satellite imagery, and DeepFake media, as shown in Figure 2. Moreover, we prioritize source diversity by including synthetic data generated through various cutting-edge architectures, i.e., GANs, Diffusion models, and Transformer-based generators, paired with authentic content from real-world contexts. Additionally, to maintain dataset relevance in the face of rapidly evolving generative technologies, we supplemented

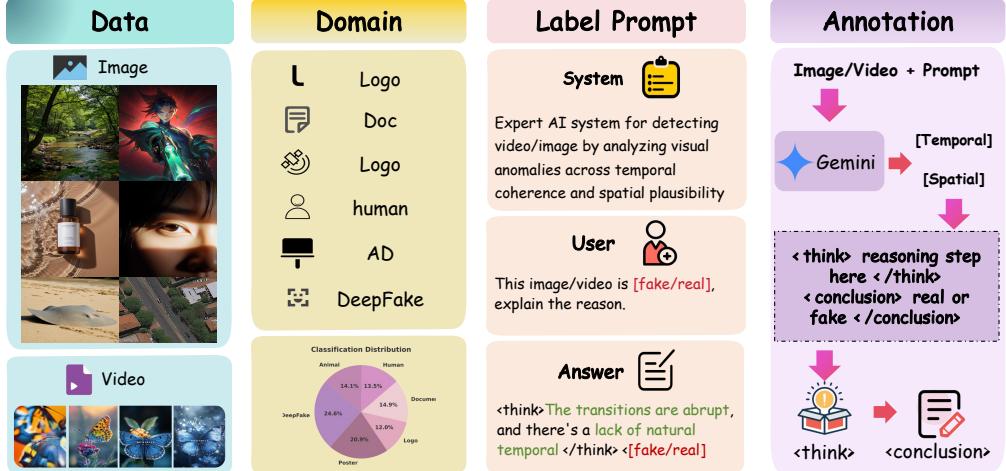


Figure 2: Overview of the proposed unified and explainable IVY-FAKE dataset. Input images or videos from diverse domains are processed alongside specific prompts by an MLLM. The model leverages temporal and spatial analysis to produce structured, explainable annotations.

publicly available data by continuously collecting the latest AIGC images and videos from online sources. This strategy ensures that IVY-FAKE remains current, comprehensive, and reflective of contemporary generation techniques. Next, we will detail how the IVY-FAKE dataset is created.

3.1 Data Collection

3.1.1 Video Dataset Construction

Public Benchmark Sources A total of approximately 110,000 video clips were collected from diverse sources to ensure comprehensive coverage of contemporary generative models. A significant portion was sourced from established public datasets, notably GenVideo (Chen et al., 2024a) and LOKI (Ye et al., 2025), which offer large collections of AI-generated videos alongside authentic counterparts. These datasets include outputs from a variety of advanced video generation frameworks, ensuring broad model diversity. Specifically, our dataset includes samples generated by state-of-the-art models such as SORA (?), Stable Video Diffusion (SVD), Runway, ModelScope, Latte, VideoCrafter, and OpenSora.

Web-Crawled Sources To supplement public datasets and enhance real-world relevance, we crawled additional video content from publicly accessible platforms. This included authentic and suspected AI-generated videos from popular sources such as YouTube and various social media platforms. This web-crawled content contributes valuable diversity and reflects a broader spectrum of naturally occurring AI-generated media.

3.1.2 Image Dataset Construction

Public Benchmark Sources Approximately 110,000 images were collected following a similar methodology to the video dataset. Publicly available datasets, including FakeClue (Wen et al., 2025) and WildFake (Hong et al., 2025), were utilized to gather a wide range of AI-generated images. These benchmarks include content produced by mainstream generative models, notably GANs (Goodfellow et al., 2014) and Diffusion models, thus ensuring the inclusion of diverse generative paradigms.

Web-Crawled Sources To further enrich the dataset and capture authentic and synthetic images prevalent in real-world scenarios, additional samples were sourced from various public websites. These included social media platforms, image-sharing sites, and other third-party repositories that host both genuine and AI-generated images.

3.2 Sampling Strategy and Dataset Balance

To ensure balanced representation across generative models and avoid potential biases, a stratified sampling strategy was employed during data collection. This approach maintained proportional inclusion of samples from each generator type, thereby preserving the diversity of visual content and facilitating fair and comprehensive evaluation across detection models.

3.3 Data Annotation

To generate explainable annotations, we employed the multimodal large language model Gemini 2.5 Pro (Team et al., 2023), leveraging a knowledge distillation process to produce structured, interpretable outputs. Each of the collected images and videos is classified as either authentic or AI-generated. The specific system prompt guiding this task is provided in Appendix. To ensure consistent and structured output, we adapted the distillation template from DeepSeek-R1-Zero (Guo et al., 2025a), which requires the model to first articulate a reasoning process before issuing a final determination, as outlined in Figure 2. This structured approach was essential to mitigate content-related biases, such as enforcing reflective reasoning or imposing predefined problem-solving patterns, thereby allowing the model’s natural reasoning process to emerge during reinforcement learning (RL). Furthermore, this format facilitated the identification and exclusion of non-compliant distillations, ensuring high data quality and consistency for subsequent model training. As noted in FakeVLM (Wen et al., 2025), directly predicting content authenticity using vision-language models exhibits performance limitations. FakeVLM addresses this by adopting a more complex output structure instead of a simple “Real”/“Fake” binary classification. In contrast, our approach introduces a structured response format using `<think></think>` and `<conclusion></conclusion>` tags. This template guides the model to first articulate its reasoning process before arriving at a final decision, which has been empirically shown to improve both detection accuracy and interpretability.

To enhance annotation accuracy and guide the generation of relevant explanations, we incorporated category-level prior knowledge during annotation. Specifically, Gemini was provided with the ground truth authenticity label (i.e., “real” or “fake”) and instructed to explain the rationale behind this classification. The user prompt was standardized as follows:

```
This {file_type} is {label}. Explain the reason.
```

In this template, `{file_type}` indicates the modality of the input—either “image” or “video”—and `{label}` represents the ground-truth label, assigned as either “real” or “fake”. The distilled explanations were further categorized along two major dimensions to facilitate structured analysis: **Spatial Features**, which comprises eight sub-dimensions and captures artifacts and inconsistencies observable within individual frames or static images. **Temporal Features**, which includes four sub-dimensions and describes anomalies associated with motion, temporal coherence, and cross-frame consistency. Since still images inherently lack temporal attributes, this category is exclusively applicable to video annotations. These categories were informed by established taxonomies of visual artifacts in generative content, as detailed in prior research (Deng et al., 2024).

3.4 Comparison with Existing Datasets

A comparative overview of IVY-FAKE and several existing AIGC detection datasets is provided in Table 1. Notably, IVY-FAKE offers unique advantages by integrating explainable annotations across both image and video modalities, addressing a significant gap in current resources.

In contrast, datasets such as FakeClue (Wen et al., 2025) and FakeBench (Li et al., 2024c) provide only limited annotated samples, thereby restricting their applicability primarily to evaluation scenarios rather than serving as robust resources for model training. Similarly, LOKI (Ye et al., 2025), while offering explainability-focused annotations, is confined to a single modality (images) and evaluates across fewer dimensions. This constraint limits its suitability for comprehensive research on multimodal AIGC detection and explainability. By offering large-scale, richly annotated datasets spanning images and videos, IVY-FAKE establishes a more holistic and versatile benchmark for advancing explainable AIGC detection in multimodal contexts.

Table 1: Comparison on the different datasets used in binary classification and interpretability tasks. Token lengths are computed using the GPT-4o tokenizer from the tiktoken library.

Dataset	Avg Token Lengths	Generator	Modality	Dataset	
				fake	real
FakeBench (Li et al., 2024c)	-	6	Image	3K	3K
VANE-Bench (Bharadwaj et al., 2024)	101	5	Image*	2K	1K
LOKI (Ye et al., 2025)	99	~16	Image+Video	3K	0
FakeClue (Ye et al., 2025)	120	26	Image	68K	36K
IVY-FAKE	530	>30	Image+Video	74K	86K

4 Methodology

This study targets at an unified benchmark of AIGC detection for both image and video content. Our preliminary investigations revealed that existing MLLMs exhibit inadequate performance on these tasks. To overcome this limitation, we propose IVY-XDETECTOR, a multimodal large language model designed explicitly for robust and explainable AIGC detection. The following subsections detail our instructional dataset construction and the progressive multimodal training framework employed in developing Ivy-Detector.

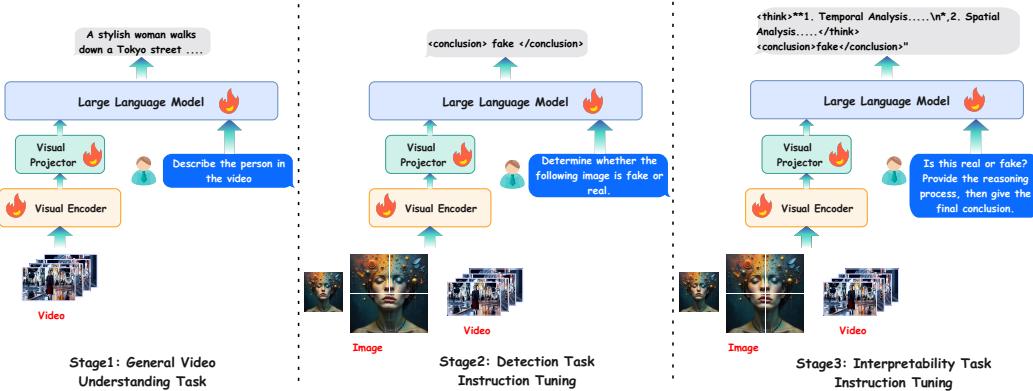


Figure 3: Overview of the three-stage training pipeline for Ivy-Detector, including general video understanding, detection instruction tuning, and interpretability instruction tuning.

4.1 Ivy-Detector model

We introduce Ivy-xDetector, an architecture that adheres to the LLaVA (Liu et al., 2023; Li et al., 2024a) paradigm and is initialized using Ivy-VL-LLaVA (Zhang et al., 2024) weights. As illustrated in Figure 3, The Ivy-xDetector framework comprises three core components: a Visual Encoder, a Visual Projector, and a Large Language Model (LLM). We employ SigLIP (Zhai et al., 2023) as the visual backbone to process input images and frames extracted from video sequences. To enable fine-grained detection, particularly for high-resolution images, we implement a dynamic resolution strategy. Input images are partitioned into multiple 384×384 sub-images, which are then collectively fed into the Visual Encoder. This approach supports an effective input resolution of up to 2304×2304 . For video inputs, individual frames are resized to 384×384 . Post-encoding by the Visual Encoder, the resulting features undergo a pooling operation, compressing the token sequence to one-quarter of its original length. Crucially, to preserve rich temporal information from video data, we eschew temporal compression of video features. Features from all frames are concatenated and subsequently processed by the LLM. Through joint training on image and video datasets, Ivy-xDetector acquires a unified understanding of both modalities, enabling it to perform explainable detection tasks for both images and videos.

4.2 Progressive Multimodal Training Framework

To effectively train IVY-XDETECTOR and fully leverage our curated instruction dataset, we devise a progressive multi-modal training framework. This framework employs a staged optimization strategy to incrementally enhance the model’s proficiency in both AI-Generated Content (AIGC) detection and the generation of associated explanations.

Stage 1: Imbuing Video Understanding Capabilities. We initialize IVY-XDETECTOR using the Ivy-VL-LLaVA model (sub-4B parameter scale), which has demonstrated state-of-the-art (SOTA) performance on image-text benchmarks but lacks prior exposure to video data. To address this, we curate a dataset of 3 million video-text pairs, aggregated from sources such as VideoChatFlash (Li et al., 2024b) and VideoLLaMA3 (Zhang et al., 2025), to equip the model with fundamental video comprehension abilities. Further details regarding these datasets are provided in Appendix.

Stage 2: AIGC Detection Fine-tuning. Previous iterations of the base model were primarily trained on authentic (non-generated) datasets. To specialize IVY-FAKE for AIGC detection, we compile a targeted dataset for instruction fine-tuning, drawing from established datasets including Demamba, FakeClue, and WildFake. This dataset undergoes rigorous preprocessing, including the removal of corrupted or low-quality samples. For video data, frames are extracted at a uniform rate of 1 frame per second (fps). The core objective of this stage is to train the Multi-modal Large Language Model (MLLM) component for binary AIGC discrimination—classifying content as either “real” or “fake.” Given the focused nature of this binary classification task, a single, consistent instruction template is utilized. This stage aims to cultivate high-accuracy classification capabilities across diverse sources and types of AIGC.

Stage 3: Joint Optimization for Instruction-Driven Detection and Explainability. This culminating stage of training is designed to endow IVY-XDETECTOR with the capacity to generate high-quality, human-understandable explanations while preserving the AIGC detection accuracy achieved in Stage 2. Our preliminary experiments revealed that sequentially fine-tuning for explainability after the detection-focused training (Stage 2) often resulted in the model struggling to adhere to detection-specific instructions. Consequently, in Stage 3, we adopt a joint training approach, concurrently fine-tuning the model on a combined dataset comprising the AIGC detection data from Stage 2 and newly introduced explainability-focused instruction data. The instructions in this phase are crafted to elicit detailed, step-by-step reasoning from the model, thereby placing greater demands on its generative and inferential capabilities. The model emerging from this stage is expected to not only accurately perform AIGC detection but also provide comprehensive and persuasive rationales for its classifications.

Through this three-stage progressive training paradigm, we aim for IVY-XDETECTOR to systematically develop a comprehensive skill set: from discerning subtle AIGC artifacts and making accurate classifications, to articulating coherent and justifiable explanations. This holistic approach is designed to ensure IVY-XDETECTOR exhibits robust and superior performance on complex AIGC detection challenges.

5 Experiments

We perform extensive experiments to assess both detection and explanation capabilities. In particular, the proposed method is evaluated on the classification (real/fake) tasks and reasoning tasks for both image and video content using the proposed unified framework. For the classification task, we test our model on both image and video content to detect the synthetic content. We report standard accuracy (Acc) and macro-averaged F1 score (F1) to assess the model’s ability to distinguish real from fake instances. For the reasoning task, we measure the similarity between the model’s reasoning process and the reference annotations using the ROUGE-L score (Lin, 2004), which captures the longest common subsequence between predicted and reference texts, reflecting token-level overlap. Since ROUGE-L may fail to fully capture the fidelity of reasoning steps, we adopt an LLM-as-a-judge evaluation paradigm (Zheng et al., 2023b), following the FakeBench protocol (Li et al., 2024c), which assesses model responses along four dimensions: (1) Completeness: It reflects the extent to which the response fully addresses all aspects of the user’s question. More complete responses should incorporate information aligning well with the “golden clues” or reference answers. Incomplete

Table 2: Comparison on the Genimage (Zhu et al., 2023b). Accuracy (%) of different detectors (rows) in detecting real and fake images from different generators (columns). The best result and the second-best result are marked in bold and underline, respectively.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Mean
CNNSpot (Wang et al., 2020)	52.80	96.30	95.90	50.10	39.80	78.60	53.40	46.80	64.21
F3Net (Qian et al., 2020)	50.10	99.90	99.90	49.90	50.00	99.90	49.90	49.90	68.69
DIRE (Wang et al., 2023)	60.20	99.90	<u>99.80</u>	50.90	55.00	<u>99.20</u>	50.10	50.20	70.66
GenDet (Zhu et al., 2023a)	89.60	96.10	96.10	58.00	78.40	92.80	66.50	75.00	81.56
PatchCraft (Zhong et al., 2023)	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.30
AIDE (Yan et al., 2025)	79.38	<u>99.74</u>	<u>99.76</u>	78.54	91.82	98.65	80.26	66.89	86.88
Ivy-Det	92.37	99.20	99.23	99.35	99.33	99.08	99.20	99.16	98.36
Ivy-xDet	<u>90.26</u>	98.28	98.09	98.58	<u>98.38</u>	98.20	<u>98.29</u>	<u>98.23</u>	<u>97.29</u>

or partially answered responses will receive lower scores. (2) Relevance: Measure how closely the content relates to the original annotation; (3) Level of Detail: Assess whether the response includes enough examples or elaborations; (4) Explanation: Verify the accuracy and consistency of explanations for any causes mentioned. Each response is scored using GPT-4o mini (Achiam et al., 2023) under a unified evaluation prompt (see Appendix 2), which instructs the model to act as an impartial judge and assign a score from 1 to 5. To ensure consistency and reduce variance, each response is rated over five independent rounds, and the final score is obtained by averaging these outputs.

5.1 Image Content Classification

We evaluated our image AIGC detector on the GenImage (Zhu et al., 2023b) and Chameleon (Yan et al., 2025) benchmarks. GenImage comprises seven subsets generated by leading models, i.e., Midjourney, Stable Diffusion v1.4 & v1.5, ADM, GLIDE, Wukong, VQDM, and BigGAN. We compared against five state-of-the-art detectors, which are CCNSpot (Wang et al., 2020), F3Net (Qian et al., 2020), DIRE (Wang et al., 2023), GenDet (Zhu et al., 2023a), PatchCraft (Zhong et al., 2023), and AIDE (Yan et al., 2025). As Table 2 shows, our method raises mean accuracy from 86.88 % to 98.36 %. For the subsets like BigGAN, we can even improve 32.27%, which shows the superiority of our new benchmark.

On Chameleon, we compared with ten detection methods, i.e., CCNSpot, F3Net, DIRE, GenDet, PatchCraft, and AIDE. Following work (Yan et al., 2025), both the overall accuracy and the separate accuracy on synthetic content (fake) and real data (real) are reported. As shown in the Table 3, our methods (last two columns), at least improved 20% when compared with any of previous best methods. This further demonstrate the superiority of our method on the image level AGIC detection.

Table 3: Comparison on the Chameleon (Yan et al., 2025). Accuracy (%) of different detectors (rows) in detecting real and fake images. For each training dataset, the first row is overall accuracy, the second row is “fake/real” class accuracy.

CNNSpot	FreDect	Fusing	GramNet	LNP	UnivFD	DIRE	PatchCraft	NPR	AIDE	Ivy-Det	Ivy-xDet
60.89 9.86/99.25	57.22 0.89/99.55	57.09 0.02/99.98	60.95 4.76/99.66	58.52 7.72/96.70	60.42 85.52/41.56	59.71 11.86/95.67	56.32 3.07/96.35	58.13 2.43/100.00	65.77 26.80/95.06	85.20 76.21/95.61	83.39 74.49/93.71

5.2 Video Content Classification

We further evaluate our unified model on video-level AIGC detection using the GenVideo dataset (Chen et al., 2024a), the largest benchmark for generated video detection, in a many-to-many generalization setting. Without modification, we apply the same architecture used for image analysis and compare against four state-of-the-art methods: F3Net (Qian et al., 2020), NPR (Tan et al., 2024), STIL (Gu et al., 2021), and DeMamba-XCLIP-FT (Chen et al., 2024a). Performance is reported in terms of recall (R), F1 score (F1), and average precision (AP).

Figure 4 shows one case for the video detection. Our model progressively conducts the spatial and temporal analysis to better capture generation content. As Table 4 shows, our model outperforms all baselines, achieving over 99 % accuracy across most generative sources. Notably, on the most challenging “HotShot” subset, we achieve a recall of 99.57 %, versus 65.43 % for the previous best method, underscoring our approach’s superior effectiveness in video-level AIGC detection.

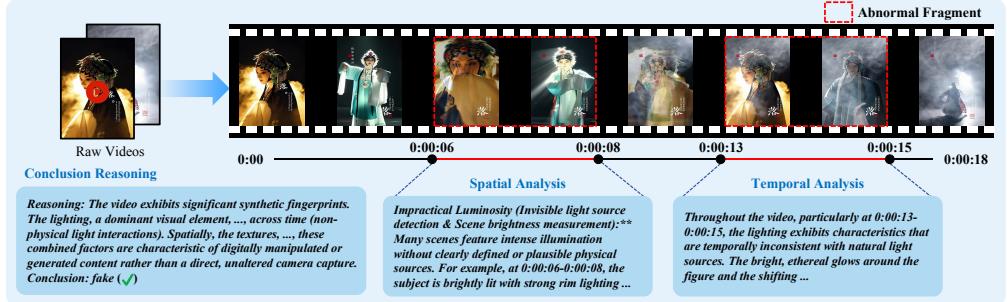


Figure 4: Showcase of the Video artifact detection.

Table 4: Comparisons to the **GenVideo**. F1 score (F1), recall score (R) and average precision (AP) on the many-to-many generalization task. “Demamba-XCLIP-FT” is abbreviated as “Demamba”.

Model	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Avg.
F3Net (Image)	R	0.8393	0.9971	0.9862	0.7757	0.5700	0.3657	0.9952	0.9971	0.8943	0.7678	0.8188
	F1	0.5000	0.9406	0.9628	0.8169	0.6988	0.4904	0.9332	0.9688	0.8873	0.8251	0.8024
	AP	0.6827	0.9989	0.9967	0.8935	0.8524	0.6317	0.9958	0.9989	0.9380	0.8841	0.8873
NPR (Image)	R	0.9107	0.9957	0.9949	0.2429	0.8964	0.5771	0.9712	0.9986	0.9429	0.8780	0.8408
	F1	0.2786	0.8441	0.9131	0.3028	0.8627	0.5944	0.8170	0.9164	0.8184	0.8163	0.7164
	AP	0.6717	0.9914	0.9920	0.2276	0.9391	0.6176	0.9633	0.9972	0.9415	0.9040	0.8245
STIL (Video)	R	0.7857	0.9814	0.9804	0.7600	0.6179	0.5329	0.9936	0.9736	0.9457	0.6501	0.8222
	F1	0.3805	0.9068	0.9458	0.7824	0.7232	0.6217	0.9039	0.9433	0.8884	0.7257	0.7823
	AP	0.5721	0.9908	0.9932	0.8619	0.8224	0.7043	0.9925	0.9896	0.9718	0.8132	0.8712
DeMamba (Video)	R	0.9812	1.0000	0.9986	0.6543	0.9486	0.9886	1.0000	1.0000	0.9286	0.8909	0.9302
	F1	0.6407	0.9602	0.9790	0.7539	0.9537	0.9551	0.9557	0.9797	0.9240	0.9120	0.9020
	AP	0.9332	1.0000	0.9997	0.8555	0.9897	0.9960	0.9998	1.0000	0.9777	0.9575	0.9710
Ivy-Det	R	1.0000	0.9986	0.9993	0.9957	0.9943	0.9943	1.0000	1.0000	0.9986	0.9601	0.9948
	F1	1.0000	0.9993	0.9996	0.9979	0.9971	0.9971	1.0000	1.0000	0.9240	0.9796	0.9974
	AP	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Ivy-xDet	R	0.9821	1.0000	0.9935	0.9771	0.9943	0.9900	1.0000	0.9993	0.9871	0.8690	0.9839
	F1	0.9910	1.0000	0.9967	0.9884	0.9971	0.9950	1.0000	0.9996	0.9935	0.9299	0.9919
	AP	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

5.3 Image and Video generation content Reasoning

This section evaluates MLLMs’ reasoning over image- and video-generated content, in addition to their detection performance across both modalities. We benchmark our model against four leading LMMs on the full IVY-FAKE: two open-source (Qwen2.5-7B (Bai et al., 2025) and InternVL2.5-8B (Chen et al., 2024b,c)) and two proprietary (GPT-4V (Achiam et al., 2023) and Gemini 2.5 Pro (Team et al., 2023)). Detailed reasoning results are provided in the Appendix. Our approach not only achieves superior accuracy but also delivers more transparent explanations than any baseline. Notably, unlike methods such as LOKI (Ye et al., 2025), which express interpretability via artifact coordinates or heatmaps, our model generates natural-language descriptions of visual artifacts that are immediately accessible to human readers.

6 Conclusion

We introduced IVY-FAKE, the first unified, large-scale dataset for explainable AIGC detection across both images and videos, featuring over 150 000 richly annotated training samples and 18 700 evaluation examples with natural-language reasoning, and proposed Ivy Explainable Detector (IVY-XDETECTOR), a vision-language architecture that jointly detects and explains synthetic content. Our model sets new state-of-the-art benchmarks in AIGC detection and explainability, and our publicly released resources provide a robust foundation for transparent, trustworthy multimodal analysis.

Limitations: Future work should optimize spatial modeling efficiency and strengthen temporal consistency. Currently, the high spatial token load (729 tokens) forces aggressive temporal downsampling, which can degrade temporal coherence and reduce accuracy in detecting subtle temporal artifacts.

Broader impacts: The data and model can be helped with the detection of fake visual content, help to uncover fact. Potentially, this can also be used to train stronger generative model to synthesize more realistic visual content.

References

- Achiam Josh, Adler Steven, Agarwal Sandhini, Ahmad Lama, Akkaya Ilge, Aleman Florencia Leoni, Almeida Diogo, Altenschmidt Janko, Altman Sam, Anadkat Shyamal, others . Gpt-4 technical report // arXiv preprint arXiv:2303.08774. 2023.*
- Bai Jianfa, Lin Man, Cao Gang, Lou Zijie. AI-Generated Video Detection via Spatial-Temporal Anomaly Learning // Chinese Conference on Pattern Recognition and Computer Vision (PRCV). 2024. 460–470.*
- Bai Shuai, Chen Keqin, Liu Xuejing, Wang Jialin, Ge Wenbin, Song Sibo, Dang Kai, Wang Peng, Wang Shijie, Tang Jun, others . Qwen2.5-VL Technical Report // arXiv:2502.13923. 2025.*
- Betker James, Goh Gabriel, Jing Li, Brooks Tim, Wang Jianfeng, Li Linjie, Ouyang Long, Zhuang Juntang, Lee Joyce, Guo Yufei, others . Improving image generation with better captions // Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf. 2023. 2, 3, 8.*
- Bharadwaj Rohit, Gani Hanan, Naseer Muzammal, Khan Fahad Shahbaz, Khan Salman. VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs. 2024.*
- Brock Andrew, Donahue Jeff, Simonyan Karen. Large scale GAN training for high fidelity natural image synthesis // arXiv preprint arXiv:1809.11096. 2018.*
- Brooks Tim, Peebles Bill, Holmes Connor, DePue Will, Guo Yufei, Jing Li, Schnurr David, Taylor Joe, Luhman Troy, Luhman Eric, others . Video generation models as world simulators // OpenAI Blog. 2024. 1, 8.*
- Cao Bin, Yuan Jianhao, Liu Yexin, Li Jian, Sun Shuyang, Liu Jing, Zhao Bo. Synartifact: Classifying and alleviating artifacts in synthetic images via vision-language model // arXiv preprint arXiv:2402.18068. 2024.*
- Chen Haoxing, Hong Yan, Huang Zizheng, Xu Zhuoer, Gu Zhangxuan, Li Yaohui, Lan Jun, Zhu Huijia, Zhang Jianfu, Wang Weiqiang, Li Huaxiong. DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark // arXiv preprint arXiv:2405.19707. 2024a.*
- Chen Zhe, Wang Weiyun, Tian Hao, Ye Shenglong, Gao Zhangwei, Cui Erfei, Tong Wenwen, Hu Kongzhi, Luo Jiapeng, Ma Zheng, others . How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites // arXiv:2404.16821. 2024b.*
- Chen Zhe, Wu Jiannan, Wang Wenhui, Su Weijie, Chen Guo, Xing Sen, Zhong Muyan, Zhang Qinglong, Zhu Xizhou, Lu Lewei, others . Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024c. 24185–24198.*
- Deng Jingyi, Lin Chenhao, Zhao Zhengyu, Liu Shuai, Wang Qian, Shen Chao. A survey of defenses against ai-generated visual media: Detection, disruption, and authentication // arXiv preprint arXiv:2407.10575. 2024.*
- Dhariwal Prafulla, Nichol Alexander. Diffusion models beat gans on image synthesis // Advances in neural information processing systems. 2021. 34. 8780–8794.*
- Dong Shichao, Wang Jin, Liang Jiajun, Fan Haoqiang, Ji Renhe. Explaining deepfake detection by analysing image matching // European conference on computer vision. 2022. 18–35.*
- Goodfellow Ian J, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, Bengio Yoshua. Generative adversarial nets // Advances in neural information processing systems. 2014. 27.*

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

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

Guo Yanan, Dong Wenhui, Song Jun, Zhu Shiding, Zhang Xuan, Yang Hanqing, Wang Yingbo, Du Yang, Chen Xianing, Zheng Bo. FiLA-Video: Spatio-Temporal Compression for Fine-Grained Long Video Understanding // arXiv preprint arXiv:2504.20384. 2025b.

Hertz Amir, Mokady Ron, Tenenbaum Jay, Aberman Kfir, Pritch Yael, Cohen-Or Daniel. Prompt-to-prompt image editing with cross attention control // arXiv preprint arXiv:2208.01626. 2022.

Ho Jonathan, Jain Ajay, Abbeel Pieter. Denoising diffusion probabilistic models // Advances in neural information processing systems. 2020. 33. 6840–6851.

Hong Yan, Feng Jianming, Chen Haoxing, Lan Jun, Zhu Huijia, Wang Weiqiang, Zhang Jianfu. WildFake: A Large-Scale and Hierarchical Dataset for AI-Generated Images Detection // Proceedings of the AAAI Conference on Artificial Intelligence. Apr. 2025. 39, 4. 3500–3508.

Keita Mamadou, Hamidouche Wassim, Bougueffa Eutamene Hessen, Taleb-Ahmed Abdelmalik, Camacho David, Hadid Abdenour. Bi-LORA: A Vision-Language Approach for Synthetic Image Detection // Expert Systems. 2025. 42, 2. e13829.

Li Bo, Zhang Yuanhan, Guo Dong, Zhang Renrui, Li Feng, Zhang Hao, Zhang Kaichen, Zhang Peiyuan, Li Yanwei, Liu Ziwei, others . Llava-onevision: Easy visual task transfer // arXiv preprint arXiv:2408.03326. 2024a.

Li Xinhao, Wang Yi, Yu Jiahuo, Zeng Xiangyu, Zhu Yuhan, Huang Haian, Gao Jianfei, Li Kunchang, He Yinan, Wang Chenting, others . Videochat-flash: Hierarchical compression for long-context video modeling // arXiv preprint arXiv:2501.00574. 2024b.

Li Yixuan, Liu Xuelin, Wang Xiaoyang, Lee Bu Sung, Wang Shiqi, Rocha Anderson, Lin Weisi. FakeBench: Probing Explainable Fake Image Detection via Large Multimodal Models // arXiv preprint arXiv:2404.13306. 2024c.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries // Text summarization branches out. 2004. 74–81.

Liu Haotian, Li Chunyuan, Wu Qingyang, Lee Yong Jae. Visual instruction tuning // Advances in neural information processing systems. 2023. 36. 34892–34916.

Nichol Alex, Dhariwal Prafulla, Ramesh Aditya, Shyam Pranav, Mishkin Pamela, McGrew Bob, Sutskever Ilya, Chen Mark. Glide: Towards photorealistic image generation and editing with text-guided diffusion models // arXiv preprint arXiv:2112.10741. 2021.

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

Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer Björn. High-resolution image synthesis with latent diffusion models // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022a. 10684–10695.

Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer Björn. High-resolution image synthesis with latent diffusion models // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022b. 10684–10695.

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

Saharia Chitwan, Chan William, Saxena Saurabh, Li Lala, Whang Jay, Denton Emily L, Ghasemipour Kamyar, Gontijo Lopes Raphael, Karagol Ayan Burcu, Salimans Tim, others. Photorealistic text-to-image diffusion models with deep language understanding // Advances in neural information processing systems. 2022. 35. 36479–36494.

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

Team Gemini, Anil Rohan, Borgeaud Sebastian, Alayrac Jean-Baptiste, Yu Jiahui, Soricut Radu, Schalkwyk Johan, Dai Andrew M, Hauth Anja, Millican Katie, others. Gemini: a family of highly capable multimodal models // arXiv preprint arXiv:2312.11805. 2023.

Wang Sheng-Yu, Wang Oliver, Zhang Richard, Owens Andrew, Efros Alexei A. CNN-generated images are surprisingly easy to spot... for now // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. 8695–8704.

Wang Zhendong, Bao Jianmin, Zhou Wengang, Wang Weilun, Hu Hezhen, Chen Hong, Li Houqiang. Dire for diffusion-generated image detection // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. 22445–22455.

Wen Siwei, Ye Junyan, Feng Peilin, Kang Hengrui, Wen Zichen, Chen Yize, Wu Jiang, Wu Wenjun, He Conghui, Li Weijia. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation // arXiv preprint arXiv:2503.14905. 2025.

Yan Shilin, Li Ouxiang, Cai Jiayin, Hao Yanbin, Jiang Xiaolong, Hu Yao, Xie Weidi. A Sanity Check for AI-generated Image Detection // The Thirteenth International Conference on Learning Representations. 2025.

Ye Junyan, Zhou Baichuan, Huang Zilong, Zhang Junan, Bai Tianyi, Kang Hengrui, He Jun, Lin Honglin, Wang Zihao, Wu Tong, others. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models // ICLR. 2025.

Zhai Xiaohua, Mustafa Basil, Kolesnikov Alexander, Beyer Lucas. Sigmoid loss for language image pre-training // Proceedings of the IEEE/CVF international conference on computer vision. 2023. 11975–11986.

Zhang Boqiang, Li Kehan, Cheng Zesen, Hu Zhiqiang, Yuan Yuqian, Chen Guanzheng, Leng Sicong, Jiang Yuming, Zhang Hang, Li Xin, others. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding // arXiv preprint arXiv:2501.13106. 2025.

Zhang Ivy, Peng Wei, N Jenny, Yu Theresa, Qiu David. Ivy-VL:Compact Vision-Language Models Achieving SOTA with Optimal Data. December 2024.

Zhang Lingzhi, Xu Zhengjie, Barnes Connelly, Zhou Yuqian, Liu Qing, Zhang He, Amirghodsi Sohrab, Lin Zhe, Shechtman Eli, Shi Jianbo. Perceptual artifacts localization for image synthesis tasks // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. 7579–7590.

Zheng Bo, Song Jun, Dong Wenhui, Wang Yingbo, Zhu Shiding, Guo Yanan. HyViLM: Enhancing Fine-Grained Recognition with a Hybrid Encoder for Vision-Language Models. 2024.

Zheng Dehua, Dong Wenhui, Hu Hailin, Chen Xinghao, Wang Yunhe. Less is More: Focus Attention for Efficient DETR // ArXiv. 2023a. abs/2307.12612.

Zheng Lianmin, Chiang Wei-Lin, Sheng Ying, Zhuang Siyuan, Wu Zhanghao, Zhuang Yonghao, Lin Zi, Li Zhuohan, Li Dacheng, Xing Eric, others. Judging llm-as-a-judge with mt-bench and chatbot arena // Advances in Neural Information Processing Systems. 2023b. 36. 46595–46623.

Zhong Nan, Xu Yiran, Li Sheng, Qian Zhenxing, Zhang Xinpeng. Patchcraft: Exploring texture patch for efficient ai-generated image detection // arXiv preprint arXiv:2311.12397. 2023.

Zhu Jun-Yan, Park Taesung, Isola Phillip, Efros Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks // Proceedings of the IEEE international conference on computer vision. 2017. 2223–2232.

Zhu Mingjian, Chen Hanting, Huang Mouxiao, Li Wei, Hu Hailin, Hu Jie, Wang Yunhe. Gendet: Towards good generalizations for ai-generated image detection // arXiv preprint arXiv:2312.08880. 2023a.

Zhu Mingjian, Chen Hanting, Yan Qiangyu, Huang Xudong, Lin Guanyu, Li Wei, Tu Zhijun, Hu Hailin, Hu Jie, Wang Yunhe. GenImage: a million-scale benchmark for detecting AI-generated image // Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023b. (NIPS '23).

A Performance Evaluation of Explainability Datasets

We conduct experiments to evaluate the explanation capabilities of various models. Specifically, we assess the similarity between model-generated reasoning and reference annotations using the ROUGE-L score (Lin, 2004), which measures the longest common subsequence and reflects token-level overlap. Additionally, we adopt the LLM-as-a-judge paradigm (Zheng et al., 2023b), evaluating responses across four dimensions: Completeness, Relevance, Level of Detail, and Explanation Quality. Each response is scored using GPT-4o mini (Achiam et al., 2023) under a unified prompt that instructs the model to act as an impartial judge, assigning scores from 1 to 5. To ensure reliability, each response is rated across five independent rounds, with the final score computed as the average.

Model	Image								Video							
	Auto Metrics				GPT Assisted				Auto Metrics				GPT Assisted			
	Acc	F1	ROUGE-L	SIM	Com./Rel./Det./Exp.	AVG	Acc	F1	ROUGE-L	SIM	Com./Rel./Det./Exp.	AVG	Com./Rel./Det./Exp.	AVG	Com./Rel./Det./Exp.	AVG
<i>Closed-source MLLMs</i>																
GPT-4o	0.766	0.759	0.155	0.683	3.69/3.79/3.67/3.80	3.74	0.828	0.813	0.150	0.682	4.02/4.13/3.99/4.12	4.07				
Gemini 2.5 flash	0.668	0.656	0.223	0.709	3.29/3.33/3.27/3.33	3.30	0.787	0.784	0.188	0.678	3.85/3.93/3.82/3.92	3.88				
GPT4o-mini	0.653	0.650	0.149	0.645	3.21/3.26/3.21/3.26	3.23	0.685	0.672	0.134	0.645	3.38/3.42/3.37/3.42	3.40				
<i>Open-source MLLMs</i>																
InternVL3-8B	0.614	0.605	0.159	0.680	3.04/3.07/3.04/3.07	3.05	0.632	0.616	0.165	0.629	3.13/3.16/3.12/3.16	3.14				
Qwen2.5-VL-7B	0.556	0.516	0.199	0.660	2.73/2.77/2.73/2.77	2.75	0.589	0.527	0.207	0.621	2.89/2.94/2.89/2.89	2.91				
Phi-3.5-Vision	0.560	0.555	0.092	0.366	2.66/2.72/2.66/2.72	2.69	0.559	0.479	0.001	0.046	2.78/2.79/2.78/2.79	2.79				
Ivy-xDet	0.805	0.802	0.271	0.767	4.39/4.21/4.33/4.54	4.40	0.945	0.945	0.303	0.776	3.76/4.00/3.71/3.97	3.86				

Table 5: Performance comparison of models on image and video tasks. “Auto Metrics” include Acc, F1, ROUGE-L, and SIM scores. “GPT Assisted” includes four subjective criteria: Completeness, Relevance, Detail, and Explanation, as well as their average.

As shown in Table 5, we compare our model with several leading multimodal large language models (LMMs), including Qwen2.5-7B (Bai et al., 2025), InternVL2.5-8B (Chen et al., 2024b,c), Phi-3.5-Vision, GPT-4V (Achiam et al., 2023), and Gemini 2.5 (Team et al., 2023). The results demonstrate that our approach not only achieves higher accuracy but also provides more transparent and informative explanations than all baseline models.

B Video Understanding Models and Evaluation on General Benchmarks

We further evaluate the proposed model on several video understanding benchmarks to assess its generalization capability. Specifically, we compare it against five lightweight, general-purpose video understanding models, which are VideoLLaMA3, Qwen2-VL 2B, Qwen2.5-VL-3B, InternVL2.5-2B, and InternVL3-2B, across four benchmarks, i.e., MLVU (dev), PerceptionTest, LongVideo, and VideoMME.

As shown in Table 6, our model, Ivy-Video-3B, consistently outperforms all competing methods across these benchmarks. These results highlight the strong generalization ability of our model, which, although designed for AGIC detection, achieves high accuracy across a diverse set of general-purpose video understanding tasks.

Task	VideoLLaMA3	Qwen2-VL 2B	Qwen2.5-VL-3B	InternVL2.5-2B	InternVL3-2B	Ivy-Video-3B
MLVU (dev)	65.4	62.7	68.2	58.9	64.2	68.87
PerceptionTest	68	53.9	66.9	66.3	-	72.73
LongVideo	54.3	44.7	50.2	48.7	51.2	55.25
VideoMME	59.6	55.6	61.5	51.9	58.9	62.3

Table 6: Transposed performance table of various video understanding models on general video QA benchmarks.

Image Authenticity Analyst Assistant (User Prompt)

Is this image real or fake? Provide the reasoning process, then give the final conclusion.

Figure 5: User Prompt Template for Image Data Distillation

Video Authenticity Analyst Assistant (User Prompt)

Is this video real or fake? Provide the reasoning process, then give the final conclusion.

Figure 6: User Prompt Template for Video Data Distillation

C Effect of Incorporating Human-Annotated Labels via `gemini 2.5 pro` on Accuracy

To assess the impact of human-annotated labels on model performance, we compare the accuracy of final conclusion predictions under two settings: (i) with labels incorporated via the `gemini 2.5 pro`, and (ii) without labels. The evaluation was conducted on about 1,000 examples from the test set.

Annotation Setting	Accuracy (Acc)
With Label	1.000
Without Label	0.785

Table 7: Accuracy of conclusion prediction with and without incorporating labels.

As shown in Table 7, incorporating ground-truth labels results in a substantial performance gain, yielding perfect accuracy (1.000), compared to 0.785 without labels. The drastic performance gap suggests potential limitations in label-free or weakly supervised setups when applied to tasks requiring fine-grained semantic understanding.

D Prompts

Here we provide the prompts that are mainly used in this study. As illustrated by the following figures, there are five distillation prompts distillation we used in this paper that mainly can be divided into the following three parts:

1. Prompt Template for Image Data Distillation: Since image data consists of a single frame, it can be treated as a static instance. Therefore, AIGC detection mainly focuses on identifying spatial anomalies. Detail prompt can be found in Figure 5 and 7.

2. Prompt Template for Video Data Distillation: Compared to images, video inputs provide continuous multi-frame context. This allows for detection along both spatial and temporal anomaly dimensions. Detail prompt can be found in 6 and Figure 8.

3. GPT Assisted Evaluation Prompt: To assess the quality of model outputs, we design a GPT-based evaluator prompt that scores responses across four dimensions: Completeness, Relevance, Level of Detail, and Explanation. The evaluator receives a structured pair of `GroundTruth` and `ModelOutput`, each containing a `<think>` section (reasoning) and a `<conclusion>` (final judgment). The model must return a structured JSON object with integer scores (1–5) for each dimension. The prompt is provided in Figure 9.

E Case Study: Qualitative Comparison of Methods

According to Figures 10, 11, 12, and 13, **OurMethod** consistently demonstrates superior performance in detecting both spatial and temporal anomalies, providing stronger generalization and robustness compared to existing baselines.

Image Authenticity Analyst Assistant (System Prompt)

```
## Role
Expert AI system for detecting image by analyzing visual anomalies across spatial plausibility.

## Analysis Dimensions
### Spatial Features: static anomaly detection
- Impractical Luminosity
  - Scene brightness measurement
  - Invisible light source detection (physical validation)
- Localized Blur
  - Focus distribution mapping (sharpness gradient)
  - Artificial depth-of-field identification (algorithmic artifacts)
- Illegible Letters
  - OCR text extraction
  - Character structural integrity (stroke continuity)
- Distorted Components
  - Anatomical/proportional accuracy (biological/object logic)
  - Physics compliance (material/gravity validation)
- Omitted Components
  - Object completeness check (edge/detail absence)
  - Partial rendering artifact detection (AI-generated traces)
- Spatial Relationships
  - Contextual object placement (scene plausibility)
  - Perspective consistency (geometric projection)
- Chromatic Irregularity
  - Color database comparison (natural distribution)
  - Unnatural hue detection (oversaturation/abrupt gradients)
- Abnormal Texture
  - Surface pattern regularity (texture repetition)
  - Material property coherence (reflectance/roughness validation)

## Reasoning Step
1. Spatial Analysis
  - Analyze static features (e.g., lighting, text, objects)
2. Conclusion: Only real or fake.
  - real: Contains verifiable capture device signatures and natural physical imperfections.
  - fake: Exhibits synthetic fingerprints including but not limited to over-regularized textures and non-physical light interactions.
```

The assistant first thinks about the reasoning step in the mind and then provides the user with the reason. The reasoning step and conclusion are enclosed within `<think>` `</think>` and `<conclusion>` `</conclusion>` tags, respectively, i.e., `<think>` reasoning step here `</think>` `<conclusion>` real or fake `</conclusion>`. `<conclusion>` content must strictly align with the user-provided authenticity label (real/fake) in both value and semantic context.

Figure 7: System Prompt Template for Image Data Distillation

Video Authenticity Analyst Assistant (System Prompt)

```
## Role
Expert AI system for detecting videos by analyzing visual anomalies across temporal coherence (inter-frame dynamics) and spatial plausibility (intra-frame logic).

## Analysis Dimensions
### 1. Temporal Features: Multi-frame dynamic anomaly detection - Luminance Discrepancy
- Shadow direction consistency (cross-frame comparison)
- Light source coordination (temporal validation)
- Awkward Facial Expression
- Facial muscle motion continuity (expression dynamics)
- Emotion-context alignment (temporal coherence)
- Duplicated Components
- Repeating element pattern recognition (cross-frame tracking)
- Natural variation analysis (sequence validation)
- Non-Spatial Relationships
- Object interaction physics (motion trajectory validation)
- Fusion/penetration anomalies (temporal detection)

### 2. Spatial Features: Single-frame static anomaly detection
- Impractical Luminosity
- Scene brightness measurement (single-frame analysis)
- Invisible light source detection (physical validation)
- Localized Blur
- Focus distribution mapping (sharpness gradient)
- Artificial depth-of-field identification (algorithmic artifacts)
- Illegible Letters
- OCR text extraction (single-frame recognition)
- Character structural integrity (stroke continuity)
- Distorted Components
- Anatomical/proportional accuracy (biological/object logic)
- Physics compliance (material/gravity validation)
- Omitted Components
- Object completeness check (edge/detail absence)
- Partial rendering artifact detection (AI-generated traces)
- Spatial Relationships
- Contextual object placement (scene plausibility)
- Perspective consistency (geometric projection)
- Chromatic Irregularity
- Color database comparison (natural distribution)
- Unnatural hue detection (oversaturation/abrupt gradients)
- Abnormal Texture
- Surface pattern regularity (texture repetition)
- Material property coherence (reflectance/roughness validation)

## Reasoning Step
1. Temporal Analysis
- Track dynamic features across frames (e.g., shadows, expressions)
2. Spatial Analysis
- Analyze static features per frame (e.g., lighting, text, objects)
3. Conclusion: Only real or fake.
- real: Contains verifiable capture device signatures and natural physical imperfections.
- fake: Exhibits synthetic fingerprints including but not limited to over-regularized textures and non-physical light interactions.
```

The assistant first thinks about the reasoning step in the mind and then provides the user with the reason. The reasoning step and conclusion are enclosed within <think> </think> and <conclusion> </conclusion> tags, respectively, i.e., <think> reasoning step here </think> <conclusion> real or fake </conclusion>. <conclusion> content must strictly align with the user-provided authenticity label (real/fake) in both value and semantic context.

Figure 8: System Prompt Template for Video Data Distillation

GPT Assisted Evaluation Prompt

Role

You are an impartial evaluator. Your task is to assess whether a model-generated response accurately and coherently matches a human-annotated reference answer.

Each input contains two structured components:

- <think>: the reasoning or analytical explanation
- <conclusion>: the final judgment (e.g., real or fake)

Evaluation Dimensions

You should compare the **ModelOutput** to the **GroundTruth**, and assign integer scores from 1 to 5 (no decimals) for the following four dimensions:

1. Completeness

- Does the ModelOutput address all aspects covered in the GroundTruth?
- More complete responses should include all relevant information, especially key golden clues.
- Incomplete or partially aligned answers should receive lower scores.

2. Relevance

- Does the ModelOutput discuss the same detection dimensions as in the GroundTruth?
- Temporal features include:
 - Luminance discrepancy
 - Duplicated components
 - Awkward facial expressions
 - Motion inconsistency
- Spatial features include:
 - Abnormal texture
 - Distorted or omitted components
 - Chromatic irregularity
 - Impractical luminosity
 - Localized blur, etc.
- Penalize if irrelevant aspects are introduced or relevant ones are missing.

3. Level of Detail

- Does the ModelOutput describe fine-grained visual cues in each dimension?
- High scores require specific subcomponent elaboration, not just general terms.
- Penalize vague or generic responses that lack specific observations.

4. Explanation

- Is the reasoning in <think> logically consistent with the <conclusion>?
- The explanation should provide clear, causally-linked justifications.
- Penalize if the conclusion contradicts the reasoning or lacks support.

Figure 9: GPT Assisted Evaluation Prompt

	    		Conclusion Validity
Ivy-xDetector	<think>1. **Spatial Analysis**: The lighting ... expected from real materials. The image is uniformly sharp ... seem generally correct, but the perfection of forms (e.g., identical wine bottles, perfectly aligned shelves) leans towards artificiality ... appear geometrically sound and plausible for an interior design concept ... lacking the unique variations found in natural wood. 2. **Conclusion**: The image exhibits synthetic fingerprints ... lacking the imperfections and variations inherent in real-world photography.</think>	Conclusion fake ✓	
GPT-4o	<think>Upon analyzing the image of the interior space: 1. **Illumination Inconsistency**: The lighting appears consistent ... light directions are evident . 2. **Unnatural Content Placement**: Furniture placement ... dining table, and seating. 3. **Unrealistic Physical Properties**: Material properties ... object interactions or physical impossibilities. 4. **Digital Manipulation Signatures**: Edges are ... with the intended function. **Overall**: there are no signs of anomalies ... appears to capture a real . </think>	Conclusion real ✗	
Gemini Gemini-2.5	<think>The image displays consistent lighting and natural shadows ... without signs of partial rendering. The spatial relationships between objects and the perspective are consistent with ... indicative of repetitive patterns found in synthetic images . Based on the absence ... generated imagery, the image appears to be a real photograph.</think>	Conclusion real ✗	
InternVL InternVL3-8B	<think>The image appears to be a modern, well-lit room with a sleek design. **Impractical Luminosity**: The lighting seems evenly distributed, but there's no visible light source ... have concealed lighting. **Localized Blur**: The image is sharp throughout, with no noticeable blur, suggesting high-quality rendering. ... **Abnormal Texture**: The textures ... seems to have high-quality, realistic features without obvious anomalies .</think>	Conclusion real ✗	
Qwen2.5-VL-7B	<think>Let's analyze the image step by step. **Impractical Luminosity**: The lighting appears consistent and well-distributed across the room ... **Spatial Relationships**: - The placement of objects like the wine rack, table, chairs, and sofa is logical and follows typical spatial relationships found in a living space ... that would suggest the image is real . The image appears to be a real photograph of a well-designed interior space.</think>	Conclusion real ✗	

Figure 10: **Image** example 1 where **Ivy-xDetector** successfully detects subtle spatial anomalies missed by baselines.

	    		Conclusion Validity
Ivy-xDetector	<think>1. **Spatial Analysis**: ... the transition between the sharp foreground (tiger) and the blurred background isn't perfectly natural , particularly around the fur edges. ... **Spatial Relationships**: ... The texture of the tiger's fur, especially on the back and shoulder, appears overly smooth and lacks the fine, distinct detail expected in high-resolution photography. Similarly, the snow texture looks somewhat uniform and lacks the complex, random variations of real snow. ... 2. **Conclusion**: The image exhibits synthetic fingerprints , which lack natural complexity and detail , appearing overly smooth and regularized.	Conclusion fake ✓	
GPT-4o	<think>1. **Illumination Inconsistency**: The lighting appears coherent across the tiger and the snow. Shadows align properly with a consistent light source direction. 2. **Facial Feature Anomalies**: The tiger's features ... appear natural and well-integrated 4. **Unrealistic Physical Properties**: The interaction between the tiger's paws and the snow appears convincing with visible snow compression. ... 6. **Semiotic Plausibility**: The scene is semantically coherent —a tiger in a snowy environment reflects natural habitat possibilities. ... the image does not exhibit significant anomalies associated with synthetic generation or manipulations.</think>	Conclusion real ✗	
Gemini Gemini-2.5	<think>1. **Spatial Analysis**: The lighting seems consistent with natural light sources (sunlight) casting shadows and highlights on the tiger and snow. No abrupt or unnatural blur regions are detected. ... **Omitted Components**: The tiger is depicted completely within the image. No signs of missing details or partial rendering artifacts are apparent. ... **Spatial Relationships**: The tiger is situated in a snowy environment, which is a plausible habitat for some tiger subspecies. ... **Chromatic Irregularity**: The colors of the tiger's fur, eyes, mouth, and the surrounding snow and background exhibit natural hues ... No unnatural saturation or gradients are detected. **Abnormal Texture**: The textures of the tiger's fur and the snow appear detailed and naturally irregular ... with a real photograph and shows no significant spatial anomalies that would indicate it is fake.</think>	Conclusion real ✗	
InternVL InternVL3-8B	<think>Let's analyze the spatial features: 1. **Impractical Luminosity**: The lighting appears natural , with no obvious signs of an invisible light source. ... 4. **Spatial Relationships**: The tiger's position in the snow and the background context seem plausible . 5. **Chromatic Irregularity**: The colors are natural , with no unnatural hues or saturation. 6. **Abnormal Texture**: The snow and tiger's fur textures appear realistic. Considering these factors, the image seems to have natural features without obvious synthetic artifacts.</think>	Conclusion real ✗	
Qwen2.5-VL-7B	<think>To determine if the image is real or fake, we need to analyze various spatial features: 1. **Impractical Luminosity**: There's no visible light source that would explain the brightness, suggesting it ... like a zoo or wildlife sanctuary. 2. **Localized Blur**: The tiger's fur and facial features appear sharp ... 6. **Spatial Relationships**: ...with no unusual perspective or positioning ... There are no unnatural hues or color gradients . 8. **Abnormal Texture**: The texture of the tiger's fur and the snow appears natural and consistent with real-life observations. ... It appears to be a real photograph of a tiger in a snowy environment.</think>	Conclusion real ✗	

Figure 11: **Image** example 2 illustrating improved robustness of **Ivy-xDetector** against visually deceptive artifacts.

	 	Conclusion Validity
 Ivy-xDetector	<think> 1. **Temporal Analysis**: ... the interaction between the person in the blue shirt and the beige chair. The trajectory ... by a person. The way ... without proper support or effort from the person suggests manipulation. 2. **Spatial Analysis**: While the overall scene ... the chair's motion overrides most spatial plausibility checks. No specific ... but the unnatural movement highlights the artificiality of the interaction. 3. **Conclusion**: The video exhibits synthetic fingerprints, ..., indicating manipulation.</think>	Conclusion fake 
 GPT-4o	<think> **Temporal Features**: sand and objects looks ... natural. **Spatial Features**: The chair exhibits ... no evidence of visual anomalies, ... properties ... for the environment. ...The physical interactions, ... with a genuine scenario.</think>	Conclusion real 
 Gemini Gemini-2.5	<think> The video shows ... spatial anomalies. **Temporal Analysis**: The shadows and lighting seem consistent across the frames as people move and interact with the chair ... follow expected physics. **Spatial Analysis**: Lighting is consistent with a single, natural light source ... without over-regularization., the video appears consistent with a real recording.</think>	Conclusion real 
 InternVL InternVL3-8B	<think> Okay, let's analyze this video frame ... it's real or fake. First, I'll look ... no signs of awkwardness or discontinuity. Next, I'll examine the spatial features. The scene is set in ... no obvious signs of unnatural blur, chromatic irregularities, or distorted components that would indicate digital manipulation. Considering both temporal and spatial analysis, the video doesn't exhibit any clear signs of temporal or spatial anomalies. The actions, lighting, and object placements ... video appears to be authentic.</think>	Conclusion real 
 Qwen2.5-VL-7B	<think> ... spatial features. **Temporal Analysis**: The shadows cast by the chair and the person appear ... that the lighting is natural and not artificially manipulated. The interaction between the person and the chair ... is a logical action given the context. **Spatial Analysis**: The scene appears well-lit ... focus is sharp ... colors are natural and there are no unnatural hues. The sand and the chair looks realistic and there are no irregularities ... no signs of artificial manipulation or synthetic fingerprints.</think>	Conclusion real 

Figure 12: Video example 1 demonstrating that **Ivy-xDetector** effectively captures temporal inconsistencies overlooked by baseline models.

	 	Conclusion Validity
 Ivy-xDetector	<think> 1. **Temporal Analysis**: *Luminance Discrepancy*: The lighting appears consistent across the scene ... are not discernible for analysis. 2. **Spatial Analysis**: *Impractical Luminosity*: The overall brightness of the scene is plausible for a sunny day ... *Spatial Relationships*: The placement of buildings, people ... perspective seems generally consistent. *Abnormal Texture*: This is the most notable area. The texture of the stone walls, particularly on the right-hand structure ... lacking the intricate detail of real waves and ripples under sunlight. 3. **Conclusion**: The video exhibits synthetic fingerprints, primarily identified through spatial analysis. The textures, especially on the stone structures, lack the natural complexity and variation expected in real-world materials, which is characteristic of AI-generated content.</think>	Conclusion fake 
 GPT-4o	<think> To determine ... both temporal and spatial features: 1. **Temporal Analysis**: The movement of the water shows natural fluid dynamics ... people move naturally across the frames. The human motions (walking, gathering, posture changes) are smooth and align with physical plausibility. 2. **Spatial Analysis**: The architectural elements ... no signs of over-regularized patterns or artifacts are detected. The sea exhibits natural interaction with the rocks ... Details such as human attire, proportions, and gestures appear accurate and vary naturally. Both temporal coherence and spatial plausibility align with the characteristics of a real scene. No signs of AI-generation artifacts.</think>	Conclusion real 
 Gemini Gemini-2.5	<think> I will analyze both temporal and spatial features for potential anomalies. **Temporal Analysis**: People are moving naturally across the plaza and stairs. Their motion appears continuous and realistic frame-to-frame. ... No impractical luminosity is detected. The scene's focus seems uniform and natural; no localized blurring or artificial depth of field is apparent. ... Spatial relationships between elements (buildings relative to cliff and plaza, rocks in the sea) are logical and adhere to perspective rules. There are no detectable temporal coherence ... with synthetic media in these frames.</think>	Conclusion real 
 InternVL InternVL3-8B	<think> Upon analyzing the video, I observe the following key points: 1. **Temporal Features**: The shadows and lighting appear consistent across frames, indicating natural lighting conditions ... **Non-Spatial Relationships**: The objects and people move naturally, with no unnatural interactions. 2. **Spatial Features**: The scene lighting is consistent with natural daylight. ... **Spatial Relationships**: The placement of objects and people is contextually appropriate. **Abnormal Texture**: The textures of the buildings and natural elements look realistic. Given these observations, the video does not exhibit any obvious signs of manipulation or synthetic generation. The elements within the frames align with what one would expect in a real-world setting.</think>	Conclusion real 
 Qwen2.5-VL-7B	<think> To determine if the video is real or fake, we need to analyze it using the provided dimensions of temporal and spatial features. Temporal Analysis:... indicate a real video ... do not show any repeating patterns... The objects ... interact naturally ... ### Spatial Analysis: The scene appears well-lit with natural sunlight, which is consistent with an outdoor setting. There are no signs of artificial lighting that would be inconsistent with the environment. ... **Abnormal Texture**: The textures of the buildings and the sea appear natural and coherent ... that would suggest the video is fake. The video appears to be a genuine capture of a real location.</think>	Conclusion real 

Figure 13: Video example 2 showcasing **Ivy-xDetector**'s superior ability to detect subtle cross-frame temporal artifacts.