Ivy-Fake: A Unified Explainable Framework and Benchmark for Image and Video AIGC Detection Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

In this section, we present additional experiments to thoroughly evaluate the proposed approach, along with implementation details for the distillation prompts. The experiments specifically focus on the reasoning capabilities of the AGIC detection task. Rather than merely providing final real or fake predictions, the methods here also generate detailed explanations describing the rationale behind each prediction.

The evaluation is primarily conducted on the proposed explainability benchmark, the Ivy-Fake dataset. Additionally, we assess the generalization capability of our method on a broader video understanding dataset.

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., 2023), 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
Closed-source MLLMs												
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
Open-source MLLMs												
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	0.901	0.894	0.213	0.710	4.39/4.21/4.33/4.54	4.40	0.945	0.945	0.303	0.776	4.50/4.71/4.42/4.68	4.58

Table 1: 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: Comprehensiveness, Relevance, Detail, and Explanation, as well as their average.

As shown in Table 1, 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., 2024a,b), 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 2, 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 2: Transposed performance table of various video understanding models on general video QA benchmarks.

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 3: Accuracy of conclusion prediction with and without incorporating labels.

As shown in Table 3, 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 1 and 3.
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. etail prompt can be found in 2 and Figure 4.

Image Authenticity Analyst Assistant (User Prompt)
Is this Image real or fake? Provide the reasoning process, then give the final conclusion.

Figure 1: User Prompt Template for Image Data Distillation

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 5.

Video Authenticity Analyst Assistant (User Prompt)
Is this video real or fake? Provide the reasoning process, then give the final conclusion.

Figure 2: User Prompt Template for Video Data Distillation

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 3: 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 4: 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 g̈olden clues̈.
- 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 5: GPT Assisted Evaluation Prompt

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 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.

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. 2024a.

Chen Zhe, Wu Jiannan, Wang Wenhai, 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. 2024b. 24185–24198.

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

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.

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. 2023. 36. 46595–46623.