



Differences That Matter: Auditing Models for Capability Gap Discovery and Rectification

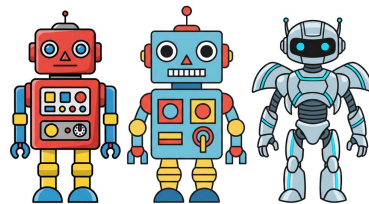
Qihao Liu^{1,2}, Chengzhi Mao¹, Yaojie Liu¹, Alan Yuille², Wen-Sheng Chu²

¹ Google ² Johns Hopkins University



Motivation

Despite steady improvements on public benchmarks, selecting the proper models for real-world deployment remains challenging, as conventional evaluations often obscure how models truly differ.

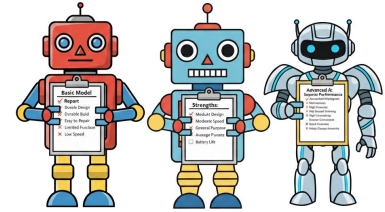


We may know who tops the leaderboard, but we still can't answer:

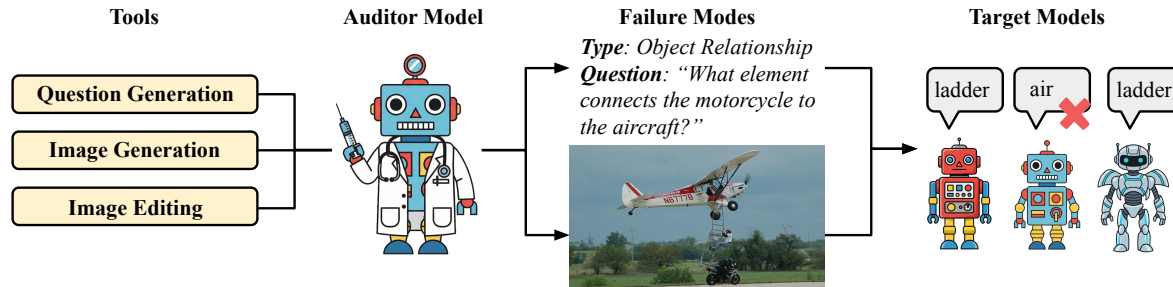
- “What’s different, and why?”
- “Where (and how) does it break?”
- “Is 3B enough for my task?”
- “What’s still missing from the latest model?”
-

Motivation

- Limitations of current evaluation methods:
 - Closed-set: fixed scope, inevitable blind spots.
 - Score-compressed: They reduce complex behavior to single metrics, hiding slice-level shifts.

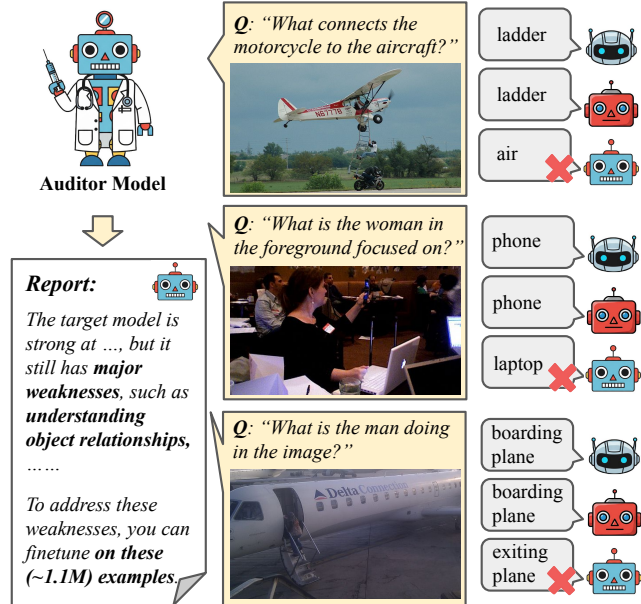


- Our solution: Model Auditing



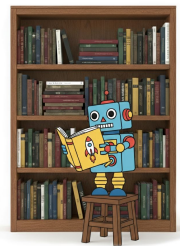
Motivation

- We propose AuditDM, which
 - (1) systematically discovers capability gaps and produces interpretable weakness summaries,

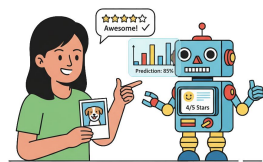


Motivation

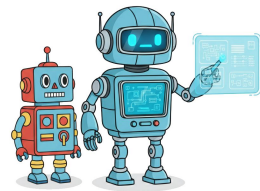
- We propose AuditDM, which
 - (1) systematically discovers capability gaps and produces interpretable weakness summaries, and
 - (2) delivers actionable feedback that guides fixes and model improvement.



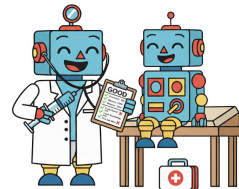
(a) Learning from Data



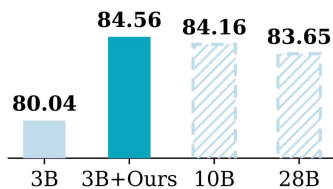
(b) Learning from Feedback



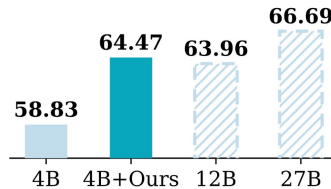
(c) Learning from Other Models



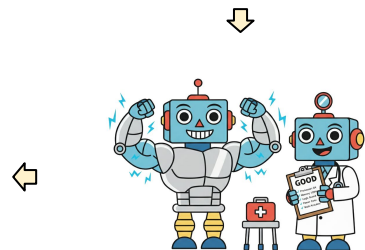
(d) Auditing (Ours)



Improving PaliGemma2



Improving Gemma3



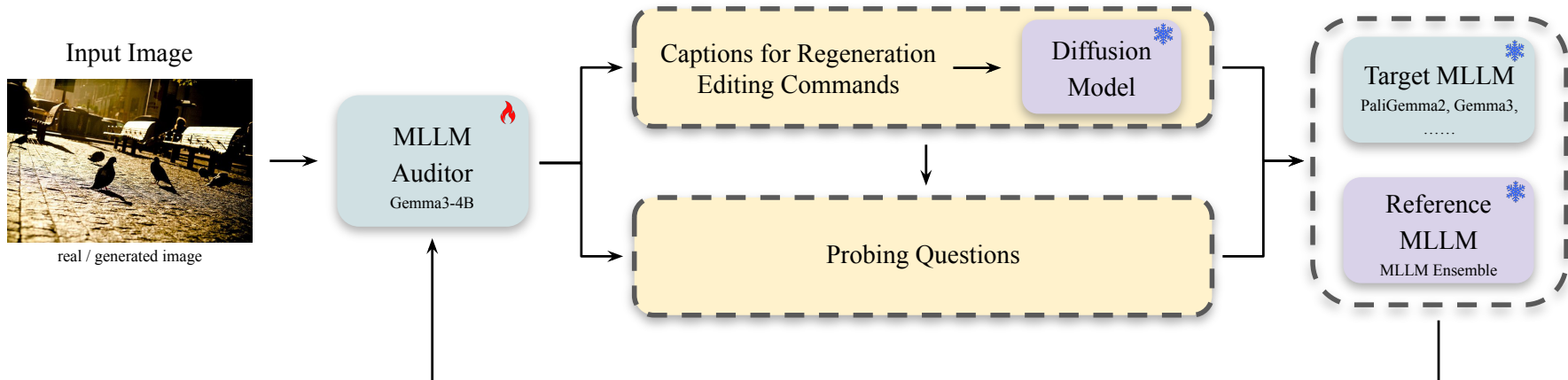
Self-Improvement with Auditor

Related work

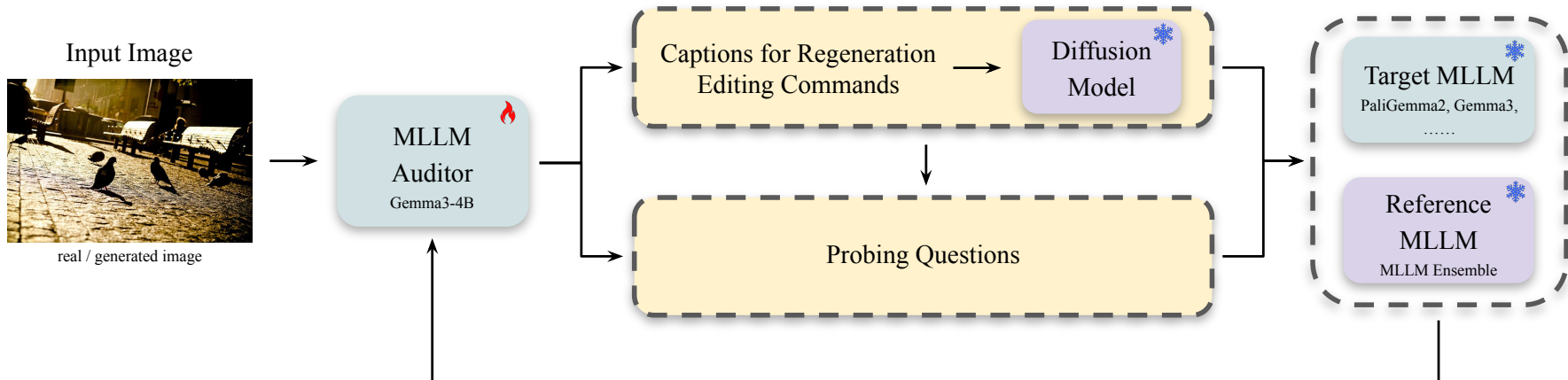
Comparison of related work on finding and fixing MLLM failures

Method	Data scale	Weaknesses Seeking	Image Weaknesses	Text Weaknesses	Failure Interpretability	Failure Rectification
Conventional evaluation [3, 34, 52, 80]	fixed set	limited	-	-	-	-
Attacks	Visual adversarial attacks [24]	open-ended	active	✓	-	adversarial only
	Jailbreak attacks [51, 87, 91]	open-ended	active	-	✓	-
Data synthesis	Caption generation [14]	fixed set	no	-	-	-
	Prompt rewriting [19]	open-ended	no	-	-	-
	Image synth/render [53]	open-ended	no	-	-	✓
	Concept perturbation [11, 27]	fixed set	passive	limited	limited	✓
AuditDM (ours)	open-ended	active	✓	✓	✓	✓

AuditDM: **A**udit the **D**ifferences that **M**atter



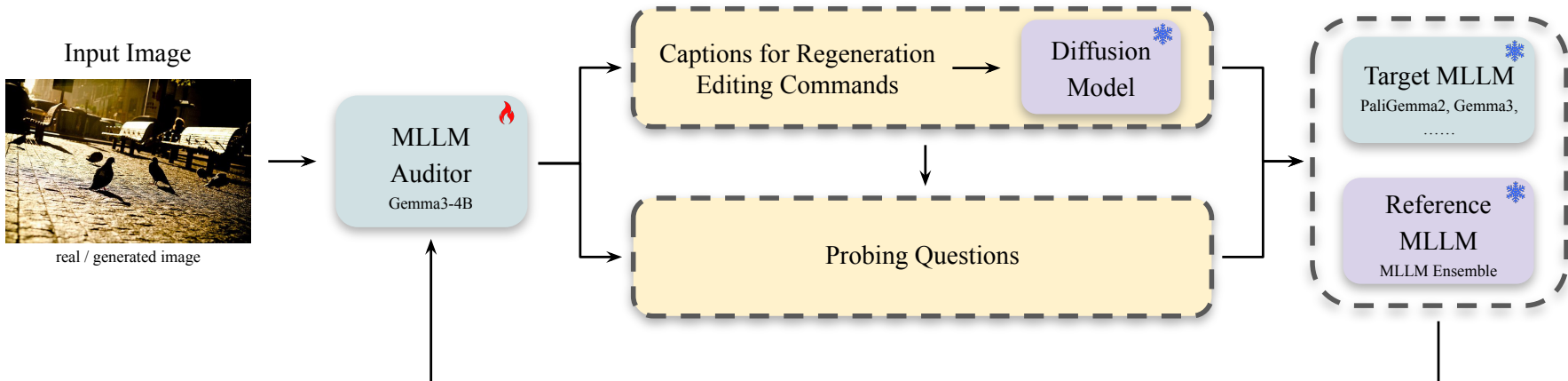
AuditDM: **Audit** the **Differences** that **Matter**



Question-Image Pair Generation:

- Text instruction for image generation;
- Text instruction for image editing;
- Direct question generation

AuditDM: Audit the Differences that Matter



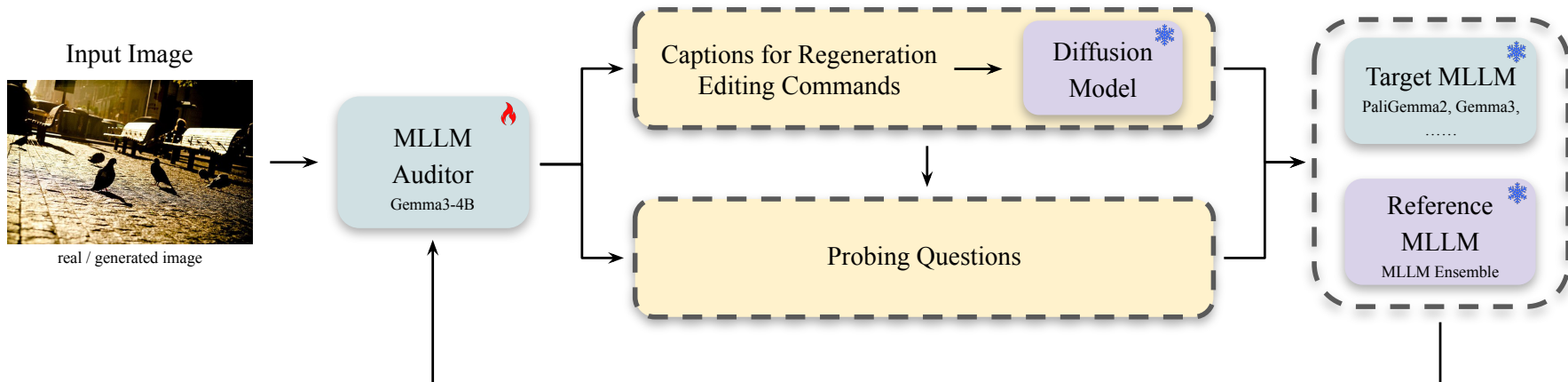
Training:

- Maximize the divergence of Models
- RL-based Auditor Training

$$s(Q^*, I^*) = D(\mathcal{M}_{tar}(Q^*, I^*), \mathcal{M}_{ref}(Q^*, I^*)),$$

$$\hat{A}^k(Q^*, I^*) = \frac{s^k(Q^*, I^*) - \text{mean}_j[s^j(Q^*, I^*)]}{\text{std}_j[s^j(Q^*, I^*)] + \epsilon}.$$

AuditDM: **Audit** the **Differences** that **Matter**



Failure Mode Rectification:

(After training the auditor)










- Augmenting labeled data
- Bootstrapping unlabeled data

Results: AuditDM for Model Failure Detection

Table 2. Effectiveness of finding model weaknesses.

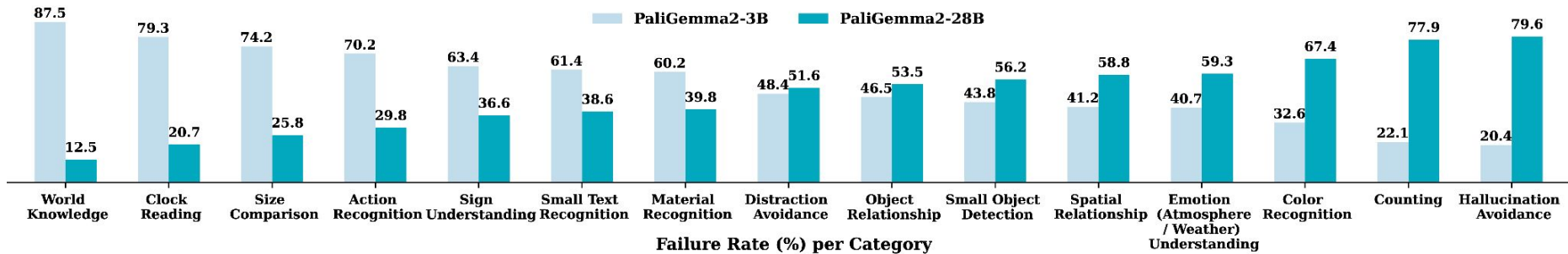
	Baseline	AuditDM (Ours)
Search Success Rate	21.4%	91.1%

Results: AuditDM for Model Failure Detection

World Knowledge	Clock Reading	Size Comparison	Action Recognition	Sign Understanding
				
<i>Q: What company could be the source of most items within the photograph?</i>	<i>Q: What time is displayed on the clock?</i>	<i>Q: Which appliance is smaller?</i>	<i>Q: What activity could the people be engaged in?</i>	<i>Q: What is displayed on the sign?</i>
3B 28B	3B 28B	3B 28B	3B 28B	3B 28B
microsoft ✗ apple ✓	12:50 ✗ 6:10 ✓	toaster oven ✗ microwave ✓	swimming ✗ talking on phone ✓	horn ✗ no horns ✓
Small Text Recognition	Material Recognition	Distraction Avoidance	Object Relationship	Small Object Detection
				
<i>Q: What is the number on the side of the bus?</i>	<i>Q: What does the fence appear to be made of?</i>	<i>Q: What is the man doing?</i>	<i>Q: What is the woman in the foreground focused on?</i>	<i>Q: What is displayed in the picture frame?</i>
3B 28B	3B 28B	3B 28B	3B 28B	3B 28B
305 ✗ 3105 ✓	wood ✗ metal ✓	walking dog ✓ fishing ✗	phone ✓ laptop ✗	pig ✓ cow ✗
Spatial Relationship	Emotion (Atmosphere / Weather) Understanding	Color Recognition	Counting	Hallucination Avoidance
				
<i>Q: What is behind the books?</i>	<i>Q: What is the apparent mood or expression of the individuals?</i>	<i>Q: What is the overall color scheme of the image?</i>	<i>Q: How many traffic lights in the image?</i>	<i>Q: What type of vehicle is parked next to the bus?</i>
3B 28B	3B 28B	3B 28B	3B 28B	3B 28B
fan ✓ cat ✗	happy ✓ confused ✗	yellow ✓ black and white ✗	4 ✓ 2 ✗	none ✓ truck ✗

Weakness Analysis of PaliGemma2

Results: AuditDM for Model Failure Detection



Weakness Analysis of PaliGemma2

Results: AuditDM for Model Failure Detection

“Replace the television with a large monitor displaying a vibrant, looping animation of coral reef life.”



Q: How many lamps are in the room?

3B 28B 3B 28B

2 ✓ 2 ✓ 2 ✓ 1 ✗

“Modify the tennis player in the red outfit to be wearing a brightly colored, patterned tracksuit instead.”



Q: Is the person, in black, serving?

3B 28B 3B 28B

no ✓ no ✓ no ✓ yes ✗

“Replace the pair of scissors with a small, sharp embroidery needle.”



Q: How many baskets?

3B 28B 3B 28B

1 ✓ 1 ✓ 1 ✓ 3 ✗

“Imagine the Phone displaying a live, high-resolution street view map of a bustling Tokyo intersection.”

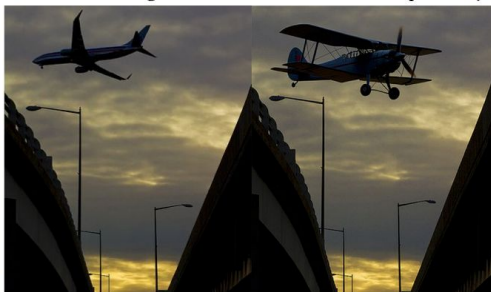


Q: Which hand is the person holding the phone in?

3B 28B 3B 28B

left ✓ left ✓ left ✓ right ✗

“Replace the airplane with a vintage biplane in flight, maintaining the same dramatic sunset backdrop.”



Q: Is it night?

3B 28B 3B 28B

yes ✓ yes ✓ yes ✓ no ✗

“Replace the man's backpack with a vintage leather camera bag.”



Q: Where is he standing?

3B 28B 3B 28B

forest ✓ forest ✓ forest ✓ outside ?

Weakness Analysis of PaliGemma2

Results: AuditDM for Model Improvement

Table 3. Improving PaliGemma2.

Model	VQAv2	GQA	OK-VQA	AI2D	DocVQA	ChartQA	RefCOCO	COCOCap
PaliGemma2-10B 448px ²	85.8	68.3	68.6	84.4	76.6	66.4	78.2	145.0
PaliGemma2-28B 448px ²	85.8	68.3	70.6	84.6	76.1	61.3	77.3	145.2
PaliGemma2-3B 448px ²	84.8	68.1	64.1	76.0	73.6	54.0	76.3	143.4
PaliGemma2-3B 448px ² + AuditDM (Ours)	86.7 (+1.9)	71.1 (+3.0)	69.2 (+5.1)	85.3 (+9.3)	77.5 (+3.9)	63.8 (+9.8)	77.8 (+1.5)	145.1 (+1.7)

Table 4. Improving Gemma3.

Model	MMBench-v1.1	MMTBench	Seed-Bench-IMG	MME	MMMU	MMStar	RealWorldQA	POPE
Gemma3-12B	73.8	58.5	70.6	1517.3	44.8	55.7	58.3	86.0
Gemma3-27B	78.3	59.2	73.2	1526.6	49.7	58.7	62.5	85.2
Gemma3-4B	67.6	53.2	65.7	1376.0	39.6	46.1	54.5	85.1
Gemma3-4B + AuditDM (Ours)	75.0 (+7.4)	58.9 (+5.7)	72.9 (+7.2)	1450.3 (+74.3)	45.2 (+5.6)	52.4 (+6.3)	61.4 (+6.9)	85.5 (+0.4)

Ablations

Table 5. Ablation on different auditing components.

	GQA	RefCOCO	AI2D
Baseline	66.2	73.4	74.7
Probing question	68.5	-	78.2
Image generation	66.9	-	-
Image editing	67.2	74.6	76.3
Best Combination	69.8	74.6	79.4

Table 6. Ablations on training images.

Model	MMBench-v1.1	MMTBench	Seed-Bench-IMG	MME	MMMU	MMStar	RealWorldQA	POPE
Gemma3-4B	67.6	53.2	65.7	1376.0	39.6	46.1	54.5	85.1
Gemma3-4B (with the same images)	69.6 (+2.0)	54.8 (+1.6)	66.9 (+1.2)	1321.6 (-54.4)	40.5 (+0.9)	45.6 (-0.5)	56.3 (+1.8)	84.2 (-0.9)
Gemma3-4B + AuditDM (Ours)	75.0 (+7.4)	58.9 (+5.7)	72.9 (+7.2)	1450.3 (+74.3)	45.2 (+5.6)	52.4 (+6.3)	61.4 (+6.9)	85.5 (+0.4)

Thank you