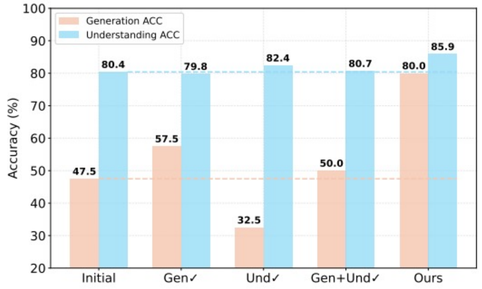
Understanding vs. Generation: Navigating Optimization Dilemma in Multimodal Models

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, Han Hu

OVERVIEW

Motivation

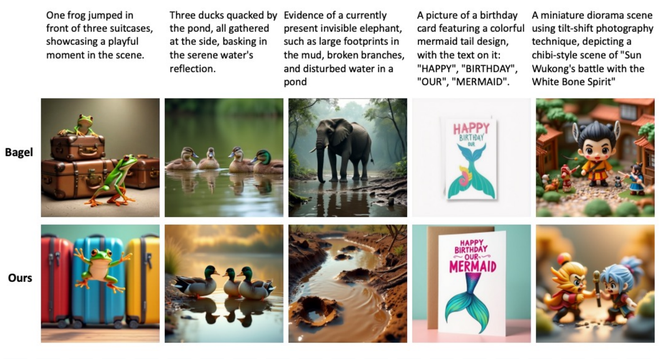
Current multimodal models face a key challenge: enhancing generative capabilities often comes at the expense of understanding, and vice versa. We analyze this trade-off and identify the primary cause: the potential conflict between generation and understanding objectives creates a competitive dynamic within the model. This optimization dilemma calls for a new paradigm that aligns the two tasks rather than treating them independently.



Contribution

(1) Systematic analysis of the conflict between generation and understanding, identifying its root cause in competing optimization objectives.
(2) Propose the Reason-Reflect-Refine (R3) framework that decomposes generation into a structured generate-understand-regenerate process.
(3) Demonstrate that R3 achieves stronger generation while simultaneously preserving understanding ability.

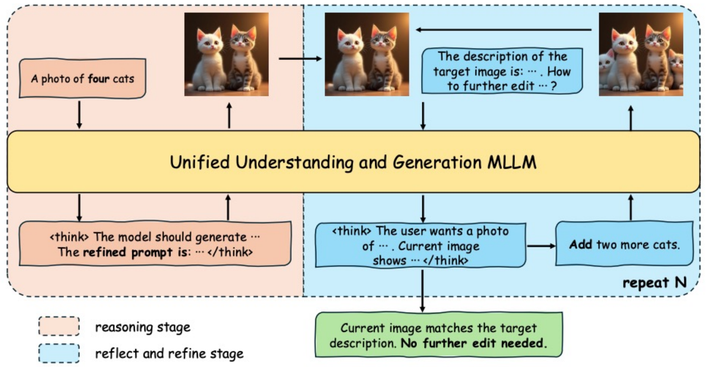
Visualization



METHOD

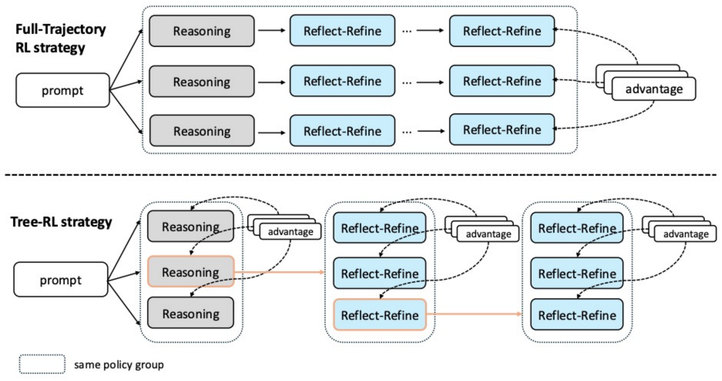
Reason-Reflect-Refine (R3) Framework

Reason: The model analyzes the user's intent, enriches the initial prompt with fine-grained details, and synthesizes an initial image.
Reflect: The model evaluates its output against the original prompt. If aligned, terminate; otherwise identify discrepancies.
Refine: Execute corrective editing instructions to modify the image. The loop continues until satisfactory alignment is reached.



Tree-RL Training Strategy

We split the trajectory into Reason and Reflect-Refine stages with separate optimization. A tree-based rollout strategy provides clear supervision for each intermediate step, overcoming error accumulation and low training efficiency of end-to-end RL.



Stage-wise Reward Design

Different reward models evaluate each stage: image-prompt alignment for diffusion generation, and a correctness metric for reflection that rewards measurable improvement or correct termination.

EXPERIMENTS

Generation Results

Table 1: Instruction-following generation ability on the GenEval++ benchmark, evaluated by GPT-4.1. Bold indicates the best result. †indicates our framework with only the reasoning stage. Green arrows indicate improvement over the BAGEL baseline.

Method	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
GPT-4o (OpenAI, 2025)	0.900	0.675	0.725	0.625	0.600	0.800	0.850	0.739
FLUX.1-Kontext (Labs, 2024)	0.425	0.500	0.200	0.250	0.300	0.400	0.325	0.343
FLUX.1-dev (Labs, 2024)	0.350	0.625	0.150	0.275	0.200	0.375	0.225	0.314
Janus-Pro (Chen et al., 2025)	0.450	0.300	0.125	0.300	0.075	0.350	0.125	0.246
T2I-R1 (Jiang et al., 2025)	0.675	0.325	0.200	0.350	0.075	0.250	0.300	0.311
Echo-4o (Ye et al., 2025)	**0.800**	0.575	0.550	**0.775**	0.625	**0.800**	0.625	0.679
BAGEL (Deng et al., 2025)	0.325	0.600	0.250	0.325	0.250	0.475	0.375	0.371
BAGEL + Ours†	0.500	0.650	0.600	0.650	0.550	0.600	0.600	0.593 ↑0.22
BAGEL + Ours	0.675	**0.725**	**0.575**	0.725	**0.750**	0.575	**0.800**	**0.689** ↑0.32

Understanding Results

Table 2: Evaluation of understanding capabilities on our proposed ITA benchmarks. All scores are reported as accuracy (%). †indicates our framework with only the reasoning stage. Green arrows indicate improvement over the BAGEL baseline.

ITA	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
BAGEL	60.63	58.54	45.42	63.54	53.96	80.83	61.50	60.60
BAGEL + Ours†	60.42	59.38	47.71	63.75	55.63	81.46	63.96	61.76 ↑1.16
BAGEL + Ours	**69.58**	**67.50**	**69.79**	**72.29**	**76.04**	**83.33**	**75.00**	**73.37** ↑12.77

Table 3: Evaluation of VQA capabilities. All scores are reported as accuracy (%). †indicates our framework with only the reasoning stage. Green arrows indicate improvement over the BAGEL baseline.

VQA	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
BAGEL	91.74	79.30	88.28	77.99	82.93	85.10	92.45	86.48
BAGEL + Ours†	91.67	76.12	88.76	78.71	83.29	84.98	93.45	86.72 ↑0.24
BAGEL + Ours	**93.95**	**84.63**	**91.15**	**84.09**	**86.06**	**86.54**	**94.50**	**89.63** ↑3.15

Ablation Studies

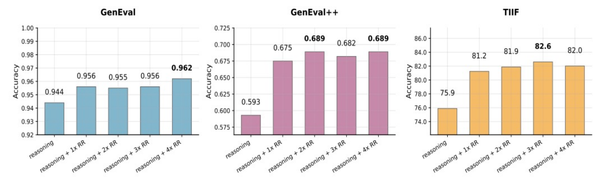


Figure 6: Inference-time scaling effect across the GenEval, GenEval++, and TIIF benchmarks (left to right). Performance is shown as a function of the maximum allowed reflection-refinement turns.

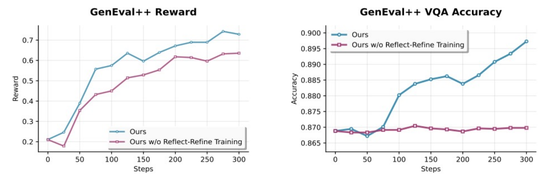


Figure 7: The evolution of generation and understanding abilities in the training process. The left figure shows the generation accuracy (measured by Qwen-2.5-VL-72B.) and the right figure shows the VQA accuracy.