

THEMIS: Towards Holistic Evaluation of MLLMs for Scientific Paper Fraud Forensics

Tzu-Yen Ma*, Bo Zhang*, Zichen Tang, Junpeng Ding, Haolin Tian, Yuanze Li, Zhuodi Hao, Zixin Ding, Zirui Wang, Xinyu Yu, Shiyao Peng, Yizhuo Zhao, Ruomeng Jiang, Yiling Huang, Peizhi Zhao, Jiayuan Chen, Weisheng Tan, Haocheng Gao, Yang Liu, Jiacheng Liu, Zhongjun Yang, Jiayu Huang, Haihong E†

* Equal Contribution † Corresponding Author

Reasoning Lab, Beijing University of Posts and Telecommunications

**BUPT
Reasoning Lab**



Motivation & Contribution

Real-World Scenarios & Complexity

Stained Micrograph

Diagram

Micrograph

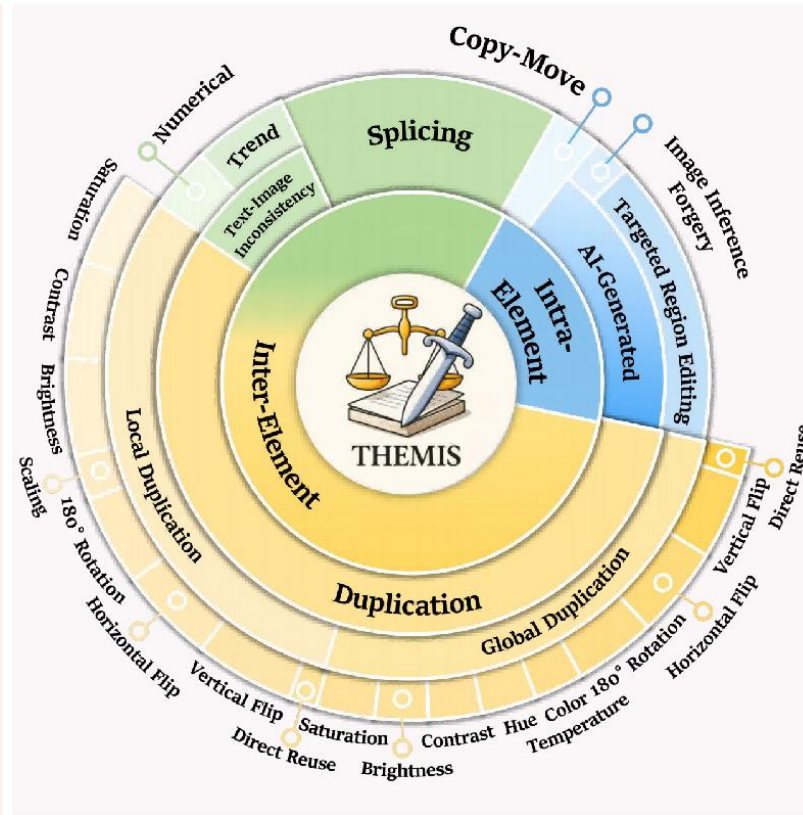
Physical Object

Medical Imaging

Chart

Others

Fraud-Type Diversity & Granularity



Multi-Dimensional Capability Evaluation

■ Splicing
 ■ Copy-Move
 ■ AI-Generated
 ■ Duplication
 ■ Text-Image Inconsistency

Expert Knowledge Utilization

Comparative Reasoning

Visual Recognition

Region Localization

Spatial Reasoning

Leading MLLMs

GPT Gemini

LLaVA Gemma

Doubao Qwen Llama

Metrics

ACC IoU

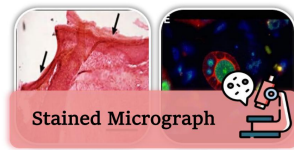
F1 EM

SPC ...

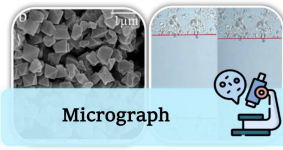
Benchmark



Medical Imaging



Stained Micrograph



Micrograph



Physical Object

Medical Imaging

Physical Object

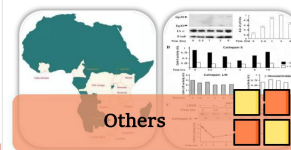
Diagram

Stained Micrograph

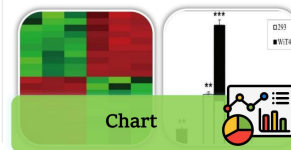
Micrograph

Others

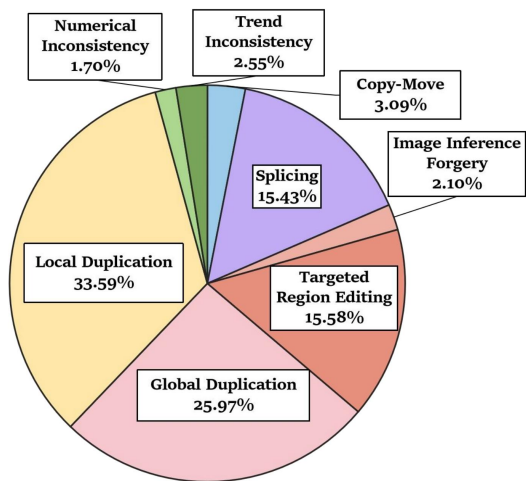
Chart



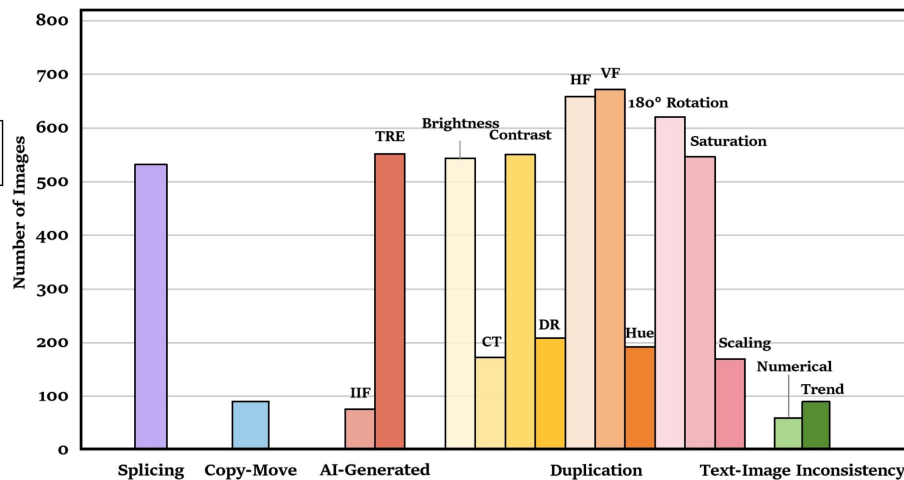
Others



Chart



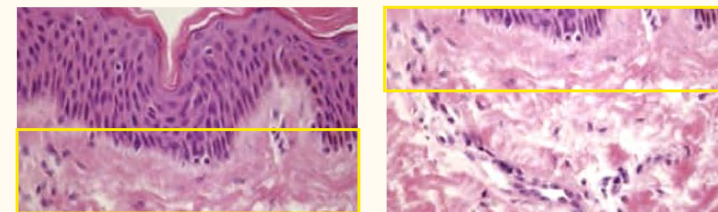
(a) Distribution of fraud methods.



(b) Distribution of manipulation operations (synthetic data).



Duplication: Stained Micrograph



Question

In this task, the model is required to perform cross-panel duplication forensics. Each input consists of multiple sub-figures (panels) that originate from the same scientific figure. The model must determine whether any panels are duplicated, and if so, identify the specific type(s) of duplication.

Options

- A) Direct: near-identical content across panels without geometric transform; many keypoints align at similar scale/position.
- B) Scaling: content in one panel appears in another after zoom-in/zoom-out.
- C) Rotated: one panel is a rotated version of another (e.g., 90°/180°).
- D) Flipped: one panel is a mirrored (horizontal/vertical) version of another.
- E) Parameter-Modified: duplication exists but with image-level adjustments (e.g., brightness, contrast, and color).
- F) No Duplication: there is no convincing duplication across panels.
- G) Not Sure: evidence is insufficient or ambiguous.

Issue E) Parameter-Modified



(GPT-5) Model Answer: B) Scaling & E) Parameter-modified X

Model Explanation

- The two panels share the same histologic landmarks (matching curved "Y"-shaped pale bands and clusters of small circular lumina) at the same relative arrangement, but the right panel appears slightly zoomed in and shows stronger exposure/contrast with a magenta shift. This indicates duplication with scaling and image-level adjustments.

Analysis

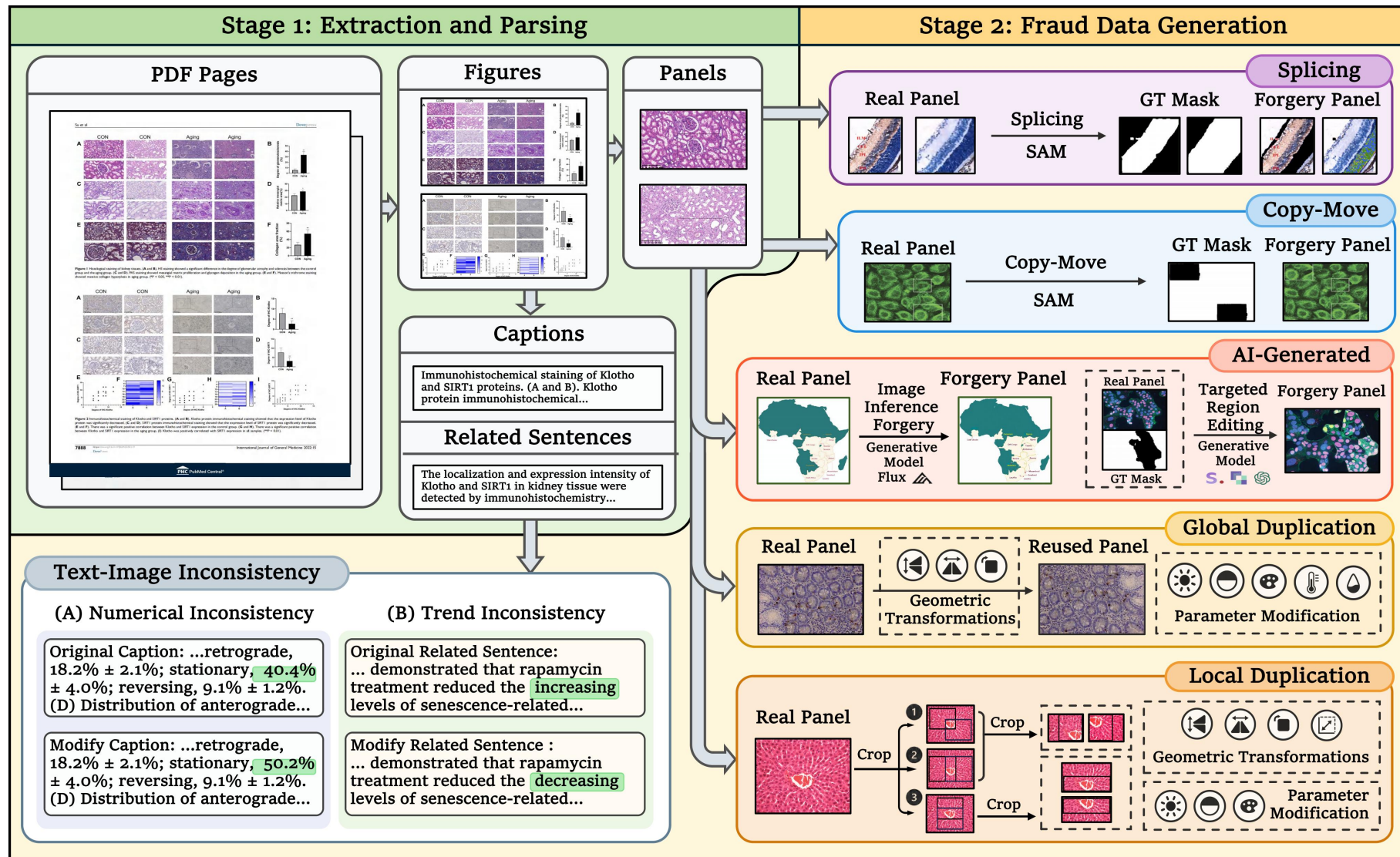
Error Analysis

- The duplicated region marked by the yellow boxes occupies different proportions in the two panels, appearing more compact in one image and more dispersed in the other. The model incorrectly attributed this change in proportion to scaling rather than to differences in background tissue that had been cropped or obscured.
- Seeking a more complex explanation, it inferred a geometric transformation because the duplicated fibers and nuclear orientations remained consistent, thus mistakenly applying a "scaled" interpretation instead of the simpler case of direct duplication with exposure adjustment.

The Lack of Five Core Reasoning Capabilities

- Visual Recognition, Spatial Reasoning, Comparative Reasoning

Dataset Curation



Task Design

Single-Mode Forgery Identification and Localization

Identification Score:

$$ACC_{sm} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i)$$

number of samples $\rightarrow N$

predicted forgery type $\rightarrow \hat{y}_i$

ground-truth label $\rightarrow y_i$

Localization Score:

$$IoU = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|}$$

ground-truth mask $\rightarrow M$

predicted mask $\rightarrow \hat{M}$

AI-Generated	Splicing	Copy-Move
<p>Task Definition Panel wholly or partially AI-generated; cues: synthetic textures, implausible details, inconsistent semantics.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition Different sources of content combined in panel; cues: unnatural seams, boundary mismatches, inconsistent textures.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition A region copied within the panel; cues: duplicated patterns/objects with same noise/texture but shifted/rotated/scaled.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>
<p>Task Definition Repeated content reused across panels; cues: identical or near-identical regions, consistent textures/patterns, matching structures possibly under geometric or parametric transformations.</p> <p>Question Identify every duplication type that appears.</p> <p>A) Direct B) Scaling C) Rotated D) Flipped E) Parameter-Modified F) No Duplication G) Not Sure</p>	<p>Task Definition Determine whether the accompanying text is consistent with the figure by identifying if contradictions arise in numerical values or in described trends.</p> <p>Question Detect whether the caption and related sentences are inconsistent with the corresponding figure and localize ALL problematic parts.</p> <p>Caption: Tissue penetration of Humabodies with and without albumin ... in the maximum signal (p = 0.85). *, p < 0.05.</p> <p>Related Sentence: Given the importance of tissue ... lowest molecular weight (~30 kDa) distributed the most evenly.</p> <p>A) Numerical Inconsistent B) Trend Inconsistent C) Consistent D) Not Sure</p> <p><SENTENCE>...signal (p = 0.85)...</SENTENCE></p>	

Task Design

Composite Manipulation Operations Identification

Identification Score:

predicted set of operations → ground-truth set of operations

$$\text{Precision} = \frac{|\hat{S} \cap S|}{|\hat{S}|}, \quad \text{Recall} = \frac{|\hat{S} \cap S|}{|S|}$$

$$F1_{\text{set}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

AI-Generated	Splicing	Copy-Move
<p>Task Definition Panel wholly or partially AI-generated; cues: synthetic textures, implausible details, inconsistent semantics.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition Different sources of content combined in panel; cues: unnatural seams, boundary mismatches, inconsistent textures.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition A region copied within the panel; cues: duplicated patterns/objects with same noise/texture but shifted/rotated/scaled.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>
<p>Duplication</p> <p>Task Definition Repeated content reused across panels; cues: identical or near-identical regions, consistent textures/patterns, matching structures possibly under geometric or parametric transformations.</p> <p>Question Identify every duplication type that appears.</p> <p>A) Direct B) Scaling C) Rotated D) Flipped E) Parameter-Modified F) No Duplication G) Not Sure</p>		<p>Text-Image Inconsistency</p> <p>Task Definition Determine whether the accompanying text is consistent with the figure by identifying if contradictions arise in numerical values or in described trends.</p> <p>Question Detect whether the caption and related sentences are inconsistent with the corresponding figure and localize ALL problematic parts.</p> <p>Caption: Tissue penetration of Humabodies with and without albumin ... in the maximum signal (p = 0.85). *, p < 0.05.</p> <p>Related Sentence: Given the importance of tissue ... lowest molecular weight (~30 kDa) distributed the most evenly.</p> <p>A) Numerical Inconsistent B) Trend Inconsistent C) Consistent D) Not Sure</p> <p><SENTENCE>...signal (p = 0.85)...</SENTENCE></p>

Task Design

Cross-Modal Inconsistency Identification and Localization

Identification Score:
predicted forgery type

number of samples

$$ACC_{sm} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i)$$

ground-truth label

Localization Score:

ground-truth text span predicted text span

$$F1_{text} = \frac{2 \cdot |\hat{T} \cap T|}{|\hat{T}| + |T|}$$

length of the span

AI-Generated	Splicing	Copy-Move
<p>Task Definition Panel wholly or partially AI-generated; cues: synthetic textures, implausible details, inconsistent semantics.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition Different sources of content combined in panel; cues: unnatural seams, boundary mismatches, inconsistent textures.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>	<p>Task Definition A region copied within the panel; cues: duplicated patterns/objects with same noise/texture but shifted/rotated/scaled.</p> <p>Question Detect image forgery type and localize forged regions.</p> <p>A) Splicing B) Copy-Move C) AI-Generated D) No Forgery E) Not Sure</p> <p><MASK>comma-separated region numbers</MASK></p>
<p>Task Definition Repeated content reused across panels; cues: identical or near-identical regions, consistent textures/patterns, matching structures possibly under geometric or parametric transformations.</p> <p>Question Identify every duplication type that appears.</p> <p>A) Direct B) Scaling C) Rotated D) Flipped E) Parameter-Modified F) No Duplication G) Not Sure</p>	<p>Task Definition Determine whether the accompanying text is consistent with the figure by identifying if contradictions arise in numerical values or in described trends.</p> <p>Question Detect whether the caption and related sentences are inconsistent with the corresponding figure and localize ALL problematic parts.</p> <p>Caption: Tissue penetration of Humabodies with and without albumin ... in the maximum signal (p = 0.85). *, p < 0.05.</p> <p>Related Sentence: Given the importance of tissue ... lowest molecular weight (~30 kDa) distributed the most evenly.</p> <p>A) Numerical Inconsistent B) Trend Inconsistent C) Consistent D) Not Sure</p> <p><SENTENCE>...signal (p = 0.85)...</SENTENCE></p>	

Task Design

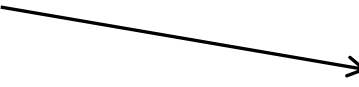
Balanced Robutness Index (BRI)

$$\mathbf{BRI}_i = \mu_i - \lambda \cdot \Delta_i,$$

mean performance $\mu_i = \frac{1}{5} \sum_{j=1}^5 s_{i,j},$

stability penalty $\Delta_i = \max_j(s_{i,j}) - \min_j(s_{i,j}).$

raw score of model i on task demension j


$$s_{i,j} = \frac{R_{i,j} - \min_{k \in \{1 \dots M\}} R_{k,j}}{\max_{k \in \{1 \dots M\}} R_{k,j} - \min_{k \in \{1 \dots M\}} R_{k,j}}$$

$$\mathbf{s}_i = [s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4}, s_{i,5}]$$

Main Results

- Systemic Reasoning Gap: SOTA GPT-5 plateaus at **56.15% BRI**.
- Localization Struggle: Sharp **55% drop** in grounding accuracy for top models.
- Competitive open models: Qwen-2.5-VL rivals proprietary MLLMs in specialized tasks.

Model	Single-Mode Forgery Identification (Id Score)				Single-Mode Forgery Localization (Loc Score)				Composite Manipulation Operations Identification (Id Score)	Cross-Modal Inconsistency Identification & Localization		BRI
	SPL (807)	CM (242)	AIG (897)	Avg.	SPL (807)	CM (242)	AIG (897)	Avg.	DUP (2,079)	TII (300) Id Score	Loc Score	
<i>Proprietary MLLMs</i>												
GPT-5	43.51	72.73	44.26	53.50	16.67	36.41	19.33	24.14	33.32	60.67	27.44	56.15
OpenAI o4-mini-high	41.49	77.89	35.67	51.68	10.44	29.78	19.79	20.00	30.34	66.33	32.22	52.34
Qwen-VL-Max	30.37	87.40	35.43	51.07	40.07	48.34	35.85	41.42	23.33	56.00	15.36	49.83
Gemini 2.5 Flash	63.39	67.56	35.45	55.47	56.72	46.98	38.87	47.52	24.96	36.33	28.24	44.70
Doubao-Seed-1.6-thinking	35.67	74.17	36.57	48.80	12.09	13.58	15.19	13.62	20.22	60.00	31.71	37.14
Doubao-Seed-1.6-vision	20.84	61.78	27.54	36.72	31.42	26.97	29.90	29.43	45.13	37.00	30.43	33.47
Gemini 2.5 Pro	30.60	47.46	14.90	30.99	46.65	46.61	47.20	46.82	21.49	44.67	37.31	31.97
GLM-4.5V	29.23	58.43	54.51	47.39	8.54	22.23	10.63	13.80	21.84	53.67	22.69	31.57
Claude Sonnet 4.5	29.71	75.66	30.43	45.27	14.95	44.12	20.98	26.68	21.86	35.67	27.42	29.96
<i>Open-Source MLLMs</i>												
Qwen2.5-VL-72B	36.40	77.27	51.06	54.91	51.66	55.16	35.57	47.46	16.75	61.33	12.32	47.16
InternVL3.5-8B	30.97	58.42	67.00	52.13	43.38	40.24	33.19	38.94	33.28	55.00	4.78	38.73
Llama 4 Maverick	23.37	51.24	58.19	44.27	17.97	28.50	19.71	22.06	19.01	54.00	16.08	34.78
LLaVA-Interleave-7B	41.80	47.55	46.93	45.43	35.97	25.22	32.06	31.08	8.01	50.00	12.57	23.59
LLaVA-NeXT-34B	32.00	84.00	34.00	50.00	50.70	45.25	34.01	43.32	10.73	41.33	4.28	18.40
Qwen2.5-VL-32B	30.31	57.48	39.24	42.34	5.91	18.93	7.45	10.76	15.26	58.33	17.94	18.22
Gemma 3 27B	25.39	28.87	34.23	29.50	27.78	32.80	26.44	29.01	12.20	31.67	21.41	9.59

Real World Data Evaluation

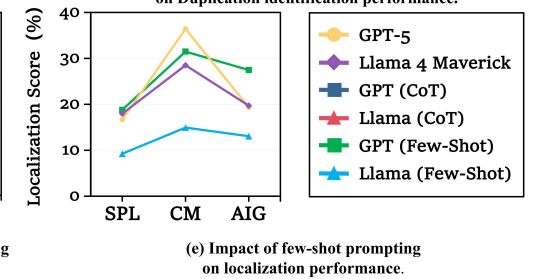
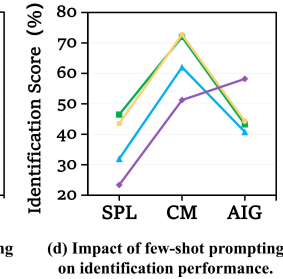
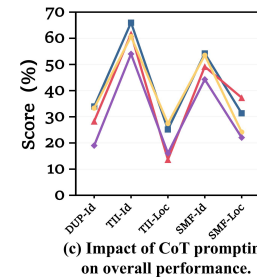
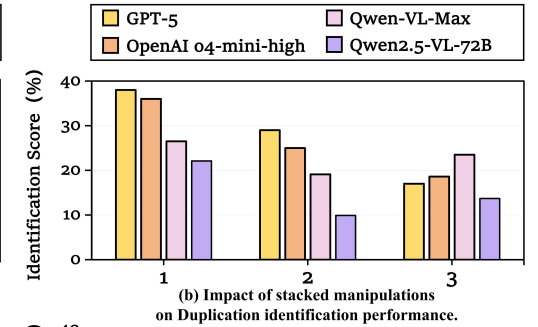
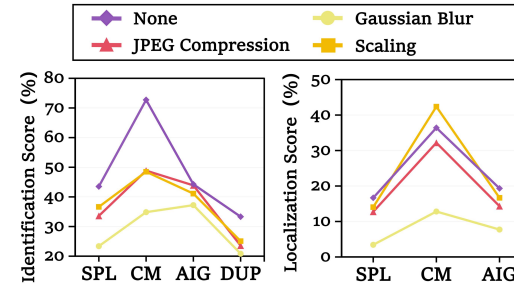
- **Fidelity:** Synthetic data exerts **comparable reasoning pressure** to real cases.
- **Complexity:** Higher difficulty in **Duplication** via compounded transformations.
- **Reliability:** Synthetic performance **highly correlates** with real-world capability.

Model	Data Type	Splicing		Copy-Move		Duplication
		Id Score	Loc Score	Id Score	Loc Score	Id Score
<i>Proprietary MLLMs</i>						
GPT-5	Synthetic	43.51	16.67	72.73	36.41	33.32
	Real	27.27	23.96	60.00	33.39	23.22
OpenAI o4-mini-high	Synthetic	41.49	10.44	77.89	29.78	30.34
	Real	9.09	11.99	60.00	31.35	36.04
Qwen-VL-Max	Synthetic	30.37	40.07	87.40	48.34	23.33
	Real	0.00	15.15	13.33	3.70	20.95
Gemini 2.5 Flash	Synthetic	63.39	56.72	67.56	46.98	24.96
	Real	9.09	38.55	80.00	47.58	49.58
<i>Open-Source MLLMs</i>						
Qwen2.5-VL-72B	Synthetic	36.40	51.66	77.27	55.16	16.75
	Real	0.00	54.54	100.00	36.27	22.39
Llama 4 Maverick	Synthetic	23.37	17.97	51.24	28.50	19.01
	Real	0.00	8.66	6.67	2.22	39.75
Gemma 3 27B	Synthetic	25.39	27.78	28.87	32.80	12.20
	Real	0.00	15.25	40.00	17.04	42.18
Avg.	Synthetic	37.70	31.62	66.14	39.71	22.84
	Real	6.49	24.01	51.43	24.51	33.44

Finegrained Results

Model	Post-Processing	SPL		CM		AIG		DUP
		Id Score	Loc Score	Id Score	Loc Score	Id Score	Loc Score	Id Score
GPT-5	None	43.51	16.67	72.73	36.41	44.26	19.33	33.32
	Gauss	23.37 (-20.14)	3.43 (-13.24)	34.85 (-37.88)	12.81 (-23.60)	37.22 (-7.04)	7.73 (-11.60)	20.84 (-12.48)
	JPEG	33.59 (-9.92)	12.74 (-3.93)	48.86 (-23.87)	32.19 (-4.22)	43.89 (-0.37)	14.32 (-5.01)	23.56 (-9.76)
	Scaling	36.63 (-6.88)	14.05 (-2.62)	48.48 (-24.25)	42.37 (+5.96)	41.11 (-3.15)	16.67 (-2.66)	25.08 (-8.24)
Gemini	None	63.39	56.72	67.56	46.98	35.45	38.87	24.96
	Gauss	48.13 (-15.26)	33.63 (-23.09)	50.38 (-17.18)	39.74 (-7.24)	31.11 (-4.34)	23.20 (-15.67)	25.84 (+0.88)
	JPEG	64.82 (+1.43)	60.01 (+3.29)	54.93 (-12.63)	42.46 (-4.52)	33.90 (-1.55)	39.93 (+1.06)	25.56 (+0.60)
	Scaling	61.37 (-2.02)	57.46 (+0.74)	46.97 (-20.59)	43.86 (-3.12)	34.44 (-1.01)	40.75 (+1.88)	25.31 (+0.35)

Model	Prompting Strategy	SPL		CM		AIG		DUP	TII	
		Id Score	Loc Score	Id Score	Loc Score	Id Score	Loc Score	Id Score	Id Score	Loc Score
GPT-5	None	43.51	16.67	72.73	36.41	44.26	19.33	33.32	60.67	27.44
	CoT	44.85 (+1.34)	25.16 (+8.49)	74.46 (+1.73)	38.74 (+2.33)	43.36 (-0.90)	30.10 (+10.77)	33.95 (+0.63)	66.00 (+5.33)	25.16 (-2.28)
	Few-Shot	46.47 (+2.96)	18.80 (+2.13)	71.97 (-0.76)	31.47 (-4.94)	43.26 (-1.00)	27.46 (+8.13)	-	-	-
Llama	None	23.37	17.97	51.24	28.50	58.19	19.71	19.01	54.00	16.08
	CoT	28.24 (+4.87)	49.34 (+31.37)	60.21 (+8.97)	28.47 (-0.03)	59.08 (+0.89)	34.06 (+14.35)	28.29 (+9.28)	61.66 (+7.66)	13.64 (-2.44)
	Few-Shot	31.97 (+8.60)	9.25 (-8.72)	62.03 (+10.79)	14.93 (-13.57)	40.80 (-17.39)	13.06 (-6.65)	-	-	-



Thank you for attention and
see you at the conference !



Our research group



Openreview



Project