

## Summary

- Goal:** Evaluate real downstream OOD benchmark gains, not policy reward values
- AlphaGRPO:** a RL framework for multimodal interleaved generation in AR-Diffusion UMMs.
  - Focusing on reasoning text-to-image generation and self-reflective refinement
- DVReward** (Compositional Verifiable Reward): training-free rubric reward method. Open-source LLMs to generate rubrics and MLLMs to verify. Superior to trained reward models.

## How to activate the ability to latent reasoning?

**User:** A hand is holding a pair of scissors. The hand and scissors are casting a shadow onto the wall behind them. The scene is a sunny outdoors image of a hand, scissors, and shadow on a wall.

**Verification**  
**Response:** The generated image **effectively fulfills the user's original intent**. It accurately depicts a hand holding scissors, casting a shadow on a wall in a sunny outdoor setting, with high visual quality and adherence to physical rules.

**Reflection**  
**Response:** **The image has issues with the shadow's position and shape**, which are inconsistent with the light source and the expected shadow pattern of the scissors and hand. To improve the image, **the shadow should be adjusted to be consistent with the light source and the position of the scissors**. The shadow on the wall should also be corrected to have a more natural and coherent appearance.

**Fail to find error!** (Standard generation does not reliably activate it.)

**Find error!** (Verify mode tends to confirm; Reflect mode searches for visual mismatches.)

**What this verifies**  
 The model already has a critique capability, but standard generation does not reliably activate it.

## What's the best way to apply MLLM as reward model?

**Prompt:** A tree in front partially hides a bench behind it.

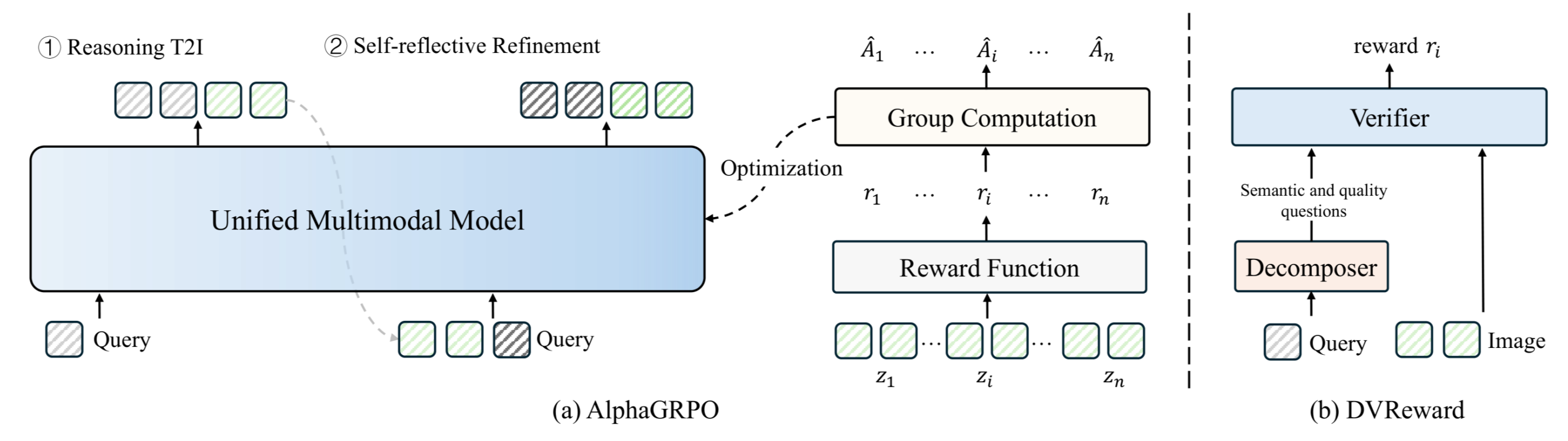
Image 1: a tree **behind** of a bench (Score-based Reward: 0.848)

Image 2: a tree in **front** of a bench (Score-based Reward: 0.848)

Question-based Reward: **0.592** (Image 1) vs **0.914** (Image 2)

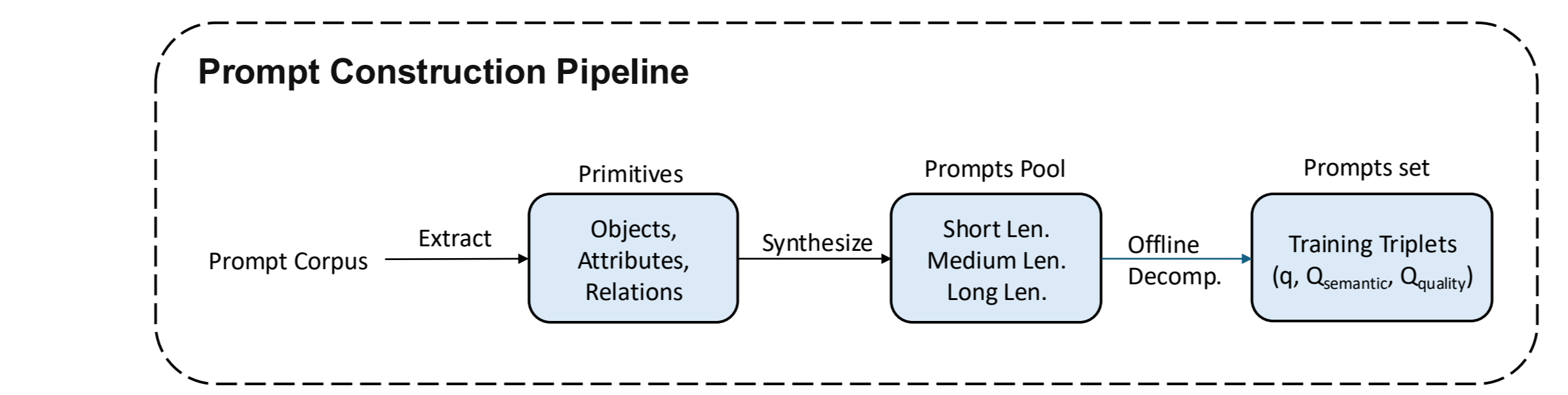
- Score-based Reward: Ask the MLLMs to give scalar score about the quality of the image.
- Question/Rubric-based Reward: Ask the yes/no answer to a specific question. Use p(yes) as reward
- Conclusion: **Question/Rubric-based Reward is more accurate and discriminative !!**

## Method



- Unified trajectory in multimodal generation tasks  
 prompt + conditional image (optional) -> think text + image
- DVReward  
 Prompt -> decomposed rubrics -> MLLM verify
- Training-free: no preference-data tuning.
  - Fully Open-source depended:
    - Qwen3-235B-A27B as Decomposer
    - Qwen3-30B-A3B as Verifier

## Dataset Construction



➢ AlphaGRPO20k: Three-level length, cover 39 T2I tasks.

(a) semantic questions: Existence (34.9%), Attribute (23.7%), Environment (15.4%), Spatial (11.4%), Count (3.4%), Action (3.1%), Style (2.6%), Negative Viewpoint (1.8%), Text (1.2%).

(b) quality questions: Physics (8.1%), Aesthetics (9.1%), Lighting (13.3%), Texture (18.2%), Geometry (29.9%), Coherence (22.4%), Legibility (1.8%), Anatomy (4.2%).

Question distribution.

➢ 10 aspect in semantic, 8 aspects in quality

Distribution of number of questions.

## Qualitative Examples

Image Generation: Prompt: "A metallic rose that is not fully bloomed is higher than a blooming fabric rose." Models: BAGEL, AlphaGRPO (RT2I), AlphaGRPO, AlphaGRPO+ Inf. SRR.

Image Editing: Prompt: "Generate a cyberpunk-style photo." Models: Original Image, BAGEL, AlphaGRPO (RT2I), AlphaGRPO.

Benchmarks: TIIFBench-Short, TIIFBench-Long, GEdit-Bench, GenEval, DPG-Bench.

## Downstream Performance

Model	TIIF Bench ↑				WIS ↑	DPG-Bench ↑	Geneval ↑
	Basic S	Basic L	Advanced S	Advanced L			
<i>Generation Only Models</i>							
SD3 Medium (Esser et al., 2024)	78.3	77.8	61.5	59.6	63.2	67.3	64.8
FLUX.1 dev (Labs, 2024)	83.1	78.7	65.8	68.5	70.7	71.5	71.8
<i>Unified Multimodal Models</i>							
Show-o (Xie et al., 2024)	73.1	75.8	55.0	50.9	53.7	50.4	59.7
JanusPro (Chen et al., 2025b)	79.3	78.3	59.7	58.8	65.8	60.3	66.5
<i>Inference on 512 Resolution</i>							
BAGEL	81.7	86.1	73.7	77.6	84.7	82.1	75.2
AlphaGRPO (RT2I)	85.5	84.2	77.4	78.9	84.3	86.6	78.9
AlphaGRPO + Inf. Self-Reflective Refinement	<b>89.8</b>	<b>88.8</b>	88.2	<b>88.7</b>	<b>84.7</b>	83.2	<b>83.9</b>
<i>Inference on 1024 Resolution</i>							
BAGEL	83.4	83.7	75.2	76.7	79.8	73.5	76.4
AlphaGRPO (RT2I)	84.8	85.9	79.9	78.4	80.2	80.2	79.5
AlphaGRPO + Inf. Self-Reflective Refinement	85.4	82.9	75.6	77.5	81.7	84.3	77.7
Overall	82.2	81.4	-	-	-	-	-

Type	Model	GEdit-Bench ↑			Model	Reward	TIIF ↑		Geneval ↑
		G_SC	G_PQ	G_O			Short	Long	
Private	Gemini 2.0 (Gemini2, 2025)	6.73	6.61	6.32	SD3.5M	-	74.0	73.2	79.8
	GPT-4o (OpenAI, 2025)	<b>7.85</b>	<b>7.62</b>	<b>7.53</b>	SD3.5M	PickScore	77.6	76.3	80.4
Open-source	Instruct-Pix2Pix (Isola et al., 2017)	3.58	5.49	3.68	SD3.5M	VIEScore	76.2	72.9	82.9
	MagicBrush (Zhang et al., 2023)	4.68	5.66	4.52	SD3.5M	DVReward	<b>79.1</b>	<b>77.7</b>	<b>86.0</b>
	AnyEdit (Yu et al., 2025)	3.18	5.82	3.21	BAGEL	-	75.2	78.6	84.0
	OmniGen2 (Wu et al., 2025b)	7.16	6.77	6.41	BAGEL	HPSv3	78.5	77.1	83.4
	Step1X-Edit (Liu et al., 2025b)	7.09	6.76	6.70	BAGEL	UnifiedReward	<b>79.2</b>	77.3	83.7
Ours	BAGEL (Deng et al., 2025)	7.36	6.83	6.52	BAGEL	VIEScore	79.1	77.9	81.7
	AlphaGRPO (RT2I)	7.54	7.18	6.85	BAGEL	DVReward	78.9	<b>79.5</b>	<b>85.1</b>
	AlphaGRPO	<b>7.67</b>	<b>7.46</b>	<b>7.08</b>					

- OOD gains across T2I and Editing.
- Low-Resolution Training Generalizes to 1024px.
- SRR Training Matches RT2I on OOD T2I, Excels on Editing.
- Inference-time SRR further improves.