LOKI: A COMPREHENSIVE SYNTHETIC DATA DETECTION BENCHMARK USING LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

Figure 1: **Overview of LOKI benchmark.** LOKI possesses four key characteristics: 1) Diverse modalities (video, image, 3D, text and audio); 2) Heterogeneous categories (26 detailed subcategories); 3) Multi-level annotations; 4) Multimodal synthetic data evaluation framework.

ABSTRACT

With the rapid development of AI-generated content, the future internet may be inundated with synthetic data, making the discrimination of authentic and credible multimodal data increasingly challenging. Synthetic data detection has thus garnered widespread attention, and the performance of large multimodal models (LMMs) in this task has attracted significant interest. LMMs can provide natural language explanations for their authenticity judgments, enhancing the explainability of synthetic content detection. Simultaneously, the task of distinguishing between real and synthetic data effectively tests the perception, knowledge, and reasoning capabilities of LMMs. In response, we introduce LOKI, a novel benchmark designed to evaluate the ability of LMMs to detect synthetic data across multiple modalities. LOKI encompasses video, image, 3D, text, and audio modalities, comprising 18K carefully curated questions across 26 subcategories with clear difficulty levels. The benchmark includes coarse-grained judgment and multiple-choice questions, as well as fine-grained anomaly selection and explanation tasks, allowing for a comprehensive analysis of LMMs. We evaluated 22 open-source LMMs and 6 closed-source models on LOKI, highlighting their potential as synthetic data detectors and also revealing some limitations in the development of LMM capabilities. More information about LOKI can be found at `https://loki102.github.io/LOKI.github.io/`.

1 INTRODUCTION

With the rapid development of diffusion models (Rombach et al., 2022; Dhariwal & Nichol, 2021b) and large language models (Abdullah et al., 2022; Brown, 2020), AI-generated content (AIGC) technology has increasingly integrated synthetic multimodal data into our daily lives. For instance, tools like SORA (Brooks et al., 2024) can produce highly realistic video, while Suno (Shulman et al., 2022) enables the creation of music at a level comparable to professional artists. However, synthetic multimodal data also brings significant risks, including potential misuse and societal disruption (Cooke et al., 2024; Ju et al., 2022). For example, the risks include generating fake news using large language models (LLMs), synthesizing fraudulent faces with diffusion models for scams, and potential contamination of internet training data. Due to the convenience of artificial intelligence synthesis, the future Internet may be saturated with AI-generated content, making the task of discerning the authenticity and trustworthiness of multimodal data increasingly challenging.

To address such threats, the field of synthetic data detection has garnered widespread attention in recent years (Barni et al., 2020; Frank et al., 2020; Gragnaniello et al., 2021; Shao et al., 2023; 2024). However, most current synthetic data detection methods are primarily focused on authenticity evaluation, with certain limitations regarding the human interpretability of the prediction results (Li et al., 2024b). The recent rapid advancement of large multimodal models (LMMs) has sparked curiosity about their performance in detecting synthetic multimodal data (Ku et al., 2023; Wu et al., 2024b). On one hand, for synthetic data detection tasks, LMMs can provide reasoning behind authenticity judgments in natural language, paving the way for enhanced explainability. On the other hand, the task of distinguishing between real and synthetic data involves the perception, knowledge, and reasoning abilities of multimodal data, serving as an excellent test of LMM capabilities. Therefore, the focus of this paper is to evaluate the performance of LMMs in synthetic data detection tasks.

However, traditional synthetic data detection benchmarks, such as Fake2M (Lu et al., 2023b) and ASVSpoof 2019 (Wang et al., 2020b), primarily assess conventional detection methods, and evaluations of LMMs in detecting multimodal synthetic data are still lacking. These benchmarks often miss fine-grained anomaly annotations represented in natural human language, making it difficult to transparently analyze the explainability capabilities of LMMs. FakeBench (Li et al., 2024a) aligns more closely with our objectives, but it only evaluates the performance of LMMs within a single standard image modality, lacking both breadth and depth. Specifically, FakeBench fails to explore other modalities such as audio and 3D data, focusing primarily on general image types and not conducting thorough tests on expert domain images like satellite imagery. To bridge this gap, we introduce LOKI, a comprehensive benchmark for evaluating the performance of LMMs on synthetic data detection. The key highlights of the LOKI benchmark include:

- *Diverse Modalities.* LOKI includes high-quality multimodal data generated by recent popular synthetic models, covering video, image, 3D data, text, and audio.
- *Heterogeneous Categories.* Our collected dataset includes 26 detailed categories across different modalities, such as specialized satellite and medical images; texts like philosophy and ancient Chinese; and audio data like singing voices, environmental sound and music.
- *Multi-level Annotations.* LOKI includes basic "Synthetic or Real" labels, suitable for fundamental question settings like true/false and multiple-choice questions. It also incorporates fine-grained anomalies for inferential explanations, , enabling tasks like abnormal detail selection and abnormal explanation, to explore LMMs' capabilities in explainable synthetic data detection.
- *Multimodal Synthetic Evaluation Framework.* We propose a comprehensive evaluation framework that supports inputs of various data formats and over 25 mainstream multimodal models.

On the LOKI benchmark, we evaluated 22 open-source LMMs, 6 advanced proprietary LMMs, and several expert synthetic detection models. Our key findings are summarized as follows:

For *synthetic data detection tasks* we find: (1) LMMs exhibit moderate capabilities in synthetic data detection tasks, with certain levels of explainability and generalization, but there is still a gap compared to human performance; (2) Compared to expert synthetic detection models, LMMs exhibit greater explainability and, compared to humans, can detect features invisible to the naked eye, demonstrating promising developmental prospects.

For *LMMs capabilities* we find: (1) Most LMMs exhibit certain model biases, tending to favor synthetic or real data in their responses; (2) LMMs lack of expert domain knowledge, performing poorly

on specialized image types like satellite and medical images; (3) Current LMMs show unbalanced multimodal capabilities, excelling in image and text tasks but underperforming in 3D and audio tasks; (4) Chain-of-thought prompting enhances LMMs' performance in synthetic data detection, whereas simple few-shot prompting falls short of providing the necessary reasoning support.

These findings highlight the challenging and comprehensive nature of the LOKI task and the promising future of LMMs in synthetic data detection tasks.

2 Related Work

2.1 Synthetic Data Detection

Currently, synthetic data detection has garnered widespread attention to prevent the misuse of multimedia synthetic data (Gragnaniello et al., 2021; Hou et al., 2023). The detection of synthetic data in image and audio has long been a popular research (Barni et al., 2020; Frank et al., 2020), while methods for synthetic video detection have recently emerged, such as DuB3D(Ji et al., 2024) and AIGVDet(Bai et al., 2024a). However, most work primarily focuses on the binary distinction between authentic and synthetic data, resulting in poor interpretability. Some studies aim to enhance the interpretability of synthetic detection by providing latent representations(Dong et al., 2022), feature explanations(Chai et al., 2020), and artifact localization(Zhang et al., 2023a; Shao et al., 2023; 2024) ; however, most research remains limited to the interpretability of abstract symbols, leaving a significant gap in alignment with human understanding. In practice, current AI-generated synthetic data still exhibits noticeable flaws, such as discontinuities in synthetic videos and insufficient geometric accuracy in 3D data. These shortcomings can be effectively captured and perceived by human users(Tariang et al., 2024), who can provide reasonable explanations. However, existing expert synthetic data detection methods fail to provide human-interpretable bases for their judgments.

2.2 Large Multimodal Models

Recently, the rapid development of multimodal large models (LMMs) has been notable, with models like GPT-4o (OpenAI, 2024) and Claude 3.5 (Anthropic, 2024) excelling in various tasks such as scientific questioning (Lu et al., 2022; Yue et al., 2024) and commonsense reasoning (Talmor et al., 2018), showcasing exceptional perceptual and reasoning abilities (Bai et al., 2024b). Research has also applied LMMs to evaluate AIGC synthetic results, utilizing GPT to assess the quality of generated images (Ku et al., 2023; Peng et al., 2024) and 3D models (Wu et al., 2024b), providing scores that align with human preferences along with interpretable justifications. Consequently, in synthetic data detection, LMMs can offer reasons for determining authenticity in natural language, paving the way for enhanced interpretability in synthetic detection. Moreover, LMMs can access features invisible to human users, such as deep image and spectral features, demonstrating their potential to exceed human detection capabilities. Furthermore, synthetic data detection involves multimodal data perception and complex logical reasoning, making it an excellent task to assess the capabilities of LMMs. This task also provides quantitative evaluation metrics like accuracy, allowing for a more direct assessment of model performance compared to more qualitative scoring tasks.

2.3 Synthetic Data Detection Benchmark

Currently, there are numerous datasets corresponding to synthetic data detection tasks, including those designed for traditional detection methods and those tailored for LMMs. For instance, traditional synthetic datasets such as Fake2M (Lu et al., 2023b), HC3 (Guo et al., 2023), and ASVSpoof 2019 (Wang et al., 2020b) have explored the performance of traditional deepfake detection methods across various modalities, but they lack assessments for LMMs models. VANE (Bharadwaj et al., 2024) evaluates the capability of LMMs in detecting video anomalies, including the detection of criminal activities in real videos and synthetic video detection, although it focuses more on video content understanding. Fakebench (Li et al., 2024b) assesses LMM performance in the image modality, yet it concentrates on a single modality and offers limited subcategories. In contrast, LOKI covers a broader range of data modalities, including video, image, 3D, text, and audio, as well as data from specialized fields such as remote sensing, medical imaging, and environmental sounds. In terms of problem design, LOKI encompasses tasks for authenticity judgment, as well as more complex challenges like Abnormal Details selection and Abnormal Explanation, which test the LMMs' ability to explain reasons in synthetic data detection.

3 DATASET

3.1 OVERVIEW OF LOKI

We introduce LOKI, a multimodal synthetic data detection benchmark, designed specifically to comprehensively assess the capabilities of LMMs in detecting synthetic data. As illustrated in Figure 2, LOKI encompasses a variety of modalities including video, image, 3D, text, and audio, with over 26 specific subcategories of data. The benchmark utilizes fine-grained anomaly annotations to construct a tiered variety of question types, including judgment questions, multiple-choice questions, abnomal detail selection and abnomal explanation questions, totaling over 18k questions.

Table 1 provides a detailed comparison of LOKI with existing datasets, including traditional synthetic detection benchmarks and those tailored for evaluating LMMs. In terms of breadth, LOKI covers a wider range of modalities and finer categories. In depth, it goes beyond binary judgment question designs to include questions that require a deep understanding and explanation of detailed anomalies. Additionally, LOKI classifies question difficulty based on human evaluation metrics.

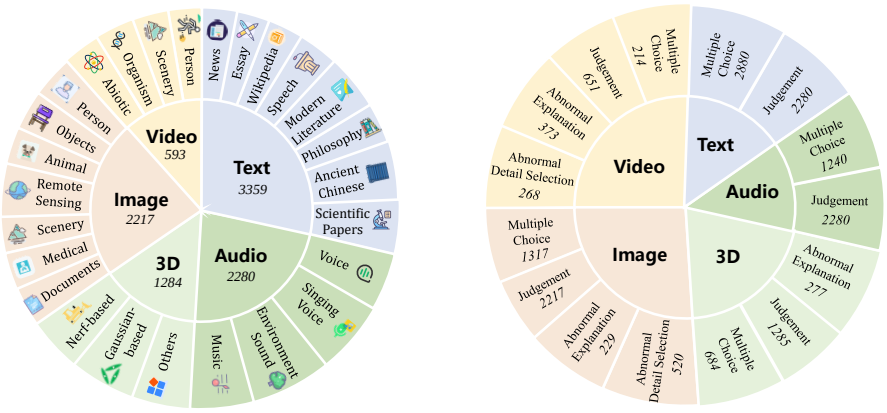


Figure 2: **Statistical information of LOKI.** The left side displays the detailed categories of each modality, while the right side presents the questions across different modalities. The inner circle numbers represent the data volume, and the outer circle numbers indicate the number of questions.

Table 1: The comparison between LOKI and other benchmarks. Answer types include JD (Judgment), MC (Multiple Choice), and OE (Open-ended). "Real paired" indicates whether real data is paired within the same domain, while "Difficulty Level" shows if questions are graded by difficulty.

Dataset	Size	Category	Data Modality					Answer			Real Paired	Difficulty Level
			Img	Vid	Txt	Aud	3D	JD	MC	OE		
FFHQ	70k	-	✓					✓			✗	✗
Fake2M	>1M	8 types	✓					✓			✓	✗
HC3	∼80K	5 types			✓			✓		✓	✓	✗
Mixset	3.6 K	5 types			✓			✓			✓	✗
ASVS2019	108K	-				✓		✓			✓	✗
Codecfake	∼1M	-				✓		✓			✓	✗
FakeBench	6K	6 types	✓					✓		✓	✗	✗
VANE	0.9K	-		✓					✓		✗	✗
LOKI	18K	26 types	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

3.2 DATA COLLECTION AND ANNOTATION

Video: We collected 593 video clips by utilizing various closed-source and open-source models such as SORA (OpenAI, 2024), Keling, and Open_sora (Zheng et al., 2024), generating high-quality text-to-video synthesis data along with corresponding real domain sample data. For the AI-generated

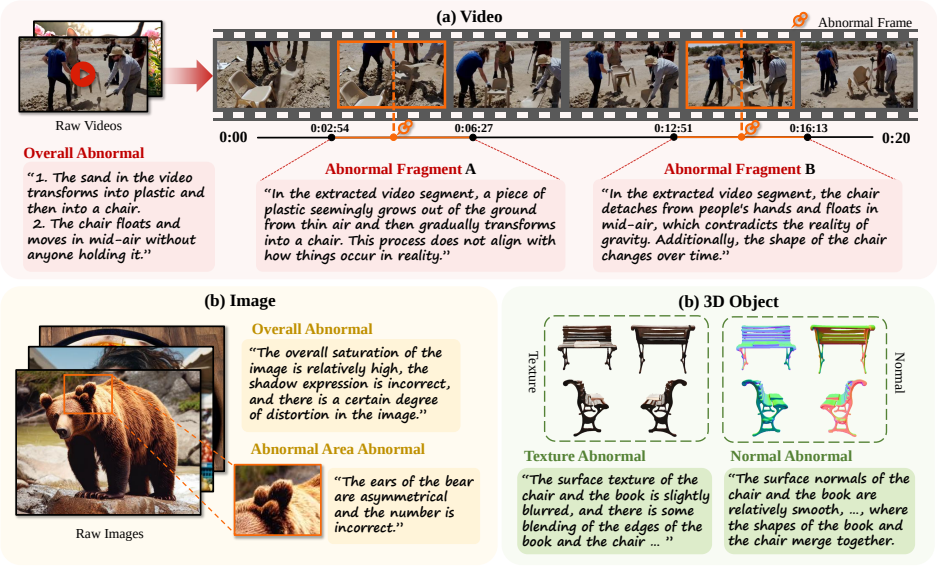


Figure 3: **Examples of Synthetic Data Annotations:** (a) Detailed annotations of video anomalies; (b) Detailed annotations of image anomalies; (c) Detailed annotations of 3D anomalies.

video clips, we employed the LabelU[1] tool to annotate anomaly details, including anomalous segments and their descriptions, anomalous key frames, and global anomaly descriptions. As shown in Figure 3 (a), anomalies in the videos, such as "violating natural physics" and "frame flickering," are also annotated globally. Additionally, the anomalous segment from 02:54 to 06:27 is highlighted, with the corresponding reasons for the anomalies explained by human annotators. Furthermore, each anomalous segment includes an anomalous key frame to facilitate subsequent LMMs in accurately reading the anomalous frames when processing video data.

Image: We have collected over 2,200 images from 7 subcategories through existing dataset extraction, internet collection, and new data synthesis. The image synthesis methods include FLUX, Midjourney (AI, 2023), Stable Diffusion (Blattmann et al., 2023), and ten other different methods to ensure high quality and diversity of the data. For the synthesized image data, in addition to overall annotations, we performed anomaly region bounding and explanations, as shown in Figure 3 (b). The region anomaly annotations allow for more fine-grained and specific labeling, which can be used for generating subsequent anomaly detail questions.

3D data: We conducted a comprehensive analysis of OmniObject3D (Wu et al., 2023), selecting scanned instances as ground truth within the same domain. By constructing prompt texts, we synthesized three Nerf models (Poole et al., 2022) and three 3D GS models (Tang et al., 2023), and supplemented them with results from the advanced commercial model Clay and some Nerf-based results from GPTEval3D (Wu et al., 2024b). We collected a total of over 1,200 3D models from ten different synthesis methods, including both synthesized and real scanned data. Additionally, as shown in Figure 3 (c), we performed texture anomaly description annotations corresponding to the RGB four views of the synthesized 3D data, as well as normal anomaly description annotations. Notably, besides the multi-view format, the 3D data also supports point clouds and panoramic videos.

Audio: We collected various categories of audio, including speech, singing voice, environmental sounds, and music. The speech and singing voice data ensured consistency in speaker timbre, sourced from the Logical Access part of ASVSpoof2019 (Wang et al., 2020b) and the CtrSVDD Benchmark, covering four generation paradigms: TTS, VC, SVS, and SVC. Environmental audio data came from DCASE 2023 Task 7, with real audio from the development set and synthetic audio generated using multiple methods from Track A. Music data were sourced from MusicCaps, with synthetic music generated based on descriptions using MusicGen (Copet et al., 2024), AudioLDM2-Music (Liu et al., 2024a), and Suno[2].

[1]LabelU: https://github.com/opendatalab/labelU
[2]Suno: https://suno.com/

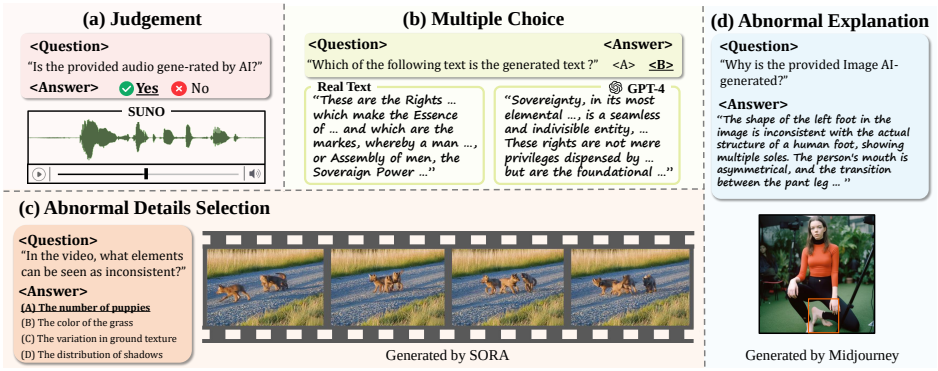


Figure 4: **Example Questions of LOKI.** LOKI includes four types of questions:(a) Judgment questions; (b) Multiple choice questions; (c) Abnormal detail selection; (d) Abnormal explanation.

Text: Based on summarization and regeneration methods, we generated counterfeit texts similar to the original texts using mainstream models such as GPT-4o, Qwen-Max, and Llama 3.1-405B. We collected eight categories of text data, pairing each sample with a real text and a model-generated similar text, totaling 3,359 text entries. Our text data were categorized by length and language, including short texts (50-100 characters), medium texts (100-200 characters), and long texts (over 200 characters), with a 1:1 ratio of Chinese to English data. More information regarding the collection and statistics of each modality can be found in Appendix B.

3.3 Question Generation

Judgment Task: This task requires large language models (LMMs) to determine whether the input data is synthetic or real. As shown in Figure 4 (a), LMMs need to answer the judgment question, "Is the provided audio generated by AI?" To minimize the influence of prompts on model judgments, questions are asked in two forms: whether the data is AI-synthesized or real, and identifying either the real or AI-synthesized data. Furthermore, we categorize the data into different difficulty levels based on human performance. If all tested human users (more than three) answer correctly, the task is classified as "easy"; if more than 50% answer incorrectly, it is classified as "hard"; all other cases fall into the "medium" category.

Multiple Choice Task: This task requires LMMs to correctly select AI-generated or real data from the provided synthetic and real data. As illustrated in Figure 4 (b), LMMs need to complete the multiple-choice question, "Which of the following texts is generated?" The design of this question benefits from our collection of real paired data within the same domain, effectively assessing LMMs' comparative analysis capabilities.

Abnormal Detail Selection: Based on fine-grained anomaly annotation data from modalities such as video, images, and 3D, we effectively design prompts and utilize GPT-4o to generate questions for Abnormal Detail Selection. As shown in Figure 4 (c), for video content's detail anomalies, we ask, "What elements can be seen as inconsistent?" By providing clear anomaly annotations, we can effectively reduce the hallucination phenomenon in GPT-4o, ensuring the quality of the questions. More details can be found in the supplementary materials.

Abnormal Explanation: Furthermore, we design open-ended abnormal explanation questions, requiring LMMs to independently identify anomalies and explain their reasons. As shown in Figure 4 (d), we ask, "Why is the provided image AI-generated?" It is worth to note that in real anomaly explanation tasks, the input does not include bounding boxes around anomalous areas. Tasks related to Abnormal Detail Selection and Abnormal Explanation can more precisely test whether LMMs genuinely perceive corresponding detail anomalies rather than guessing answers.

Quality Control: To mitigate the impact of hallucinations of GPT-4o during question generation in abnormal detail selection task, all samples in this task undergo manual reviews. Each question that involves GPT must pass through at least two rounds of verification by human users. A total of 20 users participated in the verification process, which took approximately 160 hours to complete.

4 EXPERIMENT

In this section, we evaluate various Language Model Multimodalities (LMMs) under our proposed LOKI evaluation framework, which includes both open-source and proprietary models, multimodal LMMs, Audio LMMs, and text-based LLMs. Our evaluations are conducted in a *zero-shot* setting. In the following subsections, we first introduce our evaluation models and the evaluation protocols. Next, we analyze the performance of existing LMMs in synthetic data detection tasks, comparing them with human users and expert models. We will then discuss the challenges and shortcomings faced by multimodal large models in the current task settings. Additionally, we explore the potential impact of few-shot or chain-of-thought prompting on this task.

4.1 BASELINES

LMMs. We evaluate 3 closed-source and 18 open-source LMMs across different model types and sizes. For closed-source models, we consider GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Team et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024). Given that modality alignment in multimodal LMMs may lead to a decline in LLM performance on text-based tasks (Dai et al., 2024), we also selected pure text LLMs, such as LLaMA-3.1-405B (Team, 2024), Qwen-Max (Chu et al., 2023) and Mistral-Large (Mistral, 2024), to evaluate the text modality. In the evaluation of Audio LMMs, we selected high-performing open-source models such as Qwen-Audio (Chu et al., 2023) and SALMONN-7B (Sun et al., 2024). For proprietary models, we chose Gemini-Flash (Team et al., 2023), which supports audio input.

Human Users. We invited over 50 human users, including senior university students and regular users, to participate in the judgment and multiple-choice question tests for different modalities of synthetic data. Each question was tested by at least 3 users to ensure the robustness of the results. Additionally, we designed an online platform to distribute random questionnaires, and more than 200 users participated in the testing of 15 basic questions.

Expert Models. We selected recently open-sourced expert-level synthetic data detection methods and their corresponding weights for testing, including video detection (AIGVDet (Yang et al., 2024)), image detection (AIDE (Yan et al., 2024)), text detection (RADAR-Vicuna-7B (Hu et al., 2023)), and audio detection (AASIST (Jung et al., 2022)). Due to the limited availability of 3D synthetic data detection methods, 3D was not considered. Additionally, there is no overlap between the training sets of these methods and the LOKI test data, reducing the possibility of data contamination. We selected only a small number of expert models for evaluation, primarily to serve as references, similar to the role of human references.

Evaluation Protocol. *Data Input:* For the video modality, we utilize an 8-frame video clip along with corresponding questions as input. For 3D modal data, we employ the commonly used multi-view input method. Results based on surround video and point cloud inputs are also included in the supplementary materials. For other modalities, inputs are based on textual prompts combined with corresponding images, audio, and textual materials. During the evaluation, each model independently generates responses to questions without retaining any dialogue history.

Evaluation Metric: For judgement, multiple-choice and abnormal detail selection questions, we use the average accuracy rate as a metric. In addition to accuracy, we also calculate the Normalized Bias Index (NBI) based on recall rates to assess model bias. For open-ended questions regarding anomalous details, we use the GPT-4 model to assess the score of the responses. Further details on the calculation of evaluation metrics can be found in Appendix C.2.

Evaluation Framework: To standardize the evaluation of different LMMs and various input modalities for synthetic data detection, we propose a comprehensive multimodal evaluation framework. This framework provides support for various input modalities such as 3D point clouds, videos, images, audio, and text, while unifying APIs of over 25 mainstream LMMs, ensuring both model compatibility and consistency throughout the evaluation process.

4.2 SYNTHETIC DATA DETECTION RESULTS

In this section, we provide a comprehensive analysis of the performance of various LMMs and LLMs on synthetic data detection tasks using the LOKI dataset.

Table 2: Results of different models on the LOKI for Judgment and Multiple Choice questions. (a) Multimodal evaluation of LMMs; (b) Text evaluaion of LLMs; (c) Audio evaluation of Audio LMMs; * denotes the closed-source models.

(a) Multimodal evaluation of LMMs

	Judgment					Multiple Choice				
	Video	**Image**	**3D**	**Text**	**Overall**	**Video**	**Image**	**3D**	**Text**	**Overall**
Random Choice	51.1	50.5	50.5	49.9	50.3	47.7	49.0	49.7	45.2	46.9
Human (Medium)	83.5	80.1	72.0	68.5	76.0	91.3	84.5	91,2	78.5	86.4
Expert models	53.1	63.1	-	72.1	62.8	-	-	-	-	-
Phi-3.5-Vision	56.8	52.5	50.0	49.4	52.2	58.2	44.0	59.6	42.0	50.9
MiniCPM-V-2.6	57.2	44.8	56.4	49.4	52.0	52.8	49.8	50.7	48.9	50.6
InternLM-XComposer2.5	58.4	46.4	43.9	52.6	50.3	56.3	51.0	48.0	40.5	49.0
mPLUG-Owl3-7B	55.3	45.9	49.9	53.6	51.1	60.3	52.5	49.9	50.0	53.1
LongVA-7B	60.4	46.2	49.9	48.6	51.7	57.5	51.6	61.4	48.9	52.6
Qwen2-VL-7B	59.5	47.8	**72.3**	48.9	**57.1**	64.0	65.1	55.5	46.4	57.7
LLaVA-OV-7B	56.8	49.8	68.4	53.0	57.0	59.8	51.7	53.8	48.4	53.4
Llama-3-LongVILA-8B	51.9	49.8	32.2	49.9	46.0	54.0	51.1	50.5	44.3	50.0
Idefics2-8B	54.8	45.0	38.4	47.2	46.3	55.6	51.3	54.2	37.0	49.5
Mantis-8B	55.4	**54.6**	50.0	52.0	53.0	47.9	61.5	**62.5**	48.4	55.1
InternVL2-8B	60.8	49.7	49.4	50.3	52.6	54.0	51.4	53.1	46.6	51.3
InternVL2-26B	55.0	44.3	50.4	51.1	49.9	62.4	48.5	48.8	50.3	53.2
InternVL2-40B	**62.0**	49.6	49.9	53.1	52.2	**65.7**	63.1	59.9	45.2	52.7
VILA1.5-13B	51.9	49.3	34.0	47.7	45.7	52.1	55.3	53.5	44.0	51.2
VILA1.5-40B	59.2	48.8	50.0	50.1	52.7	49.1	64.0	47.9	50.4	53.7
Qwen2-VL-72B	59.6	53.2	60.3	52.8	55.4	**65.7**	68.6	58.7	**69.7**	**65.6**
LLaVA-OV-72B	56.5	46.3	51.3	**61.2**	56.3	62.9	**70.8**	61.3	69.2	65.2
Claude-3.5-Sonnet*	61.7	53.6	58.0	**61.5**	61.6	60.5	65.5	51.9	**89.2**	**74.8**
Gemini-1.5-Pro*	58.5	43.5	55.4	55.7	53.2	66.1	67.3	60.2	57.3	62.7
GPT-4o*	**71.3**	**63.4**	**65.2**	55.9	**63.9**	**77.3**	**80.8**	**70.2**	66.6	73.7

(b) Text evaluation of LLMs

Model	**Judgment**	**Choice**	**Overall**
Human (Medium)	69.2	71.1	70.1
Expert model	69.4	-	69.4
LLaMA-3.1-405B	56.8	73.1	64.4
Mistral-Large*	52.2	69.1	57.8
Qwen-Max*	48.3	44.4	46.5
Claude-3.5-Sonnet*	**61.5**	**89.2**	**70.7**
Gemini-1.5-Pro*	55.7	57.3	56.2
GPT-4*	55.9	66.6	59.5

(c) Audio evaluation of Audio LMMs

Model	**Judgment**	**Choice**	**Overall**
Human (Medium)	69.2	71.1	70.1
Expert model	69.4	-	69.4
Qwen-Audio	49.8	50.1	49.9
SALMONN-7B	**51.2**	-	**51.2**
AnyGPT	49.8	**50.3**	50.1
OneLLM	49.9	-	49.9
LUT	44.4	-	44.4
Gemini-1.5-Flash*	49.4	49.2	49.3

Judgment and Multiple Choice. Table 2 illustrates the performance of various models on judgment and multiple-choice questions in LOKI. For the synthetic data judgment task, the closed-source model GPT-4o achieves the best results, with an overall accuracy (excluding audio) of 63.9%. When real paired data is included for comparison in the multiple-choice questions, accuracy further increases to 73.7%. In the text modality, Claude-3.5 outperform other LMMs and LLMs, achieving accuracies exceeding 70%. In the Audio LMMs category, both open-source and closed-source models show performances comparable to random selection, which is not satisfactory.

Abnormal Detail Selection and Explanation. We compared the performance of different models on the tasks of abnormal detail selection and abnormal reason explanation, as shown in Table 3. GPT-4o achieved an accuracy exceeding 75% in abnormal detail selection and a score over 70% in abnormal reason explanation. This indicates that advanced LMMs like GPT-4o has demonstrated strong detail understanding capabilities, effectively analyzing and interpreting "synthetic traces." Notably, we observe that Claude-3.5-Sonnet (Anthropic, 2024) tends to misclassify synthetic images as real, despite the primary goal of our tasks being to explain abnormalities in synthetic images. More examples of abnormal explanations can be found in Appendix F.

Table 3: Results of different models on the LOKI for Abormal Details Selection and Abnormal explanation questions.* denotes the closed-source models.

	Abnormal Details Selection			Abnormal Explanation			
	Video	Image	Overall	Video	Image	3D	Overall
LLaVA-OV-7B	76.9	18.8	43.1	46.7	68.9	71.0	62.0
Qwen2-VL-7B	**79.4**	31.5	51.5	48.4	63.8	73.4	61.9
InternVL2-8B	66.8	70.2	68.8	46.5	72.2	71.3	63.0
Gemini-1.5-Pro*	58.7	40.0	47.8	57.6	**77.1**	70.8	68.1
Claude-3.5-Sonnet*	50.9	19.8	32.8	50.1	1.7	**78.2**	45.8
GPT-4o*	74.0	**76.2**	**75.3**	**67.6**	72.9	77.0	**72.6**

Figure 5: **The multimodal large model capability assessment analysis results.** (a) Model bias assessment, where the closer the color is to red, the more the model is biased towards classifying the data as real; the closer to blue, the more it leans towards synthetic data. The size of the square also represents the degree of bias. (b) The performance of GPT-4o across different image types and its difference from human users. (c) A relative radar chart of the model's performance across various modalities, with Human benchmarks for comparison.

Comparing Humans and Expert Models. Humans exhibit an average performance of 76% in judgment tasks and 86.4% in multiple-choice questions, both 10% higher than the LMM method. Notably, if LMM tools are to be applied in production, their decision-making performance in judgment tasks must exceed 90% to be convincing. As synthesis technologies advance, the distinct "traces" of synthetic data are becoming increasingly subtle. However, LMMs capture minute details, such as image features imperceptible to the human eye, demonstrating their potential to surpass human.

LMMs demonstrate superior performance in most tasks compared to expert models. This is primarily due to the rich and diverse sources of synthetic data collected by LOKI, which significantly differ from existing data domains, resulting in suboptimal generalization performance of expert models. The accuracy of synthetic detection by expert models trained on similar data should significantly improve. Currently, LMMs perform at a moderate level in synthetic data detection but surpass expert models in generalization ability. Unlike traditional expert models, LMMs possess the capability to explain the reasons behind anomalies, highlighting their unique advantage as synthetic detectors.

4.3 Large Multimodal Models Capabilities

Model Bias. The heatmap of the Normalized Bias Index calculated based on recall rates, as shown in Figure 5 (a), is utilized for analyzing model biases. The results indicate that most models exhibit significant biases in synthetic data detection tasks, with a tendency to incorrectly categorize data as either real or synthetic. For instance, GPT-4o tends to classify textual data as real, whereas it is biased towards judging 3D data as AI-generated. Despite diverse questioning techniques implemented to minimize cueing effects, a pronounced bias is still evident across most models.

Lack of Expert Domain Knowledge. In Figure 5 (b), we present the varying performance of GPT-4o across different image subcategories. The experimental results clearly indicate that GPT exhibits strong recognition abilities on common image types such as objects and landscapes, even surpassing human users. However, GPT-4o's performance significantly deteriorates in specialized fields such

Table 4: Result decomposition across questions difficulty levels.

Difficulty Levels (Video & Image & 3D & Text)				
	Easy (2470)	Medium (1104)	Hard (3938)	Overall (7512)
LLaVA-OV-7B	60.4	47.6	39.1	47.3
InternVL2-8B	64.5	47.8	33.5	45.7
Qwen2-VL-7B	67.7	45.6	35.2	47.4
Gemini-1.5-pro	70.8	42.4	32.4	46.4
Claude-3.5-Sonnet	76.0	44.7	29.8	47.1
GPT-4o	78.8	52.3	44.4	56.8

Table 5: LMMs' performances under different prompting strategies for judgement tasks.

Prompting Strategies Performances (Image & 3D)			
	Baseline	FS Few-shot	CoT Chain-of-Thought
LLaVA-OV-7B	56.6	46.4	18.8
InternVL2-8B	49.6	46.1	50.4
Qwen2-VL-7B	56.8	52.6	59.5
Gemini-1.5-pro	47.9	41.2	51.0
Claude-3.5-Sonnet	55.2	53.7	56.4
GPT-4o	64.1	75.1	74.2

as satellite and medical imaging, and in less-trained image types like documents. This suggests that current LMMs still lack certain expert domain knowledge.

Unbalanced Multimodal Capabilities. In Figure 5 (c), we compare the performance of various LMMs across different modalities. Results indicate that current models excel in frequently trained modalities such as images and text, even surpassing human performance in some tests. However, their performance declines significantly on audio tasks, with most open-source models lacking corresponding capabilities. For future AGI to develop into a versatile assistant, it needs to possess more balanced multimodal abilities.

Model Performance across Different Levels. Based on human user performance, we categorized the difficulty levels of the questions, as shown in Table 4, which presents the performance of selected models across different difficulty levels. As the difficulty increases, the performance of LMMs gradually declines, consistent with human user performance. Under challenging conditions, GPT-4o's accuracy drops to only 44.4%, which is lower than that achieved by random selection. This indicates that LMMs have certain limitations in handling complex synthetic data detection tasks.

Prompting Strategies Impact LMMs Capabilities. In Table 5, we demonstrate the effects of different prompting strategies in LOKI's image and 3D judgement tasks, where CoT refers to the Chain of Thought prompting (Wei et al., 2022b) and FS refers to the few-shot prompting (Alayrac et al., 2022). During inference, models are prompted with two random examples that are in the same domain as the questions by different strategies. In CoT prompting, we manually craft "thought chains" with our human annotations to elicit reasoning steps out of LMMs, while in FS prompting, we simply prepend examples with answers to the questions. Interestingly, GPT-4o shows strong reasoning ability without chain-of-thought prompting, while other models rely on it for improved performance. Few-shot learning fails to support the necessary step-by-step reasoning for synthetic data detection, but GPT-4o performs well regardless, suggesting its inherent ability to reason effectively without additional reasoning guidance. However, LLaVA-OV-7B experienced significant performance drop when prompted with CoT. We conjecture that this degradation may result from a decline in its ability to understand long contexts after fine-tuning (Zhai et al., 2023). More CoT experimental results are available in Appendix E.3.

5 Conclusion

In this paper, we introduced LOKI, a multimodal benchmark designed to evaluate the performance of large multimodal models in detecting synthetic data across various modalities. We conducted a comprehensive study of LMMs' performance on video, image, 3D, audio, text, and specialized subdomains, and we also analyzed LMMs' ability to explain detailed anomalies in synthetic data. The experimental results indicate that LMMs have a certain level of competence in detecting synthetic data and a preliminary ability to explain anomalies. Synthetic data detection tasks also effectively evaluate the various capabilities of LMMs during their development. These findings highlight the challenging and comprehensive nature of the LOKI task, as well as the potential of LMMs in future synthetic data detection tasks. We aim to inspire more powerful and interpretable synthetic data detection methods through LOKI to address the potential risks posed by rapidly advancing AI synthesis technologies. Furthermore, while the relationship between synthesis and detection is adversarial, they are mutually beneficial; better and more explainable synthetic detectors will further advance AI synthesis technologies.

REFERENCES

Ai safety summit, 2023. URL https://www.aisafetysummit.gov.uk/. Hosted by the UK.

Malak Abdullah, Alia Madain, and Yaser Jararweh. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1–8. Ieee, 2022.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023. URL https://arxiv.org/abs/2301.11325.

M. AI. Midjourney: Text to image with ai art generator, 2023. URL https://www.midjourneyai.ai/en. Accessed: 2024-09-21.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Sam Altman. Openai now generates about 100 billion words per day, 2024. URL https://x.com/sama/status/1756089361609981993. Accessed: 2024-09-24.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www.anthropic.com. Accessed: 2024-09-23.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. AI-generated video detection via spatial-temporal anomaly learning. *The 7th Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2024a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*, 2024b.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. 2024. URL https://api.semanticscholar.org/CorpusID:267095113.

Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2020.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024.

Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lmms, 2024.

Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse. *arXiv preprint arXiv:2311.12202*, 2023.

Matyas Bohacek and Hany Farid. The making of an ai news anchor—and its implications. *Proceedings of the National Academy of Sciences*, 121(1):e2315678121, 2024.

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

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pp. 103–120. Springer, 2020.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22246–22256, 2023a.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 5178–5193, 2023b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147, June 2023.

Keunwoo Choi, Jaekwon Im, Laurie Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinosuke Takamichi. Foley sound synthesis at the dcase 2023 challenge. *arXiv preprint arXiv:2304.12521*, 2023.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Di Cooke, Abigail Edwards, Sophia Barkoff, and Kathryn Kelly. As good as a coin toss human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760*, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021a. URL `https://arxiv.org/abs/2105.05233`.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021b.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. *arXiv preprint arXiv:2402.07712*, 2024a.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Forty-first International Conference on Machine Learning*, 2024b.

Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European conference on computer vision*, pp. 18–35. Springer, 2022.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires reinforcement. *arXiv preprint arXiv:2406.07515*, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Liang Yu Gong and Xue Jun Li. A contemporary survey on deepfake detection: datasets, algorithms, and challenges. *Electronics*, 13(3):585, 2024.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech 2021*, pp. 571–575, 2021.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, 2021.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL `https://arxiv.org/abs/2306.11644`.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*, 2023.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016. URL `https://arxiv.org/abs/1605.01397`.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20555–20565, 2023.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations*, 2023.

Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading deepfake detectors via adversarial statistical consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12271–12280, 2023.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095, 2023.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. Distinguish any fake videos: Unleashing the power of large-scale data and motion features. *arXiv preprint arXiv:2405.15343*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363–374, 2023b.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3465–3469. IEEE, 2022.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019a.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019b. URL https://arxiv.org/abs/1812.04948.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. Fakebench: Probing explainable fake image detection via large multimodal models, 2024a. URL https://arxiv.org/abs/2404.13306.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024b.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023a.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Tianchi Liu, Lin Zhang, Rohan Kumar Das, Yi Ma, Ruijie Tao, and Haizhou Li. How do neural spoofing countermeasures detect partially spoofed audio? *arXiv preprint arXiv:2406.02483*, 2024b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023c.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023a. URL https://arxiv.org/abs/2211.01095.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Zeyu Lu, Di Huang, LEI BAI, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 25435–25447. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/505df5ea30f630661074145149274af0-Paper-Datasets_and_Benchmarks.pdf.

Zeyu Lu, Di Huang, Chunli Zhang, Chengyue Wu, Xihui Liu, Lei Bai, and Wanli Ouyang. Sentry-image leaderboard. https://github.com/Inf-imagine/Sentry, 2023c.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Towards understanding the interplay of generative artificial intelligence and the internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pp. 59–73. Springer, 2023.

Nichols Michelle. Un security council meets for the first time on ai risks. *Reuters. Last accessed*, 15, 2023.

Mistral. Large enough, 2024. URL https://mistral.ai/news/mistral-large-2407/. Accessed: 2024-09-13.

Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K Dwivedi. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154:113368, 2023.

Eliya Nachmani and Lior Wolf. Unsupervised singing voice conversion. *arXiv preprint arXiv:1904.06590*, 2019.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

OpenAI. Sora: Creating video from text, 2024. URL `https://openai.com/sora`. Accessed: 2024-09-21.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298, 2024.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345. Springer, 2020.

Maria Pawelec. Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital society*, 1(2):19, 2022.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073. IEEE, 2018.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2023.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Michael Shulman, Georg Kucsko, Martin Camacho, and Keenan Freyberg. Suno, 2022. URL https://suno.com.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157, 2020.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023.

Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Xingcheng Song, Zhiyong Wu, Yiheng Huang, Dan Su, and Helen Meng. Specswap: A simple data augmentation method for end-to-end speech recognition. In *Interspeech*, pp. 581–585, 2020b.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

svc-develop team. so-vits-svc: A github project for voice conversion. https://github.com/svc-develop-team/so-vits-svc, 2024. Accessed: 2024-10-01.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Diangarti Tariang, Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Synthetic image verification in the era of generative artificial intelligence: What works and what isn't there yet. *IEEE Security & Privacy*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Llama Team. The Llama 3 Herd of Models, 2024. URL https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

TheDataBeast. Ted talk transcripts (2006 - 2021), 2021. URL https://www.kaggle.com/datasets/thedatabeast/ted-talk-transcripts-2006-2021.

Timedomain. Ace studio, 2023. URL https://acestudio.ai/. Accessed: 2024-09-21.

Trapoom Ukarapol and Kevin Pruvost. Gradeadreamer: Enhanced text-to-3d generation using gaussian splatting and multi-view diffusion. *arXiv preprint arXiv:2406.09850*, 2024.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting text ghostwritten by large language models, 2024. URL https://arxiv.org/abs/2305.15047.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020b. URL https://arxiv.org/abs/1911.01601.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2015. doi: 10.1109/ICCV.2015.451. Acceptance rate: 30.3%.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 803–814, 2023.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024a.

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22227–22238, 2024b.

Bright Xu. Nlp chinese corpus: Large scale chinese corpus for nlp, September 2019. URL https://doi.org/10.5281/zenodo.3402023.

Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9291–9298, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

Ziqin Yang, Fuxin Xie, Jian Zhou, Yuan Yao, Cheng Hu, and Baoding Zhou. Aigdet: Altitude-information guided vehicle target detection in uav-based images. *IEEE Sensors Journal*, 2024.

Junyan Ye, Jun He, Weijia Li, Zhutao Lv, Jinhua Yu, Haote Yang, and Conghui He. Skydiffusion: Street-to-satellite image synthesis with diffusion models and bev paradigm. *arXiv preprint arXiv:2408.01812*, 2024.

Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023.

Jiquan Yuan, Xinyan Cao, Changjin Li, Fanyi Yang, Jinlong Lin, and Xixin Cao. Pku-i2iqa: An image-to-image quality assessment database for ai generated images, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, et al. Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection. *arXiv preprint arXiv:2406.02438*, 2024.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

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

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024a.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected?, 2024b.

Zangwei Zheng, Xiangyu Peng, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL `https://github.com/hpcaitech/Open-Sora`. Accessed: 2024-09-21.

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, 2023.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.

Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, and Luisa Verdoliva. M3dsynth: A dataset of medical 3d images with ai-generated local manipulations, 2024. URL `https://arxiv.org/abs/2309.07973`.

LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models

Supplementary Material

Table of Contents in Appendix

A Synthetic Data Detection **21**

A.1 Social Impact of Synthetic Data 21

A.2 Synthetic Data Contamination 21

A.3 Increasing Attention on Synthetic Data Detection 22

B Dataset Description **23**

B.1 Data Collection 23

B.2 Dataset Annotation 25

B.2.1 Annotation Guidelines 25

B.2.2 Annotator Informed Consent 26

B.3 Quality Control and Validation 26

B.4 Special data description 27

C Evaluation **28**

C.1 Evaluation Model 28

C.2 Evaluation Metric 29

D Breakdown Results on Different Modalities **30**

D.1 Video 30

D.2 Image 30

D.3 3D 30

D.4 Audio 30

D.5 Text 31

E More experiments result and discussion **37**

E.1 Compression artifact tests 37

E.2 Deepfake detection 37

E.3 More CoT experiments results 38

E.4 Performance across Different Levels and Modalities 39

F Case Study **40**

A SYNTHETIC DATA DETECTION

In this appendix, we introduce and discuss the social impacts of synthetic data, such as deepfakes, as well as data contamination introduced by synthetic data. Finally we discuss the increasing attention on synthetic data detection.

A.1 SOCIAL IMPACT OF SYNTHETIC DATA

While synthetic data generated by AIGC technology has offered numerous benefits to various aspects of society, it has also introduced significant challenges and risks. One of the most concerning risks is the potential to use synthetic data to create deepfakes. All forms of synthetic data can be leveraged to generate deepfakes, which can then be used to deceive, manipulate, or defraud individuals or organizations (see Fig.6). For instance, synthetic text data can be exploited to create fake news Papageorgiou et al. (2024), phishing emails, or manipulative advertisements. Similarly, synthetic image and 3D data can be used to generate realistic fake facesXu et al. (2023), scenes, or even content that leads to copyright violations Jiang et al. (2023b); Somepalli et al. (2023). Synthetic video data poses a threat by enabling the production of fake videos or fake news (e.g., political propaganda Pawelec (2022)), as well as deepfake video fraud calls Mustak et al. (2023). Likewise, synthetic audio data can be used for fake calls, voice cand even fake broadcasts. Furthermore, the advancements in synthetic data technologies are also impacting employment in creative industries, exemplified by the months-long strikes in the film industry Bohacek & Farid (2024).

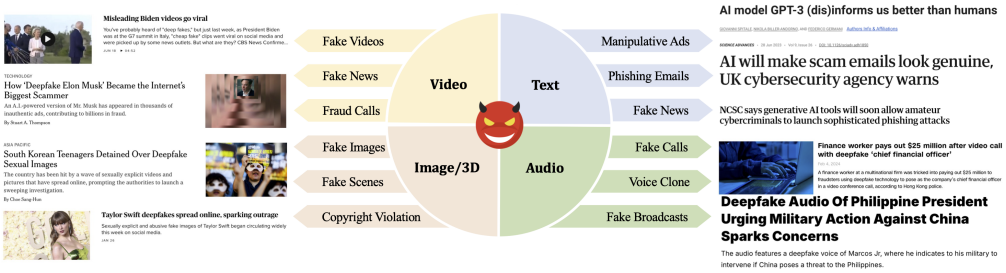


Figure 6: Soical impact of synthetic data across different modalities

A.2 SYNTHETIC DATA CONTAMINATION

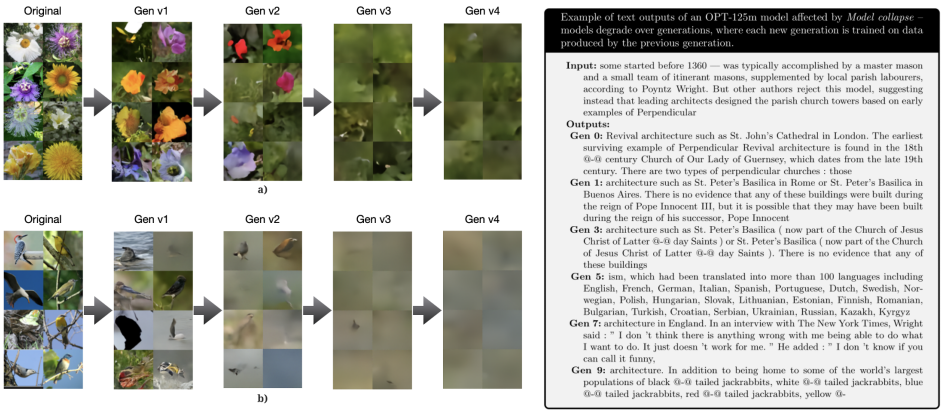


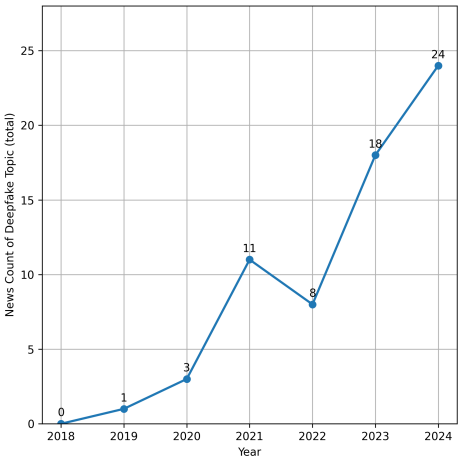
Figure 7: Model Performance Collapse Trained On Synthetic Data (Image from Martínez et al. (2023), Text from Shumailov et al. (2024))

In today's LLM era, the internet is flooded with a substantial amount of synthetic data, even existing web-scale datasets are known to contain synthetic content Schuhmann et al. (2022). According to openai Altman (2024), they now generate about 100 billion words per day, while all people on

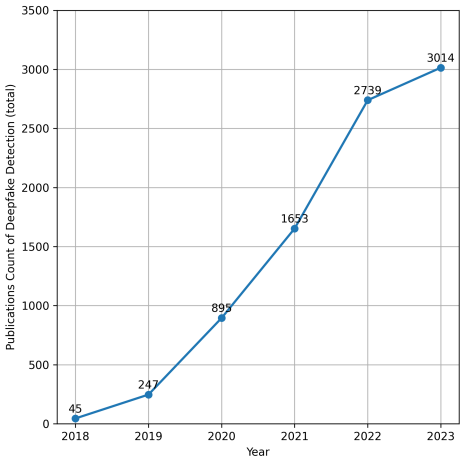
earth generate about 100 trillion words per day. All of this points to the fact that synthetic data will dominate the internet data side.

The use of synthetic data has been shown to significantly degrade the performance of deep learning models (see Fig.7), both for generation tasks and classification tasks Hataya et al. (2023); Ravuri & Vinyals (2019); Martínez et al. (2023); Shumailov et al. (2024); Bohacek & Farid (2023). Addressing the impact of synthetic data is crucial for the development of the next generation of models. There are two primary approaches to mitigating the negative effects of synthetic data. The first is exploring ways to better utilize synthetic data, proposing strategies for optimizing the integration of synthetic data into training pipelines Dohmatob et al. (2024a;b); Feng et al. (2024); Bertrand et al. (2024); He et al. (2023). The second approach involves developing methods to accurately detect synthetic data, allowing models to distinguish between real and synthetic inputs.

A.3 Increasing Attention on Synthetic Data Detection



(a) The number of BBC official news reports on deepfake topic over the years.



(b) The number of publications on deepfake detection over the years.(From Gong & Li (2024))

Figure 8: The rising concern of deepfakes in both media and academic research.

The growing prevalence of synthetic data has garnered increasing attention from society, including news reports, academic research, and government policies. The number of papers on deepfake detection has been steadily increasing, with the BBC reporting on deepfakes more frequently each year (see Fig.8). In response to the rise of synthetic data, several governments and global conferences have also introduced policies aimed at regulating the use of deepfakes and synthetic data AIS (2023); Michelle (2023); Biden (2023).

B DATASET DESCRIPTION

B.1 DATA COLLECTION

Our data primarily originates from online internet collections, reused from public datasets, and self-synthesized into new composite data, as detailed in Table 6. To ensure diversity in synthetic data, each modality incorporates more than five different synthesis methods (Figure 9 & 10). To guarantee the quality of synthetic data, we also collected samples synthesized by mature proprietary models such as Sora, Midjourney, CLAY, Suno, and GPT-4. The far-right column of the table displays the public datasets that underpin our collected synthetic or authentically paired data.

Table 6: Synthetic Methods and Public Datasets Across Modalities

Modality	Synthesis Methods	Public Datasets
Video	Sora (OpenAI, 2024), Keling, CoNo, Lumiere(Bar-Tal et al., 2024), Open-sora (Zheng et al., 2024), Runway, W.A.L.T (Gupta et al., 2023)	-
Image	FLUX, DDIM (Song et al., 2020a), Midjourney (AI, 2023), Stable Diffusion (V1.4,V1.5,V2.1) (Blattmann et al., 2023), DPM+ (Lu et al., 2023a), ADM(Dhariwal & Nichol, 2021a), Stylegan (Karras et al., 2019a), Skydiffusion (Ye et al., 2024), pix2pix (Isola et al., 2017), CUT (Park et al., 2020)	I2IQA(Yuan et al., 2023), Sentry(Lu et al., 2023c), GenImage(Zhu et al., 2023), FFHQ(Karras et al., 2019b), Stylegan3(Karras et al., 2021), CVUSA(Workman et al., 2015), ISBI 2016(Gutman et al., 2016), M3DSynth(Zingarini et al., 2024), M6Doc(Cheng et al., 2023), Deepfakeface(Song et al., 2023), VIGOR(Zhu et al., 2021)
3D	CLAY (Zhang et al., 2024a), SyncDreamer (Liu et al., 2023c), Magic3D (Lin et al., 2023), DreamFusion (Poole et al., 2022), Fantasia3D (Chen et al., 2023a), DreamGaussian (Tang et al., 2023), Wonder3D (Long et al., 2024), GaussianDreamer (Yi et al., 2023), GradeADreamer (Ukarapol & Pruvost, 2024)	OmniObject3D(Wu et al., 2023), GPTEval3D(Wu et al., 2024a)
Audio	Suno, WaveNet (Rethage et al., 2018), WaveRNN (Kalchbrenner et al., 2018), Tacotron2 (Shen et al., 2018), Hifi-GAN (Kong et al., 2020), AceSinger (Timedomain, 2023), Soft-VITS-SVC (svc-develop team, 2024), DiffSinger (Dhariwal & Nichol, 2021b), VQ-VAE (Van Den Oord et al., 2017), AudioLDM (Liu et al., 2023a), VITS (Kim et al., 2021), AudioLDM2 (Liu et al., 2024a), MusicGen (Copet et al., 2024)	ASVSpoof2019(Wang et al., 2020b), CtrSVDD(Zang et al., 2024), DCASE2023 Track 7(Choi et al., 2023), MusicCaps(Agostinelli et al., 2023)
Text	llama3.1-405B (Team, 2024) , GPT-4o (OpenAI, 2024), Qwen-Max (Bai et al., 2023), Mistral-Large (Jiang et al., 2023a), Claude-3.5-Sonnet (Anthropic, 2024), Gemini-1.5-Flash (Team et al., 2023)	TheDataBeast (TheDataBeast, 2021),Mixset(Zhang et al., 2024b), NLP_chinese_corpus(Xu, 2019), ghostbuster-data(Verma et al., 2024)

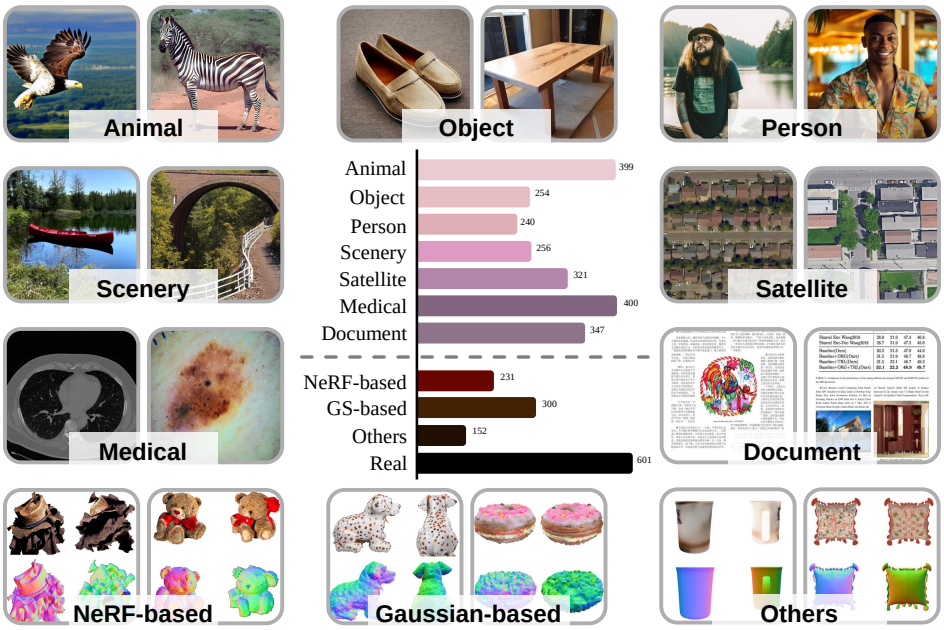


Figure 9: Examples of some 3D and image datasets, with the bar chart showing the quantity of data in different categories.

Authentic pairing data: We have collected a significant amount of authentic paired data from the internet, including sources such as arXiv, Wikipedia, Gutenberg, YouTube, TikTok, and Civitai. For data sourced from the internet, we will rigorously verify that it consists of authentic recordings or text authored by human users, rather than content synthesized using AI technology. It is important to note that our current research primarily focuses on multimedia data directly synthesized by AI, with limited consideration of methods like deepfake involving manual editing; we will continue to update our approach in future studies.

Data Availability and Social Impact: In collecting data, we strictly adhere to copyright and licensing regulations of the source websites, avoiding data acquisition from resources that prohibit copying or redistribution. For the LOKI dataset, which is open-sourced, users must submit a download request to the authors to prevent misuse of the data.

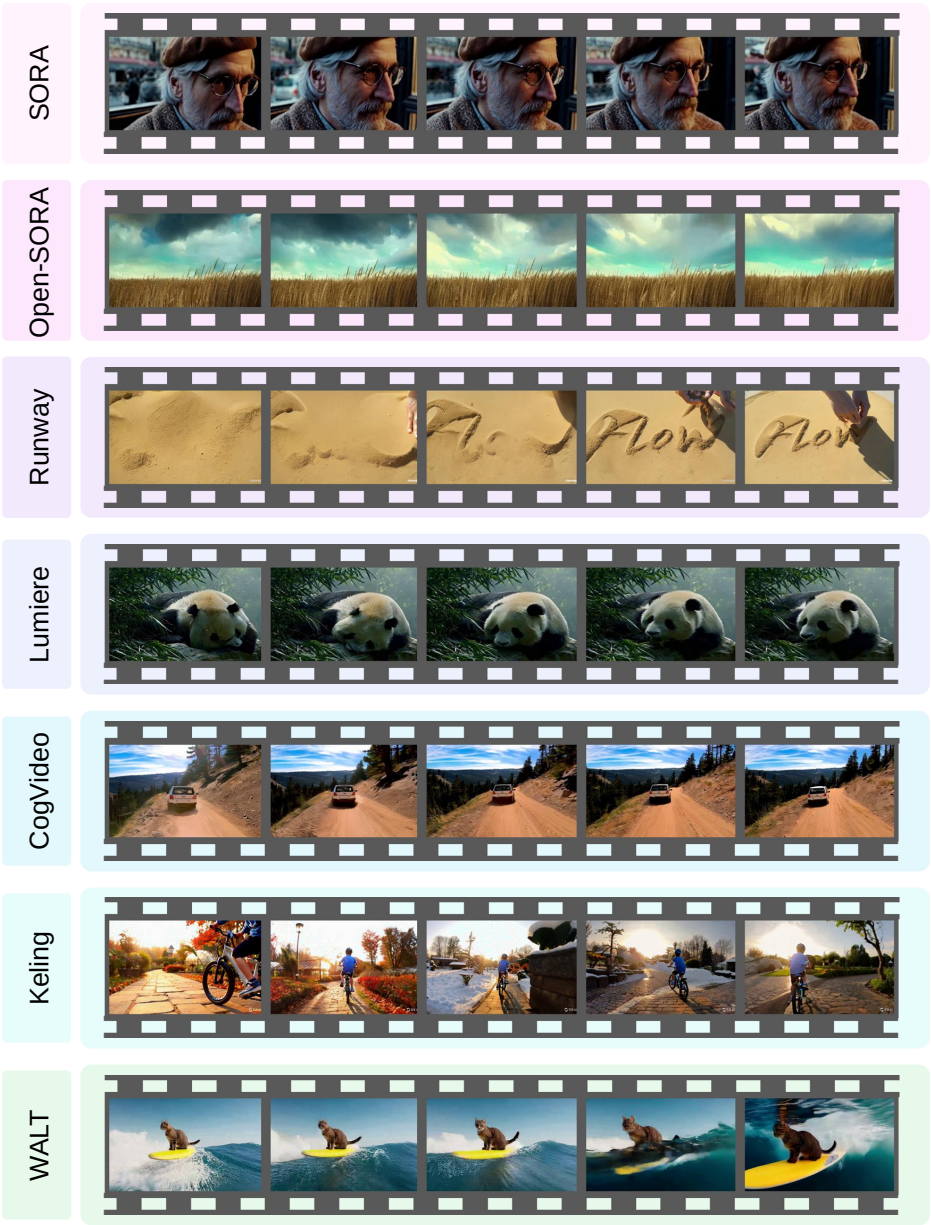


Figure 10: Examples of video data. We used 7 video generation models to obtain corresponding data for LMMs evaluation.

B.2 DATASET ANNOTATION

B.2.1 ANNOTATION GUIDELINES

Video: During the annotation of synthetic videos, we categorize the identified anomalies into two types: global anomalies and segment anomalies. Global anomalies refer to errors that persist for more than 80% of the video's duration, while segment anomalies are issues that occur for a limited portion of the video. For example, as shown in Fig. 11 (a), the anomaly of "flickering textures and distorted geometries of fences and utility poles," which is present throughout the video, is labeled as a global anomaly. In contrast, the "abnormal flames" and "basketball penetrating the hoop" in the video are classified as segment anomalies. Additionally, each identified anomaly, including both global and segment anomalies, is associated with a key frame that represents the anomaly, facilitating subsequent processing of video data by large multimodal models (LMMs).

Image: For the synthetic image data, we provide global anomaly annotations for overall image issues, as well as bounding box selections and textual descriptions for abnormal regions. The bounding boxes indicate the location and extent of the abnormal areas within the image, while the textual descriptions detail the specific anomalies present in those regions. As shown in Fig. 11 (b), the "texture quality issues" and "color distortion" in the image are annotated as global anomalies, whereas area errors such as the "texture errors" of the duck and "reflection anomalies" on the water surface are classified as region anomalies. Annotators mark these areas by drawing bounding boxes and provide textual explanations for the reasons behind the anomalies.

3D Data: Unlike video and image annotations, the annotation of 3D data involves a global-scale analysis of textures and normals. In terms of texture anomalies, we focus on assessing the authenticity, smoothness, and edge clarity of the textures. For normals, we analyze whether the model's geometric fluidity, surface smoothness, physical stability, and topological coherence are accurately represented. For instance, as shown in Fig. 11 (c), we conduct a detailed analysis of the "multiview discrepancies" and "texture blurriness" in the model's textures, while labeling issues such as "abnormal protrusions" and "asymmetrical structures" as related to normals, accompanied by appropriate textual explanations.

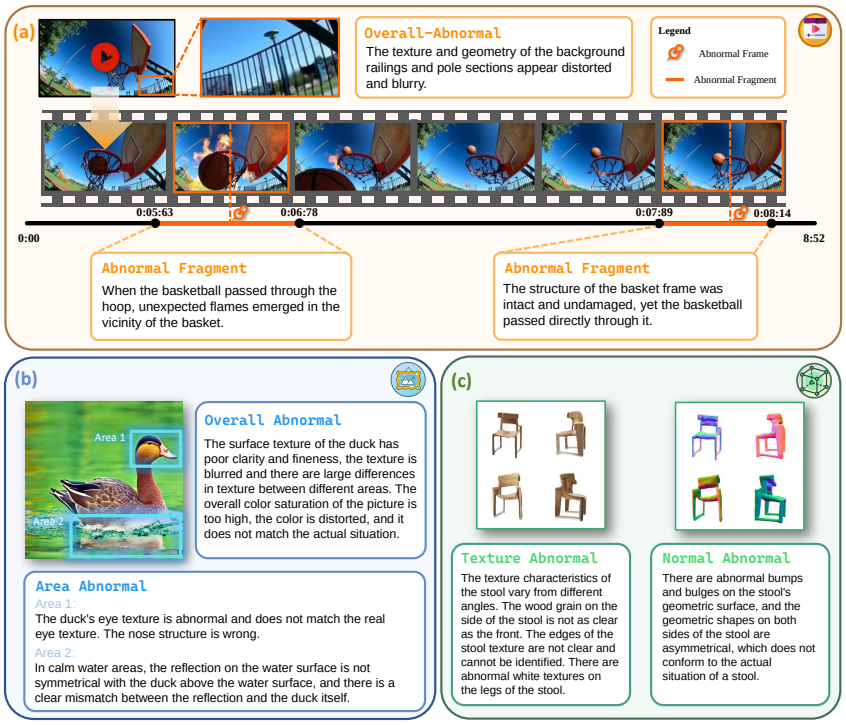


Figure 11: Examples of the synthetic data annotation process under different modalities, including (a) Video, (b) Image, (c) 3D Data.

B.2.2 ANNOTATOR INFORMED CONSENT

Before commencing the annotation process, we ensure that all participating annotators are fully informed and provide their explicit agreement to the following terms and conditions. This comprehensive informed consent is designed to promote transparency, respect their autonomy, and align with ethical standards in research. It is imperative that each annotator has a thorough understanding of the nature, purpose, and potential implications of their contributions to the labeling process. The terms are outlined as follows:

Data Usage. Annotators acknowledge and consent to the possibility that the labeled data they generate may be used in various academic and scientific contexts, including the development of research papers, presentations at conferences, and other related scholarly activities. They understand that their work may significantly contribute to advancements in research fields such as natural language processing, machine learning, and artificial intelligence. Furthermore, annotators recognize that their contributions may be referenced or cited in scientific publications, thereby playing a role in shaping future research directions and applications.

AI-Generated Content. Annotators are informed that some of the content they will be labeling may have been produced by artificial intelligence models. This includes text, images, or other data types generated by algorithms designed to simulate human-like outputs. Annotators understand that this knowledge is crucial, as it may influence their perception, judgment, and approach to the labeling task. They agree to remain mindful of the potential biases or preconceived notions that may arise from this awareness and commit to maintaining objectivity and accuracy in their work.

Potential Implications. Annotators are aware of the broader implications of their labeling activities, which extend beyond the immediate scope of data annotation. They recognize the ethical considerations inherent in AI research, particularly concerning issues such as bias, fairness, and the societal impact of deploying AI technologies. Annotators agree to reflect on these ethical dimensions and to engage in the labeling process with a conscientious approach, acknowledging that their work may contribute to both the positive advancements and challenges associated with AI development and implementation.

Commitment to Ethical Standards. By agreeing to these terms, annotators affirm their commitment to upholding high ethical standards throughout the annotation process. They understand that their participation is voluntary and that they have the right to withdraw from the project at any time, without penalty. Annotators also acknowledge their responsibility to report any concerns or issues that may arise during the labeling process, ensuring the integrity and reliability of the data they provide.

This informed consent process ensures that all annotators are equipped with a comprehensive understanding of their role and its significance. It aims to foster an environment of mutual respect and collaboration, where the contributions of annotators are valued and their rights as participants in research are protected. By clarifying the expectations and responsibilities involved, we seek to create a foundation for ethical and impactful research that benefits both the scientific community and society at large.

B.3 QUALITY CONTROL AND VALIDATION

In annotating videos, images, and 3D synthetic data for anomaly details, we maintain high standards and accuracy. All annotators possess at least a university degree and demonstrate strong decision-making and judgment skills. Before annotation, human annotators receive extensive training with numerous examples of common errors to ensure a comprehensive understanding of the synthetic data detection task. Each data instance is annotated for detailed anomalies by at least two human annotators to ensure quality. Ambiguous or unclear instances are marked for further study and annotation during team meetings. Furthermore, to avoid the impact of hallucinations from Large Language Models (LMMs) on tasks involving LMMs, all anomaly detail explanation tasks undergo manual review based on LabelLLM [3].

[3] https://github.com/opendatalab/LabelLLM

B.4 SPECIAL DATA DESCRIPTION

In the field of **document images**, we have collected synthesized images in four categories: newspapers, academic papers, magazines, and reconstructed documents. The corresponding real data for these categories come from the M6Doc dataset. Currently, document synthesis primarily follows a layout-first, content-rendering-later approach. For the first three document types, we design specific empirical rules for layout generation during the layout phase; during the content rendering phase, elements are selected from a constructed corpus to fill the structure. For the reconstructed type, we employ a restructuring algorithm on the M6Doc dataset to rearrange the content of the documents.

In the field of **remote sensing** imagery, synthetic data is primarily generated based on street-to-satellite datasets such as CVUSA and VIGOR, utilizing methods like CUT and Skydiffusion. The synthetic imagery encompasses major natural scenes such as urban and suburban environments, with the satellite remote sensing images being of high resolution.

In the field of **medical** imaging, we primarily collected two types of data: the ISIC 2016 skin dataset and the M3DSynth CT dataset. For the ISIC 2016 dataset, we utilized GAN methods for direct data synthesis. The M3DSynth dataset comprises synthetic images generated from the real-world LIDC dataset using Diffusion and GAN models. The images are categorized into two types: those with real tumors removed and those with synthetic tumors artificially inserted by the model. Each synthetic image is paired with its corresponding original image, complete with precise annotations of the tumor insertion or removal locations. Considering that most users are not medically trained, we deliberately selected more evident abnormal images to reduce the need for specialized knowledge in making decisions about synthetic data.

C EVALUATION

C.1 EVALUATION MODEL

We compared various models on the LOKI benchmark to understand their capabilities across multiple tasks. We support over ten open-source models, including InternVL2 (Chen et al., 2024), LLaVA (Liu et al., 2023b), Phi (Gunasekar et al., 2023), XComposer (Zhang et al., 2023b), Qwen2-VL (Wang et al., 2024), MiniCPM (Hu et al., 2024), and Idefics2 (Laurençon et al., 2024), as well as proprietary models such as GPT-4 (OpenAI, 2024), Gemini (Team et al., 2023), Qwen-VL-Max (Bai et al., 2023), and Claude (Anthropic, 2024). The following list details these models.

Model Family	Model Version	Parameters	Links
Close-sourced, API			
GPT4	GPT-4o	N/A	https://platform.openai.com/docs/models/gpt-4o
	GPT-4	N/A	https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
Gemini	Gemini-1.5-Pro	N/A	https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-pro
	Gemini-1.5-Flash	N/A	https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash
Claude	Claude-3.5-Sonnet	N/A	https://docs.anthropic.com/en/docs/about-claude/models
Mistral	Mistral-Large	N/A	https://docs.mistral.ai/getting-started/models/
Qwen	Qwen-Max	N/A	https://www.alibabacloud.com/help/en/model-studio/developer-reference/use-qwen-by-calling-api
Open-sourced			
LLaMA	LLaMA-3.1-405B	405B	https://huggingface.co/meta-llama/Llama-3.1-405B
InternVL2	InternVL2-8B	8B	https://huggingface.co/OpenGVLab/InternVL2-8B
	InternVL2-26B	26B	https://huggingface.co/OpenGVLab/InternVL2-26B
	InternVL2-40B	40B	https://huggingface.co/OpenGVLab/InternVL2-40B
	InternVL2-Llama3-76B	76B	https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B
LLaVA-OneVision	LLaVA-OneVision-7B	7B	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
	LLaVA-OneVision-72B	72B	https://huggingface.co/lmms-lab/llava-onevision-qwen2-72b-ov-sft
VILA	VILA-1.5-13B	13B	https://huggingface.co/Efficient-Large-Model/VILA1.5-13b
	VILA-1.5-40B	40B	https://huggingface.co/Efficient-Large-Model/VILA1.5-40b
Phi	Phi-3.5-Vision	3.5B	https://huggingface.co/microsoft/Phi-3.5-vision-instruct
Idefics2	idefics2-8b	8B	https://huggingface.co/HuggingFaceM4/idefics2-8b
Qwen2-VL	Qwen2-VL-7B	7B	https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
	Qwen2-VL-72B	72B	https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct
InternLM-XComposer	InternLM-XComposer-2d5	7B	https://huggingface.co/internlm/internlm-xcomposer2d5-7b
mPLUG-Owl3	mplug-owl3	7B	https://huggingface.co/mPLUG/mPLUG-Owl3-7B-240728
MiniCPM	MiniCPM-V2.6	8.1B	https://huggingface.co/openbmb/MiniCPM-V-2_6
LongVILA	LongVILA	8B	https://huggingface.co/Efficient-Large-Model/Llama-3-LongVILA-8B-128Frames
LongVA	LongVA-7B	7B	https://huggingface.co/lmms-lab/LongVA-7B-DPO
Qwen-Audio	Qwen-Audio-Chat	7B	https://huggingface.co/Qwen/Qwen-Audio-Chat
SALMONN	SALMONN-7B	7B	https://huggingface.co/tsinghua-ee/SALMONN-7B
AnyGPT	AnyGPT-Chat	7B	https://huggingface.co/fnlp/AnyGPT-chat
OneLLM	OneLLM-7B	7B	https://huggingface.co/csuhan/OneLLM-7B
LTU	LTU-AS-7B	7B	https://github.com/YuanGongND/ltu#pretrained-models

C.2 EVALUATION METRIC

Average accuracy: For judgment, multiple-choice, and detailed selection questions, we use the *average accuracy* as the primary metric. The accuracy rate is calculated using the following formula:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\%$$

In this context, N_{correct} is the number of correctly answered questions, and N_{total} is the total number of questions. To minimize the influence of prompts on model judgments, each question is presented in two forms: one asks whether the data is AI-synthesized or real, and the other asks the model to identify either the real or AI-synthesized data. By averaging the accuracy rates across different forms of the questions, we aim to reduce the potential bias introduced by the phrasing of prompts and ensure a fair evaluation of the model's performance.

Normalized Bias Index (NBI): To evaluate whether there is potential bias in existing models when determining authenticity on the LOKI benchmark, we introduce a metric termed the Normalized Bias Index (NBI) to quantify the performance differences of the model on natural and AI-generated data across different modalities, which is defined as follows:

$$\text{NBI} = \frac{R_{\text{natural}} - R_{\text{generated}}}{R_{\text{natural}} + R_{\text{generated}}} \in [-1, 1]$$

In this context, R_{natural} and $R_{\text{generated}}$ represent the recall rates for natural and AI-generated samples, respectively, under the corresponding modality. By normalizing the difference between the two, the model's unexpected preference in making predictions can be quantified. Specifically, a positive and larger NBI indicates that the model is more biased toward predicting samples as natural, whereas a negative and smaller NBI suggests a bias toward predicting samples as AI-generated.

GPT-Score: For open-ended questions regarding anomalous details, we use the GPT-4 model to assess the score of the responses. We adopted a rating scheme, establishing a 5-level rating system with scores ranging from 1 (poor) to 5 (excellent). The final scores are normalized to a scale of 0 to 100. We adhere to the following scoring criteria:

1) Identification: Accurately detect the globally annotated anomalies and their corresponding detailed anomalous regions specified by human annotators.

2) Explanation: Provide accurate explanations for the causes of the anomalies, ensuring consistency with the reasons outlined in the human annotations.

3) Plausibility: Avoid misclassifying authentic regions as anomalous while encouraging other reasonable explanations for anomalies.

While the scoring criteria are similar across different modalities, they are slightly adjusted according to their content characteristics; for example, the image modality is subdivided into global score and regional score, whereas 3D data is subdivided into texture score and normal score.

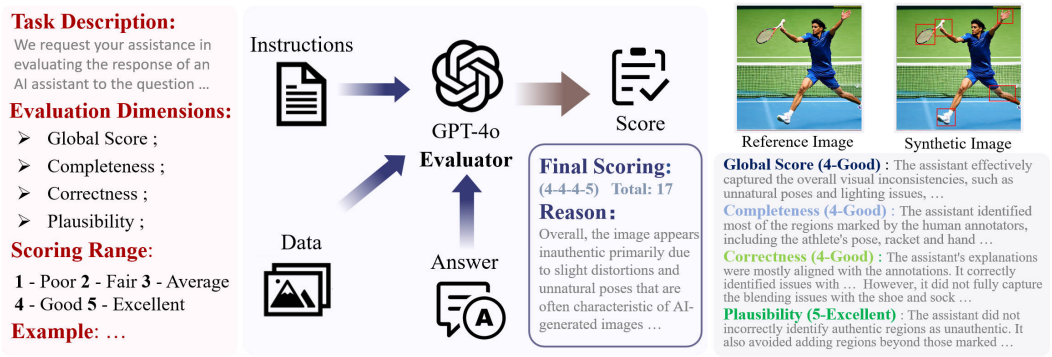


Figure 12: The overall process for automated evaluation of Abnormal explanation questions using GPT-4o.

D BREAKDOWN RESULTS ON DIFFERENT MODALITIES

In this appendix, we present detailed results of various synthesis methods or classification themes across different modalities of data.

D.1 VIDEO

Tables 7 and 8 display the evaluation results of LMMs in video modality for judgment and multiple-choice questions, respectively. In the evaluation of video modality, the proprietary model GPT-4o achieved an accuracy of 71.3% for judgment questions and 77.3% for multiple-choice questions. The open-source model InternVL2-40B recorded accuracies of 62.0% for judgment questions and 65.7% for multiple-choice questions. Table 9 presents the evaluation results of different models in the video modality for open-ended questions.

Additionally, using GPT-4's detection outcomes as a benchmark, we assessed the quality of videos synthesized by different models. It was found that videos generated by Sora had a judgment accuracy of 67.2% and a multiple-choice accuracy of 64.5%, indicating a high quality of video synthesis. Conversely, videos synthesized by Runway and W.A.L.T exhibited more noticeable synthetic traces.

D.2 IMAGE

Tables 10 and 11 present the evaluation results of LMMs in the image modality for judgment and multiple-choice questions, broken down by specific categories. The proprietary model GPT-4o scored a judgment accuracy of 63.4%, primarily impacted by specialized image types such as Satellite, Doc, and Medicine, while achieving a multiple-choice accuracy of 80.8%. Among open-source models, Qwen2-VL-72B performed exceptionally well, even surpassing GPT-4o in specialized image categories. Table 12 presents the evaluation results of different models in the Image modality for open-ended questions.

D.3 3D

Table 13 presents the evaluation results for different 3D synthesis methods. Tests on judgment questions for 3D models reveal that most open-source models perform close to a 50% random chance. In contrast, the proprietary model GPT-4o demonstrates better decision-making and judgment capabilities, particularly in identifying low-quality Nerf models, where it achieves an accuracy rate exceeding 75Assuming the performance of GPT-4o as a benchmark, we can assess the quality of 3D data synthesized using different methods. Among three synthesis methods, Gaussian synthesis exhibits higher quality compared to Nerf synthesis, while commercial models like Clay offer the best quality. Table 14 presents the evaluation results of different models in the Image modality for open-ended questions.

As shown in Table 15, we extracted 200 3D objects, with an equal ratio of real to synthetic data, and evaluated the performance of different 3D input formats, including point clouds, surround videos with eight input frames, and multi-view inputs. In comparison, for current LMMs, the performance between surround video and multi-view inputs is similar. PointLLM, which partially supports point cloud inputs, performs particularly well, indicating that the point cloud format still holds certain advantages. Additionally, our results indicate that compared to the 7B PointLLM, the 13B PointLLM is more inclined to respond with "I cannot determine whether this is a real object or one generated by AI" rather than a straightforward "Yes" or "No." We speculate that as the size of PointLLM increases, it incorporates more comprehensive measures to correct illusions, making it less likely to provide definitive answers when faced with uncertainties (Wei et al., 2022a).

D.4 AUDIO

Table 16 presents the evaluation results of various audio LMMs on the LOKI dataset. Currently, there are few open-source and proprietary LMMs that support the audio modality, and most models perform poorly on synthetic data detection tasks, showing little difference from random selection. However, the AASIST method, trained specifically for speech tasks, demonstrates relatively strong performance, although it still underperforms in other modalities. Overall, human performance in

the audio modality is comparatively weak relative to other modalities, yet it still exceeds the best-performing LMMs by 20%.

It has been shown that acoustic features, such as pauses between words and silent segments within audio, play a more critical role in detecting deepfake audio than the actual content of the audioLiu et al. (2024b). However, current audio language models are primarily focused on content comprehension rather than acoustic characteristics. From the model architecture perspective, audio language models typically initialize their audio encoders with pre-trained audio models such as WhisperRadford et al. (2023), ASTGong et al. (2021), or BEATsChen et al. (2023b). These pre-trained models are trained with various data augmentation techniques, including SpecAugmentPark et al. (2019) and SpecSwapSong et al. (2020b). The target of the augmented data remains constant, often focused on tasks like transcribing textual content, which shifts the model's attention towards understanding content rather than being sensitive to acoustic features. On the training data side, most audio-text datasets to date emphasize content comprehension in tasks like speech recognition, emotion/age recognition, audio/music captioning, sound event classification and music genre recognition. As a result, these models lack the inherent capability to distinguish between real and fake audio based on acoustic cues.

D.5 TEXT

Tables 17 and 18 present the test results across different genres in the text modality. Most models underperform in this modality, primarily due to the low distinctiveness of text data itself, coupled with the maturity of current text synthesis technologies. Notably, genres such as modern literature and philosophy may benefit from the extensive training and memory capabilities of GPT series models, showing relatively better performance compared to other categories.

In Table 19, we evaluated the detection results for texts of varying lengths. The findings indicate that longer texts are detected more frequently and are more likely to reveal flaws. Additionally, we assessed the performance across different languages and found that, except for GPT-4o, most proprietary models perform significantly better in English than in Chinese. This indicates that the multilingual capabilities of large multimodal models require further development.

Table 7: **Judgement** questions results of different models on the LOKI **Video** modality. * denotes the closed-source models.

	Overall	**Sora**	**Keling**	**CogVideoX**	**Lumiere**	**Open-sora**	**Runway**	**W.A.L.T**
Expert (AIGVDet)	52.1	39.5	61.5	65.0	54.2	56.0	37.1	61.0
MiniCPM-V-2.6	57.2	57.9	56.5	54.0	66.4	55.2	45.8	62.6
Phi-3.5-Vision	56.8	52.7	52.0	53.0	57.1	57.5	55.1	66.5
LLaVA-OneVision-7B	56.8	57.6	53.7	61.0	57.1	56.6	53.7	59.6
InternLM-XComposer2.5	58.4	55.2	61.1	63.0	60.0	57.1	49.5	62.6
mPLUG-Owl3-7B	55.3	56.7	54.3	54.0	56.4	54.2	45.4	61.3
Qwen2-VL-7B	59.5	60.1	58.5	54.0	59.3	51.4	59.7	65.9
LongVA-7B	60.4	66.2	51.4	65.0	60.0	55.7	73.6	57.7
Mantis-8B	55.4	48.8	50.0	57.0	52.9	51.9	66.7	62.4
Idefics2-8B	54.8	56.1	54.0	51.0	57.9	51.4	41.2	64.3
InternVL2-8B	60.8	61.9	56.3	69.0	59.3	61.3	56.5	64.8
Llama-3-LongVILA-8B	51.9	54.0	49.1	53.0	52.1	48.1	52.8	54.1
VILA1.5-13B	51.9	51.2	50.9	51.0	54.3	52.8	52.3	51.9
InternVL2-26B	55.0	55.2	53.4	61.0	57.9	52.4	50.0	58.2
VILA1.5-40B	59.2	54.9	57.4	66.0	61.4	59.9	54.2	64.8
InternVL2-40B	62.0	61.3	57.7	68.0	60.7	61.8	62.0	65.9
Qwen2-VL-72B	59.6	58.2	61.4	59.0	58.6	54.7	58.8	63.2
LLaVA-OneVision-72B	56.5	57.0	55.4	62.0	55.7	51.9	46.8	64.6
Gemini-1.5-Pro*	58.5	53.4	60.5	58.0	63.6	62.3	45.4	64.8
Claude-3.5-Sonnet*	61.7	60.1	57.4	59.0	68.6	64.2	49.5	71.2
GPT-4o*	71.3	66.8	67.9	70.0	68.6	72.6	76.9	75.8

Table 8: **Multiple Choice** questions results of different models on the LOKI **Video** modality. * denotes the closed-source models.

	Overall	Sora	Keling	CogVideoX	Lumiere	Open-sora	Runway	W.A.L.T
MiniCPM-V-2.6	52.8	39.5	62.1	55.6	58.8	54.0	48.0	53.7
Phi-3.5-Vision	58.2	57.9	42.4	50.0	70.6	58.0	68.0	58.5
LLaVA-OneVision-7B	59.8	55.3	62.1	63.9	66.2	56.0	64.0	54.9
InternLM-XComposer2.5	56.3	43.4	54.5	66.7	67.6	62.0	54.0	53.7
mPLUG-Owl3-7B	60.3	53.9	53.0	61.1	66.2	62.0	62.0	64.6
Qwen2-VL-7B	64.0	63.2	62.1	52.8	69.1	66.0	70.0	62.2
LongVA-7B	57.5	59.2	50.0	58.3	63.2	52.0	58.0	59.8
Mantis-8B	47.9	56.6	43.9	50.0	52.9	52.0	42.0	39.0
Idefics2-8B	55.6	56.6	54.5	55.6	57.4	58.0	52.0	54.9
InternVL2-8B	54.0	53.9	57.6	55.6	61.8	48.0	52.0	48.8
Llama-3-LongVILA-8B	54.0	50.0	47.0	55.6	61.8	54.0	52.0	57.3
VILA1.5-13B	52.1	51.3	53.0	55.6	60.3	54.0	50.0	43.9
InternVL2-26B	62.4	61.8	57.6	66.7	63.2	56.0	68.0	64.6
VILA1.5-40B	49.1	50.0	48.5	47.2	54.4	50.0	42.0	48.8
InternVL2-40B	65.7	53.9	62.1	77.8	76.5	56.0	66.0	70.7
Qwen2-VL-72B	65.7	57.9	62.1	72.2	72.1	60.0	80.0	62.2
LLaVA-OneVision-72B	62.9	52.6	48.5	66.7	69.1	74.0	78.0	61.0
Gemini-1.5-Pro*	66.1	46.1	60.6	66.7	75.0	64.0	72.0	79.3
Claude-3.5-Sonnet*	60.5	53.9	54.5	58.3	72.1	72.0	58.0	57.3
GPT-4o*	77.3	61.8	71.2	75.0	94.1	78.0	82.0	80.5

Table 9: **Abnormal Explanation** questions results of different models on the LOKI **Video** modality. * denotes the closed-source models.

Model	Overall	Correctness	Explanation	Plausibility
LLaVA-OV-7B	46.7	41.0	48.4	50.8
InternVL2-8B	46.5	38.0	48.4	53.2
Qwen2-VL-7B-Instruct	48.4	41.2	50.4	53.6
Claude-3-5-sonnet*	50.1	44.2	48.8	57.4
GPT-4o*	67.6	60.6	69.2	73.0

Table 10: **Judgment** questions results of different models on the LOKI **Image** modality. * denotes the closed-source models.

	Overall	Scene	Animal	Person	Object	Medicine	Doc	Satellite
Expert (AIDE)	63.1	-	89.9	62.5	96.5	53.4	49.7	39.3
MiniCPM-V-2.6	44.8	52.0	34.4	53.1	31.5	53.8	51.5	38.3
Phi-3.5-Vision	52.5	50.8	41.7	71.5	34.1	57.3	54.3	60.5
LLaVA-OneVision-7B	49.8	59.2	41.9	58.1	37.3	52.3	53.0	50.1
InternLM-XComposer2.5	46.4	52.7	40.0	56.7	32.5	56.1	49.8	38.2
mPLUG-Owl3-7B	45.9	52.1	37.3	52.9	31.4	55.3	53.8	38.1
Qwen2-VL-7B	47.8	54.7	38.9	57.9	30.3	56.0	59.6	36.9
LongVA-7B	46.2	57.6	37.4	52.5	34.1	54.4	49.8	39.7
Mantis-8B	54.6	54.9	52.2	54.8	53.5	53.1	51.9	63.3
Idefics2-8B	45.0	51.8	35.3	52.3	29.2	52.3	53.9	40.6
InternVL2-8B	49.7	58.8	39.4	54.4	37.8	53.9	60.2	44.2
Llama-3-LongVILA-8B	49.8	49.8	50.5	50.6	47.2	50.0	49.9	50.0
VILA1.5-13B	49.3	52.0	38.6	54.2	31.0	50.1	56.6	62.4
InternVL2-26B	44.3	51.6	35.4	50.8	28.2	51.3	54.4	37.6
VILA1.5-40B	48.8	53.7	39.3	50.0	33.4	52.5	59.9	50.6
InternVL2-40B	49.6	55.7	37.3	59.2	34.8	55.5	64.8	40.8
Qwen2-VL-72B	53.2	55.9	43.4	66.9	38.0	55.9	73.7	38.2
LLaVA-OneVision-72b	46.3	54.7	31.6	53.1	27.8	52.1	67.9	36.6
Claude-3.5-Sonnet*	53.6	51.6	51.6	55.2	51.4	51.9	59.1	50.9
Gemini-1.5-Pro*	43.5	53.7	35.7	51.5	30.3	50.0	47.2	38.1
GPT-4o*	63.4	70.1	69.7	84.4	70.3	54.3	60.1	45.0

Table 11: **Multiple Choice** questions results of different models on the LOKI **Image** modality. * denotes the closed-source models.

	Overall	Scene	Animal	Person	Object	Medicine	Doc	Satellite
MiniCPM-V-2.6	49.8	49.2	52.3	49.2	48.5	50.5	45.1	51.6
Phi-3.5-Vision	44.0	31.3	22.9	67.9	20.7	69.8	58.3	52.1
LLaVA-OneVision-7B	51.7	55.9	51.8	63.3	46.9	52.8	51.4	46.3
InternLM-XComposer2.5	51.0	48.4	49.5	46.3	50.5	57.0	52.0	51.2
mPLUG-Owl3-7B	52.5	53.1	53.5	53.8	49.5	55.5	49.4	52.3
Qwen2-VL-7B	65.1	60.9	71.3	70.4	60.5	68.0	63.1	59.6
LongVA-7B	51.6	58.2	54.9	42.9	48.2	55.3	46.6	51.6
Mantis-8B	61.5	66.0	58.1	57.1	56.9	62.3	68.3	63.6
Idefics2-8B	51.3	47.3	55.3	56.3	55.4	46.5	48.0	49.3
InternVL2-8B	51.4	56.6	51.2	54.2	48.5	48.8	50.6	52.6
Llama-3-LongVILA-8B	51.1	48.4	49.8	47.5	51.0	54.3	53.1	52.1
VILA1.5-13B	55.3	55.5	57.4	58.3	56.9	57.5	54.9	47.7
InternVL2-26B	48.5	50.4	49.1	50.8	40.3	50.8	53.1	46.7
VILA1.5-40B	64.0	59.8	71.0	58.8	53.8	71.5	62.9	63.8
InternVL2-40B	63.1	60.9	72.0	62.5	62.5	62.3	61.6	56.8
Qwen2-VL-72B	68.6	60.9	71.1	60.8	68.9	74.3	77.7	61.0
LLaVA-OneVision-72b	70.8	65.6	62.3	76.7	67.1	73.0	77.6	72.4
Claude-3.5-Sonnet	65.5	58.6	53.3	75.8	77.8	70.5	68.6	59.3
Gemini-1.5-Pro*	67.3	72.7	70.1	74.2	69.4	55.8	64.0	68.0
GPT-4o*	80.8	79.7	92.3	91.3	94.9	65.3	81.7	61.0

Table 12: **Abnormal Explanation** questions results of different models on the LOKI **Image** modality. * denotes the closed-source models.

	Overall	Global Score	Completeness	Correctness	Plausibility
Qwen2-VL-7B-Instruct	63.8	66.6	60.8	53.2	74.4
Llava-OV-7B	68.9	72.4	63.0	60.6	79.8
InternVL2-8B	72.2	70.0	67.6	61.2	85.2
Claude-3-5*	1.7	2.0	1.6	1.4	1.8
Gemini-1.5-pro*	77.1	77.6	70.8	70.0	90.2
GPT-4o*	72.9	69.8	68.6	63.4	90.0

Table 13: **Judgment & Multiple Choice** questions results of different models on the LOKI **3D** modality. * denotes the closed-source models.

	Judgment				Multiple Choice			
	Nerf	Gaussian	Other	Overall	Nerf	Gaussian	Other	Overall
MiniCPM-V-2.6	56.7	56.1	56.7	56.4	50.6	50.0	52.9	50.7
Phi-3.5-Vision	50.0	50.0	50.0	50.0	62.0	56.1	62.9	59.6
LLaVA-OneVision-7B	69.3	67.4	69.0	68.4	52.5	54.6	54.6	53.8
InternLM-XComposer2.5	42.8	44.6	44.4	43.9	41.2	51.0	55.4	48.0
mPLUG-Owl3-7B	49.1	50.5	49.8	49.9	46.6	51.7	52.5	49.9
Qwen2-VL-7B	**72.1**	**73.2**	**70.4**	**72.3**	53.2	57.6	55.0	55.5
LongVA-7B	49.6	50.0	50.0	49.9	**63.2**	59.9	61.3	61.4
Mantis-8B	50.0	50.0	50.0	50.0	51.3	**73.0**	60.4	**62.5**
Idefics2-8B	38.5	38.1	39.0	38.4	51.7	56.5	54.2	54.2
InternVL2-8B	48.8	49.8	49.6	49.4	52.7	54.3	51.3	53.1
Llama-3-LongVILA-8B	32.3	32.1	32.5	32.2	51.5	49.8	50.0	50.5
VILA1.5-13B	32.3	35.5	33.8	34.0	49.0	57.3	53.8	53.5
InternVL2-26B	50.3	50.4	50.4	50.4	44.8	51.2	51.7	48.8
VILA1.5-40B	50.0	50.0	50.0	50.0	44.1	50.2	50.4	47.9
InternVL2-40B	49.8	50.0	50.0	49.9	54.0	63.6	63.8	59.9
Qwen2-VL-72B	60.9	60.3	59.0	60.3	56.5	61.1	57.5	58.7
LLaVA-OneVision-72B	53.1	49.6	51.9	51.3	63.0	58.3	**65.0**	61.3
Claude-3.5-Sonnet*	59.5	57.1	57.1	58.0	52.3	52.0	50.8	51.9
Gemini-1.5-Pro*	60.5	50.7	56.0	55.4	64.3	56.0	61.7	60.2
GPT-4o*	**65.9**	**64.4**	**65.4**	**65.2**	**78.8**	**66.4**	**60.8**	**70.2**
Human	87.3	78.4	71.8	82.0	94.5	90.2	78.6	91.3

Table 14: **Abnormal Explanation** questions results of different models on the LOKI **3D** modality. * denotes the closed-source models.

	Overall	Texture Correctness	Normal Correctness	Texture Plausibility	Normal Plausibility
Llava-OV-7B	71.0	63.2	70.8	73.8	76.2
InternVL2-8B	71.3	62.4	69.4	76.4	76.8
Qwen2-VL-7B-Instruct	73.4	64.8	72.2	77.8	78.6
Claude-3-5-sonnet*	78.2	70.0	80.0	78.8	83.8
Gemini-1.5-pro*	70.8	62.4	67.2	77.2	76.4
GPT-4o*	77.0	69.8	64.2	82.4	81.4

Table 15: **Judgment & Multiple Choice** questions results of different models on the LOKI **3D** modality. S-Video denotes Surround Video, MV-Image denotes Multi-Views Image.

	Judgment		Multiple Choice	
	S-Video	MV-Image	S-Video	MV-Image
LLaVA-OneVision-7B	67.4	69.0	52.5	54.6
Idefics2-8B	38.1	39.0	51.7	56.5
LLaVA-OneVision-72B	49.6	51.9	63.0	58.3

	Judgment
	PointCloud
PointLLM-7B	61.9
PointLLM-13B	16.9
OneLLM	55.6

Table 16: **Overall** results of different models on the LOKI **Audio** modality.

	Judgment					Multiple Choice				
	Speech (800)	**Singing (800)**	**Audio (280)**	**Music (400)**	**Overall (2280)**	**Speech (400)**	**Singing (400)**	**Audio (140)**	**Music (300)**	**Overall (1240)**
AASIST Jung et al. (2022)	96.4	54.4	57.1	54.3	69.4	-	-	-	-	-
Qwen-Audio	50.0	49.8	49.8	49.7	49.8	50.8	48.0	58.9	48.0	50.1
SALMONN-7B	52.8	52.4	49.3	46.8	51.2	-	-	-	-	-
AnyGPT	48.5	52.4	51.4	45.9	49.8	50.8	48.0	58.2	49.0	50.3
OneLLM	50.3	50.2	50.9	47.9	49.9	-	-	-	-	-
LTU	40.2	45.3	50.5	46.9	44.4	-	-	-	-	-
Qwen-Audio2	49.4	40.0	49.6	27.5	42.3	49.3	49.3	48.6	51.3	49.7
Gemini-1.5-Flash Team et al. (2023)	50.1	48.3	47.3	49.8	49.4	48.6	49.5	48.6	49.7	49.2

Table 17: **Judgment** questions results of different models on the LOKI **text** modality. * denotes the closed-source models.

	Scientific Papers	**News**	**Essay**	**Wikipedia**	**Speech**	**Modern Literature**	**Phiosophy**	**Ancient Chinese**
Expert (RADAR-Vicuna-7B)	76.2	81.4	90.0	67.2	73.9	63.3	79.45	1.7
MiniCPM-V-2.6	51.4	47.7	49.8	50.8	49.4	50.5	49.4	43.5
Phi-3.5-Vision	48.1	50.3	47.9	47.5	48.2	53.1	50.7	50.0
LLaVA-OneVision-7B	54.5	50.3	55.1	55.8	52.7	57.0	51.2	42.1
InternLM-XComposer2.5	51.9	52.5	55.8	48.2	49.7	55.7	54.9	52.2
mPLUG-Owl3-7B	51.3	50.5	53.4	52.3	53.2	60.6	59.5	41.9
Qwen2-VL-7B	49.9	48.3	49.6	51.4	48.1	50.5	48.0	42.1
LongVA-7B	51.0	48.8	48.0	49.8	47.2	50.2	44.8	49.0
Mantis-8B	51.0	51.8	51.5	52.8	53.5	54.9	52.7	43.2
Idefics2-8b	49.2	46.4	45.5	45.4	44.0	52.4	49.4	43.1
InternVL2-8B	51.6	49.4	45.6	49.1	47.1	52.7	54.7	54.4
Llama-3-LongVILA-8B	49.4	50.6	49.1	48.9	51.2	50.0	49.8	50.0
VILA1.5-13B	50.5	48.0	44.0	47.2	42.6	55.1	45.8	48.5
InternVL2-26B	51.6	49.7	50.4	48.4	47.2	56.2	53.3	53.6
VILA1.5-40B	50.7	50.3	50.0	50.0	49.7	50.0	49.8	50.7
Qwen2-VL-72B	53.6	49.9	50.6	50.6	52.0	53.1	59.8	52.5
LLaVA-OneVision-72B	56.0	61.7	50.7	57.0	54.7	73.4	71.5	68.6
Llama-3.1-405B	54.5	53.4	49.2	57.8	56.7	57.7	67.2	58.8
Qwen-max	48.0	51.5	47.6	48.3	50.1	49.1	47.4	41.1
Mistral Large*	49.4	48.9	45.1	49.4	50.9	60.6	61.3	50.8
Claude-3.5-Sonnet*	53.2	56.0	60.2	58.9	66.1	68.3	68.2	60.1
Gemini-1.5-Pro*	55.0	52.2	49.9	52.9	56.5	62.5	59.0	58.6
GPT-4o*	50.1	54.0	48.4	52.1	55.3	64.2	64.7	61.5

Table 18: **Multiple Choice** questions results of different models on the LOKI text modality. * denotes the closed-source models.

	Scientific Papers	News	Essay	Wikipedia	Speech	Modern Literature	Phiosophy	Ancient Chinese
MiniCPM-V-2.6	48.6	42.6	47.8	48.2	48.5	50.1	56.3	50.0
Phi-3.5-Vision	45.8	39.9	41.0	41.1	36.3	42.6	44.3	47.5
LLaVA-OneVision-7B	51.5	48.2	48.6	44.9	44.9	49.6	51.1	48.3
InternLM-XComposer2.5	44.0	42.8	33.9	41.0	34.3	46.3	40.8	41.9
mPLUG-Owl3-7B	51.8	49.2	45.1	48.3	44.9	54.0	62.1	38.6
Qwen2-VL-7B	49.6	46.3	42.6	44.2	45.8	45.4	50.7	46.1
LongVA-7B	50.1	49.0	48.1	47.5	46.1	51.4	49.3	50.3
Mantis-8B	58.9	54.4	36.3	51.5	35.3	49.7	59.6	34.7
Idefics2-8B	50.1	40.6	31.5	40.1	35.8	35.0	33.6	20.8
InternVL2-8B	53.9	52.8	38.9	47.2	42.5	43.8	50.7	40.0
Llama-3-LongVILA-8B	46.4	43.5	40.1	42.2	37.4	51.4	46.7	49.2
VILA1.5-13b	50.3	40.8	39.9	44.6	40.8	43.5	48.1	44.2
InternVL2-26B	52.6	54.9	43.2	49.3	47.6	48.9	56.3	48.9
VILA1.5-40B	49.9	52.5	39.4	48.9	46.3	57.6	58.2	50.8
Qwen2-VL-72B	67.2	79.2	53.1	72.9	62.5	76.7	76.1	69.7
LLaVA-OneVision-72B	65.1	79.9	50.6	63.1	61.3	82.6	78.3	76.9
Llama-3.1-405B	56.8	61.7	74.7	54.4	69.6	74.2	86.5	83.1
Qwen-max	48.3	48.3	47.4	33.8	43.3	43.8	48.3	48.1
Mistral Large*	55.7	75.1	55.4	62.8	61.8	86.7	86.1	69.4
Claude-3.5-Sonnet*	83.9	88.3	76.4	86.9	92.8	96.9	97.5	92.2
Gemini-1.5-Pro*	51.3	56.7	33.5	46.8	60.7	70.4	71.5	77.2
GPT-4o*	62.2	62.9	49.4	59.0	66.9	81.5	79.3	75.6

Table 19: Performance Comparison of Models across Text Lengths and Languages. * denotes the closed-source models.

	Short text	Medium text	Long text	English	Chinese
MiniCPM-V-2.6	49.1	48.9	46.6	49.7	48.9
Phi-3.5-Vision	63.2	59.2	56.5	48.6	45.5
LLaVA-OneVision-7B	45.5	45.7	46.8	53.1	50.1
InternLM-XComposer2.5	46.4	46.7	47.2	48.2	49.0
mPLUG-Owl3-7B	48.8	47.6	49.0	56.3	48.9
Qwen2-VL-7B	43.8	42.4	43.5	47.6	48.4
LongVA-7B	37.7	37.0	37.5	49.5	48.0
Mantis-8B	54.7	55.5	55.6	53.7	48.2
Idefics2-8B	46.9	41.8	33.9	46.7	41.2
InternVL2-8B	45.6	44.4	42.1	49.7	48.6
Llama-3-LongVILA-8B	47.1	46.8	47.0	47.7	48.3
VILA1.5-13B	49.4	48.0	44.4	47.0	46.0
InternVL2-26B	34.9	33.9	35.4	51.0	50.7
VILA1.5-40B	50.5	50.9	50.7	49.8	50.5
Qwen2-VL-72B	40.4	41.6	47.4	60.2	56.9
LLaVA-OneVision-72B	53.5	56.6	63.1	63.8	64.0
Qwen-max	35.8	34.0	32.8	48.3	45.9
Llama-3.1-405B	49.0	49.6	56.5	63.4	61.3
Mistral-Large*	44.6	46.3	51.3	60.5	55.5
Claude-3.5-Sonnet*	56.1	60.3	67.6	73.5	68.3
Gemini-1.5-Pro*	39.7	40.3	47.0	58.4	54.3
GPT-4o*	46.9	49.4	58.2	60.0	59.0

E MORE EXPERIMENTS RESULT AND DISCUSSION

In this appendix, we provide a detailed discussion of additional experimental results, including compression artifacts tests, analyses of Deepfake category experiments, and further results from CoT experiments.

E.1 COMPRESSION ARTIFACT TESTS

To better evaluate the applicability of large multimodal models (LMMs) in real-world scenarios, we conducted additional compression artifact tests. Following previous studies(Wang et al., 2020a; Ojha et al., 2023), we conducted JPEG compression tests on judgment tasks under the image modality, based on three representative LMMs and an expert model AIDE, as previously introduced in the main text.

Table 20: **Compression artifact test**. The test results of different models under various compression ratios.

	100%	95%	90%	80%	60%
InternVL2-8B	56.6%	56.1%	56.1%	56.4%	56.6%
Qwen2-VL-7B	54.5%	54.0%	54.2%	54.6%	54.1%
GPT-4o	63.4%	63.2%	63.1%	62.8%	63.1%
AIDE	62.8%	46.4%	45.0%	44.2%	45.2%

From these results, we observe that expert models AIDE(Yan et al., 2024) are more significantly impacted by compression artifacts. In contrast, LMMs demonstrate remarkable robustness, with compression levels having little to no effect on their detection capabilities. This distinction highlights the inherent advantages of LMMs over traditional expert models. There are several possible factors that contribute to the robustness of LMMs to compressed images.

Extensive Pretraining on Diverse Data: The vast amount of data available on the internet includes images with varying levels of compression. During pretraining, LMMs inherently learn features that are robust to such artifacts by including the internet-scale data, as demonstrated in works like UniversalFakeDetect(Ojha et al., 2023) and CLIPping(Khan & Dang-Nguyen, 2024). Model Scale Advantage: Unlike traditional expert modules, which typically operate with parameter counts in the millions, LMMs are equipped with parameters in the billions. This scale significantly enhances their generalization and robustness, consistent with the emergent properties observed in large-scale models. This finding further underscores the potential of LMMs for real-world applications, particularly in scenarios involving compression artifacts. We will include the above experimental results and analysis of compression robustness in the supplementary material to provide a comprehensive view of LMM reliability and applicability. If you have additional questions, we welcome further discussion.

E.2 DEEPFAKE DETECTION

Deepfake detection is one of the most prominent synthetic data detection categories today. LOKI already includes partial Deepfake datasets from audio and image modalities. The existing data for the Deepfake category includes:

Audio Modality: In audio tasks, some samples of speech and singing are sourced from their respective audio deepfake datasets ASVSpoof2019 and DCASE2023 Track 7as mentioned in Table 6. These datasets are built on state-of-the-art voice cloning techniques, such as Text-to-Speech(Ren et al., 2019), Voice Conversion(Sisman et al., 2020), and Singing Voice Conversion(Nachmani & Wolf, 2019). These technologies address major challenges in deepfake detection, including vulnerabilities in voice authentication systems and the misuse of synthetic voice technology to generate counterfeit content, such as fake artists.

Image Modality: As shown in Table 6, our image data includes a portion of facial deepfake images, comprising partial datasets from FFHQ and Deepfakeface. These data are directly aligned with the

core challenges in image-based deepfake detection, ensuring that our benchmark comprehensively evaluates models on this critical aspect of synthetic data detection.

Table 21: The accuracy of different models on the deepfake dataset and their overall accuracy.

Model	Claude-3.5-Sonnet	InternVL2-8B	VILA1.5-40B	Llava-OV-7B	Mantis-8B	Qwen2-VL-72B	Llama-3-8B	GPT-4o
Deepfake Accuracy	0.14	0.13	0.29	0.13	0.53	0.34	0.47	0.59
Overall Accuracy	0.54	0.50	0.49	0.50	0.55	0.53	0.50	0.63

We have independently evaluated the deepfake-related components of our current dataset and presented the results below. From the results, it is evident that most models perform significantly worse on deepfake data compared to their overall accuracy. This discrepancy likely arises because the overall evaluation task includes simpler synthetic samples (e.g., basic synthetic data), which are easier for the models to distinguish, thereby inflating their overall accuracy. In contrast, deepfake data often involves more nuanced and fine-grained forgery features that are particularly challenging for general-purpose multimodal pretrained models to capture effectively. Interestingly, GPT-4o achieves the best performance on deepfake data among all evaluated models, demonstrating its relatively stronger ability to handle such complex forgeries.

E.3 More CoT experiments results

To further validate the effectiveness of CoT across various question types and modalities, we conducted additional experiments. These experiments included testing multiple question types in the image modality (e.g., multiple-choice and anomaly detection) and extending the evaluation to video and text modalities (e.g., binary classification tasks). The results are shown in Table 22.

From the results, we observe that for most models, CoT significantly improves performance across different question types and modalities. While LLaVA-OV-7B struggles under CoT settings, other models, such as InternVL2-8B, Qwen2-VL, and GPT-4o, consistently benefit from CoT prompting, with noticeable performance gains across both tasks (e.g., anomaly detection, binary classification) and modalities (e.g., image, video, text).

In conjunction with the results in Table 4, we observe that the negative impact of FS (Few-shot) prompting on LMMs may be attributed to the inherent difficulty of synthetic detection tasks, whereas CoT typically enhances the performance of most LMMs (neutral or better).

Synthetic detection requires fine-grained perception and complex reasoning to recognize artifacts, which might be out of the training scope of current LMMs. Therefore, in FS prompting, simply providing in-context examples without reasoning steps might be insufficient for most models to learn how to detect artifacts. Furthermore, adding them to their context window might interfere with their original reasoning paths, resulting in degraded performances. Therefore, in the difficult setup of synthetic detection, the effects of FS prompting on model performance are limited. We note that the strong performance of GPT-4o models after FS prompting might stem from its inherently strong reasoning abilities. Consistent with findings in other CoT prompting methods(Brown, 2020; Alayrac et al., 2022; Wei et al., 2022b), we find CoT generally improves model performance.

Table 22: More Chain of Thought prompting experiments results.

Model	**Image Multi-choice**		**Image Abnormal Selection**		**Video Judgement**		**Text Judgement**	
	w/o CoT	**w/ CoT**	**w/o CoT**	**w/ CoT**	**w/o CoT**	**w/ CoT**	**w/o CoT**	**w/ CoT**
LLaVA-OV-7B	51.7	33.7	18.8	14.4	65.0	50.5	51.6	43.1
InternVL2-8B	51.4	53.2	70.2	75.3	51.2	55.1	41.6	52.3
Qwen2-VL	65.1	67.6	31.5	36.6	62.3	69.4	45.3	54.3
GPT-4o	80.8	88.9	76.2	84.3	71.2	77.8	53.5	59.6

E.4 Performance across Different Levels and Modalities

Table 23 presents the performance of different models across various modalities and difficulty levels. Based on human user performance, we categorized the difficulty levels of the questions. As the difficulty of the questions increases, most models show a gradual decline in performance. Notably, GPT-4o maintains relatively strong performance in both the Video and Image modalities. Given the diverse tasks and modalities involved in LOKI, considering the scores across different question difficulties is beneficial for assessing the overall performance of the models.

Table 23: The performance of different models across various modalities and task difficulty levels.

Model	Video			Image			Text			3D		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
LLaVA-OV-7B	0.58	0.55	0.52	0.55	0.59	0.56	0.62	0.46	0.41	0.61	0.58	0.55
InternVL2-8B	0.64	0.56	0.49	0.54	0.58	0.53	0.56	0.50	0.39	0.42	0.32	0.25
Qwen2-VL-7B	0.63	0.52	0.52	0.55	0.55	0.54	0.62	0.46	0.37	0.67	0.55	0.45
Gemini-1.5-pro*	0.63	0.55	0.48	0.52	0.51	0.51	0.64	0.14	0.17	0.60	0.64	0.68
Claude-3.5-Sonnet*	0.65	0.61	0.54	0.54	0.53	0.51	0.72	0.15	0.27	0.55	0.53	0.58
GPT-4o*	0.74	0.65	0.63	0.70	0.69	0.66	0.64	0.18	0.29	0.58	0.51	0.58

F CASE STUDY

In this section, we delve into an analysis of GPT-4o's performance on synthetic data detection tasks, with a particular focus on the questions it answered incorrectly. This examination is crucial for understanding the model's operational capabilities and limitations. The analysis not only identifies the model's current shortcomings but also provides guidance for future design and training improvements. We randomly selected 50 erroneous instances from GPT-4o's predictions across various modalities and difficulty levels for detailed inspection. These instances included multiple-choice, true/false, and open-ended responses, and even for the structured questions, GPT-4o was required to provide the reasoning behind its decisions. These instances were analyzed by specialized annotators who identified the fundamental causes of the erroneous predictions based on their expertise and annotations.

Image: Animals **[Opened-Ended]**

Question:

This is a fake image generated by an advanced AI model.
Your Task: Please perform a detailed analysis explaining where this image seems inauthentic from the following perspectives:
1. Global Perspective: Provide an overall explanation of why the entire image seems inauthentic. Be specific and detailed in your explanation. Reference any general visual cues that contribute to the inauthenticity.
2. Regional Perspectives: Identify specific regions in the image that appear inauthentic. For each region: Describe the location or provide a reference to the region (e.g., "top-left corner," "the person's right hand," etc.). Explain in detail why the region seems inauthentic. Cite specific visual anomalies or inconsistencies observed in that region.

(GPT-4o) Answer:

Overall Explanation: ***Strings and Attachments:*** The visible strings attached to the bird are a clear indication that it is a kite. Real birds do not have strings or any external attachments. ***Material and Texture:*** The texture of the bird's body and wings appears to be made of fabric or synthetic material, lacking the natural feather detail and variation seen in real birds. ***Pose and Movement:*** The bird's pose is rigid and lacks the dynamic movement typical of a real bird in flight. The wings and body do not show the natural flexibility and motion expected in a living creature.
Regional Perspectives: ***Region 1:*** **Head and Beak Location:** The head and beak area of the bird. **Inauthenticity Explanation:** The head and beak are overly symmetrical and lack the subtle asymmetries found in real birds. The eyes appear painted or printed, lacking the depth and reflection seen in living eyes. ***Region 2:*** **Wings Location:** The wings of the bird. **Inauthenticity Explanation:** The wings have a uniform color and texture, which is unusual for real birds that typically have varied feather patterns. The edges of the wings are too smooth and lack the individual feather detail. ***Region 3:*** **Talons Location:** The talons of the bird. **Inauthenticity Explanation:** The talons are unnaturally positioned and appear to be made of a solid material rather than the articulated structure of real bird talons. They lack the sharpness and detail expected in a predator bird. ***Region 4:*** **Strings Location:** Attached to various parts of the bird, including the wings and body. **Inauthenticity Explanation:** The presence of strings is a clear indicator that this is a kite. The strings are visibly taut and connected to the bird, which is not possible in a real bird.

Ground Truth:

Overall Explanation: The texture detail of the eagle's feathers is poor, with blurred patterns, and the material does not match that of real feathers, resulting in low authenticity.
Regional Perspectives: ***Region 1: Claws*** **Inauthenticity Explanation:** The shape of the eagle's claws is overly enlarged, differing from reality. There are significant discrepancies in the lengths of the different claws, and the curvature of the claw tips varies noticeably, lacking symmetry and consistency. Additionally, the number of claws on each side is inconsistent. ***Region 2: Rope*** **Inauthenticity Explanation:** The texture of the rope on the eagle shows obvious breakage, with no clear connection to the surrounding ropes, which does not align with reality. ***Region 3: Feather*** **Inauthenticity Explanation:** The feathers on the head do not exhibit a clear structural or edge boundary relationship with the feathers on the body, which is inconsistent with the characteristics of a real eagle.

Figure 13: Abnormal detail explanation example of category **animals** in image modality, with GPT-4o's answers

Image: Object [Opened-Ended]

Question:

This is a fake image generated by an advanced AI model.
Your Task: Please perform a detailed analysis explaining where this image seems inauthentic from the following perspectives:
1. Global Perspective: Provide an overall explanation of why the entire image seems inauthentic. Be specific and detailed in your explanation. Reference any general visual cues that contribute to the inauthenticity.
2. Regional Perspectives: Identify specific regions in the image that appear inauthentic. For each region: Describe the location or provide a reference to the region (e.g., "top-left corner," "the person's right hand," etc.). Explain in detail why the region seems inauthentic. Cite specific visual anomalies or inconsistencies observed in that region.

(GPT-4o) Answer:

Overall Explanation: ***Key Labels:*** The characters on the keys do not align with any known keyboard layout. Many of them appear as nonsensical or distorted versions of letters and symbols, indicating an AI-generated image. ***Lighting:*** The lighting on the keys is overly uniform, lacking the natural variation expected from real-world lighting. ***Wear and Tear:*** There is a noticeable absence of any wear or tear, which would typically be visible on frequently used keys in a real keyboard.
Regional Perspectives: ***Region 1: Top-Left Corner*** **Location:** Top-left corner of the keyboard. **Inauthenticity Explanation:** The characters on the keys in this area are particularly distorted, with some appearing as random shapes rather than recognizable letters or symbols, highlighting a lack of attention to detail. ***Region 2: Center of the Keyboard*** **Location:** Center of the keyboard, where the alphanumeric keys are located.**Inauthenticity Explanation:** The characters do not correspond to any known language or keyboard layout, and the spacing between the keys is inconsistent, which is unusual for a manufactured product. ***Region 3: Bottom Row*** **Location:** The bottom row of the keyboard, containing the space bar and modifier keys. **Inauthenticity Explanation:** The space bar appears overly smooth, without the texture or slight curvature found in real keyboards. Additionally, the symbols on the modifier keys are not standard. ***Region 4: Key Edges and Shadows*** **Location:** The edges of the keys and the shadows they cast. **Inauthenticity Explanation:** The shadows are overly uniform and do not reflect the subtle variations of real-world lighting. The key edges are too smooth and lack the slight imperfections found in real objects.

Ground Truth:

Overall Explanation: The shapes of the keyboard keys are irregular, distorted, and the textures are unclear.
Regional Perspectives: ***Region 1: Keyboard Cap*** **Inauthenticity Explanation:** The shapes of the keys are distorted, and the characters on the keycaps are blurred, exhibiting abnormal reflections.

Figure 14: Abnormal detail explanation example of category **object** in image modality, with GPT-4o's answers

Image: Document [Multi-Choice]

Question:

You have been shown two different images. One of the images is generated by AI. Which of the above images is most likely generated by AI?
A. The first one
B. The second one

(GPT-4o) Answer: (B)

The second image features a cartoon-style illustration of a person and a robot playing chess, which is a style more commonly generated by AI or digitally created compared to the first image, which consists mainly of text and a straightforward graphic about stars and stripes on a flag.

Ground Truth: (A)

Analysis:

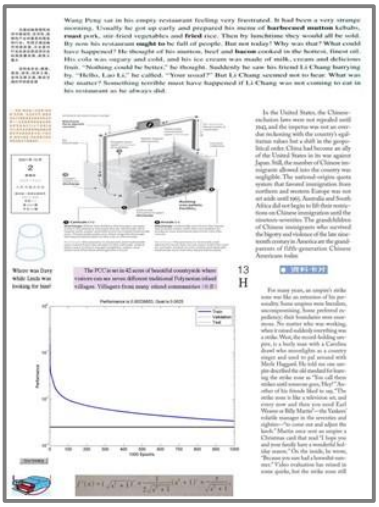
Cartoon-style illustrations cannot be used as a standard to determine whether a document image is AI-generated, as humans can also create cartoon-style images through painting. The reasoning behind this analysis is flawed.

Figure 15: Multi-choice example of category **document** in image modality, with GPT-4o's answers

Image: Document [Judgement]

Question:

You have been shown one image, possibly generated by AI. Please identify whether this image is generated by AI or not.



(GPT-4o) Answer: NO

Mixed Content with Consistent Styles: The image contains a variety of elements—such as text in different languages (English and Chinese), diagrams, graphs, and hand-drawn sketches—all seamlessly integrated. AI-generated images often struggle with producing this level of diverse, yet contextually appropriate, visual content while maintaining distinct styles across different sections.**Clear and Detailed Text**: The texts in the image appear consistent and legible, with different font styles and sizes matching their context (e.g., a passage of text, a mathematical formula, and a caption under a graph). AI models sometimes have difficulty generating lengthy, detailed passages with proper spacing and accurate alignment across various sections. **Scientific Diagrams and Charts**: There are scientific diagrams and a performance graph, including proper annotations and labels. Such precise scientific representation is challenging for AI to generate convincingly due to the need for accurate placement of elements, meaningful labels, and clear line work that conveys the correct information. **Natural Layout and Organization**: The layout of the image appears very human-like, with different sections logically arranged (e.g., text aligned with diagrams, a newspaper-like column for reading). AI-generated images can struggle to create complex layouts that appear natural and purposeful.**Small, Detailed Annotations**: There are small details like handwritten notes and tiny labels in the scientific diagrams, which have a level of human touch and imperfection that AI-generated images often miss or overcomplicate.

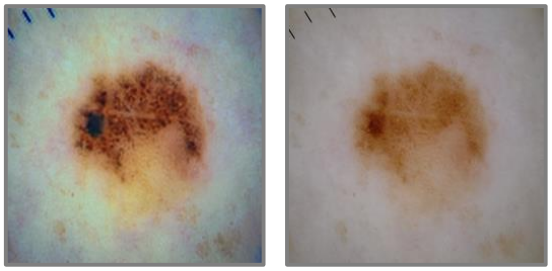
Ground Truth: NO

Figure 16: Judgement example of category **document** in image modality, with GPT-4o's answers

Image: Medical [Multi-Choice]

Question:

You have been shown two different images. One of the images is generated by AI. Which of the above images is most likely generated by AI?
A. The first one
B. The second one



(GPT-4o) Answer: (A)

The first image has more pronounced color artifacts, uneven blending, and unusual texture patterns, which are typical indicators of AI-generated imagery. The second image displays a more natural appearance with smoother transitions, realistic textures, and consistent colors, characteristic of an actual skin lesion photograph.

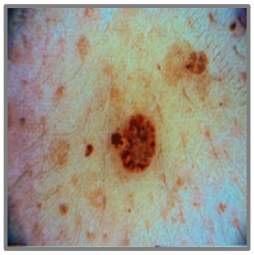
Ground Truth: (A)

Figure 17: Multi-choice example of category **medical** in image modality, with GPT-4o's answers

Image: Medical [Judgement]

Question:

You have been shown one image, possibly generated by AI. Please identify whether this image is generated by AI or not.



(GPT-4o) Answer: NO

1.Natural Variation in Texture and Color: The image displays natural variations in skin texture and pigmentation, which are often difficult for AI models to replicate accurately. Generated images might show unnaturally smooth or uniformly textured areas.
2.Absence of Artifacts or Anomalies: AI-generated images sometimes exhibit visual artifacts like distortions, unusual color blending, or pattern repetitions. This image seems to lack these common AI-generation artifacts.
3.Realistic Lighting and Shading: The way light reflects on the skin and the depth of the pigmentation appear natural, which is often challenging for AI models to replicate, especially in complex textures like human skin.
4.Organic Irregularities: The image has small, irregular details, such as the shape of the central lesion and the surrounding spots. AI models sometimes struggle to reproduce these organic irregularities accurately, often defaulting to more symmetrical or "perfect" shapes.

Ground Truth: YES

Analysis:

The image's skin texture exhibits abnormal color artifacts, with uneven color blending on the surface. This skin texture is neither smooth nor uniform. The skin surface in the image shows unusual bluish-green tones, and the pigmentation appears unnatural.

Figure 18: Judgement example of category **medical** in image modality, with GPT-4o's answers

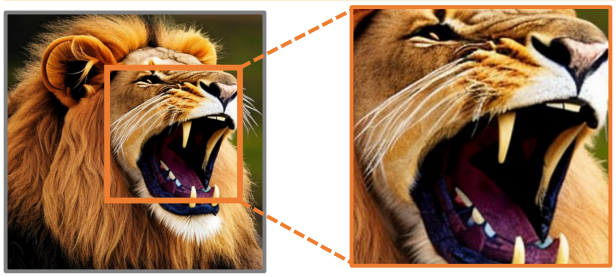
Image: Animals [VQA]

Question:

What seems inauthentic in this image?

Option:

(A) Nothing in this image seems inauthentic.
(B) The lion's eyes are purple, which is inconsistent with the normal eye color of lions.
(C) The shape of the teeth inside the lion's mouth is asymmetrical, and the texture details of the tongue and oral cavity are insufficient, with distorted colors.
(D) The lion's mane is unevenly colored, with blue stripes appearing.
(E) There is a distinct star-shaped mark on the lion's nose.



(GPT-4o) Answer: (C)

Upon inspection, the teeth appear to be asymmetrical, and the details within the mouth, such as the tongue and oral cavity, have unnatural colors and insufficient texturing. This suggests inauthenticity or manipulation in the image.

Ground Truth: (C)

Figure 19: Abnormal selection example of category **animals** in image modality, with GPT-4o's answers

Image: Object [VQA]

Question:

What seems inauthentic in this image?

Option:

(A) The text on the product packaging paper is blurry and hard to read.
(B) The pattern on the label is upside down.
(C) Nothing in this image seems inauthentic.
(D) There are unnatural light and shadow effects on the rock.
(E) The color of the bottle cap is completely different from that of the bottle body.

(GPT-4o) Answer: (A)

The image shows that the text on the labels of the bottles is distorted, blurry, and difficult to read, which is often a sign of inauthenticity or an AI-generated image, as AI models sometimes struggle with replicating clear and consistent text.

Ground Truth: (A)

Figure 20: Abnormal selection example of category **object** in image modality, with GPT-4o's answers

Image: Object [Opened-Ended]

Question:

This is a fake image generated by an advanced AI model.
Your Task: Please perform a detailed analysis explaining where this image seems inauthentic from the following perspectives:
1. Global Perspective: Provide an overall explanation of why the entire image seems inauthentic. Be specific and detailed in your explanation. Reference any general visual cues that contribute to the inauthenticity.
2. Regional Perspectives: Identify specific regions in the image that appear inauthentic. For each region: Describe the location or provide a reference to the region (e.g., "top-left corner," "the person's right hand," etc.). Explain in detail why the region seems inauthentic. Cite specific visual anomalies or inconsistencies observed in that region.

(GPT-4o) Answer:

Overall Explanation: ***Key Labels:*** The characters on the keys do not align with any known keyboard layout. Many of them appear as nonsensical or distorted versions of letters and symbols, indicating an AI-generated image. ***Lighting:*** The lighting on the keys is overly uniform, lacking the natural variation expected from real-world lighting. ***Wear and Tear:*** There is a noticeable absence of any wear or tear, which would typically be visible on frequently used keys in a real keyboard.
Regional Perspectives: ***Region 1: Top-Left Corner*** **Location:** Top-left corner of the keyboard. **Inauthenticity Explanation:** The characters on the keys in this area are particularly distorted, with some appearing as random shapes rather than recognizable letters or symbols, highlighting a lack of attention to detail. ***Region 2: Center of the Keyboard*** **Location:** Center of the keyboard, where the alphanumeric keys are located.**Inauthenticity Explanation:** The characters do not correspond to any known language or keyboard layout, and the spacing between the keys is inconsistent, which is unusual for a manufactured product. ***Region 3: Bottom Row*** **Location:** The bottom row of the keyboard, containing the space bar and modifier keys. **Inauthenticity Explanation:** The space bar appears overly smooth, without the texture or slight curvature found in real keyboards. Additionally, the symbols on the modifier keys are not standard. ***Region 4: Key Edges and Shadows*** **Location:** The edges of the keys and the shadows they cast. **Inauthenticity Explanation:** The shadows are overly uniform and do not reflect the subtle variations of real-world lighting. The key edges are too smooth and lack the slight imperfections found in real objects.

Ground Truth:

Overall Explanation: The shapes of the keyboard keys are irregular, distorted, and the textures are unclear.
Regional Perspectives: ***Region 1: Keyboard Cap*** **Inauthenticity Explanation:** The shapes of the keys are distorted, and the characters on the keycaps are blurred, exhibiting abnormal reflections.

Figure 21: Abnormal detail explanation example of category **object** in image modality, with GPT-4o's answers

Image: Person [Opened-Ended]

Question:

This is a fake image generated by an advanced AI model.
Your Task: Please perform a detailed analysis explaining where this image seems inauthentic from the following perspectives:
1. Global Perspective: Provide an overall explanation of why the entire image seems inauthentic. Be specific and detailed in your explanation. Reference any general visual cues that contribute to the inauthenticity.
2. Regional Perspectives: Identify specific regions in the image that appear inauthentic. For each region: Describe the location or provide a reference to the region (e.g., "top-left corner," "the person's right hand," etc.). Explain in detail why the region seems inauthentic. Cite specific visual anomalies or inconsistencies observed in that region.

(GPT-4o) Answer:

Overall Explanation: ***Texture and Detail*:** The skin texture appears overly smooth and lacks the natural variation found in real human skin. This uniformity is a common trait in AI-generated images. ***Lighting and Shadows*:** The lighting seems inconsistent, with shadows not aligning naturally with the light source. This can create an unnatural appearance. ***Facial Features*:** The facial features may appear slightly exaggerated or disproportionate, which is a common issue in AI-generated portraits.
Regional Perspectives: ***Region 1: Face*** **Location:** Center of the image. **Inauthenticity Explanation:** The eyes and mouth may appear slightly off in terms of alignment or proportion. The eyes might lack the depth and reflection typical in real photographs, and the mouth may not align perfectly with the facial structure. ***Region 2: Hands*** **Location:** Bottom center of the image. **Inauthenticity Explanation:** The hands might show signs of distortion or unnatural positioning. Fingers could appear fused or have an unusual number of joints, which is a common anomaly in AI-generated images. ***Region 3: T-shirt Logo*** **Location:** Center of the torso. **Inauthenticity Explanation:** The logo on the T-shirt may appear distorted or lack clarity. The text or design might not follow the natural folds and contours of the fabric, indicating a lack of realistic interaction with the clothing. ***Region 4: Background*** **Location:** Entire background. **Inauthenticity Explanation:** The background may lack depth or have an unnatural texture. The wall's surface might appear too uniform or have inconsistent lighting, which can detract from the realism of the scene.

Ground Truth:

Overall Explanation: The image exhibits a certain degree of color distortion, with noticeable color differences on both sides of the image.
Regional Perspectives: ***Region 1: Eyes*** **Inauthenticity Explanation:** There are errors in the texture of the eyes, with the visual direction of the two eyes differing, resulting in an incorrect structure. The eyes are uneven in size and asymmetrical, which does not conform to the characteristics of real eyes. ***Region 2: Hands*** **Inauthenticity Explanation:** The texture of the fingers is chaotic, with textures of different fingers blending into each other, making them difficult to distinguish. The edges between the fingers and the clothing are blurred, with an unnatural transition between the two, which does not align with real-world appearances. ***Region 3: Face*** **Inauthenticity Explanation:** The lips are asymmetrical, and there is an unnatural color transition in the nose area, with the nostrils being uneven in size on both sides. ***Region 4: Arms*** **Inauthenticity Explanation:** The edges between the sleeve and the arm are blurred, and there are incorrect results in the arm joint area, as well as in the distribution of the muscles.

Figure 22: Abnormal detail explanation example of category **person** in image modality, with GPT-4o's answers

Image: Person [VQA]

Question:

What seems inauthentic in this image?

Option:

(A) There is a non-existent pattern on the shirt.
(B) The person's ears are asymmetrical.
(C) The colors of the lights in the background are inconsistent.
(D) Nothing in this image seems inauthentic
(E) The cuff area shows an unnatural yellow texture transition, with the texture being discontinuous

(GPT-4o) Answer: (E)

Upon close inspection, the texture of the shirt near the cuff area transitions unnaturally, with a visible inconsistency in the pattern and color that is discontinuous, which might indicate that the image has been artificially generated or manipulated.

Ground Truth: (E)

Figure 23: Abnormal selection example of category **person** in image modality, with GPT-4o's answers

Image: Satellite [Multi-Choice]

Question:

You have been shown two different images. One of the images is generated by AI. Which of the above images is most likely **not** generated by AI?

A. The first one
B. The second one

(GPT-4o) Answer: (B)

The second image contains consistent and natural textures, road details, and shadows characteristic of real satellite imagery. The first image has some visual irregularities and unusual patterns that are more typical of AI-generated content.

Ground Truth: (A)

Figure 24: Multi-choice example of category **satellite** in image modality, with GPT-4o's answers

Image: Satellite [Judgement]

Question:

You have been shown one image, possibly generated by AI. Please identify whether this image is generated by AI or not.

(GPT-4o) Answer: YES

Textures and Patterns: In AI-generated images, textures can sometimes appear "off" or subtly inconsistent. In your image, while most parts look natural, there are some areas (especially around the boundaries of vegetation and buildings) where textures don't transition as smoothly as they might in a real photo. These transitions might have unusual blending or slight blurring, which can be an indicator of AI synthesis. **Sharpness and Edge Details:** Real-world images often have variations in sharpness due to factors like camera focus and lighting. AI-generated images, however, might have unnatural uniformity in sharpness or blurring. In this image, the edges of certain objects, like the road and structures, have a slightly uniform and artificial appearance, lacking the typical variation seen in real aerial photographs. **Anomalies in Vegetation:** AI models can sometimes struggle with complex textures like vegetation, producing repetitive or unnatural patterns. Looking closely at the trees and grass in your image, there seems to be a slightly repetitive and overly uniform structure, which could be an artifact of AI generation. **Lighting and Shadows:** Natural images have coherent lighting sources, and shadows are cast consistently across the scene. In some AI-generated images, shadow directions, intensities, or softness may not perfectly match. In your image, while the shadows generally align, the shading on some objects (like the small structure near the road) feels slightly mismatched with the expected light source.

Ground Truth: NO

Analysis:

In terms of texture, some edges of the textures exhibit blurriness, which is related to the resolution of the remote sensing images. The blurred edge transitions are caused by the image resolution. Due to the greater observation distance, certain high-frequency information of the observed features may be missing, resulting in repetitive and uniform textures on the ground. This is a unique characteristic of remote sensing images.

Figure 25: Judgement example of category **satellite** in image modality, with GPT-4o's answers

Image: Scenery [Opened-Ended]

Question:

This is a fake image generated by an advanced AI model.
Your Task: Please perform a detailed analysis explaining where this image seems inauthentic from the following perspectives:
1. Global Perspective: Provide an overall explanation of why the entire image seems inauthentic. Be specific and detailed in your explanation. Reference any general visual cues that contribute to the inauthenticity.
2. Regional Perspectives: Identify specific regions in the image that appear inauthentic. For each region: Describe the location or provide a reference to the region (e.g., "top-left corner," "the person's right hand," etc.). Explain in detail why the region seems inauthentic. Cite specific visual anomalies or inconsistencies observed in that region.

(GPT-4o) Answer:

Overall Explanation: ***Texture and Detail:*** **Color Saturation:** The colors, particularly the greens and oranges, are excessively saturated, which is not typical of natural environments. **Texture Uniformity:** The textures across different surfaces, such as the cliffs and water, are too consistent and lack the natural irregularities found in real-world environments. **Lighting and Shadows:** The lighting across the image is inconsistent, with shadows not aligning properly with a single light source, making the scene appear unrealistic. ***Regional Perspectives: Region 1: Cliff Face (Right Side)*** **Location:** Right side of the image. **Inauthenticity Explanation:** The cliff's texture is overly smooth, lacking the roughness and erosion marks typical of real cliffs. The transition between the vegetation and rock is too sharp and does not blend naturally, which detracts from the realism. ***Region 2: Water Surface (Bottom Left)*** **Location:** Bottom left corner of the image. **Inauthenticity Explanation:** The water appears overly regular in terms of wave and foam patterns, which lacks the natural randomness found in real water bodies. Additionally, the reflections and highlights on the water do not correspond well with the light source, further contributing to the artificial look. ***Region 3: Cliff Edge (Top Center)*** **Location:** Top center of the image. **Inauthenticity Explanation:** The cliff edge appears too sharply defined and lacks the irregularities common to natural cliff edges. The transition between the cliff and sky is abrupt, without the atmospheric blending usually seen in natural landscapes. ***Region 4: Vegetation (Right Side)*** **Location:** Right side of the cliff. **Inauthenticity Explanation:** The vegetation appears unnaturally uniform, with consistent color and texture. It does not show the diversity and natural distribution of real plant life, and the way the plants cling to the cliff is not realistic, as they do not follow the natural contours of the rock.

Ground Truth:

Overall Explanation: The overall image has excessively high saturation, and the waves and textures do not match reality. The boundary between the waves and the shore is blurred, lacking a sense of realism. The lighting in the upper part differs from that in the lower part.
Regional Perspectives: ***Region 1: Texture*** **Inauthenticity Explanation:** The color of the green-covered mountains is overly saturated, and the transition between the colors of the neighboring mountains is unnatural.

Figure 26: Abnormal detail explanation example of category **scenery** in image modality, with GPT-4o's answers

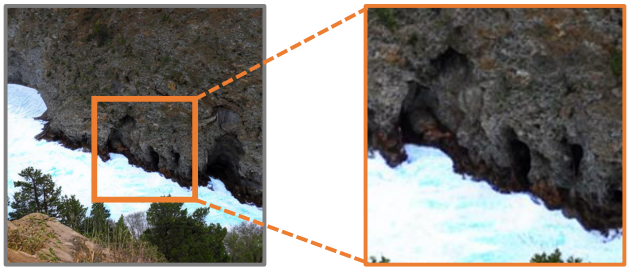
Image: Scenery [VQA]

Question:

What seems inauthentic in this image?

Option:

(A) The color of the leaves on the trees does not match the seasonal changes.
(B) There are unnatural cloud formations in the sky.
(C) The boundary between the rocks and the river is unclear, and the texture of the rocks does not match the characteristics of years of erosion.
(D) There are unnatural light and shadow effects on the rock.
(E) Nothing in this image seems inauthentic



(GPT-4o) Answer: (C)

The image shows an unclear separation between the rock formations and the water, and the rock texture lacks the natural, smooth erosion typically seen in such landscapes. These inconsistencies suggest inauthentic features in the image.

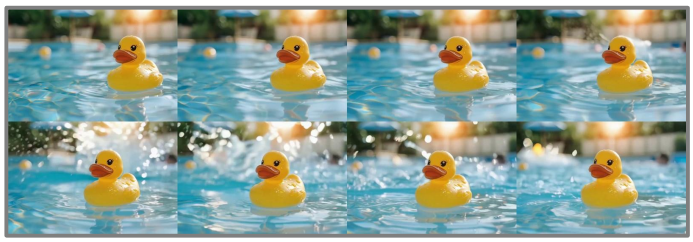
Ground Truth: (C)

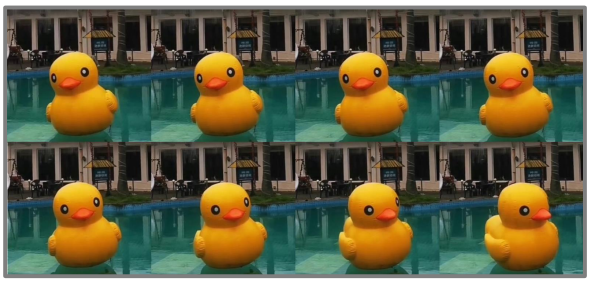
Figure 27: Abnormal selection example of category **scenery** in image modality, with GPT-4o's answers

Video: Abiotic [Multi-Choice]

Question:

You have been shown two different videos. One of the videos is taken in the real world, while the other is generated by an advanced AI model. Which of the above videos is most likely generated by AI?
A. The first one.
B. The second one.





(GPT-4o) Answer: (A)

This is due to the highly polished visuals, the realistic lighting effects, and the precision in water ripples and reflections, which are characteristic of high-quality AI-generated images or videos. In contrast, the second video has a more natural appearance with environmental elements like the palm trees and the slight imperfections typical in real-world recordings.

Ground Truth: (A)

Figure 28: Multi-choice example of category **Abiotic** in video modality, with GPT-4o's answers

Video: Animals [Multi-Choice]

Question:

You have been shown two different videos. One of the videos is taken in the real world, while the other is generated by an advanced AI model. Which of the above videos is most likely generated by AI?
A. The first one.
B. The second one.

(GPT-4o) Answer: (A)

This is based on the highly polished and uniform look, as well as the subtle but noticeable visual effects, such as perfect lighting and highly consistent textures in the environment. The second video has a more natural and imperfect feel, which is often characteristic of real-world footage.

Ground Truth: (A)

Analysis:

AI-generated videos typically exhibit refined details and appearances for the main subject, while the background and surrounding environment are often rendered in a rougher manner. In the first video, the distortion of the windmill in the background indicates that it is AI-generated. In contrast, the second video maintains overall consistency, with both the subject and the background showing the same level of detail.

Figure 29: Multi-choice example of category **Animals** in video modality, with GPT-4o's answers

Video: Animals [VQA]

Question:

What abnormal event occurs to the tentacles on the far right side of jellyfish in the video?
Option:
(A) Two tentacles on the far right side merge and then separate again.
(B) Two tentacles on the far right side disappear completely.
(C) Two tentacles on the far right side change color.
(D) Two tentacles on the far right side grow longer.

(GPT-4o) Answer: (A)

Tentacles merging and then separating again is indicative of how AI models sometimes struggle with maintaining fine details over time, especially in dynamic objects like jellyfish tentacles. The merging and separation can occur because the AI may not fully capture or consistently render the individual tentacles' paths. This behavior is less likely to happen in real-world footage but is more characteristic of AI-generated visual content, where the model might briefly combine two objects due to limitations in its rendering algorithms.

Ground Truth: (A)

Figure 30: Abnormal selection example of category **Animals** in video modality, with GPT-4o's answers

Video: Human [Opened-Ended]

Question:

These are frames of an AI-generated video. You need to use these frames as references to conduct a detailed analysis. The video illustrates various scenes and actions generated by artificial intelligence.
Your Task: Analyze and explain where this video appears inauthentic from the following perspectives:
1.Global Issues: Identify any overall problems that affect the entire video.Provide detailed explanations for why these aspects seem inauthentic.Reference specific characteristics that support your observations.
2.Keyframe Anomalies: Examine specific frames in the video for inconsistencies or errors.Explain in detail any anomalies that suggest the video is artificially generated.Use timestamps or frame numbers to reference specific parts of the video.
3. Segment Issues: Identify any problematic segments within the video where inconsistencies or artifacts are present.Provide detailed explanations for why these segments seem inauthentic.Specify the time ranges of these segments to support your analysis.

(GPT-4o) Answer:

Global Issues: Lighting and Shadows: Inconsistency: The lighting on the hands and the glass appears uniform and lacks natural variation. In a real setting, there would be subtle differences in shadow and light due to the movement of the liquid and the hand. **Explanation:** This uniformity suggests a lack of dynamic lighting, which is often a giveaway of AI-generated content. **Texture and Detail: Smoothness:** The surfaces, including the countertop and the glass, appear overly smooth and lack the fine details and imperfections found in real life. **Explanation:** AI-generated videos sometimes fail to capture the micro-textures that are present in real-world objects. **Keyframe Anomalies: Frame 1 to Frame 5: Liquid Dynamics:** The milk's flow into the glass is not consistent with real fluid dynamics. The liquid appears to pour in a straight line without the natural splash or ripple effects. **Explanation:** Real liquids have complex interactions with surfaces, which are difficult to replicate accurately in AI-generated content. **Frame 6 to Frame 8: Hand Movement:** The hand's position and grip on the glass remain almost static, which is unusual during a pouring action. Normally, there would be slight adjustments in grip and angle. **Explanation:** This lack of natural movement suggests a limitation in the AI's ability to simulate realistic human motion.**Segment Issues: Initial Pouring Segment (Frames 1-3): Inconsistency in Liquid Level:** The level of milk in the glass does not change as expected between frames. It should rise more noticeably with each frame. **Explanation:** This suggests a lack of frame-to-frame continuity, a common issue in AI-generated sequences. **Mid-Pour Segment (Frames 4-6): Liquid Texture:** The texture of the milk appears too uniform and lacks the subtle variations in opacity and color that occur in real milk. **Explanation:** AI often struggles with replicating the complex visual properties of liquids. **Final Pouring Segment (Frames 7-8): Reflection and Refraction:** The reflections on the glass and the milk do not change dynamically with the pouring action, which would be expected in a real scenario. **Explanation:** This indicates a lack of advanced rendering techniques that simulate real-world physics.

Ground Truth:

1. The milk in the cup has already begun to increase even before being poured out, which is inconsistent with reality.
2. The hand structure behind the cup is incorrect.
3. The texture of the milk in the cup differs from the material of real milk, with transparency and color not matching reality.

Figure 31: Abnormal detail explanation example of category **Human** in video modality, with GPT-4o's answers

Video: Human **[Judgement]**

Question:

You have been shown one video, which might be taken from real world or generated by an advanced AI model. Is this video taken in the real world? Answer yes if you think it is taken in the real world, and answer no otherwise.) Answer with yes or no.

(GPT-4o) Answer: YES

•**Natural Skin Texture**: The skin texture and details such as small pores, slight variations in color, and natural imperfections around the eye area seem very lifelike. AI-generated images often smooth out these details or make them unnaturally perfect, but these images exhibit the subtle imperfections typical of real skin.
•**Reflections in the Eyes**: The reflections in the eyes are consistent with a real-world lighting environment. There are complex and natural-looking reflections of the surroundings or light sources, which are harder to replicate in AI-generated models.
•**Eyelashes and Fine Details**: The eyelashes are irregular and detailed, which is more challenging for AI to recreate perfectly. In many AI-generated images, small details like hair or lashes can appear too uniform or clumped together, but here, they appear naturally varied.
•**Subtle Asymmetry**: While the images do seem highly symmetric at first glance, on closer inspection, there are slight variations in the shape and angle of the eyes, which is typical of real human features.

Ground Truth: NO

Analysis:

Regarding the reflections in the eyes, the reflections in the image are not realistic. At a certain point, the reflections in the eyes disappear along with changes in external lighting, while the overall scene shows no significant changes in the external light source. Additionally, in terms of fine structures, the eyelashes in the video are overly clustered, which does not align with the natural structure of eyelashes around a real human eye.

Figure 32: Judgement example of category **Human** in video modality, with GPT-4o's answers

Video: Human [Opened-Ended]

Question:

These are frames of an AI-generated video. You need to use these frames as references to conduct a detailed analysis. The video illustrates various scenes and actions generated by artificial intelligence.
Your Task: Analyze and explain where this video appears inauthentic from the following perspectives:
1.Global Issues: Identify any overall problems that affect the entire video.Provide detailed explanations for why these aspects seem inauthentic.Reference specific characteristics that support your observations.
2.Keyframe Anomalies: Examine specific frames in the video for inconsistencies or errors.Explain in detail any anomalies that suggest the video is artificially generated.Use timestamps or frame numbers to reference specific parts of the video.
3. Segment Issues: Identify any problematic segments within the video where inconsistencies or artifacts are present.Provide detailed explanations for why these segments seem inauthentic.Specify the time ranges of these segments to support your analysis.

(GPT-4o) Answer:

Global Issues: Unnatural Lighting and Shadows: The lighting in the frames appears overly uniform and lacks the natural variation typically seen in real-world settings. Shadows are either too soft or inconsistently placed, which can make the scenes look flat and artificial. **Lack of Motion Blur:** In a real video, especially with movement, there would be some motion blur. The frames here are too crisp, which is a common sign of AI generation. **Texture Uniformity:** The textures, especially on the skin and clothing, appear too smooth and lack the imperfections and variations found in real life. **Explanations: Lighting:** Real-world lighting varies with the environment and time of day, creating dynamic shadows and highlights. The uniformity here suggests artificial generation. **Motion Blur:** The absence of motion blur, even in action shots, indicates a lack of real-world physics simulation. **Texture:** AI often struggles with replicating the natural imperfections found in human skin and fabric, leading to an overly polished appearance.
Keyframe Anomalies: Frame 1 and 2: Hair Details: The hair appears overly detailed and perfectly styled, lacking the natural movement and randomness expected in a real setting. **Frame 3: Background Consistency:** The background buildings and street appear too perfect and symmetrical, lacking the natural irregularities of real architecture. **Frame 4 and 5: Pose and Expression:** The subject's pose and expression are too static and idealized, lacking the subtle variations and imperfections of real human movement. **Frame 6: Crowd Details:** The people in the background lack distinct features and appear as generic figures, which is a common issue in AI-generated scenes. **Explanations: Hair and Pose:** AI often generates hair and poses that are too perfect, as it lacks the ability to fully replicate the randomness of real life. **Background and Crowd:** AI-generated backgrounds and crowds often lack the complexity and detail of real environments and people. **Segment Issues: Segment: Frames 1-3 Street Scene Consistency:** The street scene remains too consistent across frames, with little variation in lighting or perspective, which is unusual for a real video. **Segment: Frames 4-5 Dance Movement:** The dance movements appear too smooth and lack the natural dynamics and imperfections of real human motion. **Segment: Frames 7-8 Walking Away Scene:** The transition from the detailed street scene to the simpler alleyway lacks depth and perspective, making it appear flat and artificial. **Explanations: Street Scene:** Real videos capture subtle changes in lighting and perspective, which are missing here. **Dance Movement:** Real human motion includes slight variations and imperfections that are difficult for AI to replicate. **Walking Scene:** The lack of depth and perspective in the transition suggests artificial generation.

Ground Truth:

Global Issues: 1. The video style leans towards painting, and the image flickers. 2. The facial features and hands of the character undergo severe distortion. **Fragment Issues:** 1.1. The facial features of the character are distorted. 1.2. The person's hands experience severe distortion, with the right hand merging with the street and the left hand merging with the shop. 2.1. The main character's facial features undergo severe distortion. 2.2. The image of the passerby is completely incorrect. 3.1. The body structure of the passerby in the background is incorrect.

Figure 33: Abnormal detail explanation example of category **Human** in video modality, with GPT-4o's answers

Video: Human [Judgement]

Question:

You have been shown one video, which might be taken from real world or generated by an advanced AI model. Is this video taken in the real world? Answer yes if you think it is taken in the real world, and answer no otherwise.) Answer with yes or no.

(GPT-4o) Answer: YES

•**Lighting and Reflections**: The complex interaction between light and shadow in the environment, particularly the sunlight streaming through the window and reflecting on the subject, looks very natural. AI-generated models often struggle to capture this level of nuanced lighting, especially in dynamic settings like the one depicted here.
•**Natural Details**: The texture of the skin, the details in the hair, and the intricacy of the tattoos all seem very detailed and realistic. These elements are often tricky for AI to replicate at this level, particularly the randomness in hair strands and the variability in skin texture.
•**Tattoo Realism**: The tattoos on the woman's arm show subtle inconsistencies and natural skin interaction (such as slight warping with the skin). AI-generated tattoos may sometimes appear overly clean or not integrate as smoothly with skin movements.
•**Depth and Focus**: The depth of field and focus transitions are handled very well. The video shows the subject clearly in the foreground, while the background is slightly out of focus. This depth effect is often challenging for AI to mimic accurately, as real-world cameras naturally create this effect due to their optical properties.

Ground Truth: NO

Analysis:

In terms of lighting, at a certain moment in the video, an unnatural light spot appears on the face. Moreover, the tattoos on the person change throughout the video, and the clothing structure on the waist does not match reality. Additionally, the environment around the person becomes distorted during camera transitions in the video.

Figure 34: Judgement example of category **Human** in video modality, with GPT-4o's answers

Video: Scenery [VQA]

Question:

What abnormality is present in the video footage of the snow-capped mountain?

Option:

(A) The colors in the video footage are distorted.
(B) The video footage is continuously flickering.
(C) The video footage is blurry and out of focus.
(D) The video footage is playing at an inconsistent frame rate.

(GPT-4o) Answer: (B)

In AI-generated videos, flickering is a common issue, particularly in scenes with complex textures like snow and mountains. This occurs because the AI model may struggle to maintain consistent rendering across frames, leading to slight variations in color, brightness, or details, which result in the flickering effect. This abnormality is typical of AI-generated content, where maintaining consistent visual quality over time can be challenging, especially in dynamic lighting conditions.

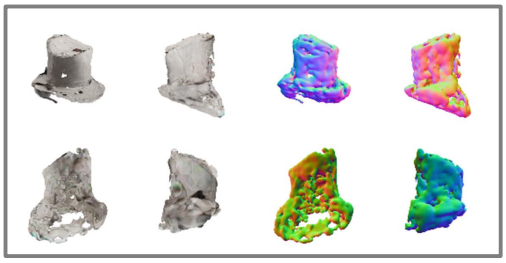
Ground Truth: (B)

Figure 35: Abnormal selection of category **Scenery** in video modality, with GPT-4o's answers

3D: 3DGS-based[Opened-Ended]

Question:

This is an AI-generated 3D object rendering. Displays four-view RGB renders on the left and four-view normal maps on the right.
Your Task: Analyze and explain the inauthentic aspects of the 3D object.
1.RGB Rendering: Identify issues related to textures, colors, and overall apearance. Provide detailed explanations for why these aspects seem inauthentic. Reference specific parts of the images to support observations.
2.Normal Rendering: Analyze the normal maps for inconsistencies or errors in surface geometry. Explain any anomalies that suggest the object is artificially generated. Use specific visual evidence from the normal maps in the explanations.



(GPT-4o) Answer:

RGB Rendering: Inconsistent Texture Quality: The textures appear uneven and lack the fine details expected in a realistic object. Some areas are overly smooth while others are rough, which is not typical of natural materials. **Color Uniformity:** The colors are flat and lack subtle variations due to lighting and material properties, making the object appear less lifelike. **Surface Imperfections:** Visible holes and irregularities in the structure are particularly noticeable in the top left and bottom left images, suggesting a lack of realistic material simulation.**Lack of Realistic Shading:**The shading does not convincingly represent how light interacts with the object. Shadows and highlights are inconsistent with a coherent light source, contributing to an artificial look.
Normal Rendering: Unnatural Surface Variations: The normal maps show abrupt changes in surface angles, seen in the top right and bottom right images, where the surface appears jagged and uneven. **Color Anomalies:** The normal maps use colors to represent surface normals, but the transitions are too sharp and lack the gradual blending typical of real objects, indicating a lack of smooth surface transitions. **Lack of Detail in Curvature:** The object's curvature is not well-defined, with many areas appearing too flat or sharply angled, suggesting the object was generated without high precision. **Inconsistent Lighting Simulation:** The normal maps do not accurately simulate how light interacts with a complex surface, exhibiting inconsistent lighting, which indicates an artificially generated model.

Ground Truth:

RGB Rendering: The texture of the hat shows color mixing on the sides, with gray and white interspersed with other colors. The color transition on the side edges is not smooth, and the edges are unclear.
Normal Rendering: The geometric surface of the hat is uneven, with numerous holes, weak topological smoothness, and poor surface smoothness. The geometric shape of the hat is asymmetrical and does not match the real shape.

Figure 36: Abnormal detail explanation example generated by **Gaussian Splatting based** methods in 3D modality, with GPT-4o's answers

3D: 3DGS-based **[Judgement]**

Question:

This is a rendering of a 3D object. The left side of the rendering shows a 4-view RGB rendering, and the right side shows a 4-view normal map, showing the surface texture and lighting interaction of the 3D model. The object may be generated by an AI model or may be a scan of a real object. Please determine whether the object was generated by an AI model. Answer with yes or no.

(GPT-4o) Answer: NO

Surface Texture Detail: The object, which appears to be a plush toy with a textured surface, exhibits highly detailed and organic variations, which are typically captured in real-world scans. AI-generated objects tend to have smoother surfaces or exhibit less random imperfections unless the AI specifically generates realistic imperfections. **Normal Maps**: The normal maps shown on the right side of the image reveal surface detail and texture interactions with light, which could be from a real-world scan. However, advanced AI models can also simulate such details, making it difficult to differentiate based on normal maps alone. **Color and Lighting**: The lighting and color seem consistent with real-world object captures. AI models, especially generative ones, often create somewhat smoother or more abstract textures, though modern models can be highly realistic.

Ground Truth: YES

Analysis:

In terms of surface texture details, the texture detail is relatively good, but the distribution of the texture is irregular and lacks symmetry. Additionally, there are noticeable differences in texture fineness across different areas, with the texture on the doll's head being finer. Regarding the normal maps, the image shows poor normal detail, with significant differences in the surface material represented by the normals, exhibiting varying degrees of protrusions and depressions.

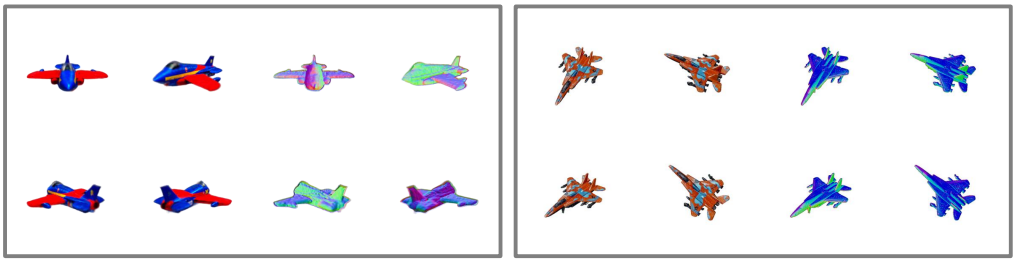
Figure 37: Judgement example generated by **Gaussian Splatting based** methods in 3D modality, with GPT-4o's answers

3D: 3DGS-based [Multi-Choice]

Question:

You have been shown two different images. These images are rendering displays four-view RGB renders on the left and four-view normal maps on the right, illustrating the surface texture and lighting interactions of the 3D model. One of the objects is not generated by AI. Which of the above objects is most likely generated by AI?

A. The first one
B. The second one



(GPT-4o) Answer: (B)

The exaggerated and vibrant color contrasts in the normal maps (on the right) suggest that it was produced by AI, as these types of extreme variations in lighting and surface texture are often characteristics of AI-generated 3D models. Additionally, the slightly smoother, less detailed texture in the RGB renders further points to AI generation.On the other hand, the second image (the orange and blue camouflaged aircraft) appears more realistic with textures and details that align with traditional 3D modeling or real-world designs.

Ground Truth: (B)

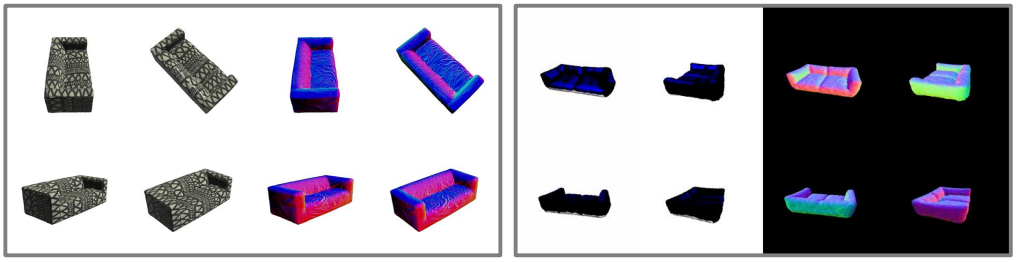
Figure 38: Multi-choice example generated by **Gaussian Splatting based** methods in 3D modality, with GPT-4o's answers

3D: 3DGS-based [Multi-Choice]

Question:

You have been shown two different images. These images are rendering displays four-view RGB renders on the left and four-view normal maps on the right, illustrating the surface texture and lighting interactions of the 3D model. One of the objects is not generated by AI. Which of the above objects is most likely generated by AI?

A. The first one
B. The second one



(GPT-4o) Answer: **(B)**

The second set of images seems to display a more stylized color distribution in the normal maps (right side of the images), with unrealistic lighting reflections and a variety of vibrant colors that aren't typical in real-world renders.The left side of the second image also exhibits more abstract, less natural shadows and forms that indicate a computer-generated or stylized model.AI-generated models often focus on abstract or highly stylized visuals, whereas the first image seems to have more natural texture and surface patterns, which might be captured through traditional 3D rendering techniques.

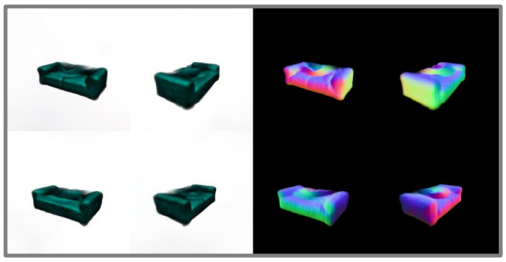
Ground Truth: **(B)**

Figure 39: Multi-choice example generated by **Gaussian Splatting based** methods in 3D modality, with GPT-4o's answers

3D: NeRF-based[Opened-Ended]

Question:

This is an AI-generated 3D object rendering. Displays four-view RGB renders on the left and four-view normal maps on the right.
Your Task: Analyze and explain the inauthentic aspects of the 3D object.
1.RGB Rendering: Identify issues related to textures, colors, and overall apearance. Provide detailed explanations for why these aspects seem inauthentic. Reference specific parts of the images to support observations.
2.Normal Rendering: Analyze the normal maps for inconsistencies or errors in surface geometry. Explain any anomalies that suggest the object is artificially generated. Use specific visual evidence from the normal maps in the explanations.



(GPT-4o) Answer:

RGB Rendering: Blurriness and Lack of Detail: The textures on the RGB renders appear blurry and lack fine detail, which is a common issue in AI-generated images. This is evident in the uniformity of the surface, which lacks the natural imperfections and variations found in real-world materials. **Unnatural Color Consistency:** The color of the object is uniformly dark green, lacking the subtle variations typically caused by lighting and material properties. Real objects often exhibit slight color shifts due to shadows, reflections, and material inconsistencies. **Lighting and Shadows:**The lighting on the object seems flat and fails to create realistic shadows or highlights. This can be observed in the absence of depth and dimension, making the object appear more like a 2D image than a 3D model. **Specific Observations:** The top-left and bottom-left images show a lack of shadowing that would typically occur on a real object, particularly around the edges and corners.
Normal Rendering: Uniform Surface Normals: The normal maps display a lack of variation in surface geometry. The colors in the normal maps are too consistent, indicating that the surface is overly smooth and lacks the natural bumps and ridges found in real objects.
Color Anomalies: The normal maps use a range of colors to represent surface angles, but the transitions between these colors are too abrupt and do not reflect the gradual changes seen in real-world surfaces. This suggests a lack of complexity in the surface geometry.
Specific Observations: In the top-right and bottom-right images, the color transitions are not smooth, indicating a simplistic approach to modeling the surface. This results in an unrealistic representation of how light interacts with the object.

Ground Truth:

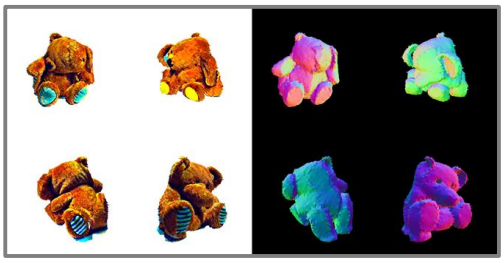
RGB Rendering: The texture of the sofa image is not clear enough, with noticeable blurriness on the backrest. The cushion and backrest appear to merge chaotically, and the edges of the sofa are quite blurry, with the edges of the cushion also appearing disorganized.
Normal Rendering: The geometric surface of the sofa has low precision, with a chaotic geometric representation of the backrest. The cushion and backrest do not conform to reality and exhibit geometric asymmetry.

Figure 40: Abnormal detail explanation example generated by **Nerf based** in 3D modality, with GPT-4o's answers

3D: Others [Opened-Ended]

Question:

This is an AI-generated 3D object rendering. Displays four-view RGB renders on the left and four-view normal maps on the right.
Your Task: Analyze and explain the inauthentic aspects of the 3D object.
1.RGB Rendering: Identify issues related to textures, colors, and overall apearance. Provide detailed explanations for why these aspects seem inauthentic. Reference specific parts of the images to support observations.
2.Normal Rendering: Analyze the normal maps for inconsistencies or errors in surface geometry. Explain any anomalies that suggest the object is artificially generated. Use specific visual evidence from the normal maps in the explanations.



(GPT-4o) Answer:

RGB Rendering: Inconsistencies in Texture: The texture of the teddy bear appears overly smooth and lacks the natural variation found in real materials. The fur does not exhibit the expected randomness and depth, making it look flat and artificial. **Color Uniformity:** The colors are too uniform and saturated, especially on the bear's body and limbs. Real objects typically have subtle variations in color due to lighting and material properties, which are missing here. **Shadows and Highlights:**The lighting seems inconsistent, with shadows and highlights not aligning with a single light source. This inconsistency can make the object appear less realistic, as natural lighting would create more coherent shadow patterns. **Ears and Feet:** The ears and feet have a uniform color and texture that do not reflect the typical wear and tear or variation seen in real plush toys. **Seams and Edges:** The seams and edges of the bear are too clean and lack the fraying or irregularities that would be present in a real stuffed animal.
Normal Rendering: Smoothness: The normal maps show an overly smooth surface, which lacks the fine details and imperfections that would be present in a real object. This smoothness can make the object appear less lifelike. **Lighting Interactions:** The way light interacts with the surface in the normal maps is too uniform. Real objects have varied surface geometry that causes light to scatter in different directions, creating a more complex interplay of light and shadow. **Limbs and Body:** The limbs and body of the bear in the normal maps do not show the expected depth and texture variations. This can make the object appear as if it is made from a single, uniform material. **Facial Features:** The facial features lack the subtle depth and contouring that would be present in a real teddy bear, making them appear flat and less defined.

Ground Truth:

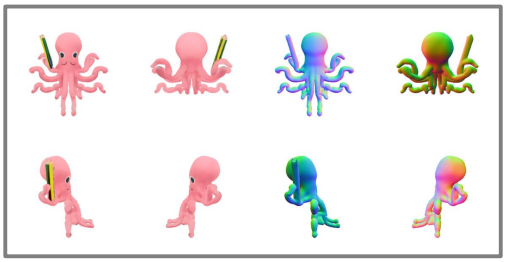
RGB Rendering: The texture resolution of the teddy bear is average, with abrupt texture transitions on the surface that do not match the surrounding colors. The textures of the bear's limbs are inconsistent from different angles.
Normal Rendering: The geometric surface of the teddy bear has poor precision, with rough surface normals and abnormal protrusions. The geometric shapes of the limbs are incorrect, and there is a lack of geometric symmetry.

Figure 41: Abnormal detail explanation example in 3D modality, with GPT-4o's answers

3D: Others [Judgement]

Question:

This is a rendering of a 3D object. The left side of the rendering shows a 4-view RGB rendering, and the right side shows a 4-view normal map, showing the surface texture and lighting interaction of the 3D model. The object may be generated by an AI model or may be a scan of a real object. Please determine whether the object was generated by AI model. Answer with yes or no.



(GPT-4o) Answer: YES

Cartoon-like Appearance: The object has a simplified, stylized design with smooth surfaces and exaggerated features (such as the large eyes and a pencil in its tentacles). These characteristics are common in digital or AI-generated 3D models, especially for animations or games. **Consistent Surface Smoothness**: The surface is highly smooth and lacks the imperfections typically seen in real-world objects or physical scans. AI-generated models often exhibit such smoothness because they are optimized for clean, stylized appearances. **Normal Maps**: The normal maps on the right side also suggest a simple, computer-generated surface with minimal texture, indicating that the lighting and surface detail are computed rather than captured from a real object. **Fantasy Design**: The octopus holding a pencil and its whimsical, playful style strongly suggest a creative, artificial design typical of 3D modeling software, not a scan of a real-world object.

Ground Truth: YES

Analysis:

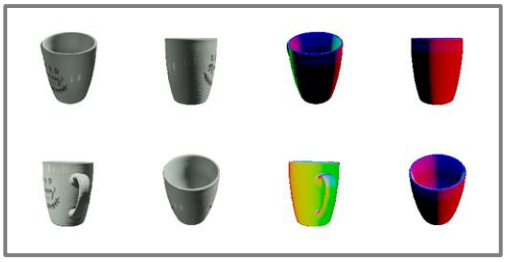
Although cartoon appearances are very common in animation and games, there are still items in the real world, such as dolls and toys, that have a cartoon appearance. These items can be scanned to obtain corresponding models with a cartoon look.

Figure 42: Judgement example in 3D modality, with GPT-4o's answers

3D: Real [Judgement]

Question:

This is a rendering of a 3D object. The left side of the rendering shows a 4-view RGB rendering, and the right side shows a 4-view normal map, showing the surface texture and lighting interaction of the 3D model. The object may be generated by an AI model or may be a scan of a real object. Please determine whether the object was scanned by a real-world camera. Answer with yes or no.



(GPT-4o) Answer: NO

Normal Maps: The right side of the image shows normal maps, which are typically generated during 3D modeling or texturing processes in computer graphics. These maps display surface orientation and how light interacts with the object, often used to simulate detailed textures without increasing the polygon count. In scans of real-world objects, you usually don't see raw normal maps in this form—they are processed into the final render.**Smooth and Clean Geometry**: The object appears to have smooth surfaces and uniform lighting behavior, which suggests it was created using a controlled environment in 3D software. Real-world scans often contain imperfections, noise, or irregularities due to scanning limitations, lighting variations, or texture distortions. **Lighting and Shading**: The consistent lighting across the different views implies the object was lit in a 3D rendering environment where lighting can be perfectly controlled. Real-world scans often have subtle lighting or reflections that are difficult to avoid. **Absence of Surface Noise**: Real-world scans tend to have some amount of surface noise or small texture distortions due to the scanning device's precision or environmental factors. The object in the image is free from discrepanciessuch imperfections, indicating a highly controlled digital environment. **Rendering Style**: The RGB rendering and normal maps presented in this specific format (side by side) are common in digital workflows for 3D model visualization, especially when developing or reviewing models in the context of gaming, simulations, or animation.

Ground Truth: YES

Analysis:

Regarding the normal maps, the results shown in the image can be visualized by collecting data through 3D scanning, rather than being generated solely through 3D modeling or texture processing. In terms of geometric shape, the smooth and clean surfaces are a result of scanning real-world objects, whereas models generated by AI often exhibit uneven surfaces. By controlling the stability of the scanning environment, consistent lighting can be achieved from multiple angles in the scanning results.

Figure 43: Judgement example in 3D modality, with GPT-4o's answers

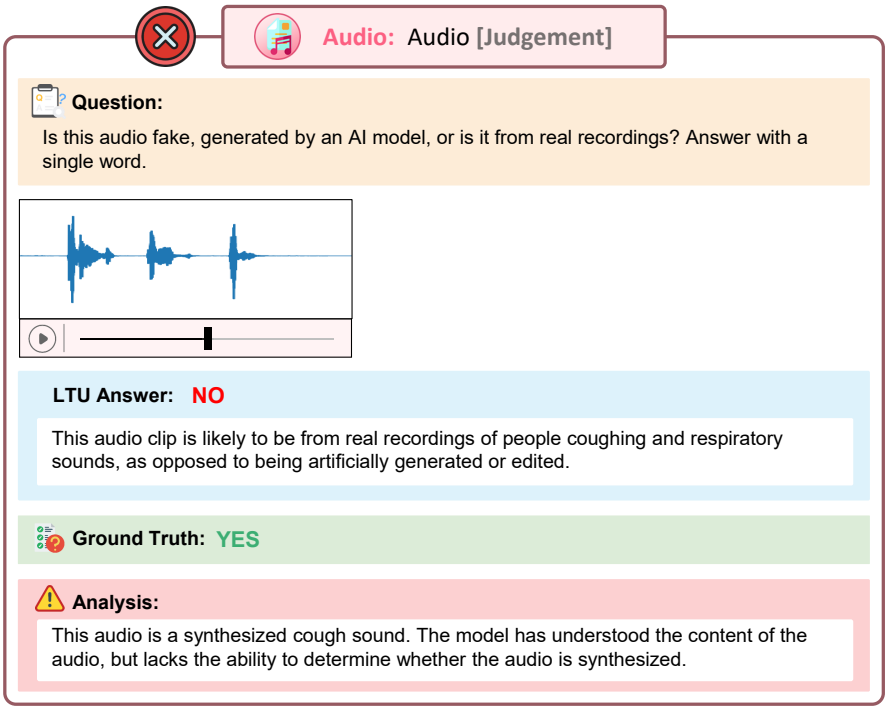


Figure 44: Judgement example of category **Audio** in audio modality, with LTU's answers

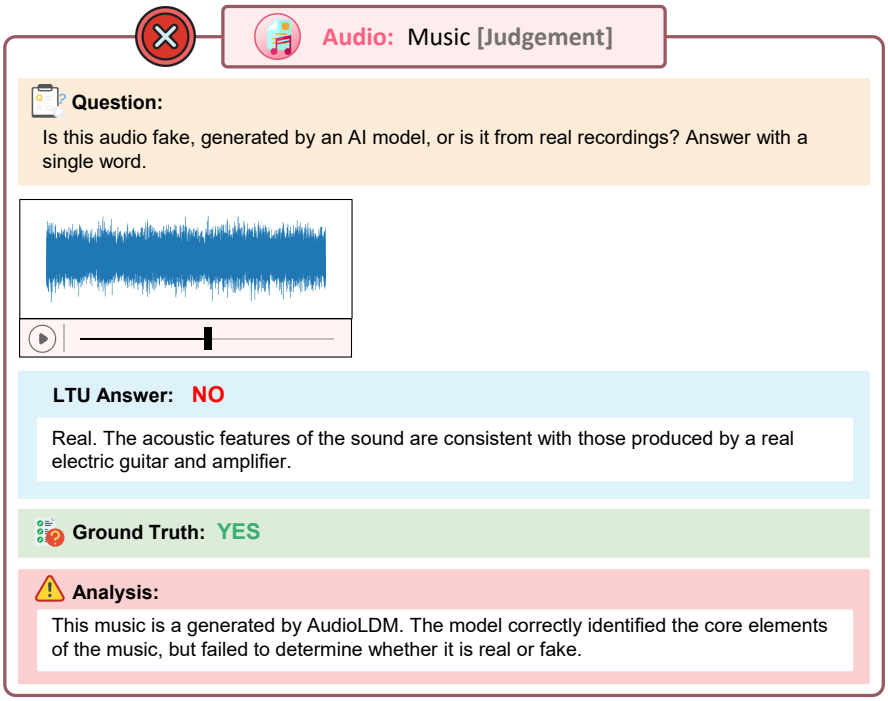


Figure 45: Multi-choice example of category **Music** in audio modality, with LTU's answers

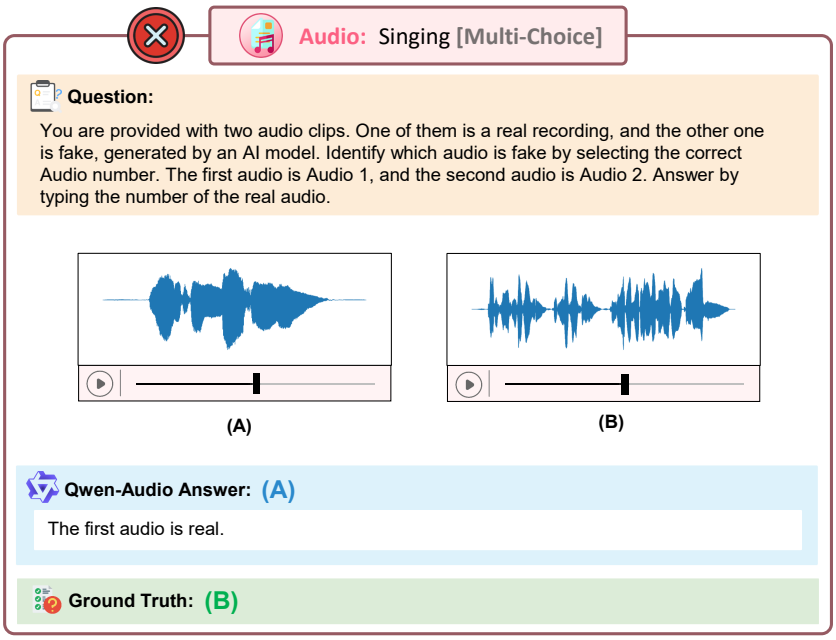


Figure 46: Multi-choice example of category **Singing** in Audio modality, with Qwen-Audio's answers

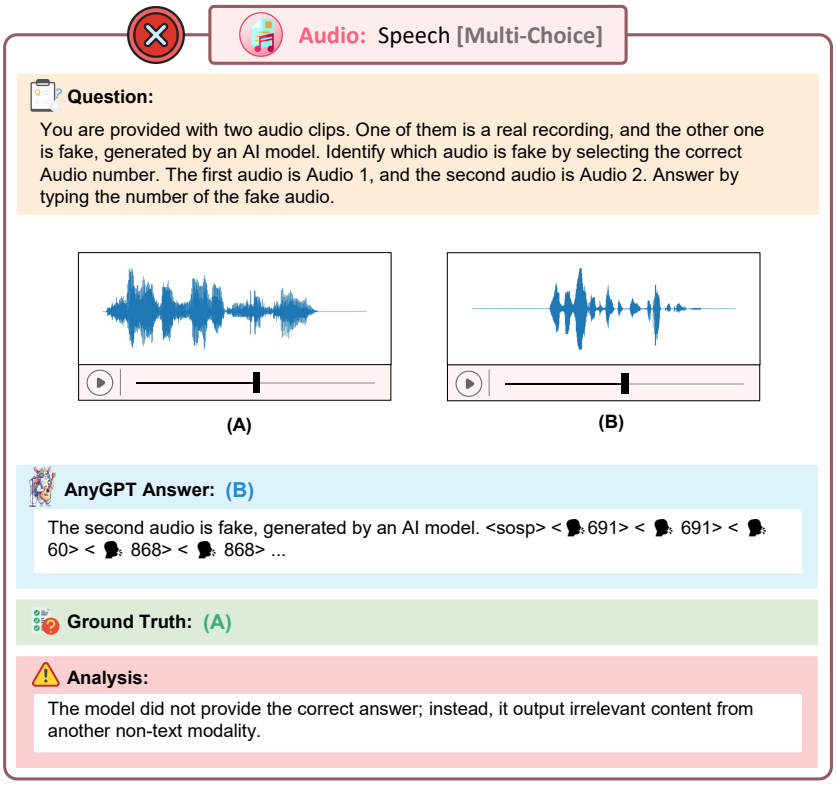


Figure 47: Multi-choice example of category **Speech** in audio modality, with AnyGPT's answers

Text: Abstract [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?

A. The first one
B. The second one

本文探讨了英国开放银行建设和国内中小银行在账户和数据开放方面的共同问题。指出这些银行因数据封闭面临经营困难，强调了数据开放对于提高银行服务质量和竞争力的重要性。

这调查也是英国开放银行建设的缘起之一,核心问题是如何做到账户和数据的开放。同样的问题也存在于国内的中小银行中,中小银行面临很多困难,账户和数据的封闭是经营中常常遇到的难题之一。

(GPT-4o) Answer: (A)

1.Polished and Concise Structure: Passage One is very concise and well-organized, presenting the problem and its significance in a straightforward way. It emphasizes the key issues without elaborating on context or background details, which is a style that AI often adopts to maintain clarity.
2.Generalization and Lack of Nuance: The phrasing in Passage One is broad and generalizes the issues faced by banks, focusing on the importance of data openness without diving deeply into the specifics. This broad overview approach is common in AI-generated content, where the focus is on summarizing key points.
3.Natural Flow and Detail in Passage Two: Passage Two has a more natural flow and includes nuanced explanations, such as referencing the "核心问题" (core problem) and how similar issues exist within domestic small and medium-sized banks. The passage is less polished and reads more like a genuine analysis of the challenges, which is typical of human-written content.
4.Less Formal and More Contextual Connections: The sentence structures in Passage Two are longer and more contextually connected, showing a more relaxed and exploratory style of writing. It discusses the challenges faced by banks in more context, which feels more authentic and less formulaic compared to the concise delivery of Passage One.

Ground Truth: (A)

Figure 48: Multi-choice example of category **Scientific Papers Abstract** in text modality, with GPT-4o's answers

Text: Abstract [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

为使电子设备之间能够相互协调、可靠地工作,电磁干扰及电磁兼容问题必须在设计中加以考虑。安检整机工作中电磁兼容测试辐射骚扰超标,其内部X射线源控制器是引起该问题的重要因素。通过对控制器外壳的屏蔽优化设计,包括对外壳材料、接缝、电缆出入口、通风孔和外接屏蔽电缆等,降低控制器辐射骚扰水平,使安检整机的电磁兼容性测试达标。

(GPT-4o) Answer: YES

1.Technical Jargon and Context: The passage uses domain-specific technical language and concepts, such as "电磁干扰" (electromagnetic interference), "电磁兼容" (electromagnetic compatibility), and "X射线源控制器" (X-ray source controller). These terms are used accurately and in a coherent context, which implies a strong understanding of the topic. Human writers familiar with this subject area typically exhibit this depth of technical knowledge.
2.Problem-Solution Structure: The passage follows a logical structure typical of technical documentation or a research summary: it starts with the problem (electromagnetic interference and compatibility issues in electronic devices), identifies the cause (X-ray source controller), and proposes a solution (shielding optimization of the controller housing). This step-by-step presentation of a technical issue and its resolution is characteristic of human-written technical analyses.
3.Concise and Objective Tone: The tone of the passage is concise, formal, and objective, which is a common trait in human-written technical literature. It is focused on delivering information efficiently without any extraneous details or stylistic embellishments. This type of writing aims to be clear and informative for a specialized audience, such as engineers or designers, rather than for general readers.
4.Contextual Use of Industry Practices: The passage describes a practical approach to solving electromagnetic interference issues by discussing specific design considerations like "外壳材料" (housing material), "接缝" (seams), "电缆出入口" (cable entry/exit points), and "通风孔" (ventilation holes). This level of detail and focus on practical, industry-standard methods is indicative of someone knowledgeable about engineering design processes and not just generating general text.
5.Focus on Compliance and Standards: The passage emphasizes achieving compliance with "电磁兼容性测试达标" (meeting electromagnetic compatibility testing standards), which is a realistic goal in engineering design and manufacturing. This emphasis on meeting specific testing standards is a sign of human-written content, as compliance with regulations is a practical concern for engineers and product designers that a human writer would highlight.
6.Clear Target Audience: The content of the passage is specifically targeted toward professionals dealing with electromagnetic compatibility in electronic devices, likely engineers or technical designers. The language, context, and details assume familiarity with technical terms and concepts, which suggests that the passage was crafted by a human with a particular audience in mind.

Ground Truth: YES

Figure 49: Judgement example of category **Scientific Papers Abstract** in text modality, with GPT-4o's answers

Text: Essay[Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

在科技浪潮席卷而来的今天，我们的生活充斥着各种电子设备和信息流。我们沉迷于手机屏幕，追逐着最新资讯，追求着最高效的工作方式。时间被切割成碎片，被各种待办事项填满，仿佛一台永不停歇的机器。然而，在追求效率的路上，我们却常常忽略了身边的美好。我们忘记了抬头看看天空的颜色，忘记了感受微风拂过脸颊的温柔，忘记了与家人朋友面对面交流的温暖。我们被困在信息的洪流中，迷失了感受自然、体验生活的能力。时间，不应被科技绑架，更不应被效率定义。掌控时间，并非将每分每秒都填满，而是要留白，要思考，要感受。放下手机，走出家门，去拥抱自然，去体验生活。在清晨的鸟鸣中醒来，在夕阳的余晖中漫步，在与亲朋好友的欢笑中感受真情。唯有掌控时间，才能真正体会世界的丰富多彩，才能在科技的浪潮中，依然保持对生活的热爱，对美好的追求。

我们消磨时间，也悔恨虚度了时间。在新时代的技术潮流中，当我们学会了用充满科技感的软件感受大千世界，便开始用无限的信息充斥着所有时间——浏览新鲜事儿，刷刷直播视频，关注好友动态……直到耗尽睡前的全部精力。当我们想要获得更多，便开始想尽办法“抢先”“争先”，不断被向前驱赶。可是，我们的世界不止有这些，我们还可以去欣赏“江间波浪兼天涌，塞上风云接地阴”的壮阔，还可以去感受“揉碎在浮躁间，沉淀着彩虹似的梦”的温柔，更可以来一场“携手藕花湖上路，一蓑黄梅细雨”的闲庭信步。只有当我们做了时间的主人才更能体会——我的世界，大于这世界。

(GPT-4o) Answer: (B)

1.Stylistic Richness and Poetic Elements: Passage Two is written with a poetic touch, using classical Chinese references and vivid imagery such as "江间波浪兼天涌" and "塞上风云接地阴." These references to traditional Chinese poetry and literature indicate a depth of cultural context and artistic flair, which is more typical of human authorship.
2.Emotional Flow and Authenticity: Passage Two has a more natural emotional flow, moving from a critical view of how technology overwhelms time to a more reflective appreciation of life's broader experiences. It uses metaphorical language and artistic expression that feels spontaneous and rich, characteristic of human creativity and introspection.
3.Nuanced Perspective on Time: Passage Two delves into a philosophical perspective on time and life, connecting different emotional experiences and views on nature. It balances criticism of technology with admiration for the beauty of life in a way that feels authentic and thought-provoking, rather than merely informative or advisory.
4.Passage One's Structured and Balanced Tone: Passage One, while well-written, maintains a clear structure and balanced tone, presenting an argument in a more straightforward manner. It focuses on the downsides of technology and the importance of taking breaks, which is delivered in a structured way without the deeper, reflective style seen in Passage Two. The use of parallel structures and emphasis on key themes (like "掌控时间") are signs of AI-generated content aiming for clarity and coherence.

Ground Truth: (A)

Figure 50: Multi-choice example of category **Essay** in text modality, with GPT-4o's answers

Text: Essay [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

南朝《唱导文》有言：“一善染心，万劫不朽；百灯旷照，千里通明”。行善利他之举正是如此，它能够驱散阴霾、照亮人心，像千盏明灯高悬，照得万里夜空一片明朗。在席卷全球的新冠疫情面前，人类的命运前所未有地紧密相联。我国作为和平崛起的东方雄狮，为构建一个开放、和平、美丽的新世界，对全球公共卫生事业尽责，与友邦患难相助，为海外源源不断地提供疫苗、医护的支援，在危机中发出了人类命运共同体的时代最强音。

(GPT-4o) Answer: YES

1.Cultural and Literary References: The passage starts with a quotation from the Southern Dynasties' "唱导文," stating "一善染心，万劫不朽；百灯旷照，千里通明" (One act of kindness can endure for eternity; a hundred lights can shine for a thousand miles). Such culturally and historically specific references indicate a deep familiarity with traditional Chinese literature and philosophy, something a human writer typically uses to provide depth and authority to their writing.
2.Elegant and Poetic Language: The passage uses highly poetic and expressive language, with metaphors and vivid imagery such as "驱散阴霾" (dispel the haze), "照亮人心" (illuminate the human heart), and "像千盏明灯高悬" (like a thousand bright lights hanging high). The elegant writing style, use of parallelism, and lyrical expressions are indicative of human authorship, aiming to evoke emotions and create a strong impression.
3.Complex Sentences and Thoughtful Structure: The structure of the passage is sophisticated, with long, complex sentences that convey multiple ideas and themes, from the idea of kindness and public health to global unity. It provides a thoughtful progression from ancient wisdom to contemporary global events, showing a deliberate organization of ideas to lead the reader through a coherent argument.
4.Strong Emotional and Moral Appeal: The passage emphasizes themes of altruism, solidarity, and global cooperation in the face of the COVID-19 pandemic, using emotive language to call for unity and collective well-being. This kind of strong moral appeal and the way it ties the pandemic to larger philosophical ideas about humanity and destiny is typical of human writing, which aims to persuade and inspire.
5.Use of Rhetorical Techniques: The passage effectively uses rhetorical devices like repetition, contrast, and parallelism to strengthen its message. For example, phrases like "驱散阴霾、照亮人心" (dispel the haze, illuminate the human heart) and "我国作为和平崛起的东方雄狮" (China as the peacefully rising lion of the East) are crafted to be memorable and impactful. This level of rhetorical sophistication is a common feature in human writing, particularly in essays or speeches meant to inspire or rally support.
6.Contextual Awareness and Real-World Relevance: The passage references the global COVID-19 pandemic and China's role in providing public health support and vaccines. The human-like perspective on China's international contributions and its framing within the context of building a "人类命运共同体" (community of shared future for humankind) suggests an understanding of contemporary geopolitical events and moral imperatives. This level of contextual awareness and its connection to real-world issues is typically found in human-authored content.

Ground Truth: YES

Figure 51: Judgement example of category **Essay** in text modality, with GPT-4o's answers

Text: News [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

The annual World Economic Forum (WEF) is underway in Davos, Switzerland, drawing over 2,000 prominent business and political leaders from around the globe. Key attendees include UK Prime Minister Tony Blair and South African President Thabo Mbeki, who will engage in discussions on pressing global issues such as China's growing economic influence and the future of Iraq following its recent elections. Unlike previous years, protests against the WEF are anticipated to be minimal, with anti-globalization activists having cancelled a planned demonstration. In a parallel event, the Brazilian city of Porto Alegre is hosting the World Social Forum, which seeks to promote alternative solutions to global problems. The WEF will tackle the world's most pressing challenges, shaping the global agenda for the year ahead..

World leaders gather to face uncertainty more than 2,000 business and political leaders from around the globe are arriving in the swiss mountain resort davos for the annual world economic forum (wef). For five days, they will discuss issues ranging from china's economic power to iraq's future after this sunday's elections. Uk prime minister tony blair and south african president thabo mbeki are among the more than 20 government leaders and heads of state leaders attending the meeting. Unlike previous years, protests against the wef are expected to be muted. Anti-globalisation campaigners have called off a demonstration planned for the weekend. The brazilian city of porto alegre will host the rival world social forum, timed to run in parallel with the wef's ritzier event in davos..

(GPT-4o) Answer: (A)

1.Clarity and Structure: Passage One is well-structured with clear sentences and logical progression of ideas. It effectively introduces the World Economic Forum (WEF), mentions the key attendees, and provides context for both the WEF and the parallel World Social Forum event in Brazil. The phrasing is polished, which is typical of professional human writing.
2.Language and Detail: Passage One uses more precise language and contextually appropriate vocabulary. Phrases like "pressing global issues," "recent elections," and "shaping the global agenda" suggest a nuanced understanding of the context, which adds depth to the content.
3.Balanced Tone and Objectivity: The passage maintains a balanced, objective tone throughout, providing an informative and neutral overview. It avoids overly dramatic or informal language, which often can be seen in AI-generated content.
4.Stylistic Consistency: Passage Two, while informative, lacks some stylistic consistency and flow. For example, "world leaders gather to face uncertainty" starts with a dramatic and somewhat ambiguous opening, which contrasts with the more factual style of the rest of the passage. Additionally, there are small grammatical inconsistencies, like the lack of capitalization in "swiss mountain resort davos" and "uk prime minister tony blair," which are less common in professional writing by humans but can occur in AI-generated texts.

Ground Truth: (B)

Figure 52: Multi-choice example of category **News** in text modality, with GPT-4o's answers

Text: News [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

World leaders gather to face uncertainty more than 2,000 business and political leaders from around the globe are arriving in the swiss mountain resort davos for the annual world economic forum (wef). For five days, they will discuss issues ranging from china's economic power to iraq's future after this sunday's elections. Uk prime minister tony blair and south african president thabo mbeki are among the more than 20 government leaders and heads of state leaders attending the meeting. Unlike previous years, protests against the wef are expected to be muted. Anti-globalisation campaigners have called off a demonstration planned for the weekend. The brazilian city of porto alegre will host the rival world social forum, timed to run in parallel with the wef's ritzier event in davos.

(GPT-4o) Answer: YES

1.Journalistic Style and Real-World Context: The passage adopts a journalistic style, covering real-world events, such as the World Economic Forum (WEF) in Davos, Switzerland, and its key attendees, including Tony Blair and Thabo Mbeki. It mentions specific details like the dates, issues discussed (China's economic power, Iraq's future), and the concurrent World Social Forum in Brazil. This kind of context-rich reporting is typically crafted by human writers to inform readers of current events.
2.Proper Structuring and Logical Flow of Ideas: The passage flows logically, introducing the main event (WEF in Davos), outlining the participants (business and political leaders, heads of state), and providing additional context (discussions about global issues and protests). The writing has a clear structure, with each sentence logically connected to the previous one, which is typical of human-written articles aiming to provide coherent and comprehensive coverage of an event.
3.Balanced and Neutral Tone: The tone of the passage is neutral and balanced, a common trait of news reporting, as it aims to present facts rather than opinions. There is no overtly biased language or exaggeration, and the focus is on providing information. This balanced presentation is a hallmark of human journalism, where objectivity is generally a goal.
4.Inclusion of Global and Specific Details: The passage includes both broad topics (global leaders discussing world issues) and specific details (the role of anti-globalization protests, concurrent forums in Porto Alegre). Human writers often include these specific details to paint a complete picture of the event, while AI-generated content may lack this level of depth and contextual accuracy.
5.Imperfect Language Use: While the passage is mostly coherent, there are minor grammatical issues, such as missing capitalizations (e.g., "swiss mountain resort davos" instead of "Swiss mountain resort Davos") and slight awkwardness in phrasing (e.g., "leaders and heads of state leaders attending"). These imperfections in grammar and style are more characteristic of human writing, which can be prone to occasional errors or rushed editing, especially in real-time reporting or drafts.

Ground Truth: YES

Figure 53: Judgement example of category **News** in text modality, with GPT-4o's answers

Text: Novel [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?

A. The first one
B. The second one

Sar stared at him in awe, as one would stare at an unbelievable monstrosity, born in a fertilely imaginative mind and portrayed in the TDs by an actor suitably deformed by the genius of make-up. But this was no TD image—this monster was real and he was standing only a few feet away, leering at her, challenging her, hurting her, condemning her..

Sar froze, her eyes widening in disbelief and terror. Before her loomed a figure torn straight from nightmares, a grotesque impossibility made real. It was close, too close, its menacing presence bearing down on her, choking her with a fear so palpable it poisoned the very air she breathed. Her heart pounded a frantic retreat, but her body remained rooted, ensnared by the grim fascination of her own private horror..

(GPT-4o) Answer: (A)

1.Stylistic Complexity and Unconventional Language: Passage One uses unique phrasing and complex, almost theatrical descriptions like "an unbelievable monstrosity, born in a fertilely imaginative mind and portrayed in the TDs." This type of creativity and unusual word choice suggests a human writer exploring language in a more original way, which is less typical of AI-generated content.
2.Narrative Depth and Tone Shifts: The passage has subtle shifts in tone and a layering of emotions that are intricately woven together, such as awe, disgust, and condemnation. This kind of multifaceted depiction of emotions often reflects a human author's nuanced grasp of storytelling, whereas AI-generated text may sometimes lean toward more predictable or straightforward emotional arcs.
3.Realistic and Natural Imperfections: There is a slightly fragmented and intense flow in Passage One, as seen in the rapid sequence of feelings Sar experiences: awe, hurt, condemnation. This flow feels more like a natural human storytelling style, with an urgency and intensity that doesn't overly structure each sentence for maximum clarity. Human writing often embraces these imperfections for stylistic effect.
4.Over-Dramatization in Passage Two: Passage Two is quite dramatic, with phrases like "a grotesque impossibility made real" and "choking her with a fear so palpable it poisoned the very air she breathed." While it paints a vivid picture, it follows a more conventional and heightened style, typical of AI-generated content attempting to convey strong emotions vividly.
5.More Fluid Imagery in Passage One: Passage One uses less explicit descriptions of terror, focusing instead on Sar's reaction and perception of the figure. This indirect approach feels more nuanced and indicative of human writing, which often leaves some details to the reader's imagination.

Ground Truth: (A)

Figure 54: Multi-choice example of category **Modern Literature** in text modality, with GPT-4o's answers

Text: Novel [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

In the dimly lit drawing room of Fitzdottrel's grand estate, the tension was palpable. Merecraft, with his sly smile and obfuscated intentions, articulated his grand scheme involving the fabled dukedom of Drownedland. Fitzdottrel, gullible and avaricious, listened with rapt attention, unaware of the deception lurking in every word Merecraft uttered. Merecraft's plot, though not the linchpin of the unfolding events, laid the groundwork for the follies soon to ensue. The real drama began when Fitzdottrel, in a fit of misguided bravado, decided to leverage the office of the Master of Dependencies, a position as ostentatious as it was functionally ambiguous, to incite a quarrel with Wittipol. Wittipol, astute and perceptive, saw through the charade but played along for the amusement it provided. The quarrel, trivial in appearance, led to an unexpected turn of events - the transfer of Fitzdottrel's estate to Manly. Manly, with a sense of justice and perhaps pity, chose to restore the estate to Mrs. Fitzdottrel, a woman undeservedly caught in the web of her husband's absurdities. The fourth act's curtain rose to reveal Wittipol in an elaborate disguise as a Spanish lady, swathed in silks and lace, embodying the very essence of foreign fashion that was both admired and ridiculed by London society. The satire was biting; Jonson's pen spared no one as it skewered the absurdity of societal follies and the blind adulation of all things foreign. Wittipol's disguise was more than a mere façade; it was a mirror held up to the audience, reflecting their own pretensions and affectations. As the play approached its conclusion, the fifth act delved into a satire on demoniacs, adding another layer to Jonson's critique of contemporary abuses. The devil-plot, with its undertones of dark humor, served as a fitting appendix to the unfolding madness. Merecraft, the architect of deceit, found himself entangled in his own schemes, while Fitzdottrel, none the wiser, continued his hapless journey through the political and social labyrinths Jonson so deftly constructed. Critics have lauded the play for its rich historical context and scathing critique of societal ills. Through a masterful fusion of character-driven drama and intricate plot development, Jonson weaved a narrative that was as entertaining as it was thought-provoking. Elements of a comedy of manners intermingled with semi-political satire, creating a tapestry of ethics and human folly. As the final act drew to a close, the audience was left to ponder the enduring relevance of Jonson's themes - the eternal dance between deception and truth, folly and wisdom, all set against the backdrop of a society that, despite the centuries, felt all too familiar.

(GPT-4o) Answer: YES

1.Complex and Contextual Language: The passage uses intricate and sophisticated language to describe the play's characters and events. Terms like "palpable tension," "obfuscated intentions," and "ostentatious as it was functionally ambiguous" reflect a nuanced grasp of vocabulary and style. While AI models can produce complex sentences, the depth and precision in the choice of words show a level of intentionality that is typical of human writing.
2.Deep Analysis of Plot and Themes: The passage provides a detailed, analytical overview of the play's characters, their motivations, and the themes within the work. It interprets the satire and moral implications of the plot and offers insightful connections between different elements of the play. This kind of thematic dissection and commentary reflects a human's capacity to draw meaning and significance from literary works, something that AI models may not execute with the same depth of interpretation.
3.Clear, Structured Flow and Transitions: The passage is carefully structured, introducing Merecraft's scheme and Fitzdottrel's gullibility before delving into each act of the play. The narrative flows naturally, with smooth transitions from one topic to the next, and builds toward an insightful conclusion. Such a structured analysis, with a strong sense of narrative progression and logic, is characteristic of well-crafted human writing.
4.Historical and Literary Contextualization: The passage contextualizes the play's satire within the broader cultural and historical context, such as London society's attitudes toward foreign fashion and the critique of societal pretensions. It situates Jonson's work within its time, while also hinting at its contemporary relevance. This context-aware discussion indicates a depth of historical and literary knowledge that aligns more closely with a human writer's perspective.
5.Balanced and Nuanced Critique: The passage not only summarizes the plot but also critically engages with the play's structure, themes, and significance. It reflects on the "scathing critique of societal ills," "fusion of character-driven drama and intricate plot development," and the "comedy of manners intermingled with semi-political satire." This kind of balanced critique—acknowledging both the entertainment value and the deeper commentary of the work—demonstrates human-like critical thinking.
6.Sophisticated Use of Metaphor and Imagery: The writer uses metaphor and imagery effectively, such as comparing Wittipol's disguise as a Spanish lady to a "mirror held up to the audience," and calling the play's satire "biting." Such metaphorical language and use of imagery add depth and color to the analysis, which is often a hallmark of skilled human writers.
7.Reflection on Relevance and Themes: The passage ends with a reflective statement on the "enduring relevance of Jonson's themes," contemplating how the dance between "deception and truth, folly and wisdom" remains pertinent. The capacity to relate a historical work to modern-day themes is a sign of human analysis that connects past and present in a meaningful way.

Ground Truth: NO

Figure 55: Judgement example of category **Modern Literature** in text modality, with GPT-4o's answers

Text: Ancient Literature [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

William Barker, a devoted servant of the late Sir John Fastolf, along with Margret Wyssetyr, testified about the unique relationship between Sir John and William Paston, a respected gentleman from Norfolk. Paston was more than just a casual acquaintance; he was a trusted kinsman to Sir John, a bond that led to him being appointed as a feoffee, or trustee, for Sir John's properties. This role was not one Paston took lightly, and despite not receiving any form of payment or reward during Sir John's lifetime, he willingly undertook substantial tasks and legal responsibilities, driven purely by goodwill and the strength of their relationship.Following Sir John's passing, the executors of his estate turned to Paston for assistance. Without hesitation, Paston continued to shoulder the considerable burden of managing extensive legal affairs, embarking on numerous journeys, and covering substantial costs to protect and defend Sir John's estate. This was not a task he took on with the expectation of personal gain; rather, it was an act of loyalty and respect for his late kinsman.The magnitude of Paston's commitment cannot be overstated. The legal matters he dealt with were complex and time-consuming, requiring a deep understanding of the law and unwavering dedication. The journeys he undertook were not merely casual trips but arduous expeditions, often fraught with the challenges and inconveniences of travel in that era. Moreover, the costs he incurred were significant, reflecting the high stakes involved in managing and defending a substantial estate like Sir John's.Paston's actions were not driven by self-interest but by a profound sense of duty and affection for Sir John. He endured great personal inconvenience and expense, demonstrating an unwavering commitment to upholding Sir John's legacy and ensuring that his estate was managed with the same care and diligence that Sir John himself would have shown.In essence, William Paston's role in Sir John Fastolf's life and posthumous affairs was not merely that of a trusted kinsman but of a steadfast ally and guardian. His actions were a testament to the depth of their relationship and the strength of his character, qualities that were evidently recognized and valued by those who knew him, including William Barker and Margret Wyssetyr. Their testimony serves as a poignant reminder of the power of loyalty and goodwill, even in the face of considerable challenges and personal sacrifices.

(GPT-4o) Answer: NO

1.Complex Structure and Detailed Context: The passage provides a well-organized historical account, describing specific individuals (William Paston, Sir John Fastolf, William Barker, and Margret Wyssetyr) and their relationships in detail. The narrative unfolds in a logical sequence, starting with background information on the characters, followed by a description of Paston's role, and ending with a reflection on the significance of his actions. Such intricate storytelling and detailed context are characteristic of human writing.
2.Emotional Depth and Character Analysis: The passage delves into the motivations and emotions of William Paston, portraying him as a character of loyalty, duty, and respect for Sir John Fastolf. It emphasizes his selflessness, sense of responsibility, and the emotional bond he had with Sir John. This type of emotional depth, character analysis, and understanding of human relationships are typical traits of human-authored biographical or historical texts, as they seek to explore underlying motivations and character virtues.
3.Sophisticated Use of Vocabulary and Style: The passage uses a rich and sophisticated vocabulary, with phrases such as "feoffee or trustee," "substantial tasks and legal responsibilities," "magnitude of Paston's commitment," and "a profound sense of duty." The formal language and varied sentence structures contribute to an elegant writing style that aligns well with human-authored prose. While AI can generate sophisticated vocabulary, the consistent use of complex phrasing throughout the passage suggests intentionality and coherence that are hallmarks of human-crafted text.
4.Balanced and Nuanced Tone: The tone of the passage is balanced and respectful, presenting an account of Paston's life with a reflective and dignified approach. The narrative refrains from sensationalism and maintains a nuanced perspective, presenting both the challenges and virtues of Paston's actions in a fair and comprehensive manner. Human writers often adopt such a balanced approach when aiming to provide an unbiased and thoughtful depiction of real historical figures.
5.Focus on Historical Context and Legacy: The passage emphasizes not just the events themselves, but also their significance and impact on Paston's legacy. It places Paston's actions within the broader context of loyalty, duty, and selflessness, making connections to societal values of the time. This focus on interpreting historical events and contemplating their meaning is a trait typical of human-written history and biography, as it requires a deeper understanding of context and implication.
6.Detailed and Realistic Depictions of Historical Challenges: The passage describes the complexities of legal matters, the challenges of traveling in that era, and the personal sacrifices involved in managing an estate. It recognizes the difficulties Paston faced and highlights the arduous nature of his role. The human touch in relating such realistic challenges, paired with an understanding of historical context, is a strong indicator that the passage is human-written.

Ground Truth: NO

Figure 56: Judgement example of category **Ancient Literature** in text modality, with GPT-4o's answers

Text: Ancient Literature [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

帝巡狩，至晋阳，省风俗，会诸侯。驻跸潞州，设牛酒之宴。外州刺史，迎驾父老、道士、僧尼之徒，跋涉而来，慰问行旅之劳。刺史赐彩四十匹，父老以下，各赐三匹。

朕巡狩晋阳，观风问俗，肆觐群后，存问百年。诸侯待于境者，抑惟故事，今停跸潞州，劳问牛酒。其外州刺史及迎驾父老道士僧尼等，远来至此，颇亦艰辛，宜并令预会。刺史赐物四十匹，父老已下各赐物三匹。

(GPT-4o) Answer: (A)

1.Polished and Balanced Structure: Passage One is structured in a polished and balanced manner, describing the events of an imperial tour with clear and concise sentences. The language is classical, but the structure is coherent and formal in a way that resembles how AI-generated text often attempts to emulate historical language without the nuanced stylistic variations found in authentic classical texts.
2.Simplified Classical Language: The passage employs classical Chinese language, but it uses a more standardized style without some of the deeper complexities or specific idiomatic expressions often found in genuine ancient texts. It describes events straightforwardly, which is typical of AI writing that aims to convey classical themes in an accessible manner.
3.Passage Two's Complex and Authentic Style: Passage Two contains more nuanced and complex phrasing, as seen in expressions like "观风问俗" and "肆觐群后," which suggest a deeper adherence to classical Chinese stylistic conventions. The language reflects a more natural and detailed royal decree style, showing the kind of linguistic texture and sophistication that is indicative of real historical writing.
4.Subtle Differences in Tone and Emphasis: Passage Two uses a tone that feels more regal and formal, with intricate wording choices that capture the authoritative voice of an imperial decree. The careful construction of the sentences to convey both command and respect for tradition is more typical of human-written historical texts.

Ground Truth: (A)

Figure 57: Multi-choice example of category **Ancient Literature** in text modality, with GPT-4o's answers

Text: Philosophical [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

In the humble recesses of my thoughts, I am keenly aware of the fractures and constraints that mar the landscape of my soul. I reach out, with a heart laden with humility, to the divine, yearning for a touch that transcends these earthly confines. My trust resides firmly in the boundless mercy and transformative power of the Almighty, that He may expand the horizons of my being, mend the broken fragments, and cleanse the stains of my faults. In this act of surrender, I find solace, for it is through His grace that I seek not just redemption, but also the fortification to resist future transgressions, and in His light, I am purged and preserved.

Narrow is the mansion of my soul; enlarge Thou it, that Thou mayest enter in. It is ruinous; repair Thou it. It has that within which must offend Thine eyes; I confess and know it. But who shall cleanse it? or to whom should I cry, save Thee? Lord, cleanse me from my secret faults, and spare Thy servant from the power of the enemy. I believe, and therefore do I speak.

(GPT-4o) Answer: (A)

1.Polished and Modern Language: Passage One employs modern, polished language that is consistent in style and formality. The phrasing is smooth and reflective, focusing on abstract concepts like "the fractures and constraints that mar the landscape of my soul." This structure is characteristic of AI-generated content, which aims for clarity and balanced expression without the nuances often found in historical or classic texts.
2.Abstract Reflections and General Tone: The language is broad and contemplative, providing a general reflection on themes like divine mercy and redemption without delving into deeply personal or historically grounded language. AI tends to create content that is widely applicable and thematically generalized, rather than tied to a specific context or individual style.
3.Classical and Personal Voice in Passage Two: Passage Two uses an older, more classical style of English, with direct appeals to the divine, such as "Lord, cleanse me from my secret faults." The structure is intimate and direct, showing a personal plea for divine intervention. The language is less polished but more deeply felt, which is often characteristic of human authorship in historical or religious contexts.
4.Direct and Emphatic Rhetoric in Passage Two: The passage contains phrases like "Narrow is the mansion of my soul; enlarge Thou it," showing a powerful and direct use of metaphor and prayerful expression. This more urgent and emotive style is typical of human-written devotional or confessional works, emphasizing the personal relationship between the individual and the divine.

Ground Truth: (A)

Figure 58: Multi-choice example of category **Philosophy** in text modality, with GPT-4o's answers

Text: Philosophical [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

在人际关系的复杂网络中，互相依赖和自身利益是两个不可忽视的关键因素。这两者不仅决定了人与人之间的合作与矛盾，也深刻影响着社会的和谐与动荡。首先，让我们考察父母与子女的关系。这是一种天然的血缘关系，通常被视为最亲密和无私的关系之一。然而，即使在这样的关系中，利益的冲突也会导致矛盾的产生。当子女的成长和发展与父母的期望和利益相一致时，家庭和谐、关系融洽；但当子女的追求与父母的期望产生冲突时，即使是这样亲密的关系，也可能出现分离和埋怨。其次，我们再来看主人与雇工的关系。这是一种典型的互相依赖关系，双方各自有着明确的利益诉求。当主人与雇工的利益一致时，即使关系疏远的人也可以通过合作实现共赢；但当利益发生冲突时，即使是长期合作的关系，也可能因为利益的纠纷而破裂。由此可见，无论是亲密的家庭关系，还是疏远的职业关系，利益的一致与冲突都是决定关系性质的核心因素。人际关系的复杂性在于，利益不仅仅是物质上的，更包括情感、心理和精神上的需求。当这些利益在不同层面上达成一致时，关系便会稳定和谐；而当这些利益在任何一个层面上产生冲突时，关系便会出现裂痕和矛盾。因此，理解和处理人际关系的关键在于，如何在利益的多样性中找到共同点，如何在矛盾中寻找和解的途径。只有通过理性的沟通和妥协，才能在复杂的人际关系中找到平衡，实现真正的和谐与合作。.

(GPT-4o) Answer: YES

1.Logical Structure and Thematic Development: The passage is organized in a clear and logical manner, progressing from general concepts to specific examples, such as parent-child and employer-employee relationships. The argument builds gradually, examining different facets of interpersonal relationships and how they are affected by interests and dependencies. Such structured thematic development is typical of human-written essays or articles, aiming to provide a comprehensive perspective.
2.Nuanced Exploration of Complex Ideas: The passage delves into the complexities of interpersonal relationships, exploring not just the material but also emotional, psychological, and spiritual aspects of human interests. The way it highlights both the potential for harmony and the risks of conflict reflects a nuanced understanding of human behavior and social dynamics, which often requires in-depth reflection and consideration—traits more characteristic of human writing.
3.Use of Concrete Examples to Illustrate Abstract Concepts: The passage uses relatable examples to explain abstract ideas about relationships. It moves from familial ties (parents and children) to professional interactions (employers and employees), making the discussion more concrete and easier to understand. This is a common technique used by human writers to make philosophical or sociological content more engaging and relevant.
4.Balanced Language and Tone: The tone of the passage is reflective and balanced, providing a neutral perspective on the dynamics of interpersonal relationships. The use of terms like "利益的冲突" (conflict of interests), "合作实现共赢" (cooperate to achieve mutual benefit), and "理性的沟通和妥协" (rational communication and compromise) show a thoughtful choice of words. The language is nuanced, striking a balance between emotional and rational elements, which is typical of human-written content aiming for depth.
5.Sophisticated Use of Concepts and Analogies: The passage effectively discusses the role of mutual interests and conflicts in relationships, using appropriate analogies to illustrate how harmony or discord arises. The distinction between material and non-material interests (emotional, psychological, and spiritual) is elaborated clearly and insightfully, suggesting that the writer has a deep understanding of these concepts, something that requires nuanced thinking and is less likely to be produced by an AI model.
6.Reflections on Solutions and Harmony: The conclusion of the passage does not just present problems but reflects on finding solutions for balancing conflicting interests, suggesting the use of "理性的沟通和妥协" (rational communication and compromise) as ways to achieve harmony. This approach of proposing thoughtful resolutions and ending on a note of balance and harmony is characteristic of human writers who aim to provide a complete perspective on an issue.

Ground Truth: NO

Figure 59: Judgement example of category **Philosophy** in text modality, with GPT-4o's answers

Text: Speech [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

Imagine making an apple pie with oranges—it's a whimsical idea, much like the curiosity that drives scientific discovery. Speaking of impactful moments, in 1988, James Hansen testified about the seriousness of human-caused global warming—a pivotal moment in environmental awareness. And on a lighter note, for those planning a wedding, don't miss the latest collection of stunning wedding dresses, tailored to make your special day unforgettable..

Baking apple pie? Discount orange warehouse has you covered! A fruit's a fruit, right? It's 1988, and scientist James Hansen has just testified to the United States Congress that global warming trends are caused by human activity, and will pose an increasing threat to humanity in the future.Well, well. That's unusually prescient for a human.Looking for a wedding dress? Try a new take on a timeless classic..

(GPT-4o) Answer: (A)

1.Narrative Coherence and Tone: The first passage flows naturally and maintains a cohesive tone throughout. It connects the seemingly unrelated topics of apple pie, James Hansen's testimony, and wedding dresses with a playful, whimsical touch. This kind of narrative creativity and smooth transitioning is often characteristic of human writing.
2.Topic Juxtaposition: While both passages mention similar topics (apple pie, James Hansen's testimony, and wedding dresses), the first passage has a more sophisticated style in how it draws parallels and makes transitions. The use of "whimsical idea" and "impactful moments" gives a deeper context and reflection on the themes discussed, which suggests a more human approach to storytelling.
3.Stylistic Elements: The second passage feels a bit disjointed and abrupt, with a mix of styles that don't flow together as seamlessly. Phrases like "Well, well. That's unusually prescient for a human" give off a robotic, somewhat sarcastic tone, which is more characteristic of an AI trying to mimic humor or commentary.
4.Commercial and Promotional Language: The first passage talks about wedding dresses in a way that seems genuinely enthusiastic and subtly promotional, while the second passage feels more overt and direct in its promotion with statements like "Baking apple pie? Discount orange warehouse has you covered!" which may come off as more mechanical.

Ground Truth: (B)

Figure 60: Multi-choice example of category **Speech** in text modality, with GPT-4o's answers

Text: Speech [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

In times of stress, economic worries, personal conflicts, or loss, many of us naturally seek refuge in our favorite spots in nature. These sanctuaries offer us not just comfort and relief, but also a space to play, reflect, and find purpose. Immersing ourselves in the beauty of nature restores our perspective and heals our spirit, reminding us of the simple pleasures and profound connections that ground us in our humanity. Let us cherish and protect these natural havens.

(GPT-4o) Answer: YES

1.Generic Theme and Flowery Language: The passage discusses nature's restorative qualities in a generic way, focusing on broad themes like stress relief, reflection, and healing. The phrasing is somewhat idealistic and poetic ("Immersing ourselves in the beauty of nature restores our perspective and heals our spirit"), which is a style often found in AI-generated content meant to sound profound or uplifting without delving into specific, nuanced details.
2.Lack of Personalization or Unique Insight: While the passage speaks to familiar human experiences, it does so in a way that lacks individual perspective or distinctive storytelling. A human writer may include specific anecdotes, observations, or a more personalized tone to make the message stand out. The AI-generated content often uses widely applicable statements that can apply to many scenarios without offering a unique angle.
3.Smooth, Yet Predictable Transitions: The flow between ideas is smooth, moving from stress relief to play and purpose, but in a predictable manner. Human writers might use a variety of transition techniques, mixing complex and simple sentences, or even inserting contradictory ideas to create a more dynamic narrative. AI-generated content tends to maintain a consistent and somewhat repetitive structure.
4.Overuse of Positive Tone and Encouragement: The passage uses positive language consistently, emphasizing cherishing, protecting, and connecting. While this positivity isn't necessarily a sign of AI alone, the lack of any contrast, skepticism, or deeper exploration of issues is a typical trait in AI-generated motivational or inspirational content.
5.Balanced and Non-Specific Call to Action: The final sentence, "Let us cherish and protect these natural havens," is a generic call to action that aligns with AI models' tendencies to wrap up paragraphs with a broad moral or positive takeaway. A human writer might be more specific about what actions to take, how to protect nature, or even explore the challenges in doing so.

Ground Truth: NO

Figure 61: Multi-choice example of category **Speech** in text modality, with GPT-4o's answers

Text: Wikipedia [Multi-Choice]

Question:

You have been shown two different passages. One of the passages is generated by a LLM, while the other is written by real people. Which of the above passages is most likely written by real people?
A. The first one
B. The second one

四大元素由这种原始物质形成之后，就以土、水、气、火的次序分为四层。火使水蒸发，产生陆地，水气上升把火围在云雾的圆管里。人们眼中看见象是天体的东西，就是这些管子的洞眼，使我们能从洞眼中望见里面的火。形成了四元素的最早雏形。另一个米利都派哲学家阿那克西米尼则把气或者空气看作是原始物质，并把其他元素说成是由空气组成。空气变得稀薄后就成了火。

米利都学派是古希腊的重要哲学流派之一，致力于探讨世界的原始物质和四大元素理论。该学派认为世界由火、土、水、气四大元素组成，这些元素通过相互作用形成各种自然现象。例如，火通过水气的上升形成天体。学派中重要的哲学家阿那克西米尼提出了空气是万物的原始物质的理论，他认为其他元素是由空气通过稀薄或浓缩过程转变而成。这一理论在古希腊哲学史上具有重要影响，推动了自然哲学的发展。

(GPT-4o) Answer: (B)

1.Coherent Structure and Comprehensive Overview: Passage Two presents a structured and comprehensive overview of the philosophical ideas of the Milesian school, focusing on the concepts of the four elements and the contributions of Anaximenes. The passage naturally integrates background context, explaining the school's significance and the role of air in their cosmology. This type of clear, integrated storytelling is typical of human writing, where the emphasis is on building a complete picture for the reader.
2.Contextual Details and Explanatory Style: Passage Two includes detailed explanations, such as the way fire forms celestial bodies and how air transitions into other elements. The use of phrases like "推动了自然哲学的发展" (promoted the development of natural philosophy) adds depth and reflection on the broader impact, a nuance often found in human-written content.
3.Passage One's Fragmented and Abrupt Descriptions: Passage One is more fragmented, jumping quickly from one concept to another without extensive context or detailed explanation. The phrasing is more abrupt, such as "形成了四元素的最早雏形" (formed the earliest form of the four elements), which lacks the richer context and connective language that a human writer might include to explain concepts more thoroughly.
4.Less Polished and More Literal Language in Passage One: The language in Passage One seems more literal and somewhat disjointed, focusing primarily on listing the formation and interaction of the four elements. The passage feels like a direct translation or summary of ideas without the elaborate flow that human writers often include to make the content more accessible and engaging.

Ground Truth: (A)

Figure 62: Multi-choice example of category **Wikipedia** in text modality, with GPT-4o's answers

Text: Wikipedia [Judgement]

Question:

You have been shown one passage. This passage may be generated by a LLM or written by real people. Is this passage written by real people?

音程是在乐音体系中指两个音的高低关系，分为旋律音程和和声音程。旋律音程是指两个音符依次发声，较低的音称为根音，较高的音称为冠音。

(GPT-4o) Answer: YES

1.Technical and Accurate Musical Terminology: The passage uses precise musical terminology, such as "音程" (interval), "旋律音程" (melodic interval), and "和声音程" (harmonic interval). These terms are used correctly within the context of music theory, indicating a clear understanding of the subject matter, which aligns with human expertise in explaining specialized topics.
2.Clear Structure and Definition: The passage is structured in a way typical of textbook or educational content, starting with a broad definition of "音程" (interval) and then breaking it down into "旋律音程" (melodic interval) and "和声音程" (harmonic interval). This clear, logical approach to defining a term and providing subcategories is a natural way that humans present information, especially in educational contexts.
3.Natural Flow and Coherent Explanations: The passage maintains a coherent flow as it explains the distinction between the two types of intervals. It starts by defining the overarching concept (interval) and then moves on to provide an example and a definition ("较低的音称为根音，较高的音称为冠音" – the lower pitch is called the root, and the higher pitch is called the crown). This natural progression and layered detail are typical of human-written explanations.
4.Neutral and Informative Tone: The passage maintains a neutral, objective, and informative tone, which is common in human-written educational content. It avoids overly formal or overly casual language, providing an appropriate level of detail and depth for someone introducing musical concepts. This balanced tone is harder for AI-generated text to consistently maintain, especially in specialized domains.
5.Lack of Redundancy or Irrelevance: AI-generated content sometimes exhibits redundancy or introduces slightly off-topic details. In contrast, this passage is concise and stays strictly on topic, focusing only on defining intervals in music without veering off into irrelevant information or repeating itself unnecessarily.
6.Domain-Specific Knowledge: The passage assumes some level of familiarity with music theory, focusing directly on the definitions and relationships between notes without further elaboration for laypeople. This implies that the writer is catering to an audience with some prior knowledge, which suggests an intention to educate or clarify specific technical terms, a trait common in human-written instructional content.

Ground Truth: NO

Figure 63: Judgement example of category **Wikipedia** in text modality, with GPT-4o's answers