NEURAL INFORMATION PROCESSING SYSTEMS

IMPERIAL

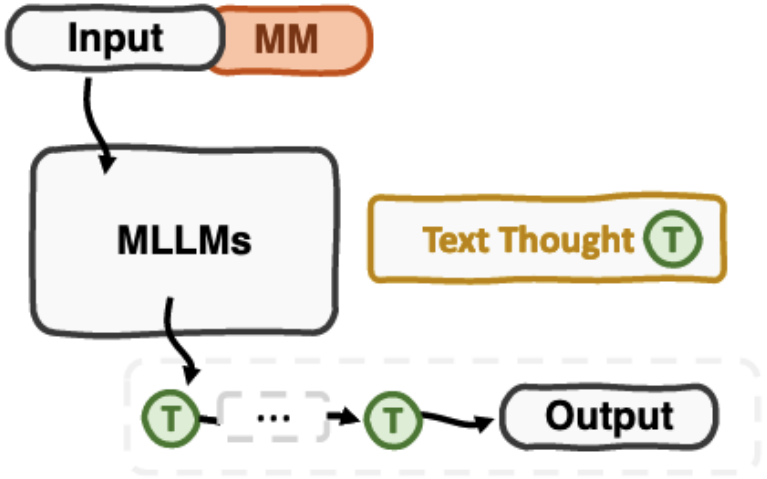
Adaptive Visual Gating for Unified Multimodal Reasoning

Wenxi Gao , Guanxi Lu , Didi Zhu , Hao Mark Chen Quan Deng , Zhican Wang , Jiankang Deng , Hongxiang Fan

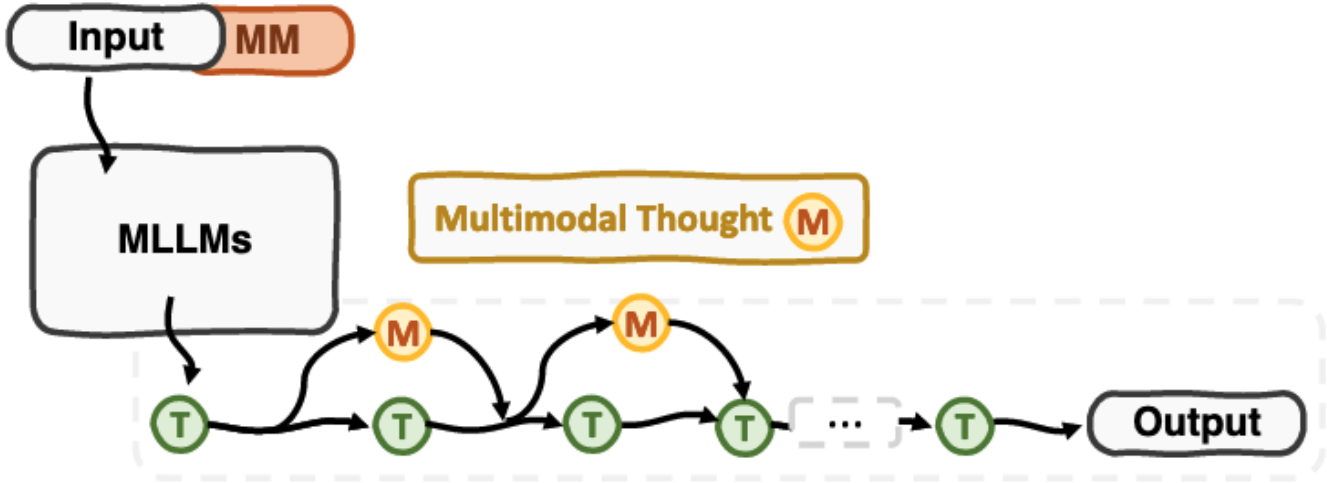


1 Background

- A comparison between **Text-based Rationale** and **Multimodal Rationale**
  - Input: Multimodal [Text & Image]
  - Reasoning: **Text** or **Multimodal** [Text & Image]
  - Output: Multimodal [Text / Image]



Text-based Rationale.



Multimodal Rationale.

Source: Wang Y, Wu S, Zhang Y, et al. Multimodal chain-of-thought reasoning: A comprehensive survey[J]. arXiv preprint arXiv:2503.12605, 2025.

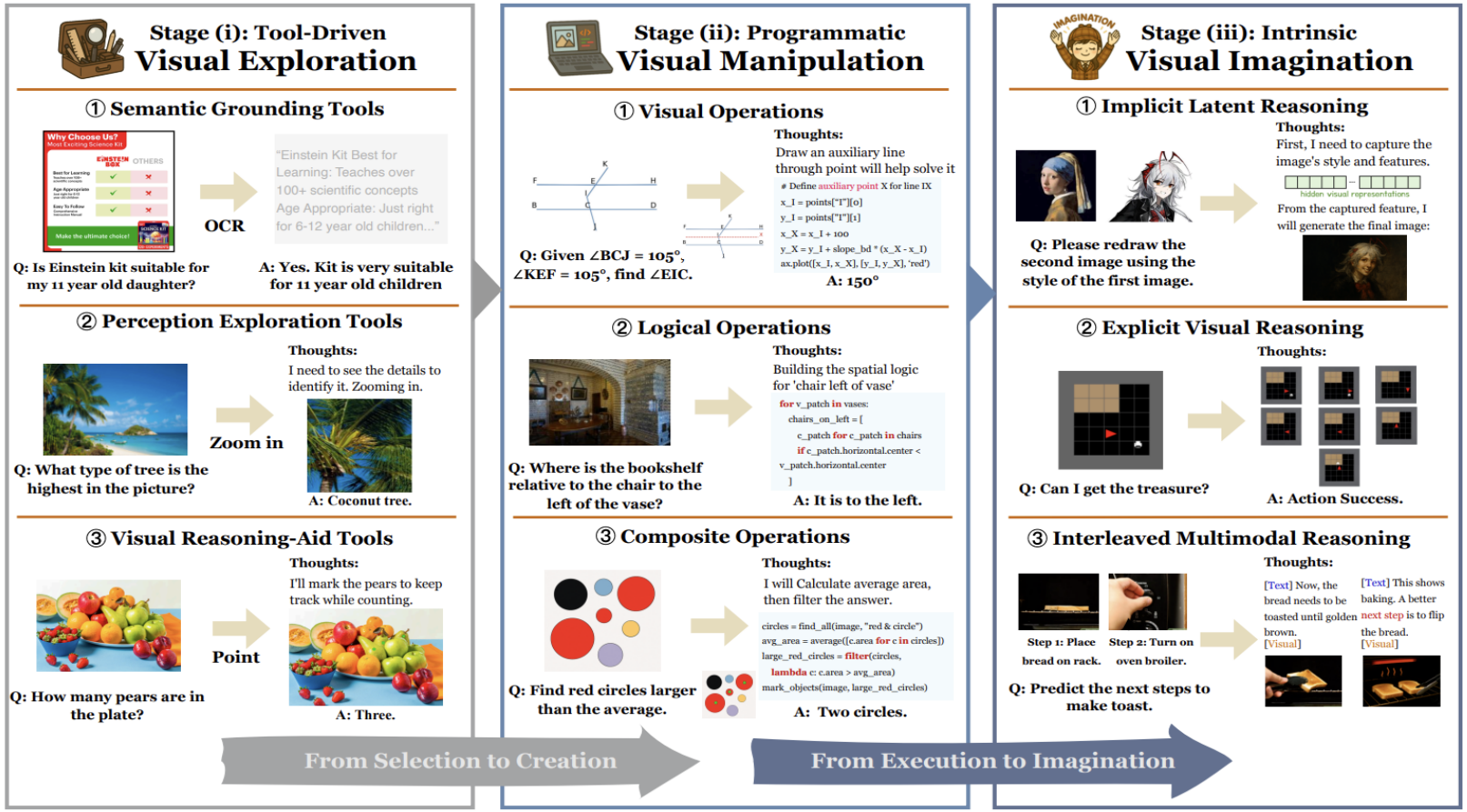
1 Background

➢ A comparison between Text-based Rationale and **Multimodal Rationale**

1. Tool-driven
eg: zoom in, crop…

2. Code-based
eg: matplotlib…

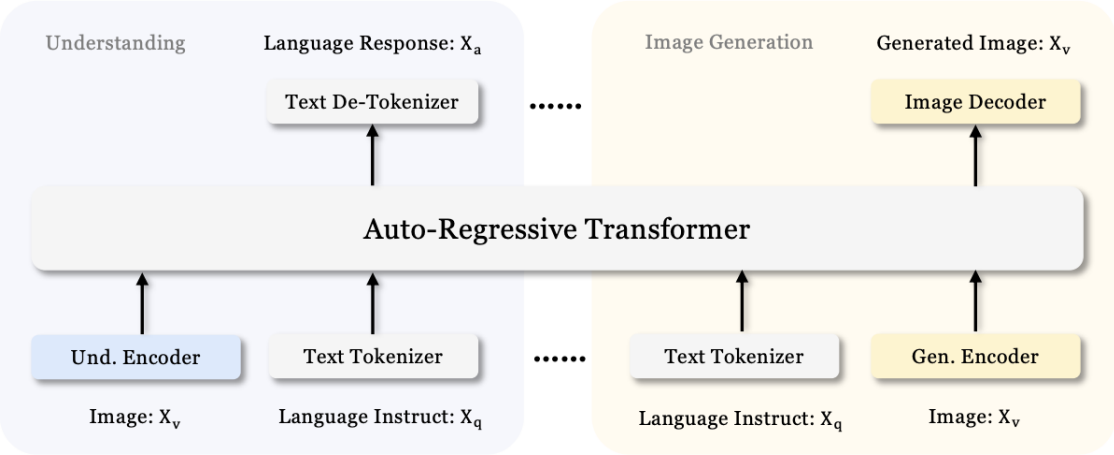
3. Unified multimodal models
eg: BAGEL, Anole…



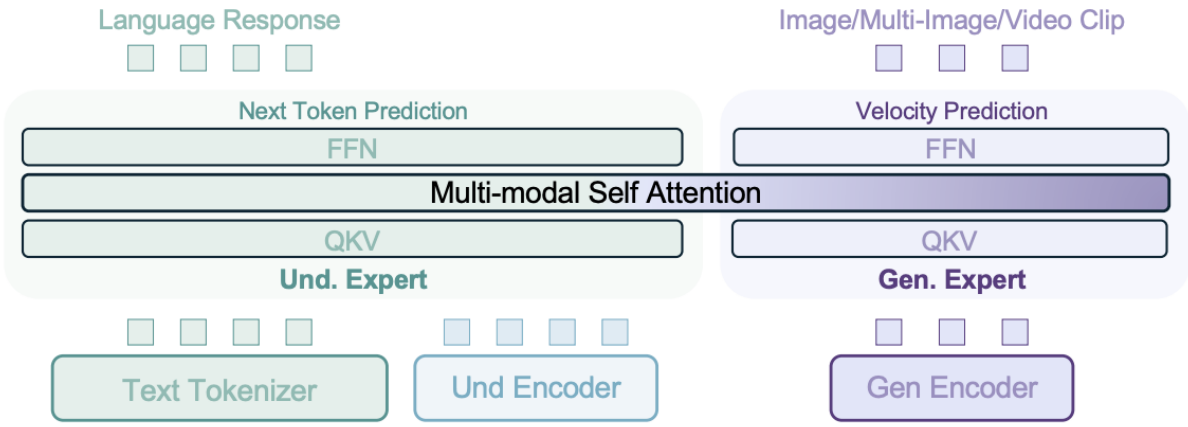
Source: Su Z, Xia P, Guo H, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers[J]. arXiv preprint arXiv:2506.23918, 2025.

1 Background

- Unified multimodal models (UMMs) interleave text reasoning with self-generated images.
  - AR-based UMMs: Text and visual tokens are decoded autoregressively.
  - Hybrid UMMs: Text tokens are decoded autoregressively while visual tokens are generated continuously by diffusion.



AR-based UMMs: Janus.

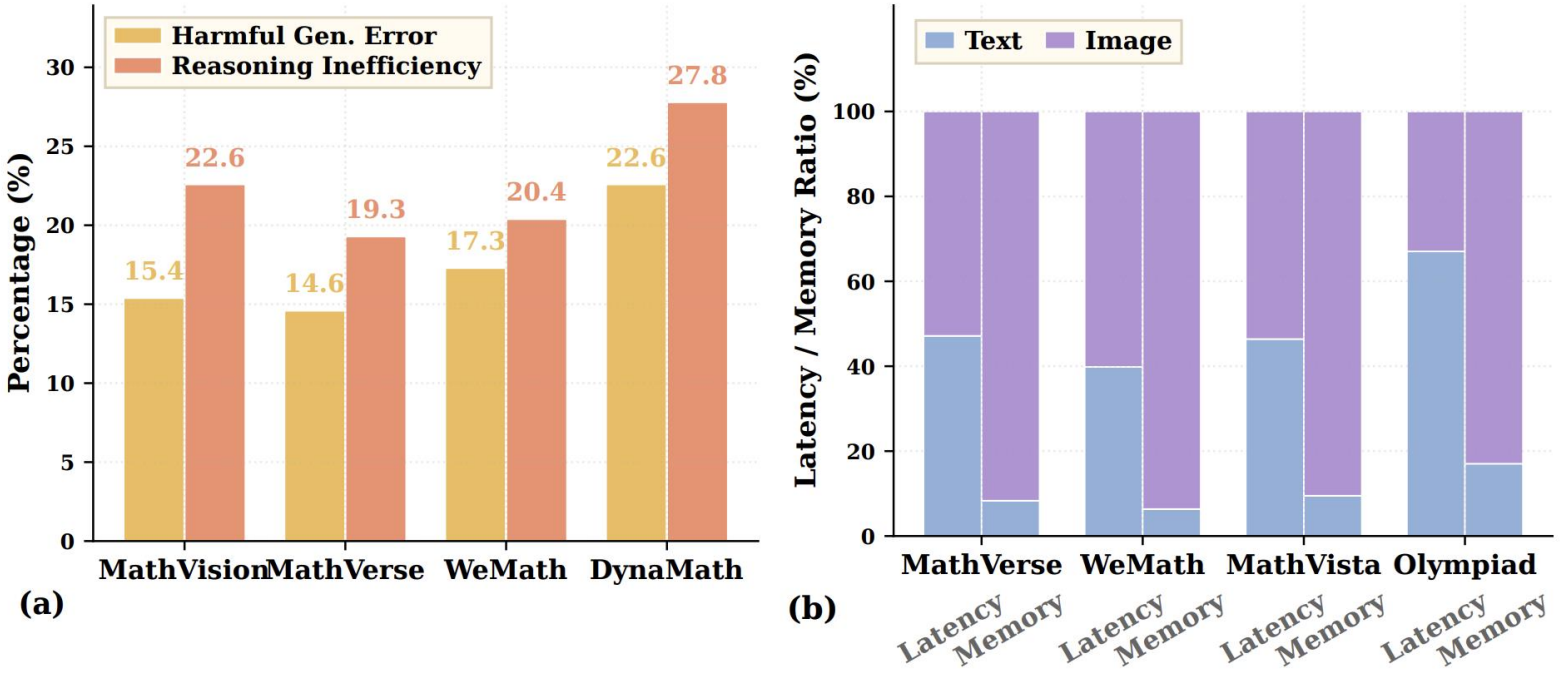


Hybrid UMMs: Bagel.

1 Background

➢ **Limitation of Fixed UMMs Reasoning**

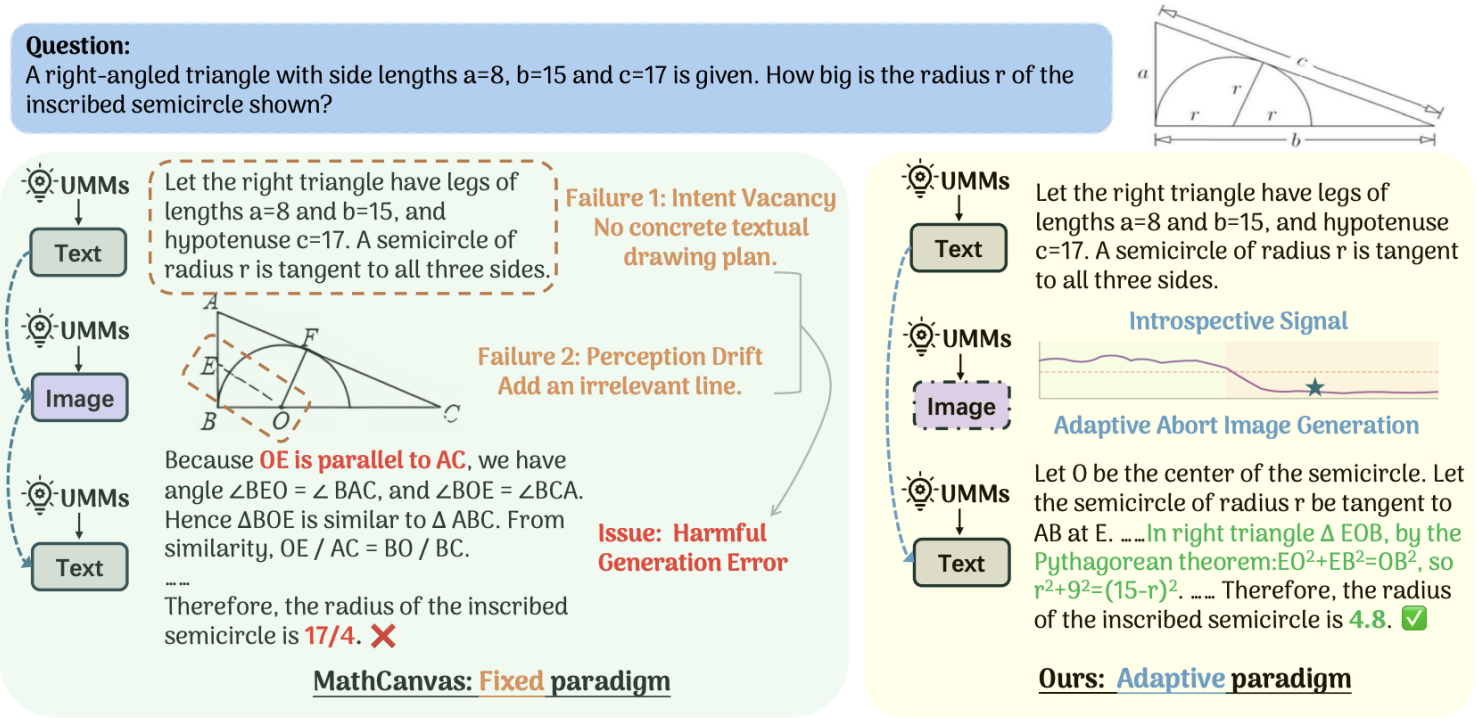
1. **Harmful Generation Error**: Self-generated visual steps in UMMs can be harmful and misleading.
2. **Reasoning Inefficiency**: Visual steps accounts for nearly 50% of overall inference latency 80% of the KV-cache memory footprint.



1 Background

➢ Why Can Self-Generated Images Be Harmful?

1. **Intent Vacancy**: The visual generation is weakly grounded in the preceding textual reasoning.
2. **Perception Drift**: The generated image deviates from the original visual input and introduces misleading evidence.



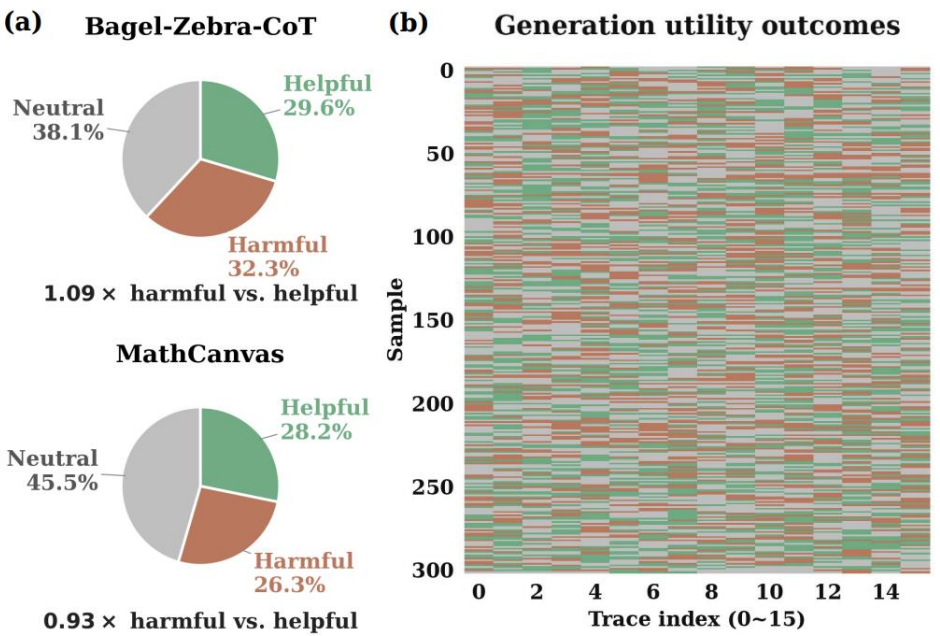
2 Motivation

1. Generation Utility Gap: Visual steps help and harm at similar rates.
2. Harmfulness emerges from the evolving reasoning trace rather than from the problem alone.

Compare two counterfactual reasoning traces:

- C_{gen} :reasoning with intrinsic image generation.
- C_{txt} :text-only counterfactual reasoning.

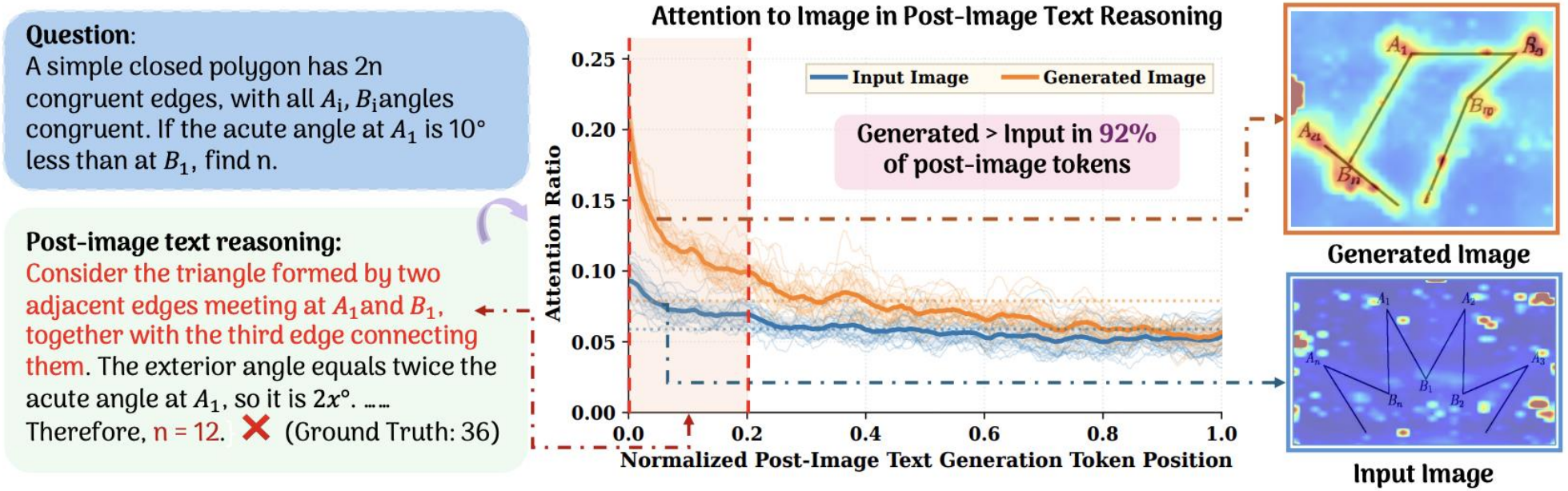
Outcome	C_{gen}	C_{txt}
Helpful	Correct	Incorrect
Harmful	Incorrect	Correct
Neutral	Correct	Correct



3 Method

➢ How Generated Images Mislead Reasoning?

- MathCanvas commits a **perception drift**, generating an irregular non-closed polygon that distorts the input.
- This induces an **attention drift**: post-image text attends more to the generated image (Orange) than to the input (Blue) in 92% of tokens, with the largest gap in the early span.



3 Method

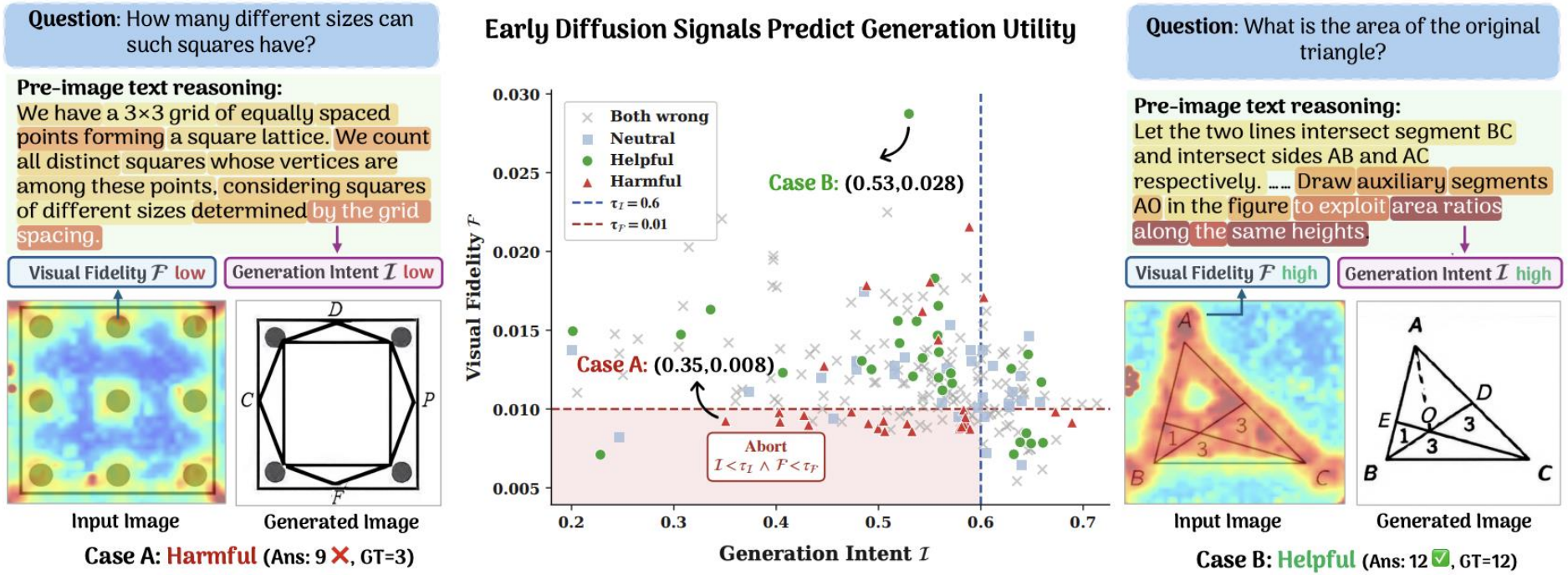
➢ **AdaViG uses early diffusion signals to predict generation utility.**

- Generation Intent I: Does the model have a concrete textual plan for what to draw?
- Visual Fidelity F: Does the generation remain grounded in the original input image?

$$\mathcal{I} = \frac{m_{t^\star}(\mathcal{Q}_{\text{vae}}, \mathcal{S}_{\text{text}})}{m_{t^\star}(\mathcal{Q}_{\text{vae}}, \mathcal{S}_{\text{text}}) + m_{t^\star}(\mathcal{Q}_{\text{vae}}, \mathcal{S}^{\text{in}})}.$$

Abort when both signals are weak

$$\mathcal{F} = m_{t^\star}\left(\mathcal{Q}_{\text{vae}}, \mathcal{S}^{\text{in}}_{\text{vit}}\right).$$



3 Experiment

➢ 1) Main Results

- **Accuracy & Efficiency gains:** AdaViG improves accuracy and image generation efficiency across models and datasets. Notably, on DynaMath with Bagel-Canvas, it improves accuracy by 5.7% and reduces image generation FLOPs by 75%.

Table 1: **Main results on math reasoning benchmarks.** The best accuracy (%) results among UMMs *w/ CoT* are in **bold**, and the second best are underlined. Orange indicates accuracy gains; blue indicates efficiency gains, measured by the reduction (%) in image-generation FLOPs.

Method		DynaMath	WeMath	MathVerse	MathVista	Olympiad	Avg.
General MLLMs							
GPT-4o		56.3	45.8	50.8	60.0	33.0	49.2
Gemini-2.5-Pro		66.8	78.0	76.9	80.9	42.1	68.9
Qwen2.5-VL-7B		49.8	35.3	45.7	72.6	22.0	45.1
InternVL3-8B		37.1	34.8	39.8	70.8	21.5	40.8
UMMs							
Emu3-Chat-8B		24.6	20.5	15.5	47.6	10.2	23.7
Janus-Pro-7B		20.3	16.3	15.9	42.8	9.7	21.0
MUSE-VL-7B		30.8	30.2	39.1	51.3	15.0	33.3
Bagel-7B		38.2	36.2	42.6	68.8	22.7	41.7
UMMs w/ CoT							
Bagel-7B *w/ CoT*		38.6	37.1	43.0	68.8	23.0	42.1
Bagel-7B *w/ ICoT*		39.5	37.0	45.8	70.0	22.7	43.0
Bagel-Zebra-CoT		39.0	34.5	48.2	81.0	16.7	43.9
w/ AdaViG (Ours)	Acc. ↑	43.5 (+4.5)	38.1 (+3.6)	51.3 (+3.1)	**83.4** (+2.4)	18.9 (+2.2)	47.0 (+3.1)
	FLOPs ↓	-45%	-91%	-75%	-34%	-40%	-57%
Bagel-Canvas		42.3	35.3	55.8	77.9	24.0	47.1
w/ AdaViG (Ours)	Acc. ↑	**48.0** (+5.7)	**38.8** (+3.5)	**58.4** (+2.6)	81.3 (+3.4)	**26.2** (+2.2)	**50.5** (+3.4)
	FLOPs ↓	-75%	-27%	-25%	-26%	-37%	-32%

3 Experiment

➢ 1) Main Results

- **Latency:** AdaViG also brings consistent end-to-end latency reductions. Bagel-Zebra-CoT achieves larger latency reductions, up to 45.6%, while Bagel-Canvas achieves 15.4%–30.8%.

Table 2: **End-to-end inference latency (seconds per sample).** The results are measured on a single H200 GPU and averaged across each benchmark. Blue indicates the reduction in latency.

Method	DynaMath	WeMath	MathVerse	MathVista	Olympiad
Bagel-Zebra-CoT	37.4	41.6	53.2	52.7	24.9
w/ AdaViG (Ours)	25.7 (-31.3%)	22.6 (-45.6%)	31.7 (-40.0%)	39.8 (-24.5%)	17.6 (-29.3%)
Bagel-Canvas	13.2	13.1	11.5	10.4	5.72
w/ AdaViG (Ours)	9.4 (-28.9%)	10.8 (-17.6%)	9.7 (-15.7%)	8.8 (-15.4%)	3.96 (-30.8%)

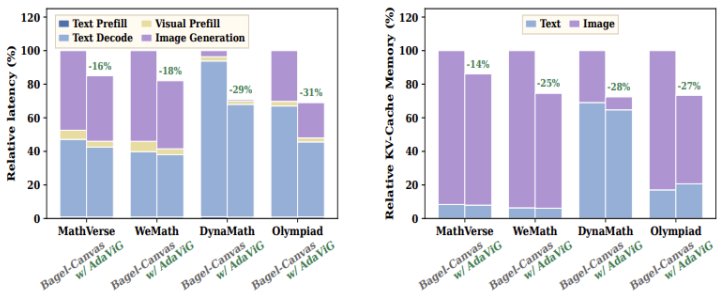
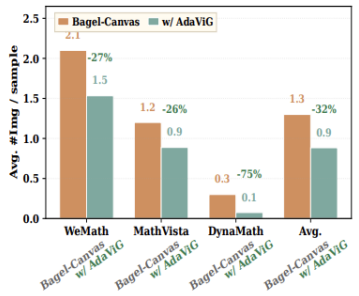


Figure 7: Relative latency and KV-cache memory breakdown across benchmarks.



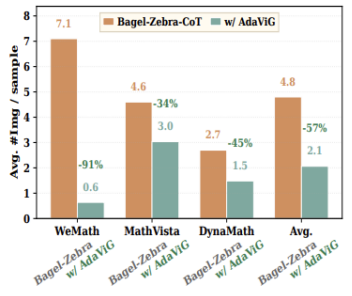


Figure 8: Reduction in the average number of generated images per sample.

➢ 2) Ablation Studies

- **Performance of different gating method:** Single signal gates improve over Bagel-Canvas but remain suboptimal, as the two signals capture different failure modes. AdaViG aborts only when both signals are weak and achieves the best performance with a moderate drop rate.

Table 3: Performance on Bagel-Canvas under gating methods with different ($\mathcal{I}$,$\mathcal{F}$) combinations.

Method	Avg. Drop	Performance		
		MathVista	WeMath	DynaMath
Bagel-Canvas	–	77.9	35.3	42.3
w/ $\mathcal{I}$ only	55%	78.8	36.2	46.1
w/ $\mathcal{F}$ only	31%	79.4	36.5	45.4
w/ $\mathcal{I} \vee \mathcal{F}$	61%	75.1	32.7	41.2
w/ ***AdaViG***	43%	**81.3**	**38.8**	**48.0**

Experiment

➢ 2) Ablation Studies

- **Comparison with alternative aborting strategies:** We compare AdaViG with random image generation aborting, text-only reasoning, and DAP-ICoT on Bagel-Canvas.

Table 4: Comparison with alternative image generation aborting strategies on Bagel-Canvas.

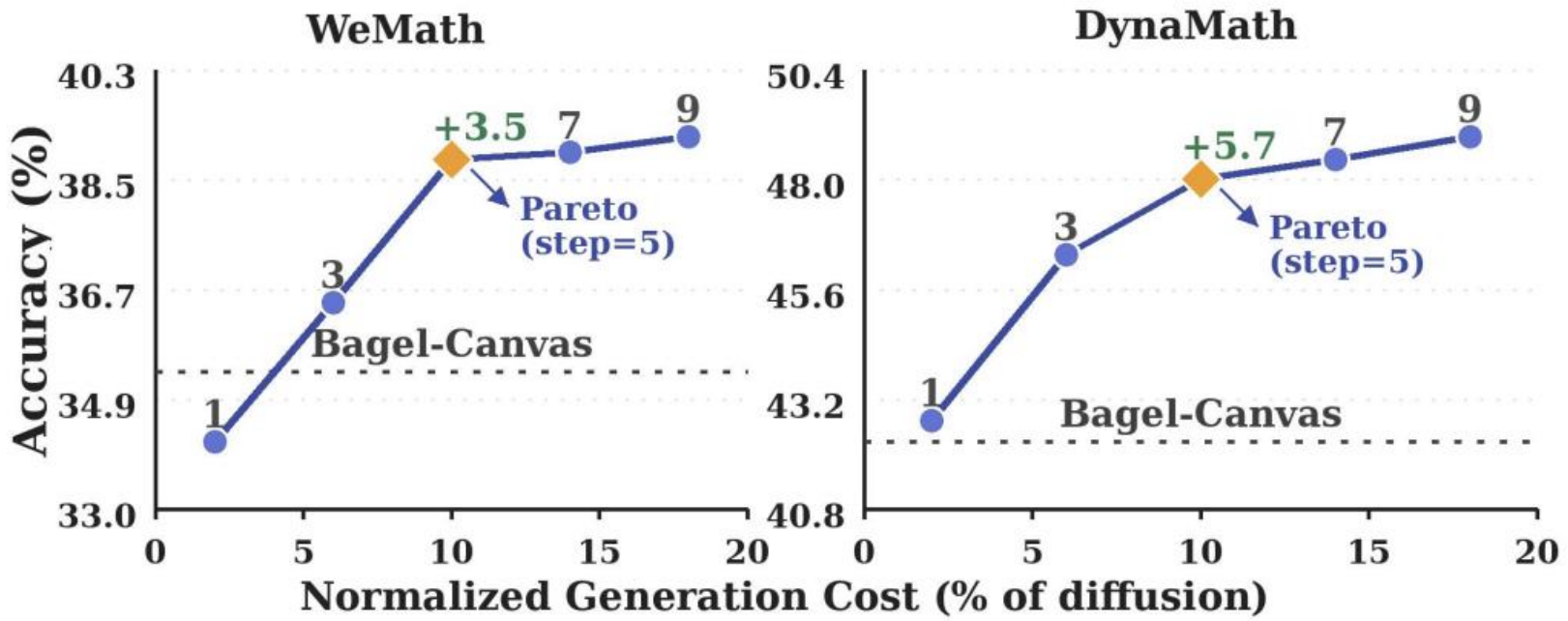
Method	Drop Rate	Performance		
		MathVista	WeMath	DynaMath
AdaViG	43%	**81.3**	**38.8**	**48.0**
Random	25%	78.2	34.6	44.1
	50%	77.4	35.6	42.9
Text-only	100%	76.9	34.1	42.7
DAP-ICoT [20]	12%	78.4	35.4	44.3

AdaViG achieves the best performance, suggesting that its improvement does not come from merely reducing image generations, but from selectively suppressing harmful ones.

Experiment

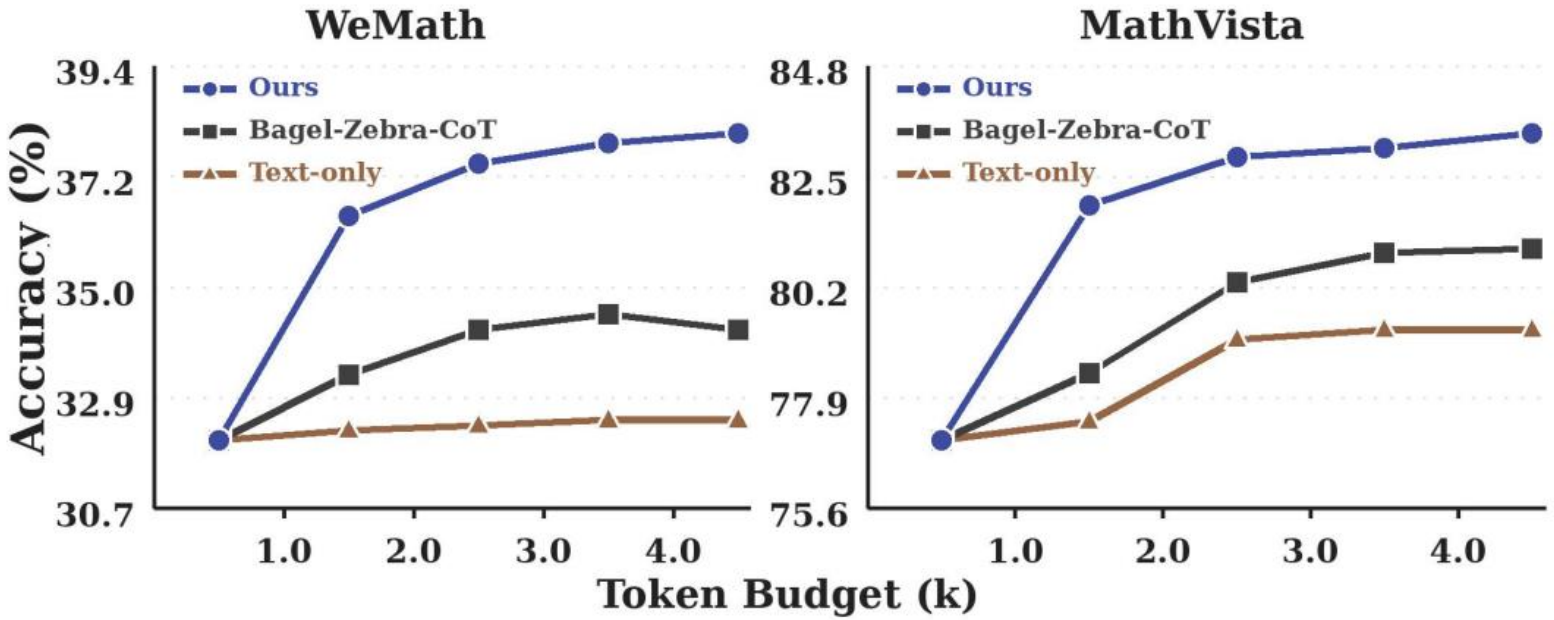
➢ 2) Ablation Studies

- **Selection of early abort step:** The results reveal a trade-off between accuracy and efficiency: decisions at later steps provide stronger gating capability but incur higher generation cost.



➢ 2) Ablation Studies

- **Performance across different token budgets:** The results suggest that visual gating is important when scaling intrinsic interleaved reasoning with larger token budgets.



4 Conclusion

- **1)** We present the generation utility gap in unified multimodal reasoning.
- **2)** We identify two introspective signals during early diffusion and use these signals to build AdaViG, a training-free adaptive visual gating method that aborts harmful generations early.
- **3)** AdaViG improves accuracy while reducing visual-generation cost across benchmarks.

Thanks !

