AEGIS: A Holistic Benchmark for Evaluating Forensic Analysis of AI-Generated Academic Images

Anonymous ACL submission

Abstract

We introduce **AEGIS**, **A** holistic benchmark for **E**valuating forensic analysis of AI-**G**enerated academic **I**mage**S**. Compared to existing benchmarks, AEGIS features three key advances: (1) **Domain-Specific Complexity**: covering 7 academic categories with 39 fine-grained subtypes, exposing intrinsic forensic difficulty, where even GPT 5.1 reaches 48.80% overall performance and expert models achieve only limited localization accuracy (IoU 30.09%); (2) **Diverse Forgery Simulations**: modeling four prevalent academic forgery strategies across 25 generative models, with 11 yielding average forensic accuracy below 50%, showing that forensics lag behind generative advances; and (3) **Multi-Dimensional Forensic Evaluation**: jointly assessing detection, reasoning, and localization, revealing complementary strengths between model families, with multimodal large language models (MLLMs) at 84.74% accuracy in textual artifact recognition and expert detectors peaking at 79.54% accuracy in binary authenticity detection. By evaluating 25 leading MLLMs, nine expert models, and one unified multimodal understanding and generation model, **AEGIS** serves as a diagnostic testbed exposing fundamental limitations in academic image forensics. Data and code are available[1].

1 Introduction

With the rapid advancement of generative models, AI-generated images have reached high visual fidelity, raising concerns about the trustworthiness of visual evidence. In response, image forensics has evolved into multiple detection methods, including frequency-domain (Corvi et al., 2022), diffusion-process (Chen et al., 2024a), patch-level (Chen et al., 2024b), and alignment-based approaches (Rajan et al., 2025). Meanwhile, leveraging visual understanding and reasoning, multimodal large language models (MLLMs) have been increasingly applied to image forgery analysis, either directly (Wen et al., 2025) or in conjunction with expert systems (Xu et al., 2025).

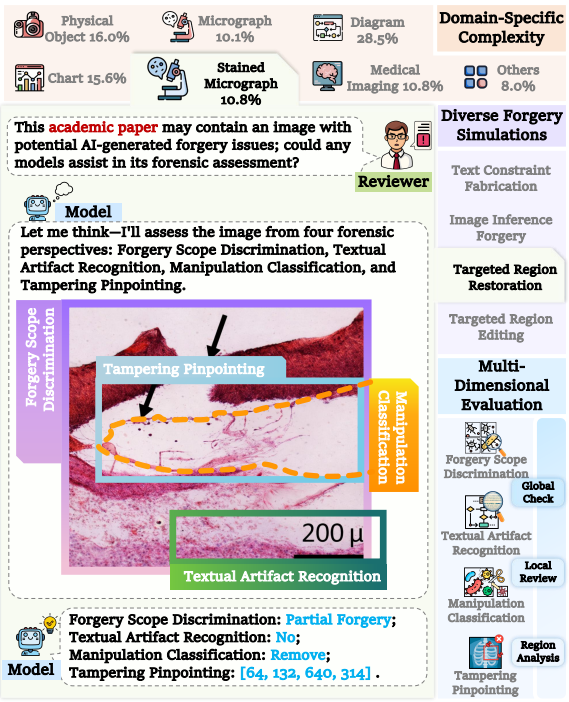


Figure 1: **AEGIS** investigates whether current models can effectively audit *AI-generated academic images* in scholarly papers by performing holistic forensic analysis across four complementary tasks.

Despite recent advances, the applicability of current models to expert-level academic review in high-stakes, domain-specific settings remains underexplored. From an expert artificial general intelligence (AGI) perspective (Figure 1), academic image forensics requires models to progress from coarse authenticity judgment to precise region attribution and spatial grounding, as academic images are often element-rich and structured. This raises the question of ***whether current models can generalize their capabilities to support the multi-stage forensic analysis required for academic review***.

[1] https://anonymous.4open.science/r/AEGIS-2E31

Benchmark	Scene	Task	# GMs	Forgery Strategies					Evaluation Tasks				Samples
				TCF	IIF	TRR	TRE	Total	FSD	TAR	MC	TP	
GenImage (NeurIPS 2023)		D	8	✔	✔	✘	✘	2	✔	✘	✘	✘	2.7M
Semi-Truths (NeurIPS 2024)		D	7	✘	✔	✔	✘	2	✔	✘	✘	✘	1M
AIGIBench (NeurIPS 2025)		D	12	✔	✔	✘	✔	3	✔	✘	✘	✘	440K
GIM (AAAI 2025)		D+L	3	✘	✘	✘	✔	1	✔	✘	✘	✔	2M
ForensicHub		D+L	10	✘	✘	✔	✔	2	✔	✔	✘	✔	5M
GRE (ACM MM 2024)		D+L	4	✘	✘	✘	✔	1	✔	✘	✔	✔	7.5M
DFBench (ACM MM 2025)		D	12	✔	✔	✔	✔	4	✔	✘	✘	✘	540k
AI GUARD (ACL 2025)		D+E	2	✔	✔	✘	✘	2	✔	✘	✘	✘	253K
AEGIS (ours)	**Academic**	**D+L**	**20**	✔	✔	✔	✔	**4**	✔	✔	✔	✔	**20K**

Table 1: **Comparison of AEGIS and other benchmarks.** : Natural; : Human; : Generic; : Life; : News; : E-commerce. **D**: Detection; **L**: Localization; **E**: Explanation. **GMs**: Number of generative models (excluding pre-2022, primarily GAN-based models). **TCF**: Text Constraint Fabrication; **IIF**: Image Inference Forgery; **TRR**: Targeted Region Restoration; **TRE**: Targeted Region Editing. **FSD**: Forgery Scope Discrimination; **TAR**: Textual Artifact Recognition; **MC**: Manipulation Classification; **TP**: Tampering Pinpointing.

As shown in Table 1, existing benchmarks fall short of supporting expert-level forensic research due to three key limitations:

- ***Generic-scene bias.*** *Academic images are characterized by fine-grained, structured, and semantically dense visual content that poses challenges for forensic interpretation.* In contrast, benchmarks like Semi-Truths (Pal et al., 2024) and ForensicHub (Du et al., 2025) focus on generic imagery such as faces, natural scenes, or social media content with coarser structures, limiting their suitability for academic forensic reasoning.
- ***Restricted forgery strategies.*** *In the context of academic imagery, forgeries exhibit diverse forms with high degrees of subtlety.* However, benchmarks such as AI GUARD (Zhang et al., 2025) and GRE (Sun et al., 2024) primarily focus on global forgery or single-type editing forgery, failing to simulate the complex and misleading manipulation behaviors.
- ***Forensic-agnostic evaluation protocols.*** *Expert academic review requires evidence-grounded forensic assessment beyond single authenticity prediction.* However, benchmarks such as GIM (Chen et al., 2025b) and DFBench (Wang et al., 2025a) focus on detection accuracy, overlooking key forensic dimensions such as manipulation scope and localization.

To fill this gap, we introduce **AEGIS**, a holistic benchmark comprising over 20k forensic questions, incorporating advanced generative models and simulating diverse forgery scenarios.

- ***Domain-Specific Complexity.*** AEGIS captures the complexity of academic imagery through systematic coverage of seven major categories and 39 fine-grained subtypes. Our taxonomy spans diverse scientific visual forms, including microscopic particles, cellular structures, and heterogeneous imaging modalities (*e.g.*, microscopy, medical, and physical imaging), enabling rigorous forensic evaluation in scholarly settings.
- ***Diverse Forgery Simulations.*** AEGIS covers four representative forgery types commonly found in academic publications, including: (1) Text Constraint Fabrication; (2) Image Inference Forgery; (3) Targeted Region Restoration; and (4) Targeted Region Editing. These forgeries are instantiated using **25 state-of-the-art generative models**, enabling realistic and adversarial simulation of academic image manipulation.
- ***Multi-Dimensional Forensic Evaluation.*** AEGIS introduces a multi-level and multi-perspective evaluation suite, including (1) Forgery Scope Discrimination, (2) Textual Artifact Recognition, (3) Manipulation Classification, and (4) Tampering Pinpointing. These tasks comprehensively assess models' capabilities in judgment accuracy, localization precision, and explainability depth, closely aligning with real-world academic review needs.

We evaluate 14 proprietary MLLMs (OpenAI, 2025b), 11 open-source MLLMs, one representative unified multimodal understanding and generation model, as well as 6 vision-only expert models

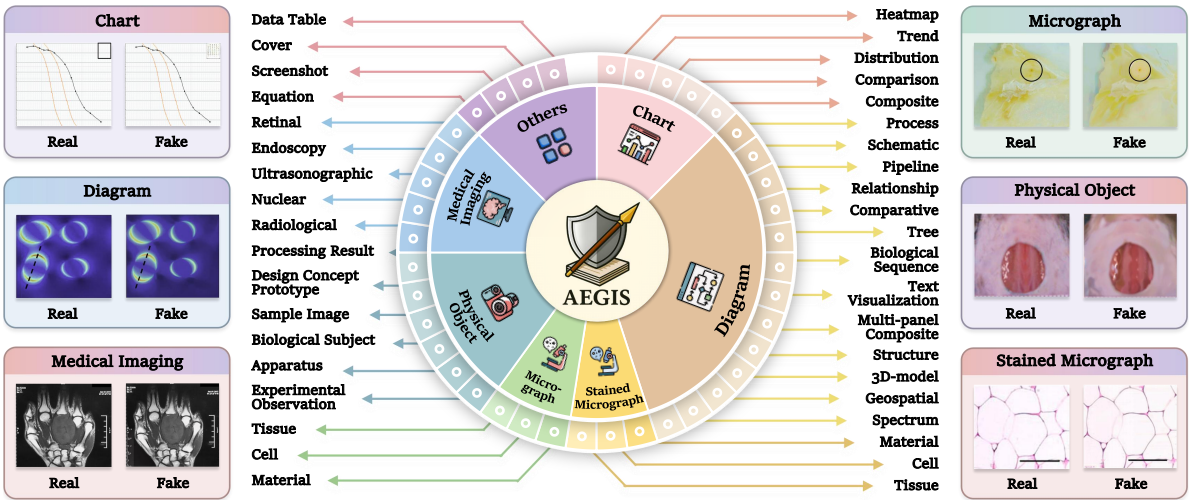


Figure 2: **Hierarchical Taxonomy of AEGIS.** We organize academic images into 7 categories and 39 fine-grained subtypes based on their structural and semantic characteristics. Real (left) and fake (right) examples are shown for comparison, where "fake" refers to AI-Generated forgeries.

and 3 hybrid MLLM-assisted expert models. The experimental results reveal three key findings:

- ***Domain-Specific Complexity Challenges Detection and Localization:*** Both MLLMs and expert models struggle on academic images, with GPT 5.1 achieving at 48.80% overall performance and expert baselines exhibiting an approximately 30% Real–Forgery F1 gap. This pattern indicates a structural bias toward geometric regularities, failing to generalize to the high-variance, texture-rich nature of scientific imagery and leading to severely limited localization accuracy (IoU 30.09%, with only one MLLM exceeding 50% region-level accuracy).
- ***Generative Diversity Highlights Forensic Vulnerabilities:*** Generative diversity exposes an adversarial asymmetry between generation and forensic analysis. Averaged across models, 11 generative models reduce accuracy below 50%, with 4 falling below 30%, indicating that existing static forensic cues struggle to generalize against the rapid evolution of generative systems.
- ***Comprehensive Forensic Reasoning Requires Complementarity:*** MLLMs excel at forensic reasoning, achieving 84.74% accuracy in textual artifact recognition and 60.07% in manipulation classification, capabilities beyond expert models. In contrast, expert detectors achieve over 70% accuracy in binary authenticity detection but are less robust to post-processing perturbations, indicating that reliable academic image scrutiny requires quantitatively complementary forensic evidence rather than any single paradigm.

2 Benchmark

2.1 Overview of AEGIS

We introduce **AEGIS**, a comprehensive benchmark for systematic evaluation of model performance in academic image forensics. As illustrated in Figure 2, **AEGIS** established a hierarchical taxonomy with 7 main categories (*i.e.* Chart, Medical Imaging, Physical Object, Micrograph, Stained Micrograph, Diagram, and Others), comprising 39 annotated subtypes. It defines four evaluation dimensions targeting distinct forensic capabilities, (*i.e.* Textual Artifact Recognition, Forgery Scope Discrimination, Manipulation Classification, and Tampering Pinpointing), with over 20k question instances. More detailed statistics of **AEGIS** are shown in Appendix A.

2.2 Dataset Construction

Data Curation Extraction and expert validation of academic panels. Details are in Appendix A.3.

- ***Paper Parsing.*** **AEGIS** collected over 4,000 high-quality academic papers from the open-access PubMed repository and performed document-level parsing to extract figures, captions, and panels, where a panel denotes the smallest indivisible visual unit in a academic image. This process yielded more than 50k candidate panels and 20k associated captions.
- ***Expert Curation and Validation.*** We conducted a two-stage quality control process involving manual screening and expert review. Panels with insufficient resolution, parsing errors, or lacking

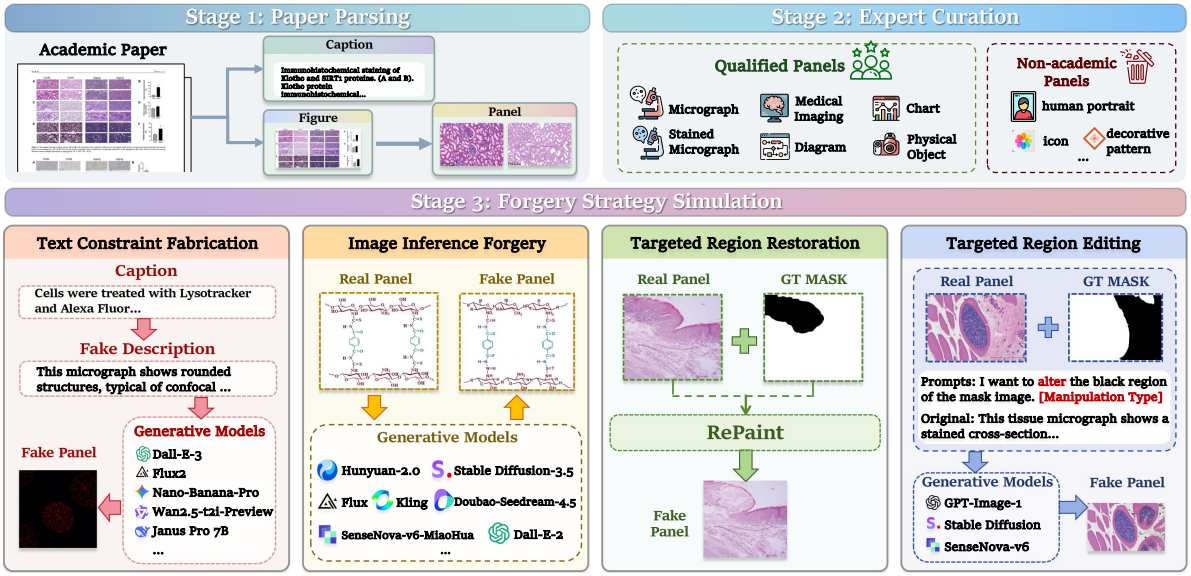


Figure 3: **Construction Pipeline of AEGIS. Stage 1: Paper Parsing** parses papers into captions, figures, and panels. **Stage 2: Expert Curation** retains qualified academic panels while excluding non-academic ones. **Stage 3: Forgery Strategy Simulation** synthesizes AI-generated academic image forgeries via four representative strategies.

academic semantics were removed, while panels consistent with the predefined hierarchical taxonomy were retained and annotated. Non-academic or non-forgeable visual elements (e.g., portraits, icons, and decorative graphics) were excluded, resulting in over 8,000 high-quality annotated panels for benchmark construction.

Generative Models To ensure representative quality and diversity for forensic evaluation, **AEGIS** incorporates 25 advanced generative models (Labs et al., 2025; Zhao et al., 2025; Esser et al., 2024) from three major architectures: (1) diffusion-based (*e.g.* Flux (Labs et al., 2025), Midjourney v6/v7, DALL-E 2/3, Kling); (2) autoregressive-based (*e.g.* CogView); and (3) hybrid (*e.g.* Wanx2.1-t2i-Turbo (Wan et al., 2025), GPT-Image-1). More Details are in Appendix A.4.1.

Forgery Strategy Simulations As illustrated in Figure 3, to simulate realistic forgery behaviors in academic scenarios, **AEGIS** defines four representative forgery strategies, inspired by common manipulation practices in scholarly imagery and aligned with the capabilities of contemporary generative models. Further details are provided in Appendix A.4.2.

- ***Text Constraint Fabrication.*** *TC models the fabrication of entire academic images from scratch under textual constraints.* We reconstruct captions of authentic academic images into semantically equivalent prompts, which are then used to guide text-to-image generative models (Labs et al., 2025), resulting in 3,121 Entire Forgeries.
- ***Image Inference Forgery.*** *II generates realistic and detailed globally forged images based on the original, simulating reference-guided.* Specifically, the model (Wang et al., 2025b) uses authentic academic images as references to synthesize 2,274 Entire Forgeries.
- ***Targeted Region Restoration.*** *Restoration simulates the forgery risks associated with reconstructing missing local content.* By applying masks to specific regions of authentic images, we instruct generative models to restore these areas, resulting in 1,650 Partial Forgeries.
- ***Targeted Region Editing*** *Edit emulates common "lightweight tampering" operations.* By providing masks or textual instructions, we perform localized edits on authentic images, including alterations, insertions, or removals, to produce 1,165 Partial Forgeries.

Data Quality Assurance To mitigate hallucinations from generative models, all AI-generated images undergo a dual-review protocol, evaluating both local manipulation plausibility and global visual fidelity. This process leads to the discard of 29% of AI-generated images, resulting in a final set of 8,210 high-quality samples. Verification, performed by five trained annotators, took approximately 200 hours. Detailed assessment is shown in Appendix A.4.3.

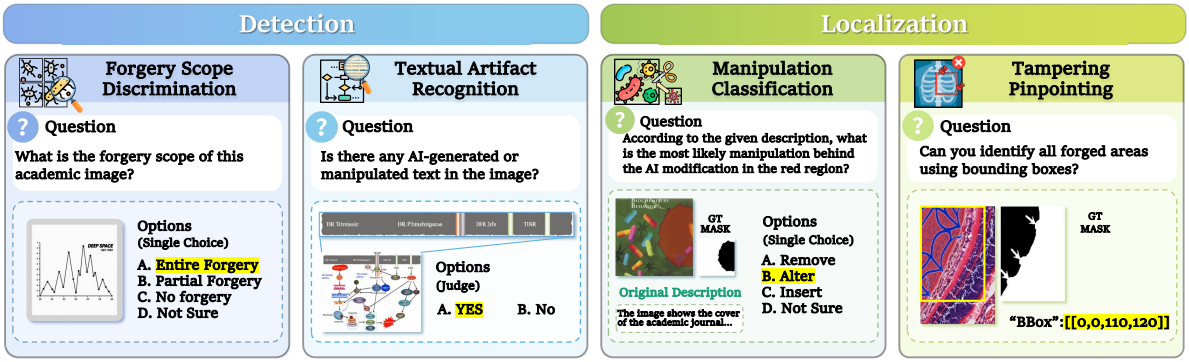


Figure 4: **Evaluation Design of AEGIS.** Four evaluation question types are designed to support staged forensic analysis, from global authenticity assessment to fine-grained region-level localization, including Forgery Scope Discrimination, Textual Artifact Recognition, Manipulation Classification, and Tampering Pinpointing.

2.3 Design of Evaluation Questions

AEGIS designs a set of progressively structured tasks to evaluate forensic capability from global authenticity judgment to fine-grained spatial localization, as shown in Figure 4. More details are provided in Appendix A.5.

- ***Forgery Scope Discrimination (FSD).*** This task evaluates coarse-grained authenticity judgment, requiring models to determine whether an input image exhibits no forensic artifacts, contains regionally constrained AI-generated traces, or represents a comprehensively synthesized fabrication.
- ***Textual Artifact Recognition (TAR).*** This task assesses fine-grained visual cue perception by requiring models to perform binary classification on whether textual regions in an input image exhibit traces of AI synthesis, relying on text region recognition and semantic consistency analysis.
- ***Manipulation Classification (MC).*** This task evaluates structure-aware reasoning, requiring models to infer the type of manipulation applied to an image. Given a red-highlighted manipulated region and the original image caption, the model analyzes the structural and contextual role of the edited area to classify the manipulation as insertion, removal, or alteration.
- ***Tampering Pinpointing (TP).*** This task focuses on fine-grained spatial grounding, evaluating the localization accuracy of partially edited regions in academic images. We adopt an adaptive granularity protocol: region-level bounding boxes for MLLMs and pixel-level masks for expert models, emphasizing spatial precision beyond global visual cues.

3 Experiments

3.1 Setup

Model. We evaluate 14 proprietary MLLMs (OpenAI, 2025a,c; DeepMind, 2025), 11 open-source MLLMs (Li et al., 2024b; Team et al., 2025; AI@Meta, 2025), a unified multimodal understanding and generation model (DeepSeek, 2025), and 9 expert models, including vision-only and hybrid systems, of which only SIDA (Huang et al., 2025) support multi-granular forgery discrimination.

Metric. We propose a new composite metric, the **Normalized Forensic Index (NFI)**, to reflect balanced forensic capability across tasks rather than peak performance on any single dimension. For localization, we employ an adaptive protocol: pixel-level evaluation for expert masks and region-level assessment for MLLM bounding boxes, using CLA to capture completeness and OLR to penalize over-localization. Further details are provided in Appendix B.

3.2 Experiment Analysis

3.2.1 Domain-Specific Complexity Challenges Current Models

We first investigate how the intrinsic complexity of academic imagery impacts forensic capability. Our analysis suggests that the high variance between structured and dense scientific visual forms poses a substantial challenge to the generalization capability of current architectures.

Holistic Capability Gap: No model meets the multifaceted forensic demands. As detailed in Table 2, models excel only at isolated tasks, a limitation quantified by the Normalized Forensic Index (NFI) where no evaluated MLLM exceeds 60%, indicating limited balanced forensic capability.

Model	Detection				Localization						NFI
	FSD		TAR		MC		TP				
							Region-Level		Pixel-Level		
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR	IoU	F1	
Human											
Human	44.20	37.83	76.14	75.02	68.01	51.11	–	–	–	–	–
Proprietary MLLMs											
GPT 5.1	50.99	46.16	76.43	76.38	**60.07**	**48.34**	46.87	12.25	–	–	**48.80**
GPT 4.1	66.34	59.98	83.57	82.74	44.55	40.05	25.93	**0.14**	–	–	43.31
OpenAI o4 Mini High	54.04	49.35	80.71	80.02	36.80	35.54	33.29	5.44	–	–	42.77
Claude Sonnet 4.5	25.36	19.62	58.76	58.70	44.80	39.70	27.41	24.32	–	–	26.83
Gemini 3 Pro Preview	64.37	53.44	**84.74**	**84.67**	48.54	45.42	39.14	20.09	–	–	45.79
Gemini 2.5 Flash	60.29	53.78	81.96	80.97	44.80	37.28	47.42	14.07	–	–	47.02
Doubao Seed 1.6	41.31	35.38	77.23	76.90	50.05	37.08	36.38	0.28	–	–	41.73
Doubao Seed 1.6 Thinking	44.57	37.13	77.92	77.57	58.47	41.85	46.00	0.64	–	–	46.66
Doubao Seed 1.6 Flash	29.26	23.77	60.67	60.39	35.89	30.96	**52.36**	1.78	–	–	36.03
Doubao 1.5 Thinking Vision Pro	39.78	35.31	75.55	75.44	36.65	32.33	47.53	1.71	–	–	42.39
Doubao 1.5 Vision Pro	39.87	32.52	72.35	72.32	29.17	28.96	26.96	2.02	–	–	34.07
Qwen3 VL Plus	38.77	35.72	79.25	78.98	59.28	36.89	5.76	0.53	–	–	16.53
Qwen VL Max	43.24	37.30	73.76	73.74	52.06	36.76	40.30	1.32	–	–	43.01
Qwen VL Plus	38.20	31.81	51.09	48.23	42.99	32.23	29.37	1.32	–	–	33.89
Open-Source MLLMs											
Gemma3 27B	44.86	42.36	66.88	58.39	42.49	34.76	48.70	28.56	–	–	37.55
Qwen2.5 VL 72B	37.63	33.16	70.81	70.76	51.76	35.96	32.29	0.53	–	–	38.71
LLaMA 4 Maverick	39.66	37.00	68.37	68.32	54.62	40.18	14.56	2.88	–	–	29.15
Ministral 3 14B	27.19	25.44	62.41	62.25	53.62	33.89	9.73	12.86	–	–	19.90
LLaVA NeXT 7B	10.05	12.23	53.82	53.36	59.04	27.01	10.37	6.47	–	–	16.53
LLaVA NeXT 34B	18.57	17.01	41.71	35.16	32.65	27.19	51.01	0.78	–	–	27.74
LLaVA Interleave 7B	33.53	20.61	64.68	54.01	26.00	26.67	5.65	8.28	–	–	13.61
LLaVA Interleave DPO 7B	35.56	26.09	59.01	53.98	25.15	24.99	25.22	7.50	–	–	28.18
LLaVA OneVision 7B	39.20	38.76	50.46	50.45	37.80	31.34	3.94	9.98	–	–	11.45
LLaVA OneVision 72B	30.37	28.57	56.28	55.92	55.31	39.42	33.89	17.05	–	–	33.81
LLaVA OneVision Chat 7B	40.54	39.94	50.44	50.26	39.47	32.45	4.14	6.45	–	–	12.19
Unified Multimodal Understanding and Generation Model											
Janus Pro 7B	24.97	27.18	60.25	42.59	24.18	24.10	5.31	0.25	–	–	13.77
Vision-Only Expert Models											
DRCT (Chen et al., 2024a) *	55.05	39.55	–	–	–	–	–	–	–	–	–
DIRE (Wang et al., 2023) *	58.16	50.41	–	–	–	–	–	–	–	–	–
AIDE (Liang et al., 2025) *	**79.54**	44.74	–	–	–	–	–	–	–	–	–
AlignedForensics (Rajan et al., 2025) *	47.57	42.48	–	–	–	–	–	–	–	–	–
Denoising Trajectory (Liang et al., 2025) *	73.68	65.37	–	–	–	–	–	–	–	–	–
DDA (Chen et al., 2025a) *	74.19	**67.07**	–	–	–	–	–	–	–	–	–
Hybrid MLLM-Assisted Expert Models											
FakeShield (Xu et al., 2025) *	59.72	62.08	–	–	–	–	–	–	**30.09**	**69.11**	–
SIDA (Huang et al., 2025)	52.91	64.38	–	–	–	–	–	–	3.66	54.28	–
FakeVLM (Wen et al., 2025) *	58.73	55.39	–	–	–	–	–	–	–	–	–

Table 2: Performance across four evaluation dimensions on **AEGIS**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Tampering Pinpointing remains a major bottleneck. Across all expert models, the highest pixel-level IoU reaches only 30.09%, while among MLLMs, only a single model exceeds 50% CLA.

Visual Density Bias: Dense semantics impede generalization. Performance correlates heavily with visual density. Figure 5 (c) shows that robust generalization on structured images (e.g., *charts*) contrasts sharply with marked degradation on dense categories like *stained micrographs*. This suggests a structural bias: current models tend to over-rely on explicit geometric regularities, while struggling to capture the complex texture details.

Authenticity Modeling Deficit: Expert models exhibit a pronounced forgery bias. The challenge of modeling diverse scientific distributions predisposes expert models toward biased prediction strategies. Figure 5 (a) shows most expert baselines achieve lower Real-F1 than Forgery-F1. This imbalance suggests that expert models, unable to encompass the manifold of *authentic* scientific imagery, default to flagging unseen patterns as forged.

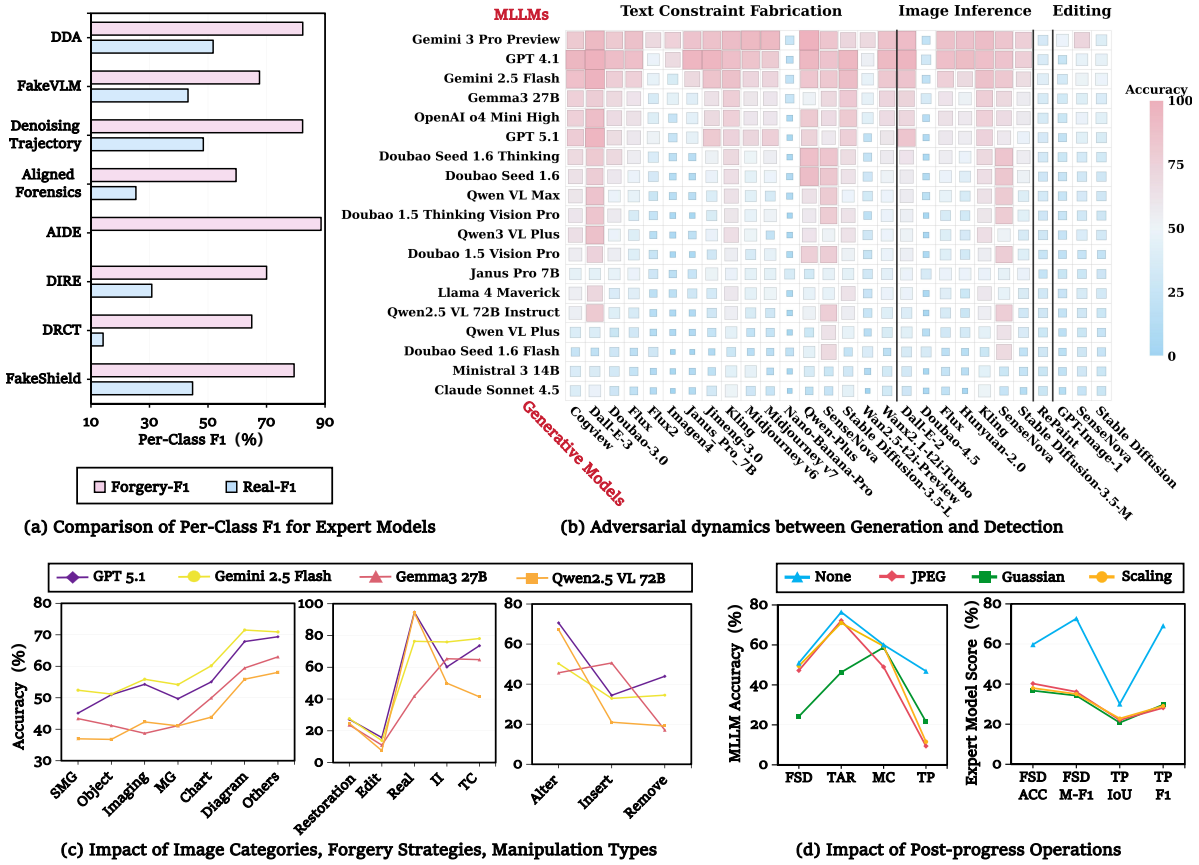


Figure 5: **Fine-grained experimental analysis of AEGIS. II**: Image Inference Forgery; **TC**: Text Constraint Fabrication; **SMG**: Stained Micrograph; **MG**: Micrograph.

3.2.2 Generative Diversity Highlights Forensic Vulnerabilities

Adversarial analysis reveals a capability gap: generative evolution outpaces forensic capability, particularly in fine-grained manipulations.

The Granularity Gap: Localized Forgeries Are Harder. Models perform reliably on globally synthesized images but show notable degradation on localized manipulations. As shown in Figure 5(c), performance on targeted restoration and editing consistently underperforms that on global forgeries, indicating the increased difficulty of localized forgery analysis.

Adversarial Asymmetry: Generation Outpaces Forensic Capabilities. Figure 5(b) reveals a clear adversarial asymmetry. Averaged across models, 4 generative methods reduce accuracy below 30%(*e.g.*, Nano-Banana-Pro), while robust defense remains limited to a few models (*e.g.*, Gemini 3 Pro Preview, GPT 4.1). This result shows that generative advances degrade performance more broadly than forensic capabilities can adapt, exposing a persistent gap in academic image forensics. More Details are in Appendix D.3.

3.2.3 Towards Expert AGI: Synergizing Detection and Reasoning

Our findings suggest that no single paradigm is sufficient; instead, the future of academic forensics likely lies in synergizing the sensitivity of experts with the reasoning of MLLMs.

Foreground–background entanglement complicates recognition. Figure 5(c) shows that *insertion* and *removal* are consistently harder to identify than *alteration*. Even the strongest MLLM, GPT 5.1, reaches 60.07% accuracy on Manipulation Classification, indicating reasonable reasoning capability but remaining challenges in separating foreground anomalies from coherent backgrounds.

Complementary Synergies: Orchestrating sensitive sensors with cognitive agents. The results reveal a functional orthogonality between expert models and MLLMs: experts act as precise but fragile visual "sensors" for authenticity discrimination, while MLLMs provide stronger semantic reasoning (e.g., Textual Artifact Recognition). This contrast is reinforced by their robustness profiles, where vision-centric experts degrade sharply under post-processing (Figure 5(d)) perturbations (e.g., JPEG),

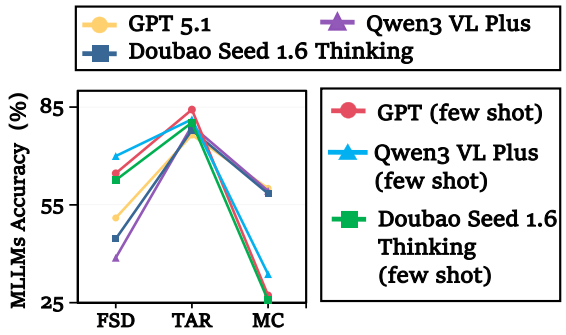


Figure 6: **Impact of Few-Shot prompting on AEGIS.**

whereas MLLMs remain comparatively stable, indicating weaker reliance on low-level visual fidelity. These findings motivate adaptive expert–MLLM orchestration for future **Expert AGI**.

3.3 Prompt Strategies

We evaluate two prompting strategies in **AEGIS**, namely **Chain-of-Thought**(CoT) (Wei et al., 2022) and **Few-Shot** prompting (Wang et al., 2020). More details are provided in Appendix C.

Few-shot prompting yields selective gains, benefiting forgery scope discrimination and textual artifact recognition but not manipulation classification.

CoT prompting yields consistent gains across most tasks, while substantially degrading manipulation classification. CoT improves detection-centric tasks, indicating that explicit reasoning helps surface more reliable authenticity cues and localization signals. In contrast, Manipulation Classification suffers a significant drop , suggesting limited capability to perform hypothetical judgments.

3.4 Error Analysis

We analyze representative error cases from GPT 5.1 by randomly sampling 100 instances for each task. Full examples and further interpretation are provided in Appendix E.

- ***Forgery Scope Discrimination (48%):*** Caused by the model's tendency to overgeneralize localized edits as entire forgeries , or to overlook subtle tampering and misclassify partial forgeries.
- ***Textual Artifact Recognition (22%):*** Failure to detect fine-grained textual anomalies introduced by text-based modifications in partial forgery.
- ***Manipulation Classification (36%):*** Stemming from difficulty in distinguishing foreground manipulations from complex background regions.
- ***Tampering Pinpointing (56%):*** Stemming from a tendency to predict fewer but overly large bboxes, resulting in imprecise localization.

4 Related Work

Existing Benchmark or Dataset. Prior efforts fall short of supporting expert-level forensic analysis for academic images, due to three structural limitations: ***Generic-scene bias***, where datasets such as Semi-Truths (Pal et al., 2024) and GenImage (Zhu et al., 2023) are dominated by faces or natural scenes rather than structured academic images; ***Restricted forgery strategies***, as benchmarks including GRE (Sun et al., 2024) and AI GUARD (Zhang et al., 2025) largely focus on global or single-type edits and fail to model subtle academic manipulations; and ***Forensic-agnostic evaluation protocols***, where datasets such as GIM (Chen et al., 2025b) and DFBench (Wang et al., 2025a) emphasize detection accuracy while overlooking manipulation scope and localization.

Forensic Models: Expert, MLLM, and Hybrid Approaches. In parallel, the literature has explored a broad spectrum of forensic models for generative forgery analysis, spanning vision-only expert models, MLLMs, and hybrid systems. Vision-only expert models cover multiple detection paradigms, including diffusion-process-based methods such as DIRE (Wang et al., 2023), DRCT (Chen et al., 2024a), and denoising-trajectory-based models (Liang et al., 2025), as well as alignment-based approaches including AlignedForensics (Rajan et al., 2025) and DDA (Chen et al., 2025a). Meanwhile, MLLMs have exhibited strong cross-modal understanding and reasoning capabilities, exemplified by OpenAI o4 Mini High (OpenAI, 2025c), Gemini 2.5 Flash (DeepMind, 2025), and the Qwen-VL series (Bai et al., 2025b), motivating their evaluation for generative image forensics in complex scenarios. Hybrid approaches integrate MLLMs with expert models either directly, as in FakeVLM (Wen et al., 2025), or in an assisted manner, as exemplified by SIDA (Huang et al., 2025) and FakeShield (Xu et al., 2025).

5 Conclusion

we introduce **AEGIS**, a holistic benchmark comprising over 20k forensic questions, designed to evaluate vision-only models, MLLMs, and hybrid systems for academic image forensics. AEGIS reveals substantial limitations in current models, particularly in fine-grained localization. We hope AEGIS provides a robust foundation for advancing trustworthy, interpretable, and generalizable multimodal forensic systems.

6 Limitations

This work introduces a comprehensive benchmark for evaluating detection, localization, and forensic reasoning capabilities of expert models and MLLMs on AI-generated academic images. By integrating representative forgery strategies with multi-dimensional evaluation tasks, **AEGIS** aims to expose systematic capability gaps that remain underexplored in existing benchmarks.

Nevertheless, several limitations remain. (1) The availability of *real-world retracted papers involving AI-generated images* remains limited. While AI-generated academic images are increasingly realistic, only a small number of confirmed real-world cases are publicly accessible, which constrains the scale at which authentic misconduct examples can be incorporated into the benchmark. As attention to generative image misuse in academic publishing continues to grow, a broader collection of real-world cases may gradually become available. (2) Although **AEGIS** evaluates a broad range of state-of-the-art expert models, including vision-only and hybrid MLLM-assisted approaches, baseline coverage is constrained by the availability of open implementations (e.g., AIGI-Holmes without released training weights). Future benchmark updates will incorporate additional baselines as more expert models become publicly accessible.

References

AI@Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-04-05.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

ByteDance. 2025a. Doubao-1.5-thinking-pro model card. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-pro. Accessed: 2025-04-15.

ByteDance. 2025b. Doubao-1.5-thinking-vision-pro model card. https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-vision-pro. Accessed: 2025-04-28.

ByteDance. 2025c. doubao-seed-1.6-flash model card. https://www.volcengine.com/docs/82379/1593704. Accessed: 2025-06-11.

ByteDance. 2025d. doubao-seed-1.6-thinking model card. https://www.volcengine.com/docs/82379/1593703. Accessed: 2025-06-11.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. 2024a. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*.

Jiaxuan Chen, Jieteng Yao, and Li Niu. 2024b. A single simple patch is all you need for ai-generated image detection. *Preprint*, arXiv:2402.01123.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. 2025a. Dual data alignment makes AI-generated image detector easier generalizable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, Wei Liu, and Jie Hu. 2025b. Gim: A million-scale benchmark for generative image manipulation detection and localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2311–2319.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2022. On the detection of synthetic images generated by diffusion models. *Preprint*, arXiv:2211.00680.

Google DeepMind. 2025. Gemini 2.5 flash and pro are now generally available, and we're introducing 2.5 flash-lite, our most cost-efficient and fastest 2.5 model yet. https://blog.google/products/gemini/gemini-2-5-model-family-expands/. Accessed: 2025-06-17.

DeepSeek. 2025. Janus-pro-7b release. https://janusai.pro/. Accessed: 2025-01-28.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Jizhe Zhou. 2025. Forensichub: A unified benchmark codebase for all-domain fake image detection and localization. *Preprint*, arXiv:2505.11003.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Gemini Team, Google. 2025. Gemini 3 pro model card. Model card, Google DeepMind. Accessed: 2026-01-02. Version 7 DT update published Dec 4, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *CVPR*, pages 28831–28841.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025. dots.ocr: Multilingual document layout parsing in a single vision-language model. *Preprint*, arXiv:2512.02498.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Fuqiang Du, Li Jingyuan, Lanchi Xie, Zhen Xu, and Weiqing Huang. 2025. Denoising trajectory biases for zero-shot AI-generated image detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11451–11461.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

OpenAI. 2025a. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-04-14.

OpenAI. 2025b. Introducing gpt-5.1 in the api. `https://openai.com/index/gpt-5-1/`. Accessed: 2025-11-12.

OpenAI. 2025c. Openai o3 and o4-mini system card. `https://openai.com/index/o3-o4-mini-system-card/`. Accessed: 2025-04-16.

Anisha Pal, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. 2024. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 118025–118051. Curran Associates, Inc.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. 2025. Aligned datasets improve detection of latent diffusion-generated images. In *The Thirteenth International Conference on Learning Representations*.

Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`. Version 0.3.0.

Zhihao Sun, Haipeng Fang, Juan Cao, Xinying Zhao, and Danding Wang. 2024. Rethinking image editing detection in the era of generative ai revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, Melbourne, VIC, Australia. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,

and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *Preprint*, arXiv:2503.20314.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Preprint*, arXiv:2207.02696.

Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. 2025a. Dfbench: Benchmarking deepfake image detection capability of large multimodal models. *CoRR*, abs/2506.03007.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, Zihan Li, Yuzhe Liang, Xiaopeng Wang, Haorui Zheng, Ming Wen, Kang Yin, Yiran Wang, Nan Li, Feng Deng, and 4 others. 2025b. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *Preprint*, arXiv:2506.19774.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3).

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In *ICCV*, pages 22388–22398.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, wenjun wu, Conghui He, and Weijia Li. 2025. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*.

Wenhua Zhang, Weicheng Li, Xuanrong Rao, Lixin Zou, Xiangyang Luo, Chubin Zhuang, Yongjie Hong, Zhen Qin, Hengyun Chang, Chenliang Li, and Bo Zheng. 2025. Aiguard: A benchmark and lightweight detection for e-commerce aigc risks. In *ACL (Findings)*, pages 12437–12450.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, and 55 others. 2025. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *Preprint*, arXiv:2501.12202.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS*.

A Dataset Annotation and Construction

This appendix provides a detailed description of the dataset annotation and construction procedures underpinning the **AEGIS** benchmark. The Appendix A is organized as follows. **subsection A.1** first discusses the impact of AI-generated images on academic integrity. **subsection A.2** summarizes key dataset statistics, including data distribution across forgery simulations and annotated subtypes. **subsection A.3** describes the data curation process, covering document parsing, panel extraction, and expert-driven annotation. **subsection A.4** details the synthetic data generation pipeline, including the generative models employed, the design of representative forgery simulations, and the data quality assurance. Finally, **subsection A.5** introduces the design of evaluation task questions, which operationalize four forensic competencies through structured and standardized assessment settings.

A.1 Impact of AI-Generated Images on Academic Integrity

Recent advances in generative models have enabled the creation and localized editing of scientific images with high visual plausibility, substantially increasing the difficulty of manual inspection during academic review. Unlike traditional image manipulation, AI-generated or AI-edited content can introduce subtle yet semantically misleading modifications while preserving overall visual coherence. Such manipulations challenge existing peer review and integrity auditing mechanisms, motivating the need for systematically constructed datasets and evaluation protocols for academic image forensics.

Further analysis of representative real-world retraction cases involving AI-generated images is provided in Appendix F.

A.2 Dataset Statistics

Statistics	Number
Total Images	10005
Total Generated Images	8210
AVG Images per Forgery Strategy	2001
AVG Images without Text per Main Category	724
AVG Images with Text per Main Category	705
AVG Generated Images per Generative Models	328
AVG Generated Images per Main Category	1173
AVG Generated Images per Subtype	228

Table 3: Basic statistics of **AEGIS**.

As shown in Figure 7 and Figure 8, we summarize the data distribution of **AEGIS**, comprising 1,795 real images and 8,210 AI-generated images spanning four forgery strategies: text constraint fabrication (3,121 images), image inference forgery (2,274 images), targeted region restoration (1,650 images), and targeted region editing (1,165 images). More detailed statistics across generative models and forgery strategies are provided in Table 5.

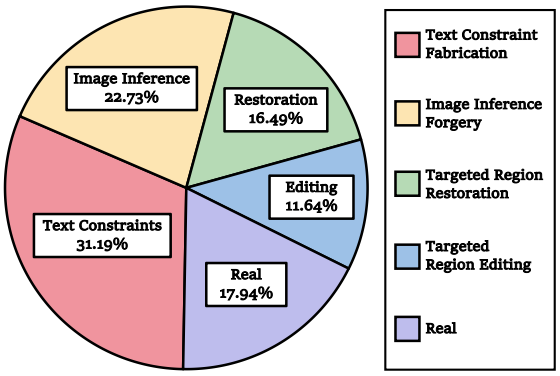


Figure 7: The distribution of real images and images generated by four types of forgery simulations.

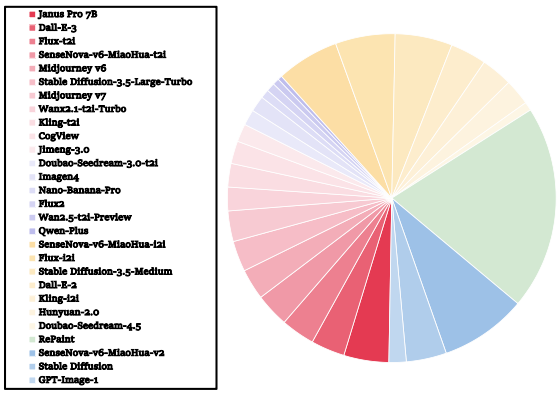


Figure 8: The distribution of generated images by 25 generative models.

In addition, as shown in Figure 9 and Table 6, the dataset statistics further cover the distribution across seven major image categories and 39 fine-grained subtypes.

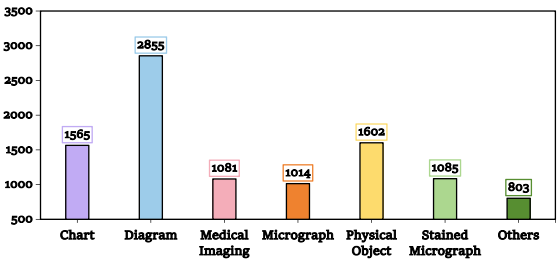


Figure 9: The distribution of seven main categories.

A.3 Data Curation

Paper Parsing **AEGIS** collects 4,362 high-quality academic papers from the open-access PubMed repository and performs document-level parsing to construct the initial visual corpus. In total, we extract 21,246 figures, 20,239 captions, and 50,063 panels. As illustrated in Figure 10, a *figure* refers to a complete figure unit paired one-to-one with its corresponding caption, whereas a *panel* denotes the smallest indivisible visual unit within an academic image. The entire extraction and parsing process required approximately 200 hours of human and computational effort.

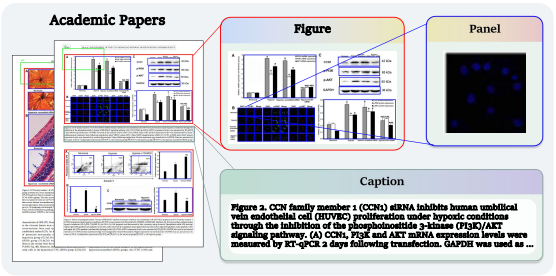


Figure 10: The extraction and parsing of academic papers. The regions highlighted with red boxes correspond to Figures, while those highlighted with blue boxes correspond to Panels. (Green boxes are used to verify whether the visualization is effective.)

The extraction and parsing pipeline consists of the following steps:

- *Step 1: Figure and Caption Extraction.* We employ the Fitz library to extract figures from PDF documents, as it consistently preserves higher visual fidelity than alternative document parsing tools, which is critical for subsequent panel-level analysis. Captions are parsed using dots.ocr (Li et al., 2025) and are further used to cross-validate the correctness of the extracted figures. Finally, three human experts with academic peer-review experience manually verify the figure–caption correspondences to ensure extraction accuracy.
- *Step 2: Panel Segmentation from Figures.* We apply YOLOv7 (Wang et al., 2022) to segment figures into panels. The YOLOv7 model is fully trained on 200,000 manually annotated academic panels and has been validated to meet production-level reliability and commercial deployment standards. In parallel, MinerU (Niu et al., 2025) is used as an auxiliary tool to cross-check segmentation results. Panels with irregular layouts, poor image quality, or lacking experimental visual content are subsequently filtered out through joint inspection by three expert reviewers and five trained graduate annotators.

Expert Curation Following automated extraction and panel segmentation, we conduct a multi-stage expert curation and validation process to ensure the academic validity, semantic clarity, and forensic relevance of the retained panels. The entire curation process involves five human experts with prior academic peer-review experience, together with three Ph.D. students and six trained Master's students, and requires approximately 260 person-hours to complete.

We first perform a manual quality screening to remove panels with insufficient spatial resolution, extremely small visual extent, or evident parsing errors introduced during figure extraction or panel segmentation. Panels exhibiting severe cropping artifacts, unreadable visual content, or incomplete semantic units are discarded at this stage to prevent low-quality samples from propagating into subsequent annotation and forgery pipelines.

Panels that pass the initial quality screening are then jointly reviewed together with their associated captions by domain experts. During this stage, each panel is evaluated against the predefined hierarchical taxonomy of AEGIS, which organizes academic imagery into seven high-level categories and 39 fine-grained subtypes. As presented in Table 8, these subtypes collectively encompass 113 concrete and domain-grounded specific examples, which serve as explicit semantic anchors during annotation. To be considered valid, each academic panel must conform to at least one specific example within its assigned subtype, ensuring that category–subtype labels are grounded in concrete visual and semantic criteria rather than abstract class names. Only panels that can be unambiguously assigned to a valid category–subtype–example tuple and whose visual content is semantically consistent with the accompanying caption are retained and labeled.

In parallel, we explicitly exclude non-academic or non-forgeable visual elements that are unlikely to constitute realistic targets of academic image misconduct. These include, but are not limited to, human portraits, personal photographs, icons, logos, decorative patterns, artistic illustrations, and purely illustrative graphics lacking experimental or scientific meaning. Such elements, while visually plausible, do not reflect manipulation behaviors commonly observed in real-world academic integrity violations and are therefore systematically

filtered out.

Through this multi-stage expert curation and taxonomy-guided validation procedure, we ultimately retain over 8,192 high-quality, fine-grained annotated academic panels. These curated panels form the foundation for subsequent forgery simulation, annotation, and benchmark construction in **AEGIS**.

A.4 Synthetic Data Generation

This subsection describes the synthetic data generation pipeline employed in **AEGIS** to construct realistic and forensic-relevant AI-generated academic image forgeries. We first introduce the diverse generative models used to support different forgery strategies. We then detail the design of representative forgery simulations. Finally, we present the data quality assurance procedures adopted to ensure visual plausibility, semantic consistency, and annotation reliability across the generated samples.

A.4.1 Generative Models

As shown in Table 7, we incorporate 25 state-of-the-art generative models spanning diffusion-based, autoregressive, and hybrid architectures to support diverse academic image forgery simulations.

A.4.2 Forgery Simulations

To systematically simulate realistic manipulation behaviors in academic imagery, **AEGIS** adopts a strategy-driven forgery simulation framework, as illustrated in Figure 13-Figure 16. Inspired by common misconduct patterns observed in scholarly publications and aligned with the capabilities of contemporary generative models, we define four representative forgery strategies: *Text Constraint Fabrication*, *Image Inference Forgery*, *Targeted Region Restoration*, and *Targeted Region Editing*. These strategies jointly cover both global and localized manipulation scenarios, ranging from semantic-level content fabrication to fine-grained region-specific edits.

Specifically, the first two strategies focus on holistic image synthesis conditioned on textual or contextual constraints, resulting in globally generated academic figures. In contrast, the latter two strategies involve localized manipulations applied to selected regions within otherwise authentic images. For these region-level forgeries, we employ the Segment Anything Model (SAM) (Kirillov et al., 2023) to automatically generate precise manipulation masks, ensuring accurate spatial control over the forged regions and enabling consistent pixel-level ground-truth annotations for subsequent localization evaluation.

A.4.3 Data Quality Assurance

To ensure the forensic validity and realism of AI-generated academic images, **AEGIS** adopts a rigorous data quality assurance protocol designed to mitigate hallucinations and implausible artifacts introduced by generative models. All synthesized images undergo a dual-review process that jointly evaluates *local manipulation plausibility* and *global visual fidelity*.

At the local level, annotators examine whether manipulated regions exhibit coherent textures, well-defined boundaries, and semantic consistency with surrounding content, while avoiding trivial artifacts such as unnatural edges, repetitive patterns, or mismatched visual semantics. At the global level, images are assessed for overall structural coherence, adherence to scientific conventions, and consistency with the corresponding academic context, including figure layout, visual hierarchy, and modality-specific characteristics.

Each AI-generated image is independently reviewed by multiple annotators following a unified set of quality-control guidelines. The review process involves five human experts with prior academic peer-review experience, together with three Ph.D. students and six trained Master's students. Images failing to satisfy either local or global criteria are discarded. Through this protocol, approximately 29% of generated samples are filtered out, resulting in a curated set of 8,210 high-quality AI-generated academic images. The verification process required approximately 400 hours in total, ensuring consistent and reliable application of quality standards across diverse image categories and forgery strategies.

In addition to human expert review, we further conduct an automated quality assessment to contextualize the realism and semantic fidelity of the synthesized images relative to datasets used by baseline forensic models. Specifically, we compare the quality scores of AEGIS-generated images against corresponding baseline datasets using three widely adopted metrics: Inception Score (IS), Fréchet Inception Distance (FID), and CLIP Score.

- *Inception Score (IS).* IS measures visual quality and content diversity of generated images based on the entropy of class predictions produced by a pretrained Inception network. Higher

Dataset	Score		
	CLIP ↑	FID ↓	IS ↑
DiffusionForensics	–	27.46	13.47 ± 0.97
DRCT-2M	75.74	22.97	34.57 ± 1.57
SID-Set (Full_Synthetic)	–	60.02	38.73 ± 0.95
SID-Set (Tampered)		30.2	32.28 ± 2.39
New-Generator	–	75.58	24.02 ± 1.35
Robust_LDM	–	99.80	13.76 ± 1.02
Fake_Club	70.95	86.59	13.30 ± 1.69
EvalGEN	–	69.60	29.67 ± 0.67
Chameleon	–	53.88	16.87 ± 0.46
ours	**78.66**	**32.22**	**13.41 ± 1.23**

Table 4: Comparison of different datasets.

scores generally indicate images that are both visually coherent and diverse. **We note that IS was originally proposed for evaluating GAN-based image generators, and its sensitivity may be reduced for modern non-GAN architectures** (e.g., diffusion-based or hybrid generative models), whose outputs often depart from the ImageNet-centric class distributions assumed by the Inception model. As a result, lower IS values in our setting should not be interpreted as diminished perceptual quality, but rather as a known limitation of the metric when applied beyond its original design scope.

- *Fréchet Inception Distance (FID).* FID (Seitzer, 2020) evaluates the distributional similarity between generated images and real academic images by computing the Fréchet distance between their feature embeddings extracted from a pretrained Inception model. Lower FID scores indicate higher visual fidelity and closer alignment with real-image statistics.

- *CLIP Score.* CLIP Score (Hessel et al., 2022) measures cross-modal semantic consistency between generated images and their corresponding textual prompts. It is computed as the cosine similarity between visual and textual embeddings extracted by a pretrained CLIP model, where higher scores reflect stronger semantic alignment.

Together, these complementary metrics provide an objective reference for evaluating perceptual fidelity, distributional realism, and semantic alignment, supporting the validity of the curated dataset alongside expert-driven quality control.

A.5 Evaluation Task Question Design

A.5.1 Forgery Scope Discrimination

Evaluation Question Design: The Forgery Scope Discrimination (FSD) task is designed to assess a model's ability to determine the scope of manipulation in an academic image. For models that support fine-grained scope reasoning, the task requires distinguishing among *real*, *global forgery*, and *local forgery*. However, we note that a subset of vision-only expert models are inherently limited to binary authenticity prediction and do not support explicit scope differentiation. To ensure fair and consistent evaluation across heterogeneous model families, we adapt the FSD task for such expert models by formulating it as a binary classification problem, where the objective is to discriminate between *real* and *forgery* images, regardless of whether the forgery is global or local.

Option Explanation: Under the binary FSD setting, answer options are restricted to *Real* and *Forgery*. Images labeled as either global or local forgeries in the ground truth are jointly mapped to the *Forgery* category. This formulation preserves the original intent of FSD as an authenticity assessment task while respecting the functional constraints of expert detectors. For models capable of multi-class outputs, fine-grained scope predictions are evaluated under the original three-way setting, whereas binary-only expert models are evaluated using the reduced two-class formulation. This design ensures comparability without artificially penalizing models for unsupported output structures.

Prompt: are shown in Figure 17.

A.5.2 Textual Artifact Recognition

Evaluation Question Design: The Textual Artifact Recognition (TAR) task is designed to evaluate the capability of multimodal large language models (MLLMs) to recognize fine-grained textual artifacts indicative of AI generation or manipulation in academic images. TAR specifically focuses on visually embedded textual regions and subtle structural textures, which often require joint visual perception and semantic interpretation. As this task relies on multimodal reasoning over both visual cues and implicit semantic consistency, it is evaluated exclusively on MLLMs. Vision-only expert models are not included, as they are typically optimized for low-level authenticity detection and do not support explicit reasoning over text-region

semantics or artifact interpretation.

The task is formulated as a binary decision problem, requiring the model to determine whether any textual artifacts associated with AI generation are present in the given academic image.

Option Explanation: Answer options for TAR are restricted to *Yes* and *No*. A response of *Yes* indicates that the model identifies one or more textual artifacts suggestive of AI-generated or manipulated content, particularly within textual regions or fine structural patterns. A response of *No* indicates that the image appears free of such artifacts and exhibits textual characteristics consistent with authentic academic imagery. This design isolates textual artifact recognition as a distinct forensic competency rooted in multimodal perception and reasoning, rather than low-level signal detection.

Prompt: are shown in Figure 18.

A.5.3 Manipulation Classification

Evaluation Question Design: The Manipulation Classification (MC) task evaluates the capability of multimodal large language models (MLLMs) to recognize the *type of manipulation operation* applied to an academic image. Different from authenticity judgment, MC requires models to distinguish *what was done* to the image content, focusing on three canonical edit operations: *Insert*, *Remove*, and *Alter*. This task is evaluated exclusively on MLLMs, as it requires fine-grained semantic interpretation of edited content and structured decision-making beyond low-level artifact detection, which is not directly supported by vision-only expert detectors.

Option Explanation: Answer options for MC include *Insert*, *Remove*, and *Alter*, along with an additional *Real* (or *No manipulation*) option for authentic images. *Insert* denotes introducing new visual elements that were not present in the original content. *Remove* denotes deleting or erasing existing elements, often leaving inpainting-like traces or semantic gaps. *Alter* denotes modifying existing content (e.g., changing shapes, textures, labels, or structures) without purely adding or deleting elements. Correct classification requires separating these operation types even when manipulations are subtle or localized within complex academic imagery.

Prompt: are shown in Figure 19.

A.5.4 Tampering Pinpointing

Evaluation Question Design: The Tampering Pinpointing (TP) task is designed to evaluate a model's ability to localize manipulated regions within an academic image. Unlike classification-oriented tasks, TP explicitly assesses spatial precision and completeness of localization. Given the heterogeneous output capabilities of different model families, TP is evaluated at two distinct granularities.

For multimodal large language models (MLLMs), which typically produce region-level localization, TP is formulated as a *region-level* task. Models are prompted to identify and output one or more bounding boxes that enclose the suspected tampered regions. These predictions are evaluated against ground-truth bounding boxes derived from annotated manipulation regions.

For vision-only expert models, which generate dense segmentation outputs, TP is formulated as a *pixel-level* task. Models are required to produce binary masks indicating manipulated areas, which are evaluated against ground-truth pixel-level masks. This setting enables fine-grained assessment of localization accuracy for models with pixel-wise prediction capability.

By aligning the TP task formulation with each model family's native output structure, this design ensures fair and meaningful evaluation of localization performance without imposing unsupported output requirements.

Prompt: are shown in Figure 20 and Figure 21.

B Experiment Details

This section provides the detailed experimental setup supporting our evaluation. subsection B.1 describes the hardware, software, and runtime configurations. subsection B.2 lists all evaluated models, including expert systems and MLLMs. subsection B.3 defines the quantitative metrics used for **AEGIS**. Finally, subsection B.4 presents the evaluation procedures and task-specific protocols.

B.1 Experiment Environment

For evaluation experiments involving two input settings, most model inferences were conducted via the OpenRouter API. Exceptions include several open-source models executed locally, as well as the Doubao and Qwen series, which were accessed through the Volcano Engine APIs and Qwen APIs, respectively.

The system configuration is summarized below:

- **CPU:** Dual-socket Intel Xeon Gold 6148 (2.40 GHz), 40 cores per socket, 80 threads total
- **GPU:** 8× NVIDIA A40, each with 48 GB VRAM
- **GPU driver:** 525.125.06, **CUDA:** 11.8
- **cuDNN:** 8.x (compiled with CUDA 11.8)
- **Operating System:** Ubuntu 20.04.6 LTS

B.2 Experimental Models

As summarized in Table 9, we evaluate nine expert models in total, including six vision-only expert models and three hybrid expert models.

As shown in Table 10, we evaluate 25 multimodal large language models (MLLMs), consisting of 14 proprietary models and 11 open-source models. The proprietary models are released by OpenAI (OpenAI, 2025a,b,c), Google (DeepMind, 2025; Gemini Team, Google, 2025), Alibaba (Bai et al., 2023, 2025a), and ByteDance (ByteDance, 2025b,a,c,d). The open-source models are released by Meta (AI@Meta, 2025), Mistral, Alibaba (Bai et al., 2025b), Google (Team et al., 2025), and the LLaVA community (Li et al., 2024a,b; Liu et al., 2023, 2024).

B.3 Evaluation Metrics

To comprehensively assess model performance across different dimensions, we design **seven core metrics**.

(1) Accuracy (ACC) measures the proportion of correctly predicted answers among all evaluated instances. All tasks evaluated with ACC are formulated as *single-choice classification problems*, where each image corresponds to exactly one ground-truth label and the model selects a single option.

Let N denote the total number of valid evaluation samples, and let $\mathbb{I}(\cdot)$ be the indicator function. Accuracy is computed as:

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big(\hat{y}_i = y_i\big), \tag{1}$$

where $\hat{y}_i$ and y_i denote the predicted and ground-truth labels of the i-th sample, respectively.

For the Forgery Scope Discrimination (FSD) task, ACC is reported under two settings depending on model capability. Models that support fine-grained scope prediction are evaluated under a multi-class setting (*Real*, *Entire Forgery*, *Partial Forgery*). Vision-only expert models that only support authenticity detection are evaluated under a binary setting (*Real*, *Forgery*), where both global and local forgeries are mapped to the *Forgery* class.

(2) Macro-F1 Score (M-F1) is adopted to measure class-balanced performance by averaging F1 scores across classes with equal weight. Similar to ACC, Macro-F1 is computed over single-choice classification outputs and mitigates the influence of class imbalance.

For each class $c \in \mathcal{C}$, precision P_c, recall R_c, and F1 score $F1_c$ are defined as:

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{2}$$

$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}, \tag{3}$$

where TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively.

Macro-F1 is then computed as:

$$\text{M-F1} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} F1_c. \tag{4}$$

For the FSD task, Macro-F1 is computed separately for the binary and multi-class settings, consistent with the corresponding ACC evaluation protocol. Binary Macro-F1 is reported for expert models distinguishing *Real* versus *Forgery*, while multi-class Macro-F1 is reported for models predicting *Real*, *Entire Forgery*, and *Partial Forgery*. Samples with abstention responses (i.e., *Not Sure*) are excluded from Macro-F1 computation.

(3)Correct Localization Accuracy (CLA) is a region-level localization metric that evaluates whether all ground-truth tampered regions are sufficiently covered by the predicted regions. CLA is computed at the *sample level* and then aggregated across samples.

For a given sample, let N_{gt} denote the number of ground-truth connected components, and let GT_i denote the i-th ground-truth component. Let Pred represent the union of all predicted regions. A sample is considered *CLA-correct* if and only if every ground-truth component is sufficiently covered by the predicted regions:

$$\mathrm{CLA}_{\mathrm{correct}} = \begin{cases} \text{False}, & \text{if } N_{\mathrm{gt}} = 0, \\ \text{True}, & \text{if } \forall i \in [1, N_{\mathrm{gt}}], \dfrac{|\mathrm{Pred} \cap \mathrm{GT}_i|}{|\mathrm{GT}_i|} \geq \tau, \\ \text{False}, & \text{otherwise}, \end{cases} \tag{5}$$

where $|\cdot|$ denotes the number of pixels in a region, and τ is a coverage threshold (set to 0.5 in all experiments).

The final CLA score is reported as the percentage of CLA-correct samples over all evaluated samples:

$$\mathrm{CLA}(\%) = \frac{\#\{\text{CLA-correct samples}\}}{\#\{\text{total samples}\}} \times 100\%. \tag{6}$$

(4)Over Localization Rate (OLR) measures whether a model produces excessively many predicted regions relative to the ground-truth, capturing the tendency of over-fragmented or overly sensitive localization.

For a given sample, let N_{gt} denote the number of ground-truth connected components and N_{pred} denote the number of predicted regions. A sample is considered *over-localized* if:

$$\mathrm{OLR}_{\mathrm{over}} = \begin{cases} \text{True}, & \text{if } N_{\mathrm{gt}} = 0 \text{ and } N_{\mathrm{pred}} > 0, \\ \text{True}, & \text{if } N_{\mathrm{gt}} > 0 \text{ and } N_{\mathrm{pred}} > N_{\mathrm{gt}} \times (1+\alpha), \\ \text{False}, & \text{otherwise}, \end{cases} \tag{7}$$

where α is an over-localization tolerance factor (set to 0.5 by default).

The final OLR score is computed as the proportion of over-localized samples among all evaluated samples:

$$\mathrm{OLR}(\%) = \frac{\#\{\text{over-localized samples}\}}{\#\{\text{total samples}\}} \times 100\%. \tag{8}$$

(5)Intersection over Union (IoU). For pixel-level localization tasks, we adopt Intersection over Union (IoU) to measure the spatial overlap between predicted tampering masks and ground-truth masks. IoU evaluates localization accuracy at the pixel granularity and is widely used for segmentation-based forensic analysis.

Let Pred and GT denote the predicted binary mask and the ground-truth binary mask for a given sample, respectively. IoU is defined as:

$$\mathrm{IoU} = \frac{|\mathrm{Pred} \cap \mathrm{GT}|}{|\mathrm{Pred} \cup \mathrm{GT}|}, \tag{9}$$

where $|\cdot|$ denotes the number of pixels in the corresponding region.

The final IoU score is reported as the average IoU over all evaluated samples.

(6)Pixel-level F1 Score (F1). To complement IoU, we further report the pixel-level F1 score, which balances precision and recall of predicted tampering masks. This metric emphasizes accurate boundary delineation and penalizes both over- and under-segmentation.

At the pixel level, true positives (TP), false positives (FP), and false negatives (FN) are defined based on pixel-wise correspondence between the predicted mask and the ground-truth mask. Precision P, recall R, and F1 score are computed as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{10}$$

$$\mathrm{F1} = \frac{2 \cdot P \cdot R}{P+R}. \tag{11}$$

The reported pixel-level F1 score is averaged across all samples in the evaluation set.

(7) Normalized Forensic Index (NFI). Beyond task-specific metrics, we define a composite score to holistically measure a model's overall forensic capability with an emphasis on balanced performance. NFI is designed to reward models that perform consistently across heterogeneous forensic tasks, while penalizing pathological localization behaviors such as excessive over-localization.

For each model i, we first collect its core forensic scores:

$$\mathcal{S}_i = \{s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4}\}, \tag{12}$$

where $\mathcal{S}_i$ includes the FSD-F1, TAR-F1, IC-F1, and CLA scores of model i. All scores are normalized to $[0,1]$ for numerical consistency.

We compute the harmonic mean to emphasize balanced performance:

$$\mathrm{HM}_i = \frac{|\mathcal{S}_i|}{\sum_{s \in \mathcal{S}_i} \frac{1}{s+\epsilon}}, \tag{13}$$

where ϵ is a small constant to ensure numerical stability.

To penalize over-localized predictions, we introduce an over-localization factor based on the Over-Localization Rate (OLR):

$$P_i = (1 - \mathrm{OLR}_i)^{\gamma}, \tag{14}$$

where $\mathrm{OLR}_i \in [0,1]$ and γ controls the penalty strength.

Finally, the Normalized Forensic Index is defined as:

$$\mathrm{NFI}_i = \mathrm{HM}_i \cdot P_i, \quad \gamma = 0.5. \tag{15}$$

A higher NFI indicates stronger and more reliable forensic capability, reflecting both balanced task performance and structurally reasonable localization behavior.

Sensitivity Analysis of NFI. To examine the robustness of the proposed Normalized Forensic Index (NFI), we conduct sensitivity analyses with respect to the OLR penalty exponent γ and the imbalance penalty weight λ, as reported in Table 11 and Table 12, respectively.

Table 11 shows that varying γ within a reasonable range ($\gamma \in \{0.3, 0.5, 0.7\}$) leads to only minor changes in absolute NFI values, while the relative ranking of top-performing models remains largely stable. In particular, leading proprietary MLLMs consistently occupy the top ranks, indicating that NFI is not overly sensitive to the exact strength of the OLR-based penalty.

Table 12 further analyzes the effect of the detection–localization imbalance penalty controlled by λ. As λ increases, models exhibiting stronger imbalance between detection-oriented and localization-oriented capabilities are more noticeably penalized, resulting in gradual rank shifts. Nevertheless, the overall ordering of strong and weak models remains coherent, suggesting that NFI provides a stable yet discriminative measure that reflects balanced forensic capability rather than peak performance on isolated tasks.

B.4 Evaluation Protocol

We uniformly adopt high-resolution PNG images for all evaluations to avoid compression artifacts introduced by lossy formats such as JPEG, which may interfere with forensic cues and bias model judgments. When an input image exceeds a model's maximum input resolution, a two-stage adaptive compression strategy is applied: (i) the long edge is first scaled to the model-specific maximum input size while preserving the aspect ratio; (ii) the total pixel count is further reduced, if necessary, through iterative 10% downscaling until the model's input constraints are satisfied.

All prompts are intentionally kept minimal, containing only task definitions and direct instructions, in order to reduce extraneous noise and better expose the intrinsic forensic capabilities of the evaluated models.

C Multi-dimensional Evaluation Analysis

C.1 Performance Analysis of Expert Models

Before presenting the detailed performance analysis, we clarify that for each expert model, we report results obtained from the publicly available checkpoint that achieves the highest detection accuracy under the authors' recommended evaluation setting. **When multiple pretrained weights or configurations are provided by the original authors, we select the variant yielding the best ACC to ensure a fair and representative comparison across models**.

As shown in Table 13, expert models exhibit substantial performance variability across forgery scope discrimination and pixel-level localization tasks.

A pronounced imbalance between Real-F1 and Forgery-F1 is observed for several expert models. In particular, AIDE attains an extremely low Real-F1 score (0.85%) despite achieving the highest ACC, suggesting a strong tendency to over-flag images as forged. Similar biases are also observed in DRCT and DIRE, where Real-F1 remains substantially lower than Entire-F1, indicating asymmetric sensitivity toward forged images at the expense of false positives on real samples. In contrast, SIDA demonstrates the opposite behavior, with a high Real-F1 (87.82%) but considerably lower Entire-F1 (62.43%), reflecting conservative detection behavior that favors real-image recognition while under-detecting forgeries.

Pixel-level localization performance further reveals systematic limitations of expert models. Among all evaluated methods, only FakeShield and SIDA support pixel-level localization, yet both exhibit weak spatial accuracy. FakeShield achieves an IoU of 30.09% with a corresponding pixel-level F1 of 69.11%, indicating coarse but incomplete localization. SIDA performs markedly worse, with an IoU of only 3.66%, suggesting that predicted masks often fail to meaningfully overlap with ground-truth tampered regions. The remaining expert models do not provide pixel-level outputs, highlighting a general lack of fine-grained localization capability across existing expert detectors.

Overall, these results indicate that while several expert models achieve relatively high binary detection accuracy, they suffer from severe class imbalance biases and limited localization precision, underscoring the gap between coarse authenticity detection and reliable, fine-grained forensic analysis in academic image scenarios.

C.2 Impact of Prompting Strategies

This subsection analyzes the impact of different prompting strategies on model performance. We first examine few-shot prompting subsubsection C.2.1, followed by chain-of-thought prompt-

ing subsubsection C.2.2, to assess how structured supervision and explicit reasoning cues influence forensic analysis.

C.2.1 Impact of Few-Shot Prompting

For few-shot prompting, we build an indexed reference library of curated examples. Given an input image, we first restrict retrieval to the same category, then compute visual similarity using VINO3 embeddings, and select the single most similar reference image as the in-context exemplar. The user prompt includes the image description, the input image, and an output-format example to guide the model's inference.

Table 14–Table 16 reveal a task-dependent impact of few-shot prompting.

For *Forgery Scope Discrimination* (Table 14), few-shot prompting consistently improves both overall accuracy and per-class F1 scores across all evaluated models, with particularly pronounced gains on the *Partial* class, indicating enhanced sensitivity to localized or subtle forgery patterns. For *Textual Artifact Recognition* (Table 15), few-shot prompting yields modest but stable improvements on most metrics, suggesting that exemplar-based context helps reinforce textual cues without substantially altering decision boundaries. In contrast, for *Manipulation Classification* (Table 16), few-shot prompting leads to notable performance degradation for all models, reflected by systematic drops in ACC and Macro-F1 despite isolated gains on specific manipulation types. This divergence suggests that while few-shot exemplars effectively support scope-level and textual-based judgments, they may introduce semantic bias or overfitting when models are required to distinguish fine-grained manipulation intents.

C.2.2 Impact of Chain-of-Thought Prompting

For Chain-of-Thought (CoT) prompting, we tailor task-specific prompts for each of the seven categories across all four tasks. The samples of CoT prompt for the *Chart* category is illustrated in Figure 25-Figure 35, while the complete set of prompts is provided at the url: https://anonymous.4open.science/r/AEGIS-2E31. The corresponding experimental results for GPT-5.1 are summarized in Table 17.

Overall, CoT prompting yields broad gains for MLLM across categories, improving detection and localization performance on all tasks except **Manipulation Classification(MC)**. Averaged over seven categories, CoT increases **Forgery Scope Discrimination(FSD)** accuracy (+4.38%) and improves **Textual Artifact Recognition(TAR)** (+3.33%), while also boosting **Tampering Pinpointing(TP)** localization quality (+4.25%), albeit with a slight rise in over-localization (+1.07%). These trends indicate that explicit reasoning steps mainly help the model surface more reliable authenticity cues and provide more correct localization hits, but may also encourage slightly more aggressive region proposals.

CoT is particularly effective for visually dense imagery (e.g. micrograph), where detection-centric tasks benefit most from structured reasoning that emphasizes semantic grounding beyond pure geometric structure. In the micrograph category, FSD increases from 39.09% to 56.78% (+17.69%), and TAR also improves from 71.98% to 73.95% (+1.97%). This suggests that the CoT prompt's instruction to check whether complex visual elements carry coherent scientific meaning instead of looking structured better activates domain knowledge and reduces superficial texture-driven errors that are common in high visually dense imagery.

In contrast, Manipulation Classification (MC) degrades substantially under CoT, implying that step-by-step reasoning can mislead the model when intent requires counterfactual reconstruction of what "should have been there" in academic figures. On average, MC accuracy drops from 60.06% to 44.66% (-15.40%) and M-F1 from 47.45% to 40.72% (-6.73%), with the largest declines appearing in Medical Imaging (-27.00%) and other dense categories such as Micrograph (-18.57%) and Physical Object (-19.07%). A plausible explanation is that the model lacks sufficient knowledge of the canonical composition of academic images making it difficult to reliably reconstruct what should have originally appeared in the edited region, thereby increasing confusion between manipulation intents and degrading intent discrimination accuracy.

C.3 Analysis of Post-Processing Results

To examine the robustness of models under common post-processing perturbations, we apply three representative image post-processing strategies. Specifically, we consider *Gaussian blurring*, *JPEG compression*, and *image rescaling*. Gaussian blurring is implemented with a fixed blur radius of 5 to suppress high-frequency details. JPEG compression is applied with a quality factor of 50, introducing moderate compression artifacts. For image

rescaling, inputs are uniformly downsampled by a factor of 0.5 using the LANCZOS resampling method to preserve interpolation quality.

As shown in Table 18, post-processing perturbations induce markedly different robustness profiles across model families. For MLLMs (*e.g.*, GPT 5.1), Gaussian blurring, JPEG compression, and image rescaling consistently degrade detection performance across Forgery Scope Discrimination, Textual; Artifact Recognition, and Manipulation Classification, with the most pronounced drops observed under Gaussian blurring, reflecting sensitivity to high-frequency visual cues. Nevertheless, the relative degradation remains moderate, and MLLMs retain non-trivial performance across tasks, suggesting that their decisions are not exclusively driven by pixel-level fidelity but are partially supported by higher-level semantic and structural reasoning.

In contrast, expert models (*e.g.*, hybrid MLLM-assisted FakeShield) exhibit substantially sharper performance collapse under all post-processing strategies. Both authenticity discrimination and pixel-level localization metrics deteriorate dramatically, particularly under JPEG compression and Gaussian blurring, indicating that the expert-driven visual sensing components remain highly dependent on fine-grained low-level statistics. Notably, pixel-level localization is more severely affected than detection accuracy, highlighting the vulnerability of dense forensic cues to even moderate image transformations, despite the presence of auxiliary MLLM components.

Together, these observations further substantiate the complementary yet asymmetric roles of the two model paradigms: expert-driven components act as highly sensitive but brittle visual sensors, while MLLMs provide comparatively robust semantic abstraction. This divergence reinforces the need for adaptive expert–MLLM orchestration, in which expert modules contribute precise low-level signals and MLLMs offer resilient reasoning under distributional shifts, aligning with the Expert AGI perspective discussed in the main text.

D Breakdown Results on Different Evaluation Tasks

D.1 Human Expert Baseline Analysis

- ***Human Expert Baseline Construction.*** The human expert baseline in **AEGIS** is constructed through a hierarchical and weighted evaluation framework. The expert panel consists of **three senior academic reviewers with over five years of experience in scientific peer review** and **two specialists with professional expertise in AI image generation and manipulation**. These five experts were responsible for defining annotation guidelines, resolving ambiguous cases, and calibrating judgment criteria across tasks.

  Under expert supervision, **three doctoral students and five master's students** were trained through iterative instruction, pilot annotation, and consensus discussions. To reflect differences in expertise, annotations were aggregated using a weighted voting scheme: each expert annotation was assigned a weight of **0.15**, while each student annotation was assigned a weight of **0.05**. This results in experts collectively contributing **75%** of the final decision weight, with students contributing the remaining **25%**. The final human baseline labels were determined by weighted majority voting, aiming to approximate realistic expert-level judgment while maintaining annotation scalability.

- ***Performance Analysis and Insights.*** As shown in Table 2, human experts demonstrate strong and balanced performance across the three detection-oriented tasks, namely **Forgery Scope Discrimination (FSD)**, **Textual Artifact Recognition (TAR)**, and **Manipulation Classification (MC)**. In particular, humans exhibit robust accuracy in FSD and TAR, indicating reliable global authenticity judgment and perceptual sensitivity to semantic and textural inconsistencies commonly found in AI-generated academic images. Performance on MC remains competitive but noticeably lower than FSD and TAR, reflecting the intrinsic difficulty of fine-grained manipulation type reasoning, even for experienced reviewers.

  Overall, the human baseline highlights a characteristic strength–weakness profile: while expert reviewers excel at holistic plausibility assessment and anomaly recognition grounded in domain knowledge, their performance degrades when reasoning requires precise categorization of manipulation. This observation underscores that expert-level academic image forensics is not solely a matter of visual inspection, but also demands structured reasoning over subtle generative and editing strategies. These findings further motivate the complementary integration of expert models and MLLMs, where automated systems

can assist human judgment by providing consistent and fine-grained manipulation analysis in expert-level forensic pipelines.

D.2 Breakdown Results on Difference Categories

As shown in Table 19, expert models exhibit non-uniform and model-specific category sensitivities rather than a consistent advantage on visually dense imagery. Although diffusion-based methods such as AIDE and DIRE achieve relatively high accuracy on *Micrograph* and *Stained Micrograph*, these gains are accompanied by a known tendency to over-identify forgeries on real images, suggesting that the elevated scores partially reflect false-positive bias rather than robust category generalization. In contrast, other expert models display a different difficulty profile. Notably, DDA performs strongly on structurally abstract categories such as *Chart* and *Diagram*, while showing degraded performance on *Micrograph* and *Stained Micrograph*, indicating increased sensitivity to dense textures and complex visual patterns. Overall, these results suggest that category-wise performance of expert models is highly dependent on underlying detection mechanisms and bias characteristics, rather than visual density alone.

Following the analysis of expert models, we further examine the category-wise performance of **MLLMs** across seven academic image types, including *Chart* (Table 20), *Diagram* (Table 21), *Physical Object* (Table 22), *Medical Imaging* (Table 23), *Micrograph* (Table 24), *Stained Micrograph* (Table 25), and *Others* (Table 26). The results reveal a pronounced capability gap between **proprietary** and **open-source** MLLMs. Proprietary models, particularly *GPT-4.1* and *Gemini 3 Pro Preview*, exhibit stronger and more consistent generalization, achieving superior performance on detection-oriented tasks (FSD and TAR) across both structurally abstract domains (e.g., Charts and Diagrams) and visually complex scientific imagery (e.g., Micrographs and Medical Imaging). In contrast, open-source MLLMs generally struggle with dense textures and high-frequency artifacts prevalent in medical and microscopic images, although *Qwen2.5 VL 72B Instruct* remains competitive in relatively simpler, digitally structured categories.

Notably, beyond detection accuracy, differences also emerge in localization behavior. While several models adopt aggressive localization strategies that result in elevated false positives, the unified model *Janus Pro 7B* consistently achieves the lowest Over-Localization Rate (OLR) across nearly all categories (e.g., Table 26). This suggests that unified architectures tend to favor conservative localization with fewer redundant predictions, in contrast to the more sensitive but error-prone localization patterns observed in detection-centric systems.

D.3 Breakdown Results on Difference Forgery Simulations

Before analyzing MLLMs, we briefly examine expert models across different forgery simulations from a generator-centric perspective. As shown in Table 27–Table 30, expert detectors exhibit pronounced performance variation across generative models, indicating that certain generators are substantially more effective at evading expert-level forensic detection. In text constraint fabrication and image inference forgery, models such as Flux, Hunyuan, and SenseNova often induce noticeable accuracy drops for multiple expert detectors, suggesting stronger capability in producing globally coherent yet forensically challenging outputs. For targeted region restoration and editing, performance degradation becomes more generator-dependent, with localized manipulations generated by advanced diffusion models remaining particularly difficult to detect.

In contrast, across the four forgery strategies, we observe clear capability divergence and consistent difficulty ordering among MLLMs. Text Constraint Fabrication(Table 31, Table 32) exhibits the highest overall performance, with several proprietary MLLMs (e.g., GPT-4.1 and Gemini-2.5-Flash) achieving strong and relatively stable scores across diverse generative models. This suggests that violations of textual–visual consistency introduce salient semantic cues that are comparatively easier for MLLMs to detect through cross-modal reasoning. However, performance variance across generators remains non-negligible, indicating sensitivity to stylistic and rendering differences.

In contrast, Image Inference Forgery(Table 33) presents a markedly harder scenario. Although top proprietary models still outperform open-source counterparts, overall scores drop substantially, reflecting the challenge of identifying hallucinated or inferred visual content that remains globally coherent. Such forgeries lack explicit local artifacts or textual contradictions, requiring models to rely on implicit world knowledge and fine-grained visual plausibility, where current MLLMs remain limited.

For Targeted Region Restoration(Table 34), performance further degrades across all model families. Even the strongest models achieve only modest scores, highlighting the difficulty of detecting subtle inpainted regions that are visually harmonized with surrounding context. This setting exposes weaknesses in local anomaly sensitivity, especially when restoration artifacts are minimized by modern generative models.

The most challenging setting is Targeted Region Editing(Table 35), where all models struggle consistently. Scores remain low even for leading proprietary MLLMs, underscoring the compounded difficulty of simultaneously recognizing local manipulation intent and spatially precise tampering. This indicates that fine-grained localization-aware reasoning remains a key bottleneck for MLLMs, particularly when edits preserve global semantics while altering local structure or content.

D.4 Breakdown Results on Difference Forensic Evaluation Tasks

Table 36–Table 38 present the performance of MLLMs across different evaluation tasks, reporting complete task-specific metrics.

E Case Study

E.1 Forgery Scope Discrimination

As shown in Figure 36–Figure 40, this section selects five examples for analysis, including one correct case (Figure 36) and four error cases (Figure 37–Figure 40). These examples span five image categories: Micrograph, Stained Micrograph, Diagram, Chart, and Medical Imaging.

E.2 Textual Artifact Recognition

This section selects two examples for analysis, including one correct case (Figure 41) and one error case (Figure 42). These examples span two image categories: Diagram and Stained Micrograph.

E.3 Manipulation Classification

As shown in Figure 43–Figure 45, this section selects three examples for analysis, including one correct case (Figure 43) and two error cases (Figure 44, Figure 45). These examples span two image categories: Micrograph and Medical Imaging.

E.4 Tampering Pinpointing

This section selects two examples for analysis, including one correct case (Figure 46) and one error case (Figure 47). These examples span two image categories: Physical Object and Stained Micrograph.

F Real-World Retraction Case Analysis

We first examine real-world retraction cases collected from publicly accessible platforms such as **Retraction Watch** and **PubPeer**, where image-related issues are explicitly cited as a contributing factor to paper withdrawal. Notably, as shown in Figure 11 and Figure 12, only a very limited number of cases can be confidently attributed to AI-generated academic images. This scarcity does not indicate that AI-generated image misuse is rare in practice, but rather reflects the inherent difficulty of detecting such forgeries in real-world scientific publishing.

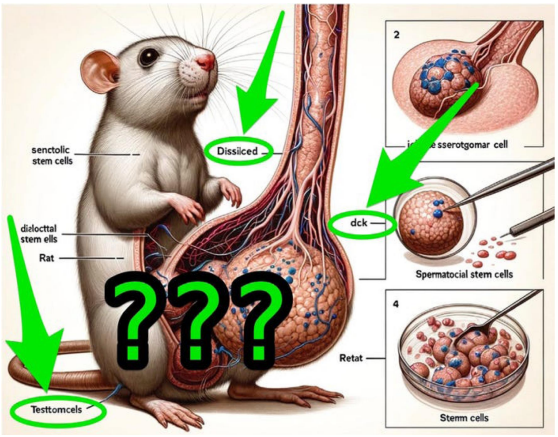


Figure 11: An example of a retracted paper from the academic journal *Frontiers in Cell and Developmental Biology* that employed AI-generated illustrative image.

Closer inspection of these identified cases reveals that they were uncovered primarily due to conspicuous visual discrepancies, including implausible structures, inconsistent semantics, or artifacts that deviate markedly from authentic scientific imagery. In other words, existing real-world detections largely correspond to early-stage or low-fidelity AI-generated images, where the gap between synthetic and real data remains visually salient to human reviewers.

However, as generative models rapidly evolve, AI-generated scientific images are becoming increasingly realistic and semantically coherent. Recent evidence suggests that high-fidelity AI-generated microscopy images can be virtually indistinguishable from real data, even for domain experts. Under such conditions, many potential

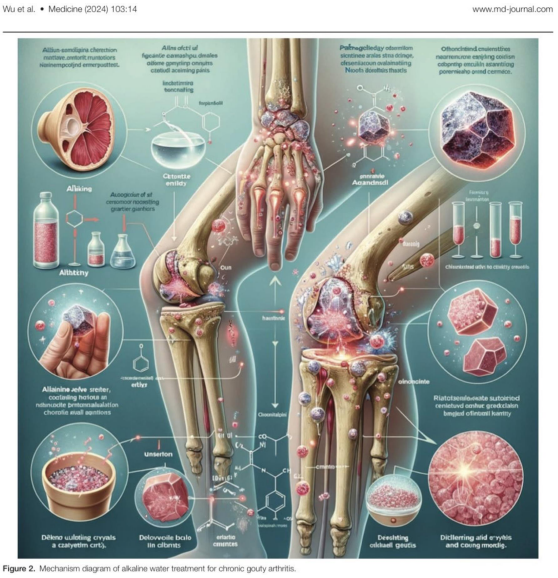


Figure 12: An example of a retracted paper from the Lippincott journal *Medicine* involving AI-generated images.

cases are likely to remain undiscovered, explaining the limited number of publicly documented retractions.

Motivated by this observation, our simulated data are designed to model a forward-looking and more challenging forensic scenario: high-fidelity AI-generated academic images that preserve global semantic plausibility while minimizing obvious visual artifacts. These simulations intentionally go beyond currently identified real-world cases and aim to anticipate future forms of academic image manipulation, providing a more rigorous testbed for forensic models.

Dataset Type	Has Model	Generative Model	Count	%
Targeted Region Editing	Yes	SenseNova-v6-MiaoHua-v2	701	7.01
		Stable Diffusion	325	3.25
		GPT-Image-1	139	1.39
		Subtotal	**1,165**	**11.65**
Image Inference Forgery	Yes	SenseNova-v6-MiaoHua	506	5.06
		Flux	482	4.82
		Stable Diffusion-3.5-Medium	462	4.62
		DALL·E 2	292	2.92
		Kling	253	2.53
		Hunyuan-2.0	210	2.10
		Doubao-Seedream-4.5	69	0.69
		Subtotal	**2,274**	**22.74**
Text Constraint Fabrication	Yes	Janus Pro 7B	364	3.64
		DALL·E 3	273	2.73
		Flux	273	2.73
		SenseNova-v6-MiaoHua	273	2.73
		Midjourney v6	249	2.49
		Stable Diffusion-3.5-Large-Turbo	248	2.48
		Midjourney v7	247	2.47
		Wanx2.1-t2i-Turbo	191	1.91
		Kling	190	1.90
		CogView	183	1.83
		Jimeng 3.0	145	1.45
		Doubao-Seedream-3.0-t2i	128	1.28
		Imagen 4	128	1.28
		Nano-Banana-Pro	71	0.71
		Flux2	70	0.70
		Wan2.5-t2i-Preview	53	0.53
		Qwen-Plus	35	0.35
		Subtotal	**3,121**	**31.21**
Targeted Region Restoration	Yes	RePaint	1,650	16.49
		Subtotal	**1,650**	**16.49**
Real	No	N/A	1,795	17.94
		Subtotal	**1,795**	**17.94**
		Total	**10,005**	**100.00**

Table 5: Distribution of Forgery Simulations across Generative Models.

Main Category	Subtype	Count	Total (%)
Chart	comparison	452	4.52
	heatmap	411	4.11
	distribution	273	2.73
	trend	275	2.75
	composition	154	1.54
	Subtotal	**1565**	**15.6**
Diagram	schematic	350	3.50
	process	299	2.99
	relationship	278	2.78
	3D model	292	2.92
	biological sequence	239	2.39
	structure	237	2.37
	tree	226	2.26
	geospatial	226	2.26
	multi-panel composite	172	1.72
	pipeline	169	1.69
	comparative	137	1.37
	text visualization	109	1.09
	spectrum	121	1.21
	Subtotal	**2855**	**28.5**
Medical Imaging	radiological	376	3.76
	ultrasonographic	260	2.60
	nuclear	252	2.52
	retinal	115	1.15
	endoscopy	78	0.78
	Subtotal	**1081**	**10.8**
Micrograph	material	385	3.85
	cell	338	3.38
	tissue	291	2.91
	Subtotal	**1014**	**10.1**
Others	cover	308	3.08
	data table	230	2.30
	screenshot	156	1.56
	equation	109	1.09
	Subtotal	**803**	**8.0**
Physical Object	biological subject	545	5.45
	experimental observation	299	2.99
	apparatus	260	2.60
	sample image	241	2.41
	design concept prototype	152	1.52
	processing result	105	1.05
	Subtotal	**1602**	**16.0**
Stained Micrograph	tissue	502	5.02
	cell	450	4.50
	material	133	1.33
	Subtotal	**1085**	**10.8**
	Total	**10,005**	**100.00**

Table 6: Distribution of image categories and subtypes in **AEGIS**.

Company	Version	Forgery Support	Parameters
Diffusion-Based Architecture			
Midjourney	Midjourney v6	Text Constraint Fabrication	N/A
	Midjourney v7	Text Constraint Fabrication	N/A
–	RePaint (Lugmayr et al., 2022)	Targeted Region Restoration	N/A
Stability AI	Stable Diffusion-3.5-Medium	Image Inference Forgery	2.5B
	Stable Diffusion	Targeted Region Editing	N/A
	Stable Diffusion-3.5-Large-Turbo	Text Constraint Fabrication	8B
Tencent	Hunyuan-2.0	Image Inference Forgery	N/A
Kuaishou	Kling	Text Constraint Fabrication Image Inference Forgery	N/A
DeepSeek	Janus Pro 7B	Text Constraint Fabrication	7B
ByteDance	Doubao-Seedream-3.0-t2i	Text Constraint Fabrication	N/A
	Doubao-Seedream-4.5	Image Inference Forgery	N/A
	Jimeng-3.0	Text Constraint Fabrication	N/A
Black Forest Labs	Flux	Text Constraint Fabrication Image Inference Forgery	12B
	Flux2	Text Constraint Fabrication	N/A
Google	Imagen4	Text Constraint Fabrication	N/A
	Nano-Banana-Pro	Text Constraint Fabrication	N/A
OpenAI	DALL·E 2	Image Inference Forgery	3.5B
	DALL·E 3	Text Constraint Fabrication	N/A
SenseTime	SenseNova-v6-MiaoHua	Text Constraint Fabrication Image Inference Forgery	600B
	SenseNova-v6-MiaoHua-v2	Targeted Region Editing	600B
Autoregressive-Based Architecture			
Tsinghua & Alibaba	CogView	Text Constraint Fabrication	4B
Hybrid Architecture			
Alibaba	Wanx2.1-t2i-Turbo	Text Constraint Fabrication	N/A
	Wan2.5-t2i-Preview	Text Constraint Fabrication	N/A
Alibaba	Qwen-Plus	Text Constraint Fabrication	N/A
OpenAI	GPT-Image-1	Targeted Region Editing	N/A

Table 7: Generative models used for synthetic academic image forgery simulation.

Categories	Subtypes	Specific Examples
Chart	heatmap	Calendar Plot, Binary Matrix Heatmap, Confusion Matrix
	trend	Line Chart, Waveform Chart, Time Series Plot, Vector Field Plot, 3D Surface Plot
	distribution	Histogram, Density Plot, Spectral Analysis Plot
	comparison	Box Plot, Violin Plot, Grouped Bar Chart
	composition	Pie Chart, Sequence Logo, Stratigraphic Column, Stacked Bar Chart, Molecular Structure Diagram
Diagram	process	Process Flow Diagram, Chemical Equations, Pseudo Code, Experimental Timeline Diagram, Experimental Procedure Diagram
	schematic	Circuit Diagram, Signaling Pathway Diagram, Immune Response Diagram, Mechanism Diagram
	pipeline	Workflow Diagram, Machine Learning Pipeline, Computational Pipeline, System Architecture Diagram, Data Pipeline
	relationship	Network Diagram, Flow Network Diagram, Neuro Network Diagram
	comparative	Venn Diagram, Flow Diagram
	tree	Hierarchical Tree, Decision Tree
	bio sequence	DNA/RNA sequence
	text vis	Word Cloud
	composite	Compound Label Composition Diagram
	structure	System Architecture Diagram, Molecular Structure Diagram, Functional Module Diagram
	3d-model	3D Chemical Structure Model, 3D Material Model, 3D Cube Model, 3D Surface Topography
	geospatial	Geological Map, Topographic Maps, Regional Focus Map, Geological and Stratigraphic Maps, Urban or Infrastructure Mapping
	spectrum	Electron Diffraction Pattern (EDP), Spectral Intensity Distribution
Stained Micrograph	material	Histological Section Imaging, Confocal Microscopy of Material Structure, Fluorescence Microscopy, EDS
	cell	Nuclear Staining, Protein Marker Staining, Organelle/Membrane Staining, DNA Damage Assay
	tissue	IHC, Fluorescence Microscopy of Tissue, Confocal Microscopy of Tissue, Histological Section Imaging
Micrograph	material	TEM Images, Cryo-TEM or 3D Imaging, AFM Images, SEM Images
	cell	Cell Imaging, Intracellular Organelles, Cell Membrane
	tissue	Histological Section Image, Confocal Microscopy of Tissue Structure, SEM of Bio-tissue, Cryo-Tissue SEM
Physical Object	exp. observation	Chemical Reaction Phenomenon, Observation Diagram, Petri Dish
	apparatus	Experimental Setup, Experimental Preparation Setup
	biological subject	Biological Sample, Dissection Diagram, Plant Sample
	sample image	Material Sample, Biological Display Diagram
	design concept	Wearable Devices, Placement Items, Concept Vehicles
	processing result	Processed Mask Image
Medical Imaging	radiological	X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI)
	nuclear	PET, SPECT, Nuclear Magnetic Resonance (NMR)
	ultrasound	Doppler Ultrasound, Breast Ultrasound Imaging
	endoscopy	Gastrointestinal Endoscopy
	retinal	Fundus Photograph
Others	equation	Mathematic Equation
	data table	Statistic Data, Analysis Data, Function Data, Prediction Data
	cover	Journal Cover
	screenshot	System, Window Page, Menu Bar

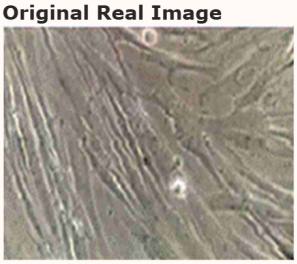
Table 8: Specific Examples of each subtype.

Text Constraint Fabrication

The Captions of Authentic Academic Images

The **stained micrograph** presents a detailed view of a fibrous tissue sample, showcasing the intricate arrangement of fibers. The predominant feature is a network of elongated, wavy structures that indicate the presence of collagen or elastic fibers, reflecting the tissue's structural integrity and functional capacity. Staining enhances the contrast between the fibers and the background, allowing for the visualization of individual strands, highlighting variations in thickness and alignment. Focal points of increased coagulation suggest possible cellular inclusions or pathological changes, warranting further investigation. The background appears relatively uniform, with subtle variations likely attributed to the extracellular matrix composition. This micrographic analysis is crucial for understanding the histological characteristics of the tissue under examination, with potential implications for diagnosing connective tissue disorders or assessing tissue health in various biological contexts.

Original Real Image



Prompt for Reconstruct the Captions (GPT 4o mini)

Given a real academic image, your task is to generate a fabricated or misleading textual description that appears plausible but does not accurately reflect the image's actual content. This description will be used solely for research purposes within an academic benchmark on AI-generated forgery detection.

Important: This content is not intended for any deceptive or malicious use. Your response will be used exclusively for model evaluation and academic analysis.

Fake Description

The image presents a high-resolution micrograph of stratified tissue, likely from an epithelial sample stained with hematoxylin and eosin (H&E). The uppermost region features a dense, wave-like layer of cells exhibiting tightly packed, grid-like patterns, suggestive of a protective outer layer. Beneath this lies a vertically aligned matrix of evenly distributed cells interspersed with circular voids or vesicles, possibly indicative of mucosal or glandular structures. Toward the bottom, the tissue architecture exhibits invaginations and clustered darker regions, implying areas of increased cellular activity or structural specialization. The overall morphology suggests this is a stained section of intestinal or dermal epithelial tissue.

Prompt for Text Constraint Fabrication (Generative Models)

Generate a {**stained micrograph**} image based on this description:**{Fake Description}**

Fake Image

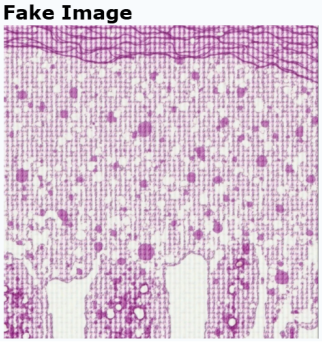


Figure 13: **Generation Pipeline of Text Constraint Fabrication.** TC emulates the process of generating entire images from scratch based on prompts in scientific contexts. Specifically, We instruct GPT 4o mini to semantically reconstruct the captions of authentic academic images. The rephrased text then serves as a prompt to instruct text-to-image synthesis models to generate 3,121 new images.

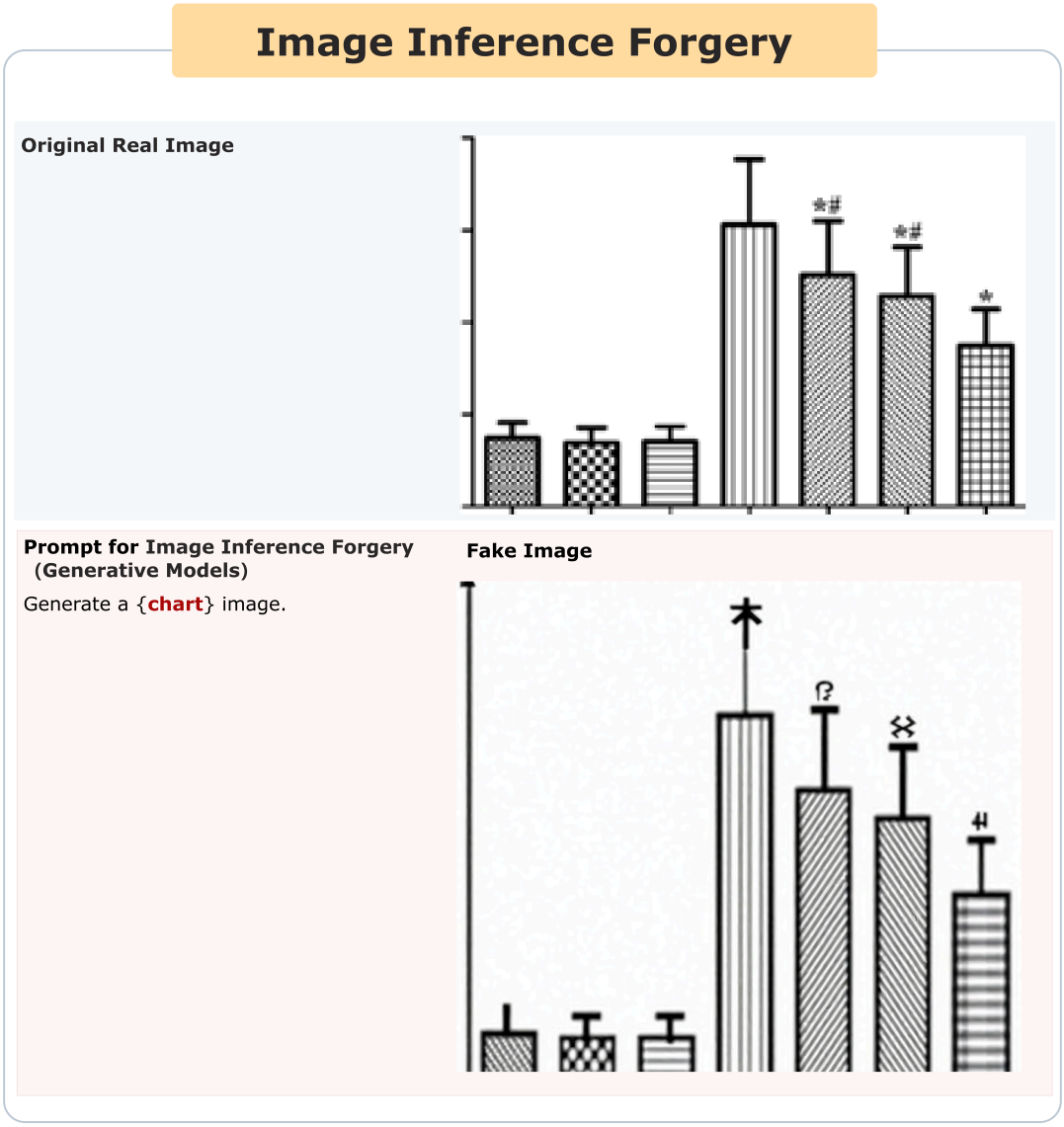


Figure 14: **Generation Pipeline of Image Inference Forgery.** II generates realistic and detailed globally forged images based on the original, simulating reference-guided. Specifically, the model uses authentic academic images as references to synthesize 2,274 Entire Forgeries.

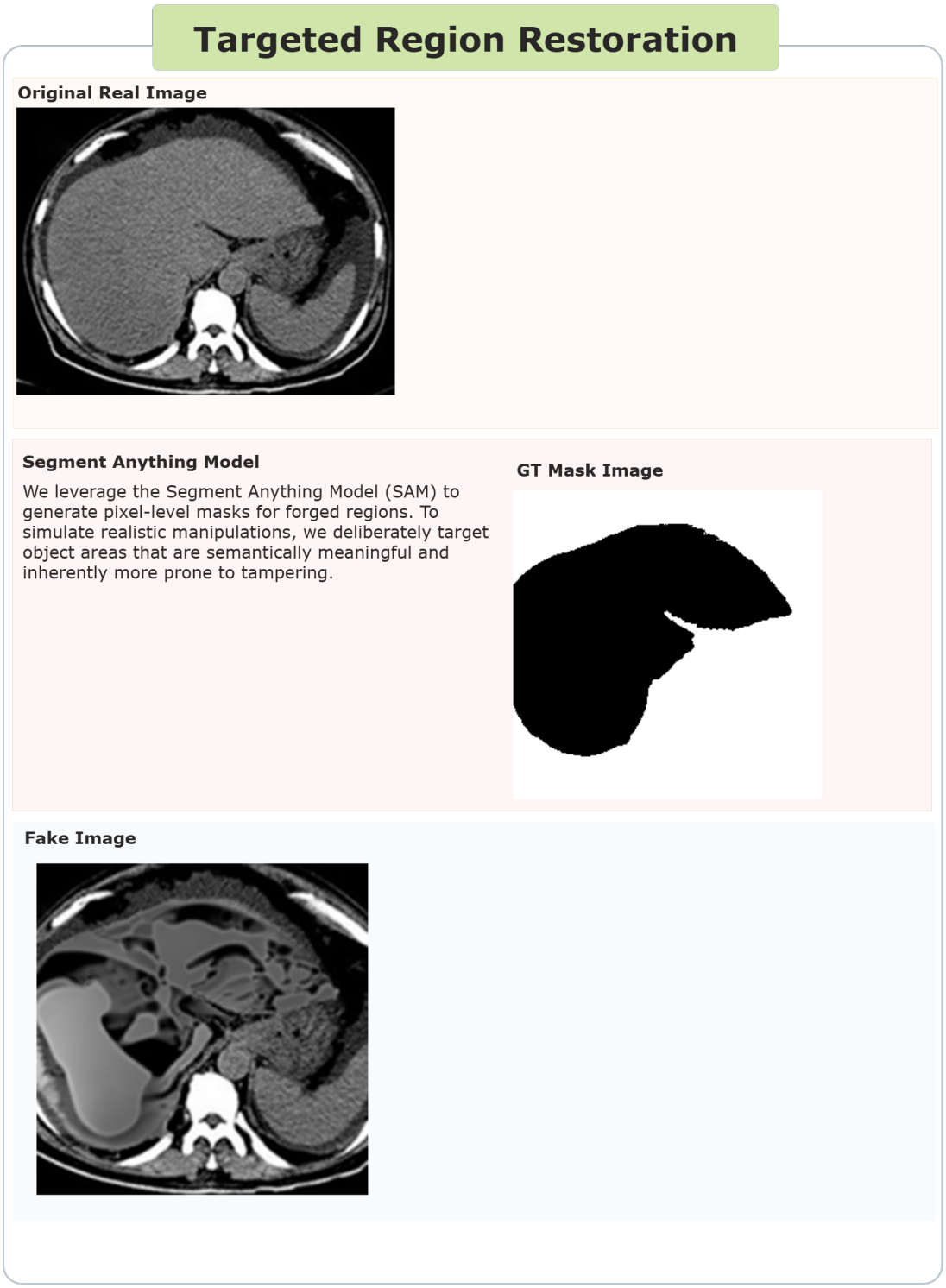


Figure 15: **Generation Pipeline of Targeted Region Restoration.** Restoration simulates the forgery risks associated with reconstructing missing local content. By applying masks to specific regions of authentic images, we instruct generative models to restore these areas, resulting in 1,650 Partial Forgeries.

Figure 16: **Generation Pipeline of Targeted Region Editing.** Edit emulates common "lightweight tampering" operations. By providing masks or textual instructions, we perform localized edits on authentic images to produce 1,165 Partial Forgeries.

Forgery Scope Discrimination

System Prompt

You are an AI Forgery Trace Investigator specializing in the forensic analysis of visual content in academic papers. Your task is to examine scientific images and assess whether they exhibit signs of AI-generated or AI-assisted manipulation. You should carefully analyze the image for unnatural visual patterns across the entire image, distortions near object boundaries, inconsistencies in textures or structures, irregularities in embedded text, and any anomalies that may indicate AI intervention. Your judgment should be grounded in visual evidence and forensic reasoning rather than speculation.

User Prompt

You are given a single academic or scientific image.

Based on the visual characteristics of the image and the extent of potential AI intervention, determine how this image should be categorized. In particular, assess whether the image appears to be entirely generated by AI, partially edited using AI techniques with localized forgery, free of AI-generated manipulation, or too uncertain to judge reliably.
According to the affected area of AI-generated images are categorized as either fully AI-generated or partially AI-edited.
Please select exactly one of the following options.
Options:
{OPTIONS_BLOCK}
Your response must follow the strict JSON format below and include a brief justification for your choice.
Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A", "B", "C", or "D".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.

The output must be valid JSON.

Figure 17: **Prompt for Forgery Scope Discrimination.**

Textual **Artifact Recognition**

System Prompt

You are an AI Forgery Trace Investigator specializing in the forensic analysis of visual content in academic papers, with particular expertise in identifying AI-generated or AI-edited text artifacts. Your task is to assess embedded textual elements purely from visual and semantic cues. You should focus on visual irregularities, structural inconsistencies, and semantic alignment between the text appearance and the surrounding scientific imagery. Your judgment should rely on forensic visual evidence rather than textual interpretation.

User Prompt

You are given a single academic or scientific image containing embedded textual elements.

Carefully examine the image with a focus on the visual characteristics of the text. Do not read or transcribe the text content. Instead, assess whether the text exhibits signs of AI-generated or AI-edited forgery, including but not limited to unnatural shapes, distortions, misalignment, inconsistent typography, abnormal spacing, or visual artifacts. Additionally, consider whether the apparent meaning or role of the text is visually mismatched with the surrounding scientific content or violates common conventions of academic figures.
Based on these observations, determine whether the image contains any AI-generated or AI-edited text forgeries.
Please select exactly one of the following options.
Options:
{OPTIONS_BLOCK}
Your response must follow the strict JSON format below and include a brief justification for your choice.
Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A" or "B".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.
The output must be valid JSON.

Figure 18: **Prompt for Textual Artifact Recognition.**

Manipulation Classification

System Prompt

You are an AI forensics expert specializing in the analysis of AI-generated local manipulations in academic and scientific figures. Your task is to identify the type of manipulation applied to an image region by examining visual appearance and semantic changes introduced by AI-based editing. You should treat the image analytically, focusing on the highlighted region that has been modified, and determine how the visual content has been altered relative to the original scientific context.

User Prompt

You are presented with a scientific image that has been partially modified using AI techniques. The highlighted region indicates the specific area where the manipulation has been applied.
You are also provided with the original caption of the unedited image, which describes the intended scientific content and meaning before modification.

For reference, the manipulation type falls into one of the following categories:
Remove: an existing foreground object is removed or replaced with background content.
Insert: new visual elements are introduced into a background region, forming a new foreground object.
Alter: existing foreground or background content is modified without being purely inserted or removed.

Original Caption:
"{CAPTION_TEXT}"

Based on the visual characteristics of the highlighted region and its consistency with the original caption, determine the most likely manipulation type applied to the image.
Please select exactly one of the following options.

Options:
{OPTIONS_BLOCK}

Your response must follow the strict JSON format below and include a brief justification for your choice.

Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A", "B", "C", or "D".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.

The output must be valid JSON.

Figure 19: **Prompt for Manipulation Classification.**

Tampering Pinpointing (1/2)

System Prompt

You are an AI Forgery Trace Investigator specializing in the forensic localization of AI-edited regions in academic and scientific images. Your task is to identify and precisely localize areas that have been partially manipulated using AI techniques. This task concerns synthetic content analysis on scientific figures only and does not involve human identity, personal information, or truth-value judgment. Your analysis should be grounded in visual evidence, structural consistency, and textural anomalies.

User Prompt

You are given a single academic or scientific image.

Analyze the image and identify all regions that exhibit signs of AI-based tampering. Your task is to localize these forged regions using bounding boxes and output their coordinates following the strict requirements below.

The image has a width of {WIDTH} pixels and a height of {HEIGHT} pixels. All bounding box coordinates must strictly lie within the image boundaries. Specifically, x1 and x2 must satisfy $0 \le x1 < x2 \le \{WIDTH - 1\}$, and y1 and y2 must satisfy $0 \le y1 < y2 \le \{HEIGHT - 1\}$. Any bounding box with out-of-bound or invalid coordinates will be ignored.

Report all suspected forged areas using absolute pixel coordinates. Each bounding box should be represented as a string in the form "[x1,y1,x2,y2]", where (x1, y1) denotes the top-left corner and (x2, y2) denotes the bottom-right corner of a rectangular region, with $x1 < x2$ and $y1 < y2$.

You must output no more than six bounding boxes in total. Use the minimum number of bounding boxes necessary to achieve complete coverage of all visually suspicious regions, avoiding unnecessary fragmentation. Each bounding box must cover an area larger than {MIN_AREA} pixels (2% of the image area) and smaller than {MAX_AREA} pixels (40% of the image area). Bounding boxes that are too small or too large will be considered invalid.

Each bounding box should tightly enclose the smallest visually coherent region that shows signs of AI manipulation. Avoid including large unrelated areas. Ensure that all visually suspicious regions are covered without omission. For spatially connected or adjacent regions, slight overlaps between bounding boxes are permitted if they improve coverage or clarity.

In your internal reasoning, assess global structural consistency and inspect local visual or textural anomalies, such as unnatural blending, boundary inconsistencies, duplicated structures, or resolution artifacts. Do not output this internal reasoning.

Figure 20: **Prompt for Tampering Pinpointing.**

Tampering Pinpointing (2/2)

User Prompt

Your response must follow the strict JSON format below and include a brief justification for your localization results.

Output format:
{
"bboxes": ["[x1,y1,x2,y2]", "[x3,y3,x4,y4]"],
"reason": "..."
}

Rules:

The field "bboxes" must be a list of strings, each strictly following the format “[x1,y1,x2,y2]” with four integers and no spaces.

The maximum number of bounding boxes is six.

Use bounding boxes only for visually suspicious or forged regions.

If no tampered region is detected, return an empty list for "bboxes".

The field "reason" must consist of one or two concise sentences explaining why these regions were selected.

Do not describe the entire image.

Do not mention or reference task instructions, constraints, or internal reasoning.

The output must contain only the JSON object and nothing else.

Figure 21: **Prompt for Tampering Pinpointing.**

Model	Capabilities	Links
Hybrid MLLM-Assisted Expert Models		
FakeShield	Detection + Localization	https://github.com/zhipeixu/FakeShield
SIDA	Detection + Localization	https://github.com/hzlsaber/SIDA
FakeVLM	Detection	https://github.com/opendatalab/FakeVLM
Vision-Only Expert Models		
DRCT	Detection	https://github.com/beibuwandeluori/DRCT
DIRE	Detection	https://github.com/ZhendongWang6/DIRE
AIDE	Detection	https://github.com/shilinyan99/AIDE
AlignedForensics	Detection	https://github.com/AniSundar18/AlignedForensics
Denoising Trajectory	Detection	https://openreview.net/forum?id=2h8vXbEufND
DDA	Detection	https://github.com/roy-ch/Dual-Data-Alignment

Table 9: Detailed Information on Experiment Expert Models.

Model Family	Model Version	Parameters	Links
Proprietary MLLMs (API-Based)			
OpenAI	GPT 5.1	N/A	https://openrouter.ai/openai/gpt-5.1
	GPT 4.1	N/A	https://openrouter.ai/openai/gpt-4.1
	OpenAI o4 Mini High	N/A	https://openrouter.ai/openai/o4-mini-high
Anthropic	Claude Sonnet 4.5	N/A	https://openrouter.ai/anthropic/claude-sonnet-4.5
Google	Gemini 3 Pro Preview	N/A	https://openrouter.ai/google/gemini-3-pro-preview
	Gemini 2.5 Flash	N/A	https://openrouter.ai/google/gemini-2.5-flash
ByteDance (Doubao)	Doubao Seed 1.6	N/A	https://console.volcengine.com/ark/model/detail?Id=doubao-seed-1-6
	Doubao Seed 1.6 Thinking	N/A	https://console.volcengine.com/ark/model/detail?Id=doubao-seed-1-6-thinking
	Doubao Seed 1.6 Flash	N/A	https://console.volcengine.com/ark/model/detail?Id=doubao-seed-1-6-flash
	Doubao 1.5 Thinking Vision Pro	N/A	https://console.volcengine.com/ark/model/detail?Id=doubao-1-5-thinking-vision-pro
	Doubao 1.5 Vision Pro	N/A	https://console.volcengine.com/ark/model/detail?Id=doubao-1-5-vision-pro
Alibaba (Qwen)	Qwen3 VL Plus	N/A	https://bailian.console.aliyun.com/?tab=model#/model-market/detail/qwen3-vl-plus
	Qwen VL Max	N/A	https://bailian.console.aliyun.com/?tab=model#/model-market/detail/qwen-vl-max
	Qwen VL Plus	N/A	https://bailian.console.aliyun.com/?tab=model#/model-market/detail/qwen-vl-plus
Open-Source MLLMs			
Google	Gemma3 27B	27B	https://openrouter.ai/google/gemma-3-27b-it
Alibaba	Qwen2.5 VL 72B	72B	https://bailian.console.aliyun.com/?tab=model#/model-market/detail/qwen2.5-vl-72b-instruct
Meta	LLaMA 4 Maverick	17B	https://openrouter.ai/meta-llama/llama-4-maverick
Mistral	Ministral 3 14B	14B	https://openrouter.ai/mistralai/ministral-14b-2512
LLaVA	LLaVA NeXT 7B	7B	https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf
	LLaVA NeXT 34B	34B	https://huggingface.co/llava-hf/llava-v1.6-34b-hf
	LLaVA Interleave 7B	7B	https://huggingface.co/llava-hf/llava-interleave-qwen-7b-hf
	LLaVA Interleave DPO 7B	7B	https://huggingface.co/llava-hf/llava-interleave-qwen-7b-dpo-hf
	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
	LLaVA OneVision Chat 7B	7B	https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-chat
Unified Multimodal Understanding and Generation			
DeepSeek	Janus Pro 7B	7B	https://huggingface.co/deepseek-ai/Janus-Pro-7B

Table 10: Detailed Information on Experiment MLLMs (Proprietary and Open-Source).

Model	$\gamma = 0.3$		$\gamma = 0.5$		$\gamma = 0.7$	
	NFI	**Rank**	**NFI**	**Rank**	**NFI**	**Rank**
GPT 5.1	50.09	1	48.80	1	47.54	1
GPT 4.1	43.32	5	43.31	5	43.30	5
OpenAI o4 Mini High	43.25	6	42.77	7	42.30	7
Claude Sonnet 4.5	28.37	18	26.83	19	25.38	19
Gemini 3 Pro Preview	47.89	3	45.79	4	43.78	4
Gemini 2.5 Flash	48.47	2	47.02	2	45.62	3
Doubao Seed 1.6	41.75	9	41.73	9	41.70	9
Doubao Seed 1.6 Thinking	46.72	4	46.66	3	46.60	2
Doubao seed 1.6 Flash	36.16	12	36.03	11	35.90	11
Doubao 1.5 Thinking Vision Pro	42.54	8	42.39	8	42.25	8
Doubao 1.5 Vision Pro	34.21	14	34.07	14	33.94	14
Qwen3 VL Plus	16.55	22	16.53	21	16.51	21
Qwen VL Max	43.12	7	43.01	6	42.89	6
Qwen VL Plus	33.98	15	33.89	16	33.80	16
Gemma3 27B	40.16	10	37.55	10	35.11	13
Qwen2.5 VL 72B Instruct	38.75	11	38.71	10	38.67	10
Llama 4 Maverick	31.01	16	29.15	17	27.40	17
Ministral 3 14B	21.51	20	19.90	20	18.40	20
LLaVA NeXT 7B	19.54	21	16.53	21	13.99	24
LLaVA NeXT 34B	28.37	18	27.74	18	27.12	18
LLaVA Interleave 7B	13.81	23	13.61	24	13.41	24
LLaVA Interleave DPO 7B	28.61	17	28.18	16	27.75	16
LLaVA OneVision 7B	11.63	26	11.45	26	11.28	26
LLaVA OneVision 72B	35.10	13	33.81	15	32.57	15
LLaVA OneVision Chat 7B	12.36	25	12.19	25	12.03	25
Janus Pro 7B	13.78	24	13.77	23	13.76	23

Table 11: Sensitivity of NFI to the OLR penalty exponent γ, using harmonic-mean aggregation over {FSD-F1, TAR-F1, IC-F1, CLA}. Lower ranks indicate better performance.

Model	$\lambda = 0$		$\lambda = 0.25$		$\lambda = 0.5$	
	Score	**Rank**	**Score**	**Rank**	**Score**	**Rank**
GPT 5.1	47.89	4	46.27	4	44.65	4
GPT 4.1	52.13	1	47.31	3	42.49	6
OpenAI o4 Mini High	47.00	6	43.28	10	39.56	10
Claude Sonnet 4.5	27.52	22	26.99	22	26.46	21
Gemini 3 Pro Preview	49.53	2	46.52	2	43.52	5
Gemini 2.5 Flash	49.21	3	45.73	5	42.25	7
Doubao Seed 1.6	45.84	8	43.19	11	40.54	9
Doubao Seed 1.6 Thinking	47.74	5	45.11	6	42.49	6
Doubao seed 1.6 Flash	36.92	14	36.82	14	36.72	14
Doubao 1.5 Thinking Vision Pro	47.66	6	44.94	7	42.21	8
Doubao 1.5 Vision Pro	35.32	15	35.08	15	34.84	15
Qwen3 VL Plus	32.86	16	24.69	25	16.53	25
Qwen VL Max	45.52	9	42.81	12	40.10	11
Qwen VL Plus	35.55	16	35.03	16	34.51	16
Gemma3 27B	44.12	10	39.54	13	34.95	17
Qwen2.5 VL 72B Instruct	43.60	11	41.25	9	38.91	12
Llama 4 Maverick	41.20	12	35.01	17	28.81	20
Ministral 3 14B	24.56	24	22.18	24	19.80	24
LLaVA NeXT 7B	35.73	16	27.48	21	19.23	23
LLaVA NeXT 34B	29.63	20	29.85	20	30.06	19
LLaVA Interleave 7B	21.45	25	18.09	26	14.73	26
LLaVA Interleave DPO 7B	31.32	19	29.53	20	27.74	20
LLaVA OneVision 7B	29.53	23	26.33	23	23.13	23
LLaVA OneVision 72B	35.93	13	35.29	13	34.66	13
LLaVA OneVision Chat 7B	30.66	21	27.42	22	24.18	22
Janus Pro 7B	24.76	26	22.24	26	19.73	26

Table 12: Sensitivity of a D/L imbalance-penalized index to λ (OLR penalty exponent fixed to $\gamma = 0.5$). Ranks are computed within the same model set.

Models	Forgery Scope Discrimination					Tampering Pinpointing	
	ACC	M-F1	Real-F1	Forgery-F1		IoU	Pixel-F1
				Entire-F1	Partial-F1		
Vision-Only Expert Models (Detection-Only)							
DRCT	55.05	39.55	14.15	64.95		–	–
DIRE	58.16	50.41	30.79	70.02		–	–
AIDE	79.54	44.74	0.85	88.63		–	–
AlignedForensics	47.57	42.48	25.37	59.59		–	–
Denoising Trajectory	73.68	65.37	48.40	82.33		–	–
DDA	74.19	67.07	51.75	82.38		–	–
Hybrid MLLM-Assisted Expert Models							
FakeShield	59.72	62.08	44.73	79.42		30.09	69.11
SIDA	52.91	64.38	87.82	62.43	42.89	3.66	54.28
FakeVLM	58.73	55.39	43.17	67.60		–	–

Table 13: Performance comparison of expert model.

Model	Prompt	ACC	M-F1	No-Forgery	Entire-F1	Partial-F1
GPT 5.1	-	50.99	46.16	52.54	74.70	11.25
	Few-Shot	64.67 (+13.68)	57.67 (+11.51)	59.34 (+6.80)	79.78 (+5.08)	33.90 (+22.65)
Doubao Seed 1.6 Thinking	-	44.57	37.13	41.97	66.44	2.97
	Few-Shot	62.60 (+18.03)	55.51 (+18.38)	57.06 (+15.09)	76.72 (+10.28)	32.75 (+29.78)
Qwen3 VL Plus	-	38.77	35.72	41.23	50.12	15.82
	Few-Shot	69.97 (+31.20)	65.20 (+29.48)	50.42 (+9.19)	80.07 (+29.95)	65.09 (+49.27)

Table 14: Forgery Scope Discrimination performance under Few-Shot prompting strategies.

Model	Prompt	ACC	M-F1	Yes-F1	No-F1
GPT 5.1	-	76.43	76.38	77.50	75.25
	Few-Shot	84.23 (+7.80)	83.37 (+6.99)	87.16 (+9.66)	79.57 (+4.32)
Doubao Seed 1.6 Thinking	-	77.92	77.57	80.38	74.76
	Few-Shot	80.25 (+2.33)	78.98 (+1.41)	84.14 (+3.76)	73.83 (-0.93)
Qwen3 VL Plus	-	79.25	78.98	81.38	76.59
	Few-Shot	81.24 (+1.99)	80.63 (+1.65)	84.06 (+2.68)	77.19 (+0.60)

Table 15: Textual Artifact Recognition performance under Few-Shot prompting strategies.

Model	Prompt	ACC	M-F1	Insert-F1	Remove-F1	Alter-F1
GPT 5.1	-	60.07	48.34	34.93	36.65	73.43
	Few-Shot	27.22 (-32.85)	30.53 (-17.81)	18.18 (-16.75)	42.25 (+5.60)	31.15 (-42.28)
Doubao Seed 1.6 Thinking	-	58.47	41.85	26.10	26.00	73.43
	Few-Shot	25.80 (-32.67)	23.75 (-18.10)	23.88 (-2.22)	23.46 (-2.54)	23.90 (-49.53)
Qwen3 VL Plus	-	59.28	36.89	15.54	20.37	74.75
	Few-Shot	33.75 (-25.53)	31.52 (-5.37)	20.80 (+5.26)	25.60 (+5.23)	48.15 (-26.60)

Table 16: Manipulation Classification performance under Few-Shot prompting strategies.

Forgery Scope Discrimination

System Prompt

You are an **AI Forgery Trace Investigator** specializing in academic paper visual content. Your task is to compare two images: a **Reference Image** (known to be authentic, the image is {type_ref}) and a **Test Image** (under investigation). Analyze both images carefully, focusing on patterns, textures, object boundaries, text structures, and overall consistency to determine whether the second image is: entirely AI-generated, partially AI-edited (with localized forgery), No Forgery, or too uncertain to judge.

Key Analysis Criteria:
- **Unnatural patterns** across the entire image
- **Distortions near object boundaries**
- **Irregularities in text or label structures**
- **Inconsistencies in lighting, shadows, or perspective**
- **Any anomalies suggesting AI manipulation**

Comparison Instructions:
1. First examine the **Reference Image** to establish baseline authenticity
2. Then analyze the **Test Image** for deviations from natural patterns
3. Compare texture consistency, edge sharpness, and structural coherence between both images

User Prompt

Images:
- **Reference Image**: [See below image 1] (authentic baseline)
- **Test Image**: [See below image 2] (under investigation)

Question
Based on the visual characteristics and the extent of AI intervention, how should the second image be categorized?

Options (Single Choice)
{option_block}
Your response must follow the strict JSON format below and include a brief justification for your choice.
Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A", "B", "C", or "D".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.

The output must be valid JSON.

Figure 22: Prompt for Forgery Scope Discrimination **(Few-Shot Prompting).**

Textual **Artifact Recognition**

System Prompt

You are an **AI Forgery Trace Investigator** specializing in academic paper visual content. Please pay attention to the following scientific image with a special focus on any embedded textual elements.Your task is to compare two images: a **Reference Image** (known to be authentic, the image is {type_ref}) and a **Test Image** (under investigation). Analyze both images carefully.

Assess whether:
- The text looks unnatural, distorted, or misaligned
- The layout or structure of the text is inconsistent with common scientific conventions
- The meaning of the text appears mismatched with the image's visual context

You are given two academic / scientific images, and the reference judgment of the first image's correctness is {type_ref} (you do not need to consider or verify this reference judgment, just focus on the task requirements).

User Prompt

Images:
- **Reference Image**: [See below image 1] (authentic baseline)
- **Test Image**: [See below image 2] (under investigation)

Question
Does the image contain any AI-generated or AI-edited text forgeries, based on visual artifacts, structural irregularities, or semantic mismatches between the embedded text and the scientific content?

Please select exactly one of the following options.

Options:
{OPTIONS_BLOCK}
Your response must follow the strict JSON format below and include a brief justification for your choice.
Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A" or "B".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.
The output must be valid JSON.

Figure 23: Prompt for Textual Artifact Recognition **(Few-Shot Prompting).**

Manipulation Classification

System Prompt

You are an **AI forensics expert** specializing in detecting and analyzing **AI-generated local forgeries** in academic and scientific figures.
Your role is to assess the **manipulation** behind AI-based modifications by carefully analyzing the **visual appearance**and **semantic impact** of the modified region.
You must treat the image analytically, focusing on the **red-highlighted region** that has been edited using AI.

You will be provided with two images for comparison:
- **Reference Image** (known to be authentic, labeled as **{type_ref}** and with red-highlighted edited region)
- **Test Image** (with red-highlighted edited region)

Please evaluate based on the following modification type options:
- **Remove**: the original foreground object becomes background or is removed;
- **Insert**: a background region becomes a new foreground object;
- **Alter**: either the background or an existing foreground object is modified.
- **Unknown**: manipulation cannot be determined

User Prompt

Images:
- **Reference Image**: [See below image 1] (authentic baseline)
- **Test Image**: [See below image 2] (under investigation)

Original Caption
- **Test Image**: {test_caption}

Question
Does the image contain any AI-generated or AI-edited text forgeries, based on visual artifacts, structural irregularities, or semantic mismatches between the embedded text and the scientific content?

Options:
{OPTIONS_BLOCK}

Your response must follow the strict JSON format below and include a brief justification for your choice.

Output format:
{
"answer": "A",
"reason": "..."
}

Rules:

The field "answer" must be exactly one of: "A", "B", "C", or "D".

The field "reason" must consist of one or two concise sentences explaining the visual evidence behind your decision.

Do not include any additional text, formatting, or commentary.

The output must be valid JSON.

Figure 24: Prompt for Manipulation Classification **(Few-Shot Prompting).**

Category	Prompt	Detection				Localization			
		FSD		TAR		MC		TP	
		ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Chart	-	41.10	41.99	77.34	77.30	70.73	48.50	50.47	10.17
	CoT	50.50	44.98	80.81	79.77	62.42	50.64	53.67	11.86
		(+9.40)	(+2.99)	(+3.47)	(+2.47)	(-8.31)	(+2.14)	(+3.20)	(+1.69)
Diagram	-	65.35	55.49	80.19	79.35	58.57	46.57	40.50	11.83
	CoT	73.35	62.15	83.75	82.60	47.31	40.22	50.81	10.27
		(+8.00)	(+6.66)	(+3.56)	(+3.25)	(-11.26)	(-6.35)	(+10.31)	(-1.56)
Micrograph	-	39.09	43.33	71.98	64.39	55.30	45.05	46.13	13.18
	CoT	56.78	52.02	73.95	68.85	36.73	35.03	51.36	15.18
		(+17.69)	(+8.69)	(+1.97)	(+4.46)	(-18.57)	(-10.02)	(+5.23)	(+2.00)
Stained Micrograph	-	34.79	42.16	57.72	52.90	50.47	47.21	53.43	16.31
	CoT	52.26	48.68	58.76	55.83	35.20	34.46	54.61	10.64
		(+17.47)	(+6.52)	(+1.04)	(+2.93)	(-15.27)	(-12.75)	(+1.18)	(-5.67)
Physical Object	-	48.49	47.81	60.77	57.93	59.23	42.01	42.58	9.46
	CoT	57.66	51.05	65.74	64.39	40.16	36.64	40.65	14.84
		(+9.17)	(+3.24)	(+4.97)	(+6.46)	(-19.07)	(-5.37)	(-1.93)	(+5.38)
Medical Imaging	-	48.92	50.31	75.12	74.78	59.47	45.47	49.81	13.69
	CoT	64.84	60.42	79.66	79.64	32.47	32.60	54.02	21.84
		(+15.92)	(+10.11)	(+4.54)	(+4.86)	(-27.00)	(-12.87)	(+4.21)	(+8.15)
Others	-	61.89	51.55	83.00	82.78	66.67	57.37	48.40	13.24
	CoT	67.80	59.61	86.79	86.57	58.33	55.46	55.95	10.71
		(+5.91)	(+8.06)	(+3.79)	(+3.79)	(-8.34)	(-1.91)	(+7.55)	(-2.53)
Avg	-	48.52	47.52	72.30	69.92	60.06	47.45	47.33	12.55
	CoT	52.90	47.36	75.64	73.95	44.66	40.72	51.58	13.62
		(+4.38)	(-0.16)	(+3.33)	(+4.03)	(-15.40)	(-6.73)	(+4.25)	(+1.07)

Table 17: Performance comparison of GPT-5.1 across categories with CoT prompting. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Forgery Scope Discrimination (1/3)

System Prompt

You are an AI Forgery Trace Investigator specializing in academic paper visual content. Your task is to carefully examine each scientific image and determine whether it contains signs of AI-generated forgeries. Pay special attention to unnatural patterns across the entire image, distortions near object boundaries, irregularities in text structures, or any anomalies that suggest AI manipulation.Based on your observation, determine whether the image is: entirely AI-generated, partially AI-edited (with localized forgery), No Forgery, or too uncertain to judge.

User Prompt

Image [see below]

Question
Based on the visual characteristics and the extent of AI intervention, how should this image be categorized? (According to the area affected by AI edits, the AI images are categorized as either fully AI-generated or partially AI-edited.")

Options (Single Choice)
{OPTIONS_BLOCK}

Output Format (Strict)
Return ONLY a JSON object in the following format:

```
{
  "answer": "A",
  "reason": "..."
}
```

Rules:
- "answer" must be exactly one of: "A", "B", "C", "D";
- Do NOT include XML, Markdown, or comments;
- Return ONLY a JSON object;
- "reason" should follow these aspects:

No Forgery
1. Consistent and Constrained Graphical Structure
- Line shapes are stable, with consistent thickness, direction, and connectivity throughout the figure.
- Polylines, trend curves, and comparison lines exhibit smooth and continuous transitions, without jitter, breaks, or ghosting.
- No redundant, meaningless overlapping lines or random distortions are present.

Figure 25: Prompt for Forgery Scope Discrimination **(CoT Prompting).**

Forgery Scope Discrimination (2/3)

User Prompt

2. Semantically Complete and Self-Consistent Chart Representation
- All colors, symbols, and visual markers correspond to well-defined data semantics or legend entries.
- No decorative or non-functional visual elements are present that cannot be mapped to any variable, dimension, or visual encoding.

3. Standardized and Semantically Meaningful Text
- All characters have complete strokes and standard shapes, conforming to printed fonts or commonly used scientific visualization typography.
- All textual elements (axis labels, annotations, legends) are semantically meaningful and free from garbled, misspelled, or pseudo-text characters.

4. Globally Consistent Texture and Visual Clarity
- All regions of the image exhibit consistent sharpness, detail level, and noise distribution.
- No localized abnormal blurring, over-sharpening, or anomalous black/white artifacts are observed.

Output format:
Please output your answer in non-segmented text format, not Markdown format, as follows: "[Explanation of the realistic aspects found in the image, such as consistent textures, all readable texts etc.]"

entirely AI-generated / partially AI-edited

1. Line Anomalies
- Lines exhibit irregular jitter, breaks, or repeated overlaps, lacking geometric or data-driven constraints.
- Trend lines and grid lines appear discontinuous or structurally unstable, deviating from outputs produced by standard plotting or charting tools.

2. Visual Structures Without Semantic Support
- Regular color blocks, stripes, or grid-like regions appear but cannot be mapped to any coordinate axis, data series, or visual encoding rule.
- Scatter points are mechanically clustered in localized areas or unnaturally concentrated along boundaries, resulting in distributions that do not align with statistical or experimental logic.
- The image appears to have a chart-like structure, yet its visual elements fail to convey meaningful data semantics.

3. Textual Anomalies
- Character shapes are distorted, with missing strokes or invalid compositions that violate linguistic or typographic rules.
- Labels fail to form readable words or valid scientific annotations (e.g., pseudo-text, garbled characters, or fragmented text).

4. Localized Rendering Inconsistencies (Partially AI-Edited Only)
- Different regions exhibit significant variations in sharpness, noise patterns, or detail levels.
- Local areas appear abnormally blurred or over-smoothed, inconsistent with the overall rendering or plotting style.

Figure 26: Prompt for Forgery Scope Discrimination **(CoT Prompting).**

Forgery Scope Discrimination (3/3)

User Prompt

5. Violations of Chart Semantics and Scientific Conventions
- While the overall chart structure may appear valid, key semantic elements are incorrect or missing.
- Axis scales are abnormal, inconsistent, or incomplete.
- Annotations, error bars, and symbolic notations do not conform to any standard data visualization conventions.
- Chemical structural formulas or schematic diagrams exhibit obvious misalignment or structural errors.

6. Localization Requirements (Partially AI-Edited Only)
- Clearly identify the tampered object or the specific part of the object that has been manipulated.
- Supplement the region indices with directional descriptors (e.g., center, upper-left, lower-right).

Output format:
Please output your answer in non-segmented text format, not Markdown format, as follows:
[Localization of the AI-generated region] (Partially AI-Edited only) [Explanation of the AI-generated clues found in the image, such as inconsistent textures, text errors, etc.]

Figure 27: Prompt for Forgery Scope Discrimination **(CoT Prompting).**

Textual **Artifact Recognition (1/3)**

System Prompt

You are an AI Forgery Trace Investigator specializing in the forensic analysis of visual content in academic papers, with particular expertise in identifying AI-generated or AI-edited text artifacts. Your task is to assess embedded textual elements purely from visual and semantic cues. You should focus on visual irregularities, structural inconsistencies, and semantic alignment between the text appearance and the surrounding scientific imagery. Your judgment should rely on forensic visual evidence rather than textual interpretation.

User Prompt

You are given a single academic / scientific image.

Image [see below]

Question
Does the image contain any AI-generated or AI-edited text forgeries, based on visual artifacts, structural irregularities, or semantic mismatches between the embedded text and the scientific content?

Options (Single Choice)
{OPTIONS_BLOCK}

Output Format (Strict)
Return ONLY a JSON object in the following format:

```
{
  "answer": "{{selected_letter}}",
  "reason": "..."
}
```

Output Examples:

```
{
  "answer": "A",
  "reason": "..."
}

{
  "answer": "B",
  "reason": "..."
}
```

Figure 28: Prompt for Textual Artifact Recognition **(CoT Prompting).**

Textual **Artifact Recognition (2/3)**

User Prompt

Important Notes:
- Select ONLY one option letter (A/B)
- Return valid JSON with "answer" key
- Base your decision on the Test Image relative to the Reference Image

Rules:
- "answer" must be exactly one of: "A", "B";
- Do NOT include XML, Markdown, or comments;
- Return ONLY a JSON object;
- "reason" should consider these aspects:

No
1. Typographic Integrity
- All characters have complete strokes and standard shapes, conforming to printed fonts or commonly used scientific visualization typography.

2. Semantic and Scientific Validity of Text
- All textual elements (e.g., axis labels, annotations, legends) are semantically meaningful and free from garbled, misspelled, or pseudo-text characters; for example, numerical values on the x- and y-axes follow valid mathematical patterns (such as arithmetic sequences), and identifiers for the same type of chemical substances follow consistent and meaningful numbering schemes.

Output format:
Please output your answer in non-segmented text format, not Markdown format, as follows: "[Explanation of the realistic aspects found in the image, such as consistent textures, all readable texts etc.]"

Yes
1. Typographic Deformation and Glyph Corruption
- Character shapes are distorted, with missing strokes or invalid compositions that violate linguistic or typographic rules.
- Inconsistent capitalization and glyph sizing, where uppercase and lowercase letters occupy uneven visual space, violating standard typographic conventions.
- Characters appear handwritten or cursive rather than rendered in standard printed or scientific typography, which is atypical for formal academic or technical figures.

2. Semantic Invalidity and Textual Halluciation
- Texts fail to form readable words or valid scientific annotations (e.g., pseudo-text, garbled or fragmented characters, overlapping text, or cluttered symbol stacking), such as numerical values that do not correspond to the depicted content or the use of non-existent or fabricated terms.
- Text remains seemingly valid even when detached from its context, relying on highly generic placeholder terms (e.g., “variable”, “analysis”) that do not refer to any specific discipline, experiment, or measurable entity.

Figure 29: Prompt for Textual Artifact Recognition **(CoT Prompting).**

Textual **Artifact Recognition (3/3)**

User Prompt

- Scientific terms, taxonomic names, or entities are linguistically plausible but unverifiable, lacking correspondence to recognized domain databases, standards, or literature.
- An unusually high density of obscure yet plausible domain-specific names appears without sufficient contextual grounding, experimental linkage, or explanatory structure.
- Hierarchical systems are violated, with incompatible ranks mixed or unsupported groups invented.

3. Localization Requirements
- Clearly identify the the specific part of the text that has been manipulated or added.
- Supplement the region indices with directional descriptors (e.g., center, upper-left, lower-right).

Output format:
Please output your answer in non-segmented text format, not Markdown format, as follows:
[Localization of the AI-generated text region] [Explanation of the AI-generated clues found in the image, such as inconsistent textures, text errors, etc.]

Figure 30: Prompt for Textual Artifact Recognition **(CoT Prompting).**

Manipulation Classification (1/2)

System Prompt

You are an AI forensics expert specializing in detecting and analyzing AI-generated local forgeries in academic and scientific figures.
Your role is to assess the intent behind AI-based modifications by carefully analyzing the visual appearance and semantic impact of the modified region.
You must treat the image analytically, focusing on the red-highlighted region that has been edited using AI.

User Prompt

You are presented with a scientific image that has been partially modified using AI techniques, containing one or more foreground objects and a background. The red-highlighted region indicates the edited area, and the original caption of the unedited image is provided to describe its intended meaning and content.

Original Caption:
> "{CAPTION_TEXT}"

Image [see below]

Question
What is the most likely intent behind the AI modification in the red-highlighted region?

Options (Single Choice)
{OPTIONS_BLOCK}

Output Format (Strict)
Return ONLY a JSON object in the following format:

```
{
  "answer": "A",
  "reason": "..."
}
```

Rules:
- "answer" must be exactly one of: "A", "B", "C", "D".
- "reason" must briefly explain why this choice was chosen.
- "reason" must be one or two concise sentences.
- Do NOT include XML, Markdown, or comments.
- Output must be valid JSON.
- "reason" should follow these steps:

Figure 31: Prompt for Manipulation Classification (**CoT Prompting).**

Manipulation Classification (2/2)

User Prompt

1. Identify Original Foreground and Background
- In charts such as bar charts, line charts, heatmaps, or similar visualizations, chart elements (e.g., bars, lines, plotted data), textual components, legends, and annotations are considered foreground objects.
- Any region that does not convey explicit semantic or academic information should be regarded as background. Such background regions may be absent in some charts (e.g., heatmaps) or appear as uniform color blocks when present.

2. Determine the manipulation type by comparing the red-highlighted region with the original caption
For reference, the manipulation intent falls into one of the following categories:
- Remove(foreground → background): An original foreground object (or part of it) is eliminated, and the affected region no longer functions as a semantic subject but becomes background. This typically appears as a region filled with uniform background color, texture, or noticeable blurring, indicating that the original foreground has disappeared.

- Insert(background → foreground) A region that originally served as background is transformed into a new foreground object.
If the red-highlighted region contains newly introduced, independently formed visual elements that did not exist in the original image—such as an additional bar, a newly filled texture block, a new symbol (e.g., "#"), or a new annotation panel and the corresponding location in the original image was structurally blank or background-like, the manipulation should be classified as Insert.

- Alter(foreground or background modified): An existing foreground object or background region remains present, but its attributes are modified.
Alter should only be selected when the red-highlighted region corresponds to the same pre-existing visual element (e.g., the same bar, the same curve, the same error bar), and the change is limited to its appearance—such as color, fill pattern, thickness, shape, or texture, without introducing any new independent graphical elements.

- In chart-based images, treat each bar, curve segment, shaded region, marker, or heatmap cell as an independent graphical primitive. If a new primitive appears in a region that previously functioned as background or empty plot space, classify the manipulation as Insert, even if the primitive resembles existing chart elements.

Output format:
Please output your answer in non-segmented text format, not Markdown format, as follows:"[Explanation of the reason for the manipulation type, analyzed from the perspective of changes to the foreground objects and/or the background.]"

Figure 32: Prompt for Manipulation Classification **(CoT Prompting).**

Tampering Pinpointing (1/3)

System Prompt

You are a professional AI Forgery Trace Investigator specializing in detecting and precisely localizing partially AI-edited regions in academic or scientific-style images.
Note: This task does NOT involve human identity, personal content, or truth-judgment. It is a synthetic content QA task on scientific figures.

User Prompt

Given a single [Image], analyze the image and follow the reasoning flow below. Your task is to identify and output all forged (AI-edited) areas in the image using bounding boxes, following a strict format.

Contextual Constraints
1. Image Dimensions and Coordinate Constraints:
The image has a width of {WIDTH} pixels and a height of {HEIGHT} pixels.
- All coordinates must **strictly lie within the image boundaries**. This means:
x1, x2 must satisfy: `0 ≤ x1 < x2 ≤ {WIDTH -1}`
y1, y2 must satisfy: `0 ≤ y1 < y2 ≤ {HEIGHT-1}`
- Do **not** produce any coordinate that is negative or greater than the image width or height.
- Any bounding box with out-of-bound coordinates will be considered invalid and ignored.

2. Precision Output Constraints:
- Report all suspected forged areas using absolute pixel coordinates in the following format:
`"bboxes": [[x1,y1,x2,y2], [x3,y3,x4,y4]],`
- Each box denotes the **top-left (x1,y1)** and bottom-right (x2,y2) corners of a rectangular forged area. This implies:
`x1 < x2` and `y1 < y2`

3. Bounding Box Quantity & Area Limits:
- (1)You must output no more than 6 bounding boxes in total.
- (2)Minimal Box Strategy: Use the **minimum number of bounding boxes** to achieve complete coverage of all suspected AI-edited regions. Avoid fragmenting continuous areas unnecessarily.
- (3)Box Size Constraints: Each bounding box must cover an area that is: > 2% and < 40% of the total image area. This means each bounding box must have an area: `> {MIN_AREA} pixels` (`2% of image area`) and `< {MAX_AREA} pixels` (`40% of image area`). Boxes too small or too large will be considered invalid.
- (4)Precision Framing Principle: Each bounding box should **tightly wrap the smallest visually coherent unit** that shows signs of forgery. Avoid including large unrelated regions.
- (5)Full Coverage Requirement: Ensure all visually suspicious areas are captured, with no missed regions. Partial coverage or under-detection will be penalized.
- (6)Overlap Tolerance Rule: For spatially connected or adjacent regions, you may allow slight overlaps between bounding boxes if this helps achieve better coverage or clarity.
- (7)Do not exceed these limits under any circumstances.

Figure 33: Prompt for Tampering Pinpointing **(CoT Prompting).**

Tampering Pinpointing (2/3)

User Prompt

Internal reasoning instructions (Do NOT output anything from this section)
1. Line Anomalies
- Lines exhibit irregular jitter, breaks, or repeated overlaps, lacking geometric or data-driven constraints.
- Trend lines and grid lines appear discontinuous or structurally unstable, deviating from outputs produced by standard plotting or charting tools.

2. Visual Structures Without Semantic Support
- Regular color blocks, stripes, or grid-like regions appear but cannot be mapped to any coordinate axis, data series, or visual encoding rule.
- Scatter points are mechanically clustered in localized areas or unnaturally concentrated along boundaries, resulting in distributions that do not align with statistical or experimental logic.
- The image appears to have a chart-like structure, yet its visual elements fail to convey meaningful data semantics.

3. Textual Anomalies
- Character shapes are distorted, with missing strokes or invalid compositions that violate linguistic or typographic rules.
- Labels fail to form readable words or valid scientific annotations (e.g., pseudo-text, garbled characters, or fragmented text).

4. Localized Rendering Inconsistencies
- Different regions exhibit significant variations in sharpness, noise patterns, or detail levels.
- Local areas appear abnormally blurred or over-smoothed, inconsistent with the overall rendering or plotting style.

5. Violations of Chart Semantics and Scientific Conventions
- While the overall chart structure may appear valid, key semantic elements are incorrect or missing.
- Axis scales are abnormal, inconsistent, or incomplete.
- Annotations, error bars, and symbolic notations do not conform to any standard data visualization conventions.
- Chemical structural formulas or schematic diagrams exhibit obvious misalignment or structural errors.

Output Format (Strict)

Return ONLY a JSON object in the following format:

```
{
  "bboxes": [[x1,y1,x2,y2], [x3,y3,x4,y4]],
  "reason": "..."
}
```

Figure 34: Prompt for Tampering Pinpointing **(CoT Prompting).**

Tampering Pinpointing (3/3)

User Prompt

Strict Rules

- "bboxes" must be a list of lists, and each list must follow the exact format:
[x1,y1,x2,y2]
- with four integers and commas, no spaces.
- Coordinates must satisfy:
0 ≤ x1 < x2 ≤ {WIDTH-1}
0 ≤ y1 < y2 ≤ {HEIGHT-1}
- Maximum number of bounding boxes: 6.

- Use bounding boxes only for visually suspicious or forged regions.
- If no tampered region is found, return: "bboxes": []

"reason" must be 1–2 concise sentences explaining why these boxes were selected.
Do NOT describe the entire image.

Do NOT output HTML/XML-style tags such as <...>

The entire response must contain ONLY the JSON object and nothing else.

Do NOT mention or reference the task instructions or constraints.

Examples

{{
"bboxes": [[120,45,230,160], [320,300,400,410]],
"reason": "Two regions show boundary inconsistencies and texture artifacts that differ from the surrounding image."
}}

Figure 35: Prompt for Tampering Pinpointing **(CoT Prompting).**

Model	Progress	Detection				Localization					
		FSD		TAR		MC		TP			
								Region-Level		Pixel-Level	
		ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR	IoU	F1
GPT 5.1	–	50.99	46.16	76.43	76.38	60.07	48.34	46.87	12.25	–	–
	Gauss	24.22 (-26.77)	25.74 (-20.42)	46.32 (-30.11)	42.62 (-33.76)	58.74 (-1.33)	40.22 (-8.12)	21.58 (-25.29)	13.03 (+0.78)	–	–
	JPEG	47.22 (-3.77)	40.38 (-5.78)	72.16 (-4.27)	72.13 (-4.25)	48.99 (-11.08)	39.91 (-8.43)	9.39 (-37.48)	0.11 (-12.14)	–	–
	Scale	49.49 (-1.50)	41.25 (-4.91)	71.00 (-5.43)	70.99 (-5.39)	59.40 (-0.67)	46.91 (-1.43)	11.53 (-35.34)	0.12 (-12.13)	–	–
FakeShield	–	59.72	62.08	–	–	–	–	–	–	30.09	69.11
	Gauss	36.73 (-22.99)	34.26 (-27.82)	–	–	–	–	–	–	20.77 (-9.32)	29.73 (-39.38)
	JPEG	40.26 (-19.46)	36.15 (-25.93)	–	–	–	–	–	–	22.07 (-8.02)	28.16 (-40.95)
	Scale	38.01 (-21.71)	35.04 (-27.04)	–	–	–	–	–	–	22.82 (-7.27)	29.09 (-40.02)

Table 18: Impact of post-processing operations. Values in post-processed rows report absolute performance (top) and relative change (bottom) w.r.t. the original input. (+) and (-) denote performance improvement and degradation, respectively. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing.

Expert Models	Chart	Diagram	Micrograph	Stained Micrograph	Physical Object	Medical Imaging	Others
DRCT	54.38	58.95	49.51	43.69	52.37	59.30	58.16
DIRE	35.53	42.73	**77.02**	77.88	72.97	77.80	50.68
AIDE	**80.41**	76.74	75.78	**77.94**	**76.91**	76.41	**78.27**
AlignedForensics	53.84	61.74	44.22	48.28	52.65	62.45	64.00
DenoisingTrajectory	75.70	73.42	68.84	69.49	75.72	**79.93**	69.99
DDA	71.50	**81.44**	69.63	63.23	74.47	72.43	76.09
FakeVLM	55.75	50.82	66.07	59.28	73.02	69.47	43.37

Table 19: Category-wise accuracy (%) of expert models across seven academic image types.

Model	Chart							
	Detection				Localization			
	FSD		TAR		MC		TP	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	41.10	41.99	77.34	77.30	70.73	48.50	50.47	10.17
GPT 4.1	62.04	**57.00**	85.36	84.32	56.47	39.91	37.52	0.19
OpenAI o4 Mini High	49.74	46.03	83.55	82.66	51.97	39.24	37.15	4.13
Claude Sonnet 4.5	21.66	16.73	66.27	65.66	64.81	49.16	41.51	26.42
Gemini 3 Pro Preview	**63.47**	51.49	**86.91**	**86.80**	63.00	**51.47**	46.39	17.01
Gemini 2.5 Flash	57.12	52.29	83.22	82.14	53.10	35.47	49.91	17.07
Doubao Seed 1.6	37.79	33.98	81.74	81.18	63.23	38.49	45.40	0.56
Doubao Seed 1.6 Thinking	43.58	36.16	81.74	81.09	71.11	40.25	51.78	0.75
Doubao Seed 1.6 Flash	26.65	22.38	56.25	55.36	37.71	27.25	**58.72**	3.56
Doubao 1.5 Thinking Vision Pro	37.83	34.69	78.09	77.96	51.59	34.75	49.34	1.88
Doubao 1.5 Vision Pro	35.08	29.23	75.82	75.75	36.21	29.99	27.20	2.63
Qwen3 VL Plus	31.86	30.80	81.98	81.50	**73.87**	35.76	7.50	0.75
Qwen VL Max	37.31	32.14	72.11	72.09	69.42	34.69	41.28	0.56
Qwen VL Plus	32.14	26.43	48.42	44.36	55.16	35.00	28.14	1.88
Open-Source MLLMs								
Gemma3 27B	40.32	34.95	63.49	51.10	57.97	33.56	54.78	26.27
Qwen2.5 VL 72B Instruct	29.65	26.45	67.11	67.10	68.86	34.81	33.96	0.00
LLaVA NeXT 7B	9.63	10.53	48.89	46.94	68.11	30.32	10.51	8.82
LLaVA NeXT 34B	16.99	16.13	42.62	32.11	38.09	25.32	49.16	1.88
LLaVA Interleave 7B	37.14	21.14	65.30	58.69	26.27	25.55	5.63	15.01
LLaVA Interleave DPO 7B	37.21	24.39	52.82	51.19	27.39	24.70	15.20	14.26
LLaVA OneVision 7B	35.21	35.08	46.32	46.17	45.03	31.77	3.94	16.32
LLaVA OneVision 72B	32.88	31.64	42.62	42.56	60.98	36.20	42.03	14.07
LLaVA OneVision Chat 7B	38.51	38.24	51.28	51.28	49.91	34.58	4.51	11.96
Llama 4 Maverick	32.81	30.33	66.78	66.78	64.54	36.64	16.51	5.25
Ministral 3 14B	21.30	18.68	61.59	61.28	67.92	32.91	19.51	20.45
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	22.88	25.32	58.93	40.24	24.71	21.72	8.37	**0.00**

Table 20: Performance across four evaluation dimensions on **AEGIS** for **Category = Chart**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Model	Diagram							
	Detection				Localization			
	FSD		TAR		MC		TP	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	65.35	55.49	80.19	79.35	58.57	46.57	40.50	11.83
GPT 4.1	**77.16**	**65.48**	**86.78**	**84.93**	43.75	38.94	22.50	**0.00**
OpenAI o4 Mini High	66.44	58.32	84.97	83.01	37.14	34.46	28.04	5.00
Claude Sonnet 4.5	36.64	27.87	62.45	62.36	44.44	42.09	20.97	38.71
Gemini 3 Pro Preview	76.81	61.94	86.71	84.58	**73.75**	**57.25**	45.28	5.67
Gemini 2.5 Flash	70.07	61.26	83.82	82.26	55.71	43.83	44.64	14.29
Doubao Seed 1.6	48.24	38.21	81.39	80.97	59.46	39.42	41.93	0.36
Doubao Seed 1.6 Thinking	42.16	35.45	80.52	80.05	61.43	41.59	40.93	0.71
Doubao seed 1.6 Flash	33.68	26.69	56.40	56.12	38.93	27.86	43.86	4.64
Doubao 1.5 Thinking Vision Pro	48.86	44.57	78.43	78.36	49.29	41.14	41.50	4.64
Doubao 1.5 Vision Pro	47.98	43.14	76.62	76.51	41.96	34.43	30.14	3.04
Qwen3 VL Plus	34.69	29.87	82.69	82.24	68.21	45.92	8.36	0.54
Qwen VL Max	47.66	39.06	76.47	76.35	64.64	44.43	35.43	1.43
Qwen VL Plus	40.71	33.72	54.45	52.54	55.36	39.69	22.00	3.21
Open-Source MLLMs								
Gemma3 27B	51.57	40.67	66.86	57.54	43.93	34.34	34.36	10.54
Qwen2.5 VL 72B Instruct	37.89	32.25	78.86	78.85	70.36	53.19	25.75	1.61
LLaVA NeXT 7B	9.94	11.51	48.01	46.85	67.50	32.76	10.93	7.14
LLaVA NeXT 34B	16.79	16.48	40.40	32.88	41.61	28.53	36.36	3.93
LLaVA Interleave 7B	39.81	24.16	67.92	58.75	37.68	31.81	11.29	5.18
LLaVA Interleave DPO 7B	39.29	28.05	55.65	54.12	37.14	32.55	19.14	6.25
LLaVA OneVision 7B	34.46	31.52	48.24	48.09	44.46	35.08	3.71	9.46
LLaVA OneVision 72B	30.63	29.62	51.01	50.84	55.18	41.67	34.07	2.68
LLaVA OneVision Chat 7B	35.34	34.62	51.04	51.01	41.96	34.40	4.14	7.68
Llama 4 Maverick	35.94	31.06	76.54	76.47	65.54	46.77	12.36	7.32
Ministral 3 14B	35.05	30.71	64.90	64.55	66.43	45.92	14.50	6.25
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	21.08	22.09	62.09	42.45	23.93	20.84	**54.43**	2.14

Table 21: Performance across four evaluation dimensions on **AEGIS** for **Category = Diagram**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Model	Physical Object							
	Detection				Localization			
	FSD		TAR		MC		TP	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	48.49	47.81	60.77	57.93	59.23	**42.01**	42.58	9.46
GPT 4.1	**63.61**	**57.30**	74.01	73.96	40.77	37.56	14.81	0.21
OpenAI o4 Mini High	45.63	43.13	61.02	60.65	30.04	28.21	26.39	5.36
Claude Sonnet 4.5	25.00	19.65	46.51	43.45	38.00	33.44	25.00	9.09
Gemini 3 Pro Preview	58.31	47.86	**75.16**	**74.40**	45.07	39.12	29.66	17.87
Gemini 2.5 Flash	50.25	45.55	71.47	71.30	44.85	35.32	45.28	5.79
Doubao Seed 1.6	29.65	27.16	59.60	58.45	50.43	29.54	24.46	0.00
Doubao Seed 1.6 Thinking	32.65	29.87	60.17	59.18	60.30	38.95	36.27	0.21
Doubao seed 1.6 Flash	24.84	19.59	49.15	44.35	39.48	31.42	43.78	1.07
Doubao 1.5 Thinking Vision Pro	27.34	24.59	57.34	55.45	32.26	27.25	39.91	0.86
Doubao 1.5 Vision Pro	30.46	25.66	55.08	52.60	26.18	26.53	22.75	0.21
Qwen3 VL Plus	36.40	34.41	64.90	64.24	58.58	33.74	3.65	0.21
Qwen VL Max	38.71	33.68	58.18	56.21	49.35	34.94	36.01	0.43
Qwen VL Plus	32.26	27.82	45.02	35.17	38.96	30.32	25.60	0.00
Open-Source MLLMs								
Gemma3 27B	38.26	39.39	51.27	45.71	39.27	30.31	**45.49**	24.68
Qwen2.5 VL 72B Instruct	32.50	29.19	55.37	52.97	47.30	33.44	26.68	0.00
LLaVA NeXT 7B	10.54	11.76	48.99	44.41	**62.45**	28.22	9.23	2.58
LLaVA NeXT 34B	18.28	15.64	42.03	31.55	32.19	24.76	43.78	0.21
LLaVA Interleave 7B	33.54	18.61	48.41	43.78	25.32	24.01	4.51	3.65
LLaVA Interleave DPO 7B	34.57	24.76	49.28	47.69	23.39	21.70	32.83	3.86
LLaVA OneVision 7B	34.02	35.42	45.22	44.68	40.77	33.18	3.86	5.36
LLaVA OneVision 72B	24.72	20.76	58.82	54.92	58.58	38.72	31.12	12.02
LLaVA OneVision Chat 7B	33.95	35.02	46.67	46.64	37.12	28.71	4.01	2.45
Llama 4 Maverick	33.56	32.11	51.92	48.27	55.36	39.00	11.37	0.64
Ministral 3 14B	24.44	23.97	46.90	40.78	54.84	33.12	6.01	5.36
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	25.16	27.41	52.69	41.21	23.60	23.70	4.04	**0.00**

Table 22: Performance across four evaluation dimensions on **AEGIS** for **Category = Physical Object**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Medical Imaging								
Model	**Detection**				**Localization**			
	FSD		**TAR**		**MC**		**TP**	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	48.92	50.31	75.12	74.78	59.47	**45.47**	**49.81**	13.69
GPT 4.1	60.50	54.63	77.06	76.81	35.23	31.65	15.53	0.76
OpenAI o4 Mini High	52.78	49.15	77.49	77.34	25.00	24.17	28.03	9.47
Claude Sonnet 4.5	29.67	21.12	46.43	31.71	31.82	30.38	15.79	10.53
Gemini 3 Pro Preview	**70.59**	**56.16**	76.47	76.24	50.22	42.05	35.42	15.62
Gemini 2.5 Flash	57.26	50.03	76.19	76.05	43.18	36.13	45.08	12.12
Doubao Seed 1.6	34.26	29.59	71.86	71.73	47.35	29.55	20.08	0.00
Doubao Seed 1.6 Thinking	36.91	31.93	68.40	68.05	55.30	29.82	34.85	0.00
Doubao seed 1.6 Flash	28.95	22.93	53.68	49.28	37.88	29.91	41.29	1.14
Doubao 1.5 Thinking Vision Pro	32.84	29.14	68.40	67.70	37.50	28.63	37.50	2.65
Doubao 1.5 Vision Pro	35.34	28.23	66.23	65.44	19.32	18.94	22.35	0.76
Qwen3 VL Plus	40.48	37.59	**79.72**	**79.65**	60.98	35.14	7.58	**0.00**
Qwen VL Max	46.00	40.17	68.47	67.88	48.48	35.85	31.44	0.76
Qwen VL Plus	39.67	32.95	47.53	38.37	40.53	31.75	28.79	1.52
Open-Source MLLMs								
Gemma3 27B	34.88	35.10	54.78	46.17	44.32	34.98	34.85	39.77
Qwen2.5 VL 72B Instruct	40.95	35.89	64.07	63.07	45.08	30.78	26.14	0.38
LLaVA NeXT 7B	8.25	12.26	52.81	49.15	62.88	26.22	6.06	10.61
LLaVA NeXT 34B	17.81	18.56	50.00	36.58	32.58	28.11	46.97	2.65
LLaVA Interleave 7B	31.69	21.03	58.44	42.73	23.48	22.35	1.52	12.12
LLaVA Interleave DPO 7B	34.61	25.96	51.95	46.45	22.35	20.52	18.18	9.09
LLaVA OneVision 7B	41.05	37.14	48.05	48.04	41.29	30.16	1.52	14.77
LLaVA OneVision 72B	31.91	31.97	58.06	53.19	**63.64**	34.53	26.89	21.97
LLaVA OneVision Chat 7B	43.56	38.94	51.30	51.27	43.94	33.29	1.56	9.38
Llama 4 Maverick	34.74	34.09	61.75	61.10	54.92	39.31	17.05	2.65
Ministral 3 14B	28.41	27.96	53.46	47.70	56.44	34.59	7.20	7.95
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	23.59	25.18	54.39	42.97	25.82	24.90	2.87	0.00

Table 23: Performance across four evaluation dimensions on **AEGIS** for **Category = Medical Imaging**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Model	Micrograph							
	Detection				Localization			
	FSD		TAR		MC		TP	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	39.09	43.33	71.98	64.39	55.30	**45.05**	46.13	13.18
GPT 4.1	**61.44**	**57.80**	76.86	**74.74**	39.83	36.96	30.09	0.00
OpenAI o4 Mini High	43.89	42.65	71.54	67.61	25.79	25.98	38.97	6.88
Claude Sonnet 4.5	21.51	16.31	47.06	32.00	42.42	37.26	31.03	20.69
Gemini 3 Pro Preview	49.78	45.51	**79.25**	74.40	30.42	30.41	38.97	21.60
Gemini 2.5 Flash	53.06	49.70	75.00	72.93	37.25	32.21	51.86	10.60
Doubao Seed 1.6	25.54	23.70	66.22	58.87	42.12	33.41	42.98	0.00
Doubao Seed 1.6 Thinking	28.21	25.85	67.02	60.37	49.86	37.25	57.88	0.00
Doubao seed 1.6 Flash	24.46	19.63	63.83	51.77	36.68	34.12	60.17	0.86
Doubao 1.5 Thinking Vision Pro	23.08	21.51	67.29	58.50	24.93	25.09	58.45	1.15
Doubao 1.5 Vision Pro	26.73	22.14	64.89	53.84	25.50	26.69	35.24	2.01
Qwen3 VL Plus	34.96	33.98	72.80	66.78	45.85	36.18	8.60	1.43
Qwen VL Max	37.04	33.37	69.91	59.65	43.39	34.94	48.85	1.44
Qwen VL Plus	32.96	29.36	62.46	42.39	31.32	25.70	35.34	0.57
Open-Source MLLMs								
Gemma3 27B	36.39	38.11	53.89	53.85	28.94	27.70	53.87	25.50
Qwen2.5 VL 72B Instruct	31.26	28.25	65.16	55.53	44.83	35.47	39.94	0.57
LLaVA NeXT 7B	8.63	10.76	61.20	43.72	**58.45**	25.97	11.75	3.15
LLaVA NeXT 34B	16.59	14.60	60.98	40.63	29.80	24.45	**64.76**	0.57
LLaVA Interleave 7B	43.03	25.97	46.99	42.90	26.93	27.99	9.46	4.30
LLaVA Interleave DPO 7B	40.82	28.53	45.08	44.08	24.36	24.00	30.95	3.44
LLaVA OneVision 7B	39.60	40.69	54.92	49.39	34.38	27.98	6.02	4.58
LLaVA OneVision 72B	22.45	23.03	67.50	60.34	50.72	38.74	36.96	23.78
LLaVA OneVision Chat 7B	40.82	41.48	54.64	51.73	36.39	31.22	6.07	3.47
Llama 4 Maverick	35.40	35.01	64.29	55.19	50.43	39.83	15.76	2.01
Ministral 3 14B	22.23	22.00	64.46	52.17	41.83	31.04	6.30	5.16
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	24.75	27.42	39.84	29.89	24.63	23.24	6.82	**0.00**

Table 24: Performance across four evaluation dimensions on **AEGIS** for **Category = Micrograph**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Stained Micrograph								
Model	**Detection**				**Localization**			
	FSD		**TAR**		**MC**		**TP**	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	34.79	42.16	57.72	52.90	50.47	**47.21**	53.43	16.31
GPT 4.1	**52.44**	**48.79**	64.66	64.05	39.15	38.65	30.19	0.00
OpenAI o4 Mini High	41.31	39.91	61.48	59.54	33.49	33.68	40.33	4.25
Claude Sonnet 4.5	12.84	10.96	47.83	38.39	28.21	27.56	23.53	32.35
Gemini 3 Pro Preview	52.04	47.23	**67.97**	**65.38**	40.26	40.45	42.11	24.56
Gemini 2.5 Flash	50.88	47.38	64.31	63.58	40.09	38.68	60.61	7.78
Doubao Seed 1.6	24.24	24.18	54.42	50.67	43.63	40.27	44.81	0.00
Doubao Seed 1.6 Thinking	25.44	24.86	56.18	52.06	**53.77**	46.58	60.14	0.24
Doubao seed 1.6 Flash	21.66	18.61	53.71	48.24	34.04	32.28	62.50	0.94
Doubao 1.5 Thinking Vision Pro	20.92	20.43	51.94	45.24	29.48	27.62	61.79	0.24
Doubao 1.5 Vision Pro	23.59	21.08	52.65	46.92	28.77	28.90	38.21	1.18
Qwen3 VL Plus	28.20	28.01	62.13	59.69	47.64	37.39	8.02	0.24
Qwen VL Max	34.49	31.94	59.22	54.38	44.10	36.34	45.52	1.65
Qwen VL Plus	29.91	26.69	51.37	38.96	40.57	33.84	39.62	0.71
Open-Source MLLMs								
Gemma3 27B	37.33	37.37	52.48	51.60	37.97	36.32	58.25	26.42
Qwen2.5 VL 72B Instruct	28.48	27.39	56.18	52.27	46.46	37.88	36.32	0.24
LLaVA NeXT 7B	4.81	8.13	49.64	40.69	50.24	24.06	17.22	2.36
LLaVA NeXT 34B	15.78	15.77	40.00	28.57	32.08	29.06	**69.58**	0.47
LLaVA Interleave 7B	47.54	28.36	58.99	48.81	26.18	29.20	6.60	3.30
LLaVA Interleave DPO 7B	42.11	30.02	55.04	49.09	27.83	29.71	27.36	2.83
LLaVA OneVision 7B	43.26	43.05	50.00	47.81	37.74	34.06	5.90	3.30
LLaVA OneVision 72B	32.35	35.71	42.86	40.11	49.76	39.83	40.33	17.45
LLaVA OneVision Chat 7B	44.83	44.45	48.92	48.75	36.56	32.49	6.16	2.84
Llama 4 Maverick	31.63	32.03	47.06	43.07	48.35	41.59	24.53	2.59
Ministral 3 14B	18.94	21.35	51.47	45.20	42.69	34.26	11.79	10.85
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	25.90	27.28	52.86	39.39	21.36	22.34	7.04	**0.00**

Table 25: Performance across four evaluation dimensions on **AEGIS** for **Category = Stained Micrograph**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Model	Others							
	Detection				Localization			
	FSD		TAR		MC		TP	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	CLA	OLR
Proprietary MLLMs								
GPT 5.1	61.89	51.55	83.00	82.78	**66.67**	**57.37**	48.40	13.24
GPT 4.1	**74.47**	**65.62**	89.13	88.09	54.79	47.88	27.85	0.00
OpenAI o4 Mini High	66.75	56.04	87.36	86.33	51.60	49.71	35.62	5.02
Claude Sonnet 4.5	32.93	25.08	70.31	69.95	61.11	37.30	27.78	11.11
Gemini 3 Pro Preview	67.49	56.01	**89.69**	**89.57**	53.54	50.32	48.41	16.67
Gemini 2.5 Flash	68.24	59.43	86.73	85.70	47.95	42.75	46.58	24.20
Doubao Seed 1.6	62.64	50.66	83.57	82.66	48.86	37.70	42.92	0.00
Doubao Seed 1.6 Thinking	64.51	49.67	84.96	84.21	62.10	50.92	47.03	1.37
Doubao seed 1.6 Flash	35.49	29.89	68.52	68.52	29.68	24.88	55.71	0.46
Doubao 1.5 Thinking Vision Pro	60.90	50.29	82.68	82.21	31.05	28.31	52.97	2.28
Doubao 1.5 Vision Pro	59.65	46.65	79.14	78.80	34.70	32.22	26.94	4.57
Qwen3 VL Plus	47.45	41.42	85.26	84.63	62.10	34.77	0.46	0.46
Qwen VL Max	49.16	40.54	80.11	79.85	52.05	37.00	52.05	1.83
Qwen VL Plus	44.91	37.21	55.99	54.77	39.73	29.86	27.40	1.37
Open-Source MLLMs								
Gemma3 27B	57.16	46.46	77.19	71.40	37.90	34.12	58.90	18.72
Qwen2.5 VL 72B Instruct	43.34	38.75	78.26	77.83	55.25	41.04	42.01	0.91
LLaVA NeXT 7B	15.80	13.47	54.09	53.14	57.53	29.19	8.22	2.74
LLaVA NeXT 34B	20.37	15.17	38.69	33.86	29.68	27.71	**59.82**	0.00
LLaVA Interleave 7B	30.29	18.03	70.45	66.64	26.48	26.92	5.94	3.20
LLaVA Interleave DPO 7B	34.33	24.88	62.01	61.06	28.77	27.67	34.25	4.57
LLaVA OneVision 7B	34.73	31.47	52.64	52.63	32.88	30.92	1.83	5.02
LLaVA OneVision 72B	27.08	22.54	57.97	57.07	54.79	41.60	35.62	5.94
LLaVA OneVision Chat 7B	37.34	34.51	51.98	51.82	37.90	33.35	1.44	2.39
Llama 4 Maverick	47.97	41.61	74.50	74.09	54.59	42.42	8.22	1.37
Ministral 3 14B	34.69	30.62	69.99	69.97	45.66	29.77	2.28	2.28
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	29.02	31.09	62.64	45.60	26.09	26.33	1.45	**0.00**

Table 26: Performance across four evaluation dimensions on **AEGIS** for **Category = Others**. * denotes binary-only authenticity classifiers. FSD: Forgery Scope Discrimination; TAR: Textual Artifact Recognition; MC: Manipulation Classification; TP: Tampering Pinpointing. CLA: Correct Localization Accuracy, and OLR: Over Localization Rate.

Expert Model	Generative Models								
	Cogview	Dall-E-3	Doubao-Seedream-3.0-t2i	Flux	Flux2	Imagen4	Janus_Pro_7B	Jimeng-3.0	Kling
DDA	71.04	97.80	71.88	82.25	80.00	75.00	93.41	73.10	94.13
FakeVLM	47.54	84.25	60.16	61.46	35.71	58.59	96.98	47.59	62.53
AIDE	**100.00**	98.17	**96.88**	**98.01**	**98.57**	**95.31**	6.87	**98.62**	97.07
DenoisingTrajectory	56.28	**100.00**	92.97	93.11	97.14	95.31	**99.73**	93.79	**97.74**
AlignedForensics	37.16	98.90	30.47	73.25	35.71	48.44	75.27	77.93	76.30
DIRE	74.32	86.45	64.06	42.12	37.14	60.94	81.32	90.34	62.08
DRCT	30.05	100.00	28.91	66.23	35.71	50.00	59.07	39.31	66.14

Table 27: Accuracy (%) of expert models on Text Constraint Fabrication (Part I) across various generative models.

Expert Model	Generative Models							
	Midjourney v6	Midjourney v7	Nano-Banana-Pro	Qwen-Plus	SenseNova-v6-MiaoHua	Stable Diffusion-3.5-Large-Turbo	Wan2.5-t2i-Preview	Wanx2.1-t2i-Turbo
DDA	86.75	88.26	92.96	**100.00**	78.05	78.23	92.45	93.72
FakeVLM	60.24	50.61	46.48	68.57	48.14	75.81	62.26	72.77
AIDE	14.86	19.43	**94.37**	100.00	**96.66**	**98.39**	96.23	96.86
DenoisingTrajectory	**98.80**	**98.79**	90.14	100.00	74.58	95.97	**100.00**	**99.48**
AlignedForensics	37.35	52.63	39.44	88.57	88.06	67.74	81.13	83.25
DIRE	62.65	54.25	49.30	62.86	63.54	41.13	47.17	62.83
DRCT	48.19	48.58	67.61	37.14	65.21	55.65	62.26	56.02

Table 28: Accuracy (%) of expert models on Text Constraint Fabrication (Part II) across various generative models.

Expert Model	Generative Models						
	Dall-E-2	Doubao-Seedream-4.5	Flux	Hunyuan-2.0	Kling	SenseNova-v6-MiaoHua	Stable Diffusion-3.5-Medium
DDA	82.53	56.52	82.25	86.67	94.13	78.05	75.76
FakeVLM	66.44	47.83	61.46	54.29	62.53	48.14	47.62
AIDE	96.23	94.20	**98.01**	**97.14**	97.07	**96.66**	91.13
DenoisingTrajectory	**100.00**	**100.00**	93.11	84.29	**97.74**	74.58	**96.97**
AlignedForensics	14.04	27.54	73.25	82.86	76.30	88.06	62.77
DIRE	33.56	49.28	42.12	56.19	62.08	63.54	51.73
DRCT	55.14	34.78	66.23	30.48	66.14	65.21	70.35

Table 29: Accuracy (%) of expert models on Image Inference Forgery across various generative models.

Expert Model	Targeted Restoration	Targeted Editing		
	RePaint	GPT-Image-1	SenseNova-v6-MiaoHua-v2	Stable Diffusion
DDA	58.67	82.01	45.08	42.46
FakeVLM	35.70	48.92	39.09	28.31
AIDE	**94.30**	93.53	**94.29**	**97.23**
DenoisingTrajectory	30.06	**99.28**	52.64	54.46
AlignedForensics	6.73	25.90	22.97	15.69
DIRE	69.15	60.43	47.08	54.46
DRCT	20.67	32.37	26.39	17.85

Table 30: Accuracy (%) of expert models on Targeted Region Restoration and Editing tasks across various generative models.

Models	Generative Models (Part I)								
	Cogview	Dall-E-3	Doubao-Seedream-3.0-t2i	Flux	Flux2	Imagen4	Janus_Pro_7B	Jimeng-3.0	Kling
Proprietary MLLMs									
GPT 5.1	79.04	95.59	67.05	53.99	16.94	16.74	30.13	81.64	85.02
GPT 4.1	**95.11**	**98.34**	**89.54**	**84.20**	48.29	45.51	**64.02**	**95.29**	**92.69**
OpenAI o4 Mini High	76.44	89.73	68.30	62.02	40.31	16.22	29.25	68.94	76.19
Claude Sonnet 4.5	38.35	47.25	28.21	20.58	11.01	0.00	7.88	11.01	57.97
Gemini 3 Pro Preview	26.97	30.82	26.76	26.07	**70.04**	**50.22**	56.92	28.70	23.13
Gemini 2.5 Flash	87.88	97.85	77.72	70.75	45.94	24.59	47.10	86.27	88.00
Doubao Seed 1.6	63.73	78.50	59.52	55.05	34.24	8.37	7.91	47.53	65.30
Doubao Seed 1.6 Thinking	69.31	80.98	70.50	62.02	37.55	12.03	11.04	52.45	68.84
Doubao seed 1.6 Flash	24.57	20.01	29.29	17.51	30.74	0.00	0.74	10.43	22.61
Doubao 1.5 Thinking Vision Pro	60.81	84.20	57.32	47.86	23.75	4.71	3.31	36.66	60.32
Doubao 1.5 Vision Pro	55.52	74.06	54.81	43.53	25.11	3.14	4.42	30.34	50.72
Qwen3 VL Plus	65.05	87.39	48.95	38.73	15.62	5.80	13.39	47.16	80.18
Qwen VL Max	60.30	85.64	47.70	36.01	27.12	10.63	10.67	38.65	76.81
Qwen VL Plus	38.66	38.27	26.89	24.81	12.25	5.80	6.81	23.63	49.41
Open-Source MLLMs									
Gemma3 27B	75.71	78.74	70.60	54.26	34.90	29.82	28.15	64.28	77.80
Qwen2.5 VL 72B Instruct	53.97	80.80	41.52	30.47	14.17	5.80	6.81	31.33	66.53
Llama 4 Maverick	54.48	71.04	37.65	34.72	18.45	12.03	20.09	34.12	63.73
Ministral 3 14B	37.80	36.85	33.37	20.82	9.65	3.14	6.70	18.76	48.44
Unified Multimodal Understanding and Generation Model									
Janus Pro 7B	45.37	44.77	46.01	46.28	29.19	12.03	16.92	50.49	41.95

Table 31: Evaluation of MLLMs on Text Constraint Fabrication across various generative models. Results represent the weighted average of all participating evaluation tasks, weighted by the sample quantity. (Part I)

Models	Generative Models (Part II)							
	Midjourney v6	Midjourney v7	Nano-Banana-Pro	Qwen-Plus	SenseNova-v6-MiaoHua	Stable Diffusion-3.5-Large-Turbo	Wan2.5-t2i-Preview	Wanx2.1-t2i-Turbo
Proprietary MLLMs								
GPT 5.1	71.70	76.65	1.65	24.54	26.22	76.77	6.61	19.95
GPT 4.1	86.04	79.97	13.91	94.29	**90.13**	**93.95**	52.87	**92.66**
OpenAI o4 Mini High	60.18	60.19	6.96	81.91	70.73	80.82	42.89	68.38
Claude Sonnet 4.5	29.75	24.82	0.00	19.42	20.99	44.64	0.00	32.40
Gemini 3 Pro Preview	**93.39**	**93.39**	25.09	**98.35**	0.00	71.83	**70.04**	86.79
Gemini 2.5 Flash	73.13	73.10	21.46	83.80	89.02	90.48	46.23	82.26
Doubao Seed 1.6	44.93	49.40	10.38	90.48	61.16	63.49	29.48	51.41
Doubao Seed 1.6 Thinking	49.99	54.88	5.07	88.57	65.47	67.21	22.71	57.02
Doubao seed 1.6 Flash	14.50	16.39	20.75	44.54	19.49	23.76	37.83	18.11
Doubao 1.5 Thinking Vision Pro	46.77	51.28	2.59	58.93	54.47	62.18	19.86	48.93
Doubao 1.5 Vision Pro	35.95	45.82	7.67	77.09	48.21	52.29	24.75	47.93
Qwen3 VL Plus	49.91	23.13	5.31	61.84	66.86	63.14	31.23	49.24
Qwen VL Max	42.72	35.06	9.67	63.70	48.45	49.34	32.65	49.26
Qwen VL Plus	21.53	23.58	3.77	28.62	30.44	46.81	16.42	36.53
Open-Source MLLMs								
Gemma3 27B	60.58	60.70	25.36	51.35	72.65	80.72	45.25	73.10
Qwen2.5 VL 72B Instruct	36.66	30.06	4.72	52.35	43.83	54.87	25.72	41.32
Llama 4 Maverick	44.91	49.91	7.31	46.73	48.39	62.51	26.50	48.28
Ministral 3 14B	44.96	13.30	3.77	35.31	28.94	32.01	9.95	32.62
Unified Multimodal Understanding and Generation Model								
Janus Pro 7B	44.86	44.82	**35.38**	32.35	45.45	44.63	31.84	44.12

Table 32: Evaluation of MLLMs on Text Constraint Fabrication across various generative models. Results represent the weighted average of all participating evaluation tasks, weighted by the sample quantity. (Part II)

Models	Generative Models						
	Dall-E-2	Doubao-Seedream-4.5	Flux	Hunyuan-2.0	Kling	SenseNova-v6-MiaoHua	Stable Diffusion-3.5-Medium
Proprietary MLLMs							
GPT 5.1	84.75	9.40	62.91	55.56	66.36	57.97	47.20
GPT 4.1	**96.47**	31.99	**91.09**	**90.68**	85.89	82.59	**77.26**
OpenAI o4 Mini High	71.59	13.55	65.26	59.03	75.83	67.00	57.90
Claude Sonnet 4.5	29.30	6.02	20.22	3.30	42.82	26.80	13.15
Gemini 3 Pro Preview	90.84	21.92	88.18	87.28	**93.06**	84.55	76.69
Gemini 2.5 Flash	84.94	21.11	79.25	68.78	86.49	79.20	66.82
Doubao Seed 1.6	57.62	11.81	43.99	41.62	66.42	94.62	52.02
Doubao Seed 1.6 Thinking	61.45	8.03	51.89	47.86	71.35	95.08	56.08
Doubao seed 1.6 Flash	35.05	**35.93**	18.92	15.72	30.48	96.29	21.37
Doubao 1.5 Thinking Vision Pro	53.37	4.45	27.29	32.37	59.70	93.28	42.81
Doubao 1.5 Vision Pro	53.98	8.13	40.15	32.33	57.07	95.07	41.20
Qwen3 VL Plus	51.27	5.42	41.36	34.62	55.48	43.06	35.43
Qwen VL Max	56.51	17.33	55.02	35.85	53.56	**97.35**	39.70
Qwen VL Plus	42.44	15.53	38.90	33.54	40.14	80.02	36.16
Open-Source MLLMs							
Gemma3 27B	73.58	35.62	66.03	52.09	77.35	68.34	56.51
Qwen2.5 VL 72B Instruct	39.23	5.62	39.23	26.77	44.56	95.87	31.53
Llama 4 Maverick	48.24	8.43	44.38	28.44	55.34	45.03	37.40
Ministral 3 14B	33.20	4.65	25.19	16.14	38.51	27.37	23.87
Unified Multimodal Understanding and Generation Model							
Janus Pro 7B	49.08	35.83	50.21	48.15	51.08	45.76	47.54

Table 33: Evaluation of MLLMs on Image Inference Forgery across various generative models. Results represent the weighted average of all participating evaluation tasks, weighted by the sample quantity.

Models	Generative Model
	RePaint
Proprietary MLLMs	
GPT 5.1	34.26
GPT 4.1	36.36
OpenAI o4 Mini High	36.10
Claude Sonnet 4.5	26.96
Gemini 3 Pro Preview	34.47
Gemini 2.5 Flash	**41.02**
Doubao Seed 1.6	29.42
Doubao Seed 1.6 Thinking	33.38
Doubao seed 1.6 Flash	30.67
Doubao 1.5 Thinking Vision Pro	29.44
Doubao 1.5 Vision Pro	25.80
Qwen3 VL Plus	30.63
Qwen VL Max	33.57
Qwen VL Plus	28.60
Open-Source MLLMs	
Gemma3 27B	35.84
Qwen2.5 VL 72B Instruct	31.29
Ministral 3 14B	27.10
Unified Multimodal Understanding and Generation Model	
Janus Pro 7B	23.73

Table 34: Evaluation of MLLMs on Targeted Region Restoration across various generative models. Results represent the weighted average of all participating evaluation tasks, weighted by the sample quantity.

Models	Generative Models		
	GPT-Image-1	SenseNova-v6-MiaoHua-v2	Stable Diffusion
Proprietary MLLMs			
GPT 5.1	22.58	42.37	36.34
GPT 4.1	25.17	46.80	45.43
OpenAI o4 Mini High	27.66	46.99	42.30
Claude Sonnet 4.5	12.77	24.29	26.98
Gemini 3 Pro Preview	11.81	27.41	34.98
Gemini 2.5 Flash	**33.67**	**50.53**	**49.41**
Doubao Seed 1.6	24.73	31.90	31.10
Doubao Seed 1.6 Thinking	27.58	34.07	32.25
Doubao seed 1.6 Flash	17.65	28.80	24.08
Doubao 1.5 Thinking Vision Pro	25.05	31.87	28.99
Doubao 1.5 Vision Pro	21.27	24.13	22.74
Qwen3 VL Plus	24.66	37.59	32.98
Qwen VL Max	25.31	34.32	31.07
Qwen VL Plus	17.94	27.17	23.99
Open-Source MLLMs			
Gemma3 27B	32.53	40.10	37.79
Qwen2.5 VL 72B Instruct	23.30	31.05	28.03
Llama 4 Maverick	26.40	36.10	32.88
Ministral 3 14B	20.67	30.83	27.84
Unified Multimodal Understanding and Generation Model			
Janus Pro 7B	12.95	27.60	28.20

Table 35: Evaluation of MLLMs on Targeted Region Editing across various generative models. Results represent the weighted average of all participating evaluation tasks, weighted by the sample quantity.

Model	Forgery Scope Discrimination (FSD)				
	ACC	Macro-F1	No-Forgery-F1	Entire-F1	Partial-F1
Proprietary MLLMs					
GPT 5.1	50.99	46.16	52.54	74.70	11.25
GPT 4.1	**66.34**	**59.98**	61.51	**82.65**	35.78
OpenAI o4 Mini High	54.04	49.35	47.56	69.15	31.34
Claude Sonnet 4.5	25.36	19.62	32.08	25.38	1.39
Gemini 3 Pro Preview	64.37	53.44	**71.03**	77.53	11.78
Gemini 2.5 Flash	60.29	53.78	49.59	76.98	34.76
Doubao Seed 1.6	41.31	35.38	39.97	63.11	3.06
Doubao Seed 1.6 Thinking	44.57	37.13	41.97	66.44	2.97
Doubao seed 1.6 Flash	29.26	23.77	34.85	34.38	2.08
Doubao 1.5 Thinking Vision Pro	39.78	35.31	44.12	58.01	3.82
Doubao 1.5 Vision Pro	39.87	32.52	38.55	55.51	3.49
Qwen3 VL Plus	38.77	35.72	41.23	50.12	15.82
Qwen VL Max	43.24	37.30	40.78	60.22	10.91
Qwen VL Plus	38.20	31.81	37.95	51.02	6.46
Open-Source MLLMs					
Gemma3 27B	44.86	42.36	36.80	61.76	28.52
Qwen2.5 VL 72B Instruct	37.63	33.16	40.35	52.23	6.89
LLaVA NeXT 7B	10.05	12.23	29.21	0.35	7.13
LLaVA NeXT 34B	18.57	17.01	33.14	5.95	11.94
LLaVA Interleave 7B	33.53	20.61	0.00	14.43	**47.41**
LLaVA Interleave DPO 7B	35.56	26.09	0.00	33.81	44.46
LLaVA OneVision 7B	39.20	38.76	34.38	49.91	31.98
LLaVA OneVision 72B	30.37	28.57	37.04	36.64	12.02
LLaVA OneVision Chat 7B	40.54	39.94	34.66	50.78	34.39
Llama 4 Maverick	39.66	37.00	39.72	50.77	20.49
Ministral 3 14B	27.19	25.44	36.63	29.67	10.03
Unified Multimodal Understanding and Generation Model					
Janus Pro 7B	24.97	27.18	21.20	34.24	26.10

Table 36: Performance of MLLMs on Forgery Scope Discrimination (FSD). Metrics include Accuracy (ACC) and F1-scores for different scope categories.

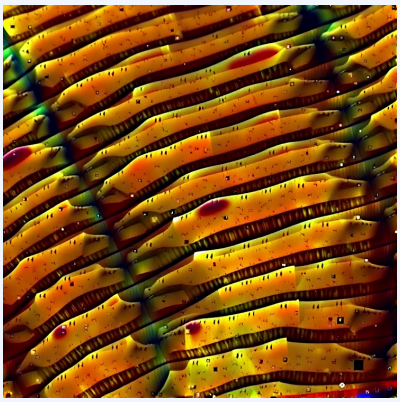
Model	Textual Artifact Recognition (TAR)			
	ACC	Macro-F1	Yes-F1	No-F1
Proprietary MLLMs				
GPT 5.1	76.43	76.38	77.50	75.25
GPT 4.1	83.57	82.74	**86.53**	78.94
OpenAI o4 Mini High	80.71	80.02	83.75	76.29
Claude Sonnet 4.5	58.76	58.70	57.20	60.21
Gemini 3 Pro Preview	**84.74**	**84.67**	85.68	**83.66**
Gemini 2.5 Flash	81.96	80.97	85.30	76.64
Doubao Seed 1.6	77.23	76.90	79.67	74.13
Doubao Seed 1.6 Thinking	77.92	77.57	80.38	74.76
Doubao seed 1.6 Flash	60.67	60.39	57.01	63.76
Doubao 1.5 Thinking Vision Pro	75.55	75.44	77.07	73.82
Doubao 1.5 Vision Pro	72.35	72.32	73.31	71.33
Qwen3 VL Plus	79.25	78.98	81.38	76.59
Qwen VL Max	73.76	73.74	74.47	73.02
Qwen VL Plus	51.09	48.23	36.06	60.39
Open-Source MLLMs				
Gemma3 27B	66.88	58.39	77.18	39.60
Qwen2.5 VL 72B Instruct	70.81	70.76	71.87	69.66
LLaVA NeXT 7B	53.82	53.36	48.72	58.00
LLaVA NeXT 34B	41.71	35.16	14.55	55.77
LLaVA Interleave 7B	64.68	54.01	76.16	31.86
LLaVA Interleave DPO 7B	59.01	53.98	69.19	38.77
LLaVA OneVision 7B	50.46	50.45	51.01	49.89
LLaVA OneVision 72B	56.28	55.92	51.95	59.88
LLaVA OneVision Chat 7B	50.44	50.26	53.29	47.23
Llama 4 Maverick	68.37	68.32	69.64	66.99
Ministral 3 14B	62.41	62.25	59.73	64.76
Unified Multimodal Understanding and Generation Model				
Janus Pro 7B	60.25	42.59	74.43	10.76

Table 37: Performance of MLLMs on Textual Artifact Recognition (TAR). Metrics include Accuracy (ACC) and F1-scores for presence (Yes) and absence (No) of artifacts.

Model	Manipulation Classification (MC)				
	ACC	Macro-F1	Insert-F1	Remove-F1	Alter-F1
Proprietary MLLMs					
GPT 5.1	**60.07**	**48.34**	34.93	36.65	73.43
GPT 4.1	44.55	40.05	30.30	34.14	55.71
OpenAI o4 Mini High	36.80	35.54	32.36	32.18	42.07
Claude Sonnet 4.5	44.80	39.70	30.77	30.40	57.93
Gemini 3 Pro Preview	48.54	45.42	**41.50**	**37.57**	57.19
Gemini 2.5 Flash	44.80	37.28	26.78	25.11	59.94
Doubao Seed 1.6	50.05	37.08	23.69	20.18	67.37
Doubao Seed 1.6 Thinking	58.47	41.85	26.10	26.00	73.43
Doubao seed 1.6 Flash	35.89	30.96	22.14	20.60	50.15
Doubao 1.5 Thinking Vision Pro	36.65	32.33	31.63	19.08	46.29
Doubao 1.5 Vision Pro	29.17	28.96	28.11	28.22	30.55
Qwen3 VL Plus	59.28	36.89	15.54	20.37	74.75
Qwen VL Max	52.06	36.76	21.31	21.22	67.74
Qwen VL Plus	42.99	32.23	25.85	13.48	57.34
Open-Source MLLMs					
Gemma3 27B	42.49	34.76	29.11	19.77	55.40
Qwen2.5 VL 72B Instruct	51.76	35.96	19.55	20.56	67.77
LLaVA NeXT 7B	59.04	27.01	1.10	4.50	**75.44**
LLaVA NeXT 34B	32.65	27.19	10.59	20.55	50.43
LLaVA Interleave 7B	26.00	26.67	23.65	19.10	37.25
LLaVA Interleave DPO 7B	25.15	24.99	20.26	17.12	37.59
LLaVA OneVision 7B	37.80	31.34	18.40	22.77	52.85
LLaVA OneVision 72B	55.31	39.42	20.16	27.40	70.71
LLaVA OneVision Chat 7B	39.47	32.45	17.20	26.08	54.08
Llama 4 Maverick	54.62	40.18	21.14	28.66	70.73
Ministral 3 14B	53.62	33.89	18.76	12.52	70.39
Unified Multimodal Understanding and Generation Model					
Janus Pro 7B	24.18	24.10	20.24	15.92	36.14

Table 38: Performance of MLLMs on Manipulation Classification (MC). Metrics include Accuracy (ACC) and F1-scores for specific manipulation types (Insertion, Removal, Alteration).

Forgery Scope Discrimination
Category: Micrograph



Question
The task requires the model to act as an image forensics expert and identify the forgery scope of the given academic image panel, strictly following the specified output format.

Options
A) Entire Forgery —— entirely AI-generated
B) Partial Forgery —— partially AI-edited (with localized forgery)
C) No Forgery
D) Not Sure —— too uncertain to judge

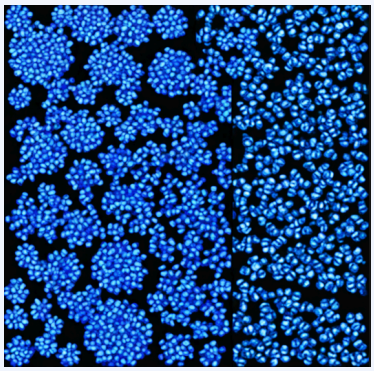
Issue A) Entire Forgery

(GPT 5.1) Model Answer: A) Entire Forgery

Model Explanation
The entire image exhibits uniform AI-art-like textures, exaggerated colors, and repetitive artifacts inconsistent with natural scientific imagery, indicating it is fully AI-generated rather than partially edited.

Figure 36: **Example 1: Forgery Scope Discrimination** correct Example by GPT 5.1. The category of this panel is Micrograph.

Forgery Scope Discrimination
Category: Stained Micrograph



Question
The task requires the model to act as an image forensics expert and identify the forgery scope of the given academic image panel, strictly following the specified output format.

Options
A) Entire Forgery —— entirely AI-generated
B) Partial Forgery —— partially AI-edited (with localized forgery)
C) No Forgery
D) Not Sure —— too uncertain to judge

Issue A) Entire Forgery

(GPT 5.1) Model Answer: D) Not Sure

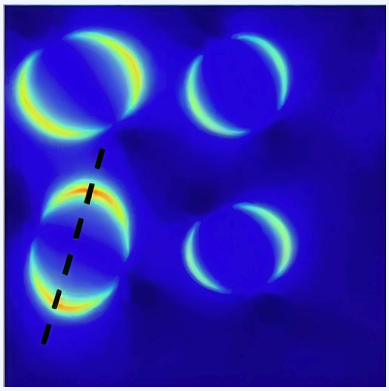
Model Explanation
The image shows plausible microscopic structures without obvious local editing artifacts or clear generative irregularities, but there are also no definitive cues to confirm or rule out full AI generation, making the level of AI intervention uncertain.

Analysis

Error Analysis
In the absence of explicit local editing artifacts, models tend to adopt a conservative abstention strategy, treating visually plausible and globally consistent structures as insufficient evidence for full-image synthesis. This indicates an over-reliance on localized forensic cues and an underutilization of global distributional signals, causing high-realism global forgeries to be underestimated when salient manipulation triggers are absent.

Figure 37: **Example 2: Forgery Scope Discrimination** error Example by GPT 5.1. The category of this panel is Stained Micrograph.

Forgery Scope Discrimination
Category: Diagram



Question
The task requires the model to act as an image forensics expert and identify the forgery scope of the given academic image panel, strictly following the specified output format.

Options
A) Entire Forgery —— entirely AI-generated
B) Partial Forgery —— partially AI-edited (with localized forgery)
C) No Forgery
D) Not Sure —— too uncertain to judge

Issue A) Entire Forgery

(GPT 5.1) Model Answer: C) No Forgery ❌

Model Explanation
The image appears to be a standard scientific heatmap/visualization with smooth gradients and clean geometric shapes, without visible artifacts or inconsistencies suggestive of AI-generated or edited regions.

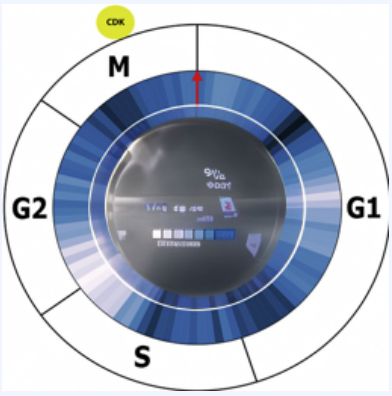
Analysis

Error Analysis
The model interprets the smooth gradients, clean geometric structures, and overall visual coherence as evidence of authenticity, conflating stylistic regularity with physical plausibility. This reflects a bias toward treating well-structured heatmap-like renderings as inherently trustworthy, while overlooking global generative cues such as overly idealized symmetry, uniform smoothness, and the absence of acquisition-related noise.

Figure 38: **Example 3: Forgery Scope Discrimination** error Example by GPT 5.1. The category of this panel is Diagram.

Forgery Scope Discrimination

Category: Chart



Question

The task requires the model to act as an image forensics expert and identify the forgery scope of the given academic image panel, strictly following the specified output format.

Options

A) Entire Forgery —— entirely AI-generated
B) Partial Forgery —— partially AI-edited (with localized forgery)
C) No Forgery
D) Not Sure —— too uncertain to judge

Issue B) Partial Forgery

(GPT 5.1) Model Answer: C) No Forgery ❌

Model Explanation

The chart and embedded image elements appear consistent with standard scientific figure design, with clean boundaries, regular text, and no obvious AI-generated artifacts or irregular patterns.

Analysis

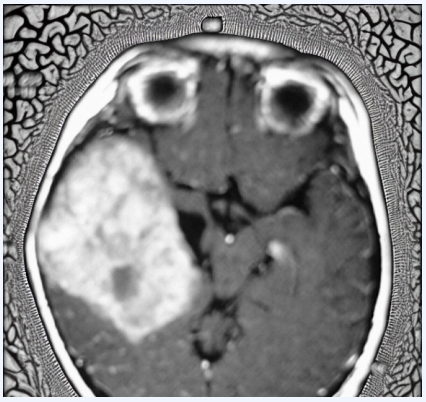
Error Analysis

The model interprets the clean layout, regular annotations, and well-aligned geometric components as indicators of authenticity, overlooking subtle inconsistencies introduced by localized manipulation. This reflects an over-reliance on global visual coherence and design regularity, causing localized forgeries embedded within otherwise well-structured scientific figures to be missed.

Figure 39: **Example 4: Forgery Scope Discrimination** error Example by GPT 5.1. The category of this panel is Chart.

Forgery Scope Discrimination

Category: Medical Imaging



Question

The task requires the model to act as an image forensics expert and identify the forgery scope of the given academic image panel, strictly following the specified output format.

Options

A) Entire Forgery —— entirely AI-generated
B) Partial Forgery —— partially AI-edited (with localized forgery)
C) No Forgery
D) Not Sure —— too uncertain to judge

Issue B) Partial Forgery

(GPT 5.1) Model Answer: A) Entire Forgery

Model Explanation

The entire image shows uniform synthetic textures, anatomically inconsistent brain structures, and repetitive decorative patterns around the border, indicating it is fully AI-generated rather than a localized edit of a real scan.

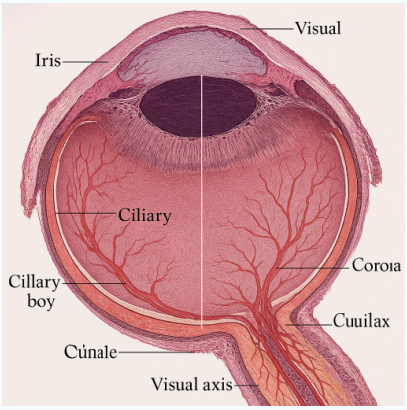
Analysis

Error Analysis

In this case, the model overemphasizes global visual irregularities—such as atypical textures, perceived anatomical inconsistencies, and decorative border patterns—and incorrectly attributes them to full-image synthesis. This leads to a misinterpretation of localized manipulations as evidence of global generation, blurring the distinction between partial edits and entirely synthetic content.

Figure 40: **Example 5: Forgery Scope Discrimination** error Example by GPT 5.1. The category of this panel is Medical Imaging.

Textual **Artifact Recognition**
Category: Diagram



Question
The task requires the model to assess whether any embedded textual elements in the given academic image exhibit signs of AI generation or AI editing, based solely on visual artifacts, structural irregularities, or semantic mismatches with the surrounding scientific content. The model should return a binary judgment (Yes/No), strictly following the specified output format.

Options
A) Yes
B) No

Issue A) Yes

(GPT 5.1) Model Answer: A) Yes

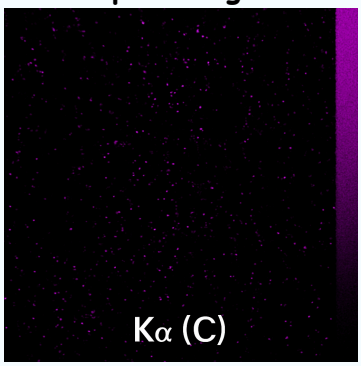
Model Explanation
Several labels (e.g., for anatomical parts of the eye) are misspelled or use nonsensical terms that do not correspond to established ocular anatomy, indicating semantic inconsistencies typical of AI-generated or altered text. The typography and label placement appear regular, but the incorrect terminology suggests the text is not authentic scientific labeling.

Figure 41: **Example 6: Textual Artifact Recognition** correct Example by GPT 5.1. The category of this panel is Diagram.

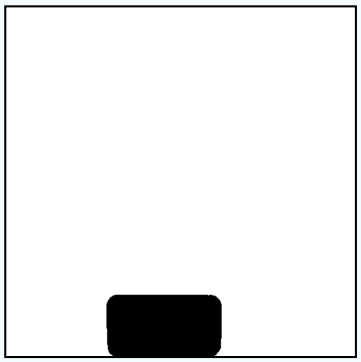
Textual Artifact Recognition

Category: Stained Micrograph

Input Image



GT Mask



Question

The task requires the model to assess whether any embedded textual elements in the given academic image exhibit signs of AI generation or AI editing, based solely on visual artifacts, structural irregularities, or semantic mismatches with the surrounding scientific content. The model should return a binary judgment (Yes/No), strictly following the specified output format.

Options

A) Yes
B) No

Issue A) Yes

(GPT 5.1) Model Answer: B) No ❌

Model Explanation

The embedded text appears clean, well-aligned, and stylistically consistent with standard scientific figure labeling, with no visual artifacts or semantic mismatches relative to the elemental map image content.

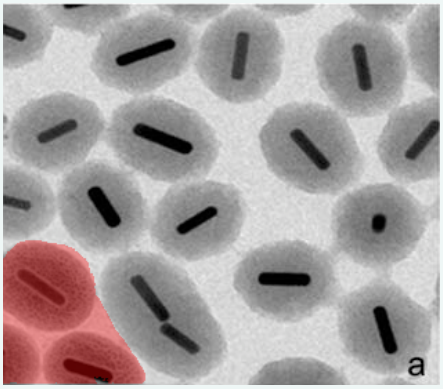
Analysis

Error Analysis

The failure stems from an over-reliance on semantic plausibility and visual neatness of the label ("Kα (C)"), while neglecting localized text-region artifacts. Specifically, the embedded text exhibits abnormal contrast and blending behavior against the dark elemental map background, with unnatural edge smoothness and luminance consistency that deviate from typical staining or elemental mapping annotations. The model's explanation focuses on global alignment and stylistic cleanliness, overlooking text–background inconsistency cues that are critical for texture-level forgery detection in scientific micrographs.

Figure 42: **Example 7: Textual Artifact Recognition** error example by GPT 5.1. The panel belongs to the Stained Micrograph category and contains the localized forgery generated via *Targeted Region Editing*.

Manipulation Classification
Category: Micrograph



Question
The task requires the model to identify the manipulation type of an AI-edited local region based on observable visual changes and structural inconsistencies, classifying the modification as Remove, Insert, or Alter, strictly following the specified output format.

Original Description
...high-magnification electron microscope image showing what appears to be bacterial cells or possibly microscopic particles. The image displays multiple oval or round-shaped structures with light gray borders and darker gray interiors. Within each of these oval structures is a distinct dark rod-shaped element...

Options
A) Insert
B) Remove
C) Alter
D) Not Sure

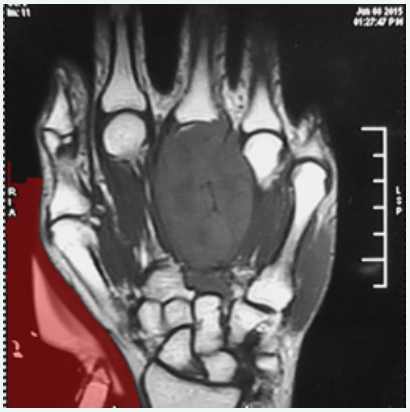
Issue C) Alter

(GPT 5.1) Model Answer: C) Alter

Model Explanation
The red-highlighted area still contains similar cell-like structures but with altered shape/contrast compared with the surrounding ones, indicating modification of existing objects rather than adding new ones or removing them.

Figure 43: **Example 8: Manipulation Classification** correct Example by GPT 5.1. The panel belongs to the Micrograph category and contains the localized forgery generated via *Targeted Region Editing*.

Manipulation Classification
Category: Medical Imaging



Question
The task requires the model to identify the manipulation type of an AI-edited local region based on observable visual changes and structural inconsistencies, classifying the modification as Remove, Insert, or Alter, strictly following the specified output format.

Original Description
...MRI scan of a hand in coronal (frontal) view. The scan clearly displays the bony structures of the hand including multiple metacarpal bones, phalanges, and carpal bones at the wrist. There is a prominent dark (hypointense) mass visible in the central palm region, likely representing a soft tissue mass or cyst. The mass appears well-circumscribed and is pushing adjacent structures aside...

Options
A) Insert
B) Remove
C) Alter
D) Not Sure

Issue A) Insert

(GPT 5.1) Model Answer: B) Remove ✗

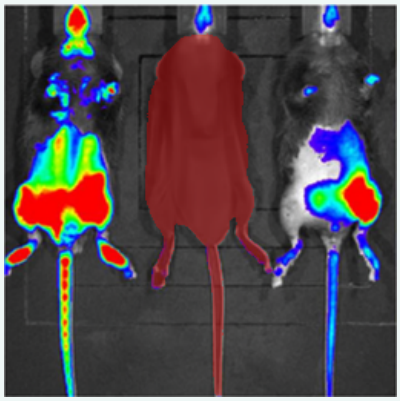
Model Explanation
The red-highlighted region overlays and occludes part of the original MRI anatomy, effectively removing existing image content rather than adding or modifying anatomical structures.

Analysis

Error Analysis
This error arises from a misinterpretation of occlusion as deletion, where the model assumes that the red-highlighted region merely covers or removes existing anatomical structures. In reality, the forged region introduces a new soft-tissue–like mass that was not present in the original image, thereby constituting an insertion rather than a removal. The model's explanation focuses on the apparent obstruction of underlying anatomy, overlooking the novel semantic content and abnormal anatomical presence introduced by the edited region. This case highlights the model's difficulty in distinguishing inserted foreground structures from removals in medical imaging scenarios.

Figure 44: **Example 9: Manipulation Classification** error example by GPT 5.1. The panel belongs to the Medical Imaging category and contains the localized forgery generated via *Targeted Region Restoration*.

Manipulation Classification
Category: Medical Imaging



Question
The task requires the model to identify the manipulation type of an AI-edited local region based on observable visual changes and structural inconsistencies, classifying the modification as Remove, Insert, or Alter, strictly following the specified output format.

Original Description
...fluorescence imaging study involving three murine subjects, aligned in a horizontal arrangement. Each specimen exhibits varying intensity of fluorescent signals, depicted through a color spectrum ranging from red (indicating high signal intensity) to blue (indicating low signal intensity), the leftmost mouse displays pronounced red fluorescence across the abdomen and proximal limbs, suggesting a significant accumulation of the fluorescent marker...

Options
A) Insert
B) Remove
C) Alter
D) Not Sure

Issue B) Remove

(GPT 5.1) Model Answer: A) Alter ❌

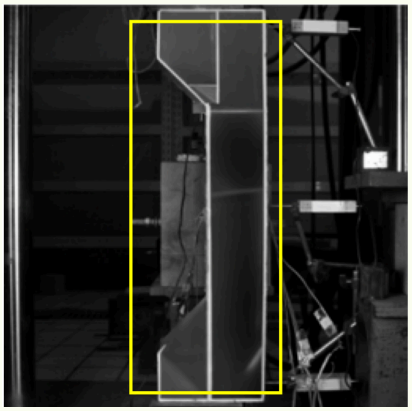
Model Explanation
The central mouse, originally present as a foreground subject with distinct fluorescence patterns, has been visually modified and overpainted in the highlighted region, changing its appearance rather than removing it or adding a new object.

Analysis

Error Analysis
This error stems from confusing content erasure with appearance modification. In the red-highlighted region, the central mouse subject has been entirely eliminated, leaving a visually plausible but semantically empty or background-like region, which constitutes a removal rather than an alteration. However, the model interprets the uniform overlay and smooth intensity transition as evidence of repainting or overpainting, leading it to infer an alteration of existing content. This case reveals the model's tendency to over-attribute surface-level visual continuity to modification, while failing to recognize the absence of an expected foreground object.

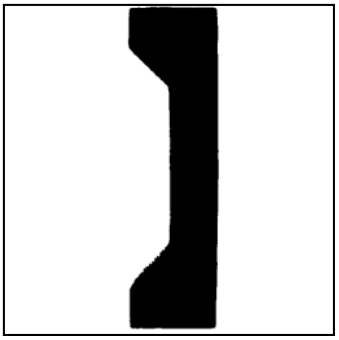
Figure 45: **Example 10: Manipulation Classification** error example by GPT 5.1. The panel belongs to the Medical Imaging category and contains the localized forgery generated via *Targeted Region Restoration*.

Tampering Pinpointing
Category: Physical Object



Question
The task requires the model to precisely localize all AI-edited regions in the given image by outputting bounding boxes that tightly enclose each forged area, strictly following the specified output format.

Forged Region (GT Mask)

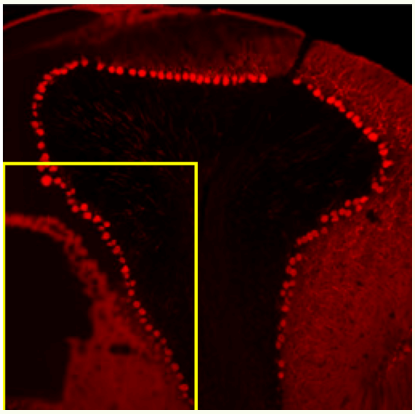


(GPT 5.1) Model Answer: [[80, 10, 175, 245]]
The predicted region is highlighted with the **yellow bounding box** in the image.

Model Explanation
The central tall, glowing structural shape has unnaturally sharp uniform borders and smooth interior gradients that differ from the surrounding photographic background, indicating it was digitally inserted or modified.

Figure 46: **Example 11: Tampering Pinpointing** correct Example by GPT 5.1. The panel belongs to the Physical Object category and contains the localized forgery generated via *Targeted Region Restoration*.

Tampering Pinpointing
Category: Stained Micrograph



Question
The task requires the model to precisely localize all AI-edited regions in the given image by outputting bounding boxes that tightly enclose each forged area, strictly following the specified output format.

Forged Region (GT Mask)

(GPT 5.1) Model Answer: [[0, 100, 120, 255]]
The predicted region is highlighted with the **yellow bounding box** in the image.

Model Explanation
The lower-left region has an unnaturally sharp rectangular boundary and a flat dark fill that contrasts with the organic curved structures and textured signal of the rest of the image, indicating likely AI editing or compositing.

Analysis

Error Analysis
The model fails to accurately localize the forged region, producing a single coarse bounding box that only partially overlaps with the ground-truth mask. This error is primarily caused by an overly conservative and axis-aligned localization strategy, where the model prioritizes a rectangular region with high visual contrast rather than tightly following the irregular forged area. Although the model correctly detects the presence of abnormal visual cues (e.g., sharp boundaries and flat dark fills), it lacks the ability to adapt bounding box geometry to complex biological structures, resulting in under-coverage of curved and branching regions. This case highlights a key limitation of current MLLMs in fine-grained spatial grounding for non-rigid anatomical patterns commonly found in stained micrographs.

Figure 47: **Example 12: Tampering Pinpointing** error Example by GPT 5.1. The panel belongs to the Stained Micrograph category and contains the localized forgery generated via *Targeted Region Restoration*.