Unified Multimodal Models as Auto-Encoders

Zhiyuan Yan[1,2◇,⋆], Kaiqing Lin[1◇], Zongjian Li[1,3◇], Junyan Ye[4◇], Hui Han[1], Haochen Wang[2,6⋆], Zhendong Wang[5], Bin Lin[1,3], Hao Li[1], Xinyan Xiao[2], Jingdong Wang[2], Haifeng Wang[2], Li Yuan[1†]

[1]Shenzhen Graduate School, Peking University
[2]Baidu, [3]Rabbitpre AI, [4]SYSU, [5]USTC, [6]CASIA

zhiyuanyan@stu.pku.edu.cn

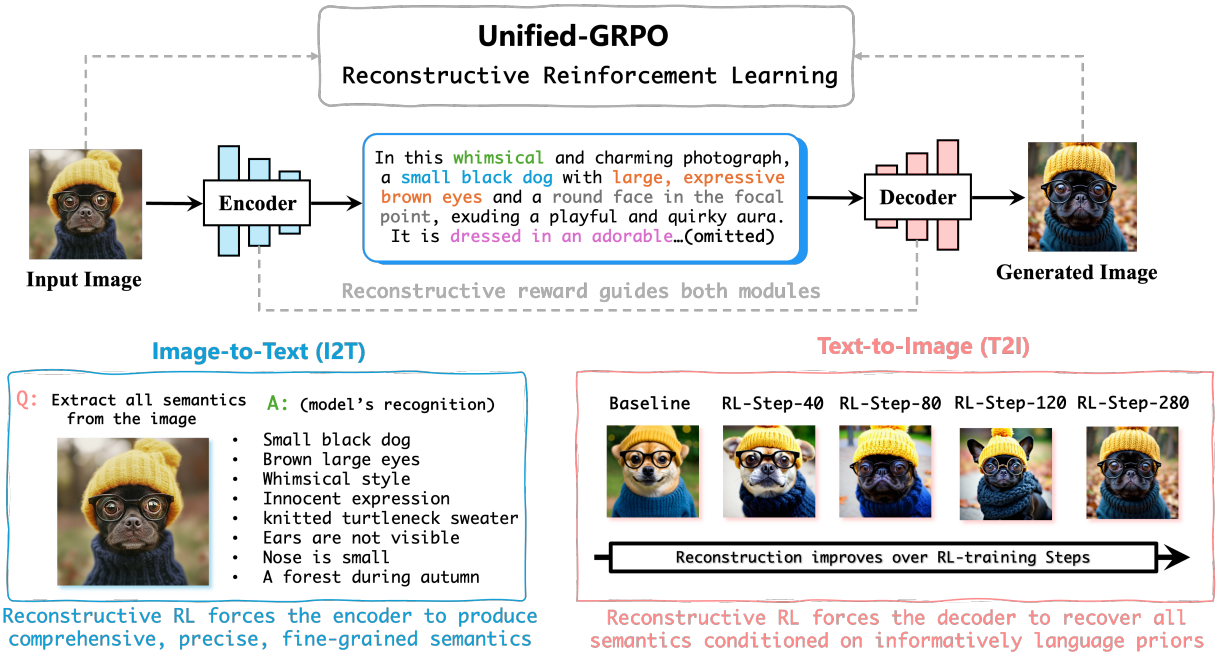

Figure 1. Illustration of the key insight of our **UAE**, unifies image-to-text understanding and text-to-image generation under a reconstructive auto-encoding perspective. By optimizing reconstruction similarity through RL, the encoder (und. module) is trained to learn richer semantic representations while the decoder (gen. module) becomes better at recovering all semantics. Illustrations show strengthened fine-grained visual perception and improved complex instruction-following generation capability across RL training steps.

Abstract

Image-to-text (I2T) understanding and text-to-image (T2I) generation are two fundamental, important yet traditionally isolated multimodal tasks. Despite their intrinsic connection, existing approaches typically optimize them independently, missing the opportunity for mutual enhancement. In this paper, we argue that the both tasks can be connected under a shared Auto-Encoder perspective, where text serves as the intermediate latent representation bridging the two directions — encoding images into textual semantics (I2T) and decoding text back into images (T2I). Our key insight is that if the encoder truly "understands" the image, it should capture all essential structure, and if the decoder truly "understands" the text, it should recover that structure faithfully. Building upon this principle, we propose Unified-GRPO, a post-training method based on reinforcement learning that jointly optimizes both modules through reconstructive rewards, maximizing the semantic consistency between the input and the generated images. Under this reconstruction objective, the encoder is encouraged to extract as much accu-

◇ Equal Contribution, ⋆ Work done during an internship at the Baidu Star Program, † Corresponding Author

rate and comprehensive semantic information from the input image to maximize reconstruction quality, while the decoder is simultaneously optimized to generate conditioned on the encoder's prior, enabling a self-evolving improvement.

Empirically, we find that using text as the intermediate representation and training under a reconstructive RL paradigm effectively benefits both I2T and T2I. The I2T module gains stronger fine-grained visual perception, such as small-object recognition, grounding, etc, while its dense embeddings and language priors, in turn, provide richer semantic signals that improve T2I fidelity and complex instruction following. These results demonstrate that the reconstructive RL establishes a mutually reinforcing cross-modal synergy within the auto-encoding framework.

1. Introduction and Motivation

Unified multimodal models (UMMs) that support both generation and understanding have recently gained increasing popularity in both academia and industry [3, 9, 18, 29, 31, 36, 41, 47]. However, directly combining the understanding and generation models together leads to a sub-optimal result, as most existing arts on UMMs [2, 18, 31] suggest that optimizing diffusion-based generative objectives negatively degrade the understanding capability and learned representations (and conversely), making joint training brittle.

Consequently, some existing works decouple the UMM problem [21, 31], training understanding and generation modules separately, and missing out on potential cross-task mutual benefits. These design choices and empirical observations have dampened confidence in truly unified systems: absent demonstrable mutual gains, "unification" collapses into training two large components side by side.

In this work, we revisit the relationship between I2T and T2I from a conceptual standpoint and argue that a more principled linkage can be established by viewing them through a unified Auto-Encoder (AE) perspective. Under this view, text acts as an intermediate latent representation: the encoder extracts a semantic description from the input image (I2T), and the decoder reconstructs an image from this semantic representation (T2I). This perspective offers a natural and powerful unifying principle: **if the encoder genuinely understands the image, it should capture all essential visual structure; if the decoder genuinely understands the text, it should faithfully recover that structure.** Thus, high-quality reconstruction becomes a proxy for enhancing both tasks simultaneously, revealing a pathway toward bidirectional synergy.

Building upon this insight, we introduce **Unified-GRPO**, a reinforcement-learning-based post-training method that jointly optimizes the encoder and decoder through reconstructive rewards. Unified-GRPO maximizes the semantic consistency between the original and reconstructed images, encouraging the encoder to produce richer and more accurate textual semantics, while guiding the decoder to generate images that better adhere to the encoder's descriptions. Through this cross-modal feedback loop, the two modules co-evolve: **the encoder learns to encode more informative and comprehensive representations, and the decoder learns to generate more faithful and semantically grounded images, creating a self-reinforcing improvement cycle.**

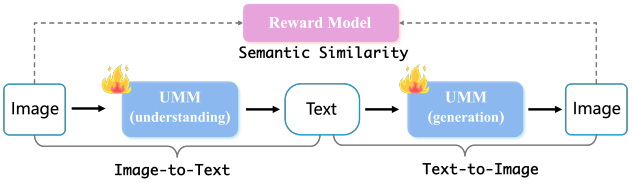

Figure 2. **The overall workflow of our method.** Our post-training method, Unified-GRPO, utilizes the reconstruction objective for improved unified multimodal models (UMMs).

We conduct extensive experiments on visual understanding, generation, and unification tasks across a broad suite of benchmarks to verify that our post-training strategy with our core AE insight can improve the UMMs [3, 13]. Specifically, our method achieves significant improvement on image generation (*e.g.*, from 0.73→0.86 on GenEval [7] and 0.296→0.475 on GenEval++ [42]), and largely improved fine-grained visual recognition and perception capability, *e.g.*, from 0.05→0.45 on small object detection and from 0.15→0.75 on Person ReID of the MMT-Bench [43], consistent with the findings reported by Ross [27]) while maintaining the overall performance across visual understanding tasks. Furthermore, results on the proposed Unified-Bench show that our post-training method can largely improve the unification, resulting in a more coherent information flow between encoding and decoding.

In summary, our work makes the following contributions:

- **A unified Auto-Encoder perspective linking I2T and T2I:** We propose a principled formulation where text serves as the intermediate representation connecting image encoding and decoding, offering a coherent bridge between multimodal understanding and generation.
- **Unified-GRPO, an RL-based post-training framework for cross-modal self-evolution:** Through reconstructive rewards, our method jointly optimizes the encoder and decoder, enabling mutual reinforcement: richer semantic encoding improves generation, and more faithful generation strengthens fine-grained visual perception.
- **Broad applicability and consistent empirical gains:** Unified-GRPO applies to various encoder–decoder multimodal systems, consistently improving text-to-image generation and enhancing fine-grained understanding (e.g., grounding, small-object recognition), while revealing interpretable trade-offs in OCR-heavy scenarios.

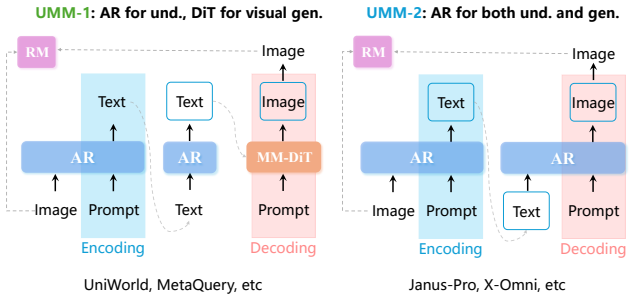

Figure 3. **Illustration of how Unified-GRPO integrates into two representative UMM architectures.** UMM-1 uses an AR model for understanding and an MM-DiT for image generation, while UMM-2 employs a single AR backbone for both understanding and generation. In the diagram, "text" or "image" inside a rectangle denotes latent tokens, whereas those without a rectangle represent raw data inputs. Unified-GRPO can be applied to the LLM backbone in both architectures to provide reconstructive RL.

2. UAE Methodology

Our goal is to unify image-to-text (I2T) understanding and text-to-image (T2I) generation within a single auto-encoding perspective, where text serves as the intermediate latent representation connecting the two directions. Given an input image x, a unified multimodal model (UMM) first produces a semantic description y (I2T), and another UMM reconstructs an image $\hat{x}$ from y (T2I). We then adopt reconstructive reinforcement learning to maximize the semantic similarity between x and $\hat{x}$, enabling mutual improvement between understanding and generation.

2.1. Unified-GRPO

We propose **Unified-GRPO**, a reconstructive reinforcement learning method designed to *unify* image-to-text (I2T) understanding and text-to-image (T2I) generation by training the model to maximize reconstruction fidelity.

To apply this framework to existing unified multimodal models (UMMs), we consider the two dominant architectural families shown in Fig. 3: (1) **UMM-1**, where an autoregressive language model (LLM) is responsible for multimodal understanding and provides language priors for a diffusion transformer (MM-DiT) used in image generation (e.g., UniWorld [13], MetaQuery [18], etc); and (2) **UMM-2**, where a single autoregressive model handles both visual understanding and visual generation in a shared token space (e.g., Janus-Pro [3], X-Omni [6], etc). In both families, the LLM plays a central role—either as the core understanding module (UMM-1) or as the backbone of both understanding and generation (UMM-2). Since GRPO and related RL algorithms have proven highly effective for training LLMs, we extend this idea to UMMs and employ GRPO to optimize the LLM components toward improved cross-modal reconstruction.

Applying Unified-GRPO to UMM-1. For UMM-1, the autoregressive LLM π_ϕ is trained using reconstructive RL, while the diffusion transformer p_θ remains *frozen* and acts as part of the reward environment (together with a CLIP encoder). Given an input image x, we sample a group of G caption sequences $\{y^{(i)}\}_{i=1}^{G}$ from the old policy $\pi_{\phi_{\text{old}}}$. For each caption $y^{(i)}$, we extract its last hidden state $h_T^{(i)}$ and project it into a diffusion condition $c^{(i)} = g(h_T^{(i)})$, which is then used to synthesize a reconstructed image $\tilde{x}^{(i)} \sim p_\theta(\cdot \mid c^{(i)})$. The LLM is updated via the GRPO objective in Eq. (**??**), where each trajectory o_i corresponds to the token sequence of $y^{(i)}$, and the probability ratio is

$$r_t^{(i)}(\phi) = \frac{\pi_\phi(y_t^{(i)} \mid x, y_{<t}^{(i)})}{\pi_{\phi_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})}.$$

This process encourages the LLM to emit hidden representations that maximize the diffusion's reconstruction quality.

Applying Unified-GRPO to UMM-2. For UMM-2, the same autoregressive model performs both I2T and T2I. Unified-GRPO is applied in an identical manner, except that the decoder D_ϕ is now autoregressive rather than diffusion-based. The RL pipeline becomes: $x \xrightarrow{\pi_\phi} y, y \xrightarrow{\pi_\phi} \tilde{x}$, with reconstruction reward $\mathcal{R}(x, \tilde{x}) = \cos(f_{\text{CLIP}}(x), f_{\text{CLIP}}(\tilde{x}))$. This enables a fully AR model to co-evolve its understanding and generation abilities within a single shared token space. The specific implementation here is similar to previous work that incentivizes AR for improved image generation (such as T2I-R1 [10] and AR-GRPO [44]), and the key difference is that we use the reconstruction reward between the input and the generated image embeddings, enabling it to optimize both understanding and generation modules jointly.

2.2. Unified-Bench: A Benchmark tailored for Evaluating the Unified Models

Motivation. As illustrated in Fig. 1, we view *understanding* (I→T) and *generation* (T→I) as a closed loop whose two halves should *mutually enhance* each other. Judging image realism alone or caption fidelity alone cannot reveal whether a system is truly *unified*. We introduce a reconstruction-based similarity, *i.e.*, **Unified-Score**, to directly test whether the semantics extracted during understanding are sufficient for faithful regeneration, and whether regeneration in turn validates the completeness of the understanding.

***Protocol-1*: Evaluation of the unified score from the reconstruction similarity.** To quantify the unified score, we start from 100 diverse source images. The prompt, used to allow the model to generate cpation, is detailed in Supplementary. The same model then synthesizes an image from its *own* caption. We compute unified scores between the reconstruction and the source using four widely adopted vision backbones, CLIP [22], LongCLIP [45], DINO-v2 [17], and

Table 1. **Ablation study on the proposed post-training on understanding, generation, and unification benchmarks.** We apply our method to the two typical unified multimodal models and show the clear improvement.

Model	Understanding		Generation		Unification
	MMB	MMMU	GenEval	DPGBench	Unified-Score
UniWorld	83.5	58.6	84.0	81.2	79.0
w/ Ours	84.8	58.2	89.0	86.4	86.1
vs. Baseline	**+1.3%**	-0.4%	**+5%**	**+5.2%**	**+7.1%**
Janus-pro	79.2	41.0	80.0	84.2	82.8
w/ Ours	80.3	41.6	84.3	88.9	89.1
vs. Baseline	**+1.1%**	**+0.6%**	**+4.3%**	**+4.7%**	**+6.3%**

Table 2. **Protocol-1 of Unified-Bench**: comparing of unified score of different methods on Unified-Bench, the tailored benchmark for evaluating the unification between understanding and generation models in the UMMs.

Method	CLIP	LongCLIP	DINO-v2	DINO-v3	Overall
GPT-4o-Image [16]	90.42	**94.37**	81.74	77.27	85.95
BAGEL [4]	88.97	93.35	78.55	73.05	83.48
BLIP-3o [2]	84.84	90.24	68.31	62.86	76.56
Janus-Pro [3]	88.72	93.45	78.30	70.61	82.77
OmniGen2 [32]	88.36	93.11	77.70	74.07	83.31
Show-o [35]	80.18	86.75	58.20	51.51	69.16
UniWorld-V1 [13]	85.49	91.53	72.12	66.83	78.99
UAE	**90.50**	94.35	**81.98**	**77.54**	**86.09**

Table 3. **Evaluating how "friendly" the output caption is for image generation**. We use the data from Unified-Bench to assess the quality of the captions produced by the understanding model for better text-to-image generation. **Bold** indicates the best result.

Method	CLIP	LongCLIP	DINO-v2	DINO-v3	Overall
Qwen-2.5-VL-3B [1]	88.34	92.62	73.91	70.02	80.85
Qwen-2.5-VL-7B [1]	88.26	92.89	76.12	70.96	81.92
UAE	**90.50**	**94.35**	**81.98**	**77.54**	**86.09**

DINO-v3 [24], and report per-backbone similarities and an overall summary.

***Protocol-2*: Quality Evaluation of the model's *output caption* for reconstruction.** We further evaluate caption quality through pairwise comparisons against various baselines, using four commercial LLM judges: Claude-4.1, GPT-4o, Grok-4, and o4-mini. The prompting strategy is detailed in Supplementary. For evaluation, we use pairwise winning rate (%), the percentage of times our model is preferred over baselines as the main metric.

3. Experiments

3.1. Ablation on Unified-GRPO

To comprehensively evaluate the effectiveness of the proposed Unified-GRPO, we implement our method on the two typical unified multimodal models: UniWorld [13] and Janus-Pro [3], among the understanding, generation, and unification benchmarks. Tab. 1 shows that applying Unified-GRPO to both representative UMM architectures consistently improves their performance. The gains are most significant on *generation* and *unification* metrics, where reconstruction is directly optimized, yielding improvements of $4\sim5\%$ on generation and over 6% on unified reconstruction quality. Understanding performance exhibits only modest gains, which we attribute to the limited capacity of current generation models: imperfect reconstructions can introduce negative feedback to the encoder. Nevertheless, as we will show later (Sec. 3.5), Unified-GRPO can notably enhance fine-grained perceptual abilities, particularly in tasks involving subtle difference recognition and visual grounding via our reconstructive RL training. Since the UniWorld-based model demonstrates stronger performance in both generation and understanding compared to Janus, we adopt this architecture as the primary backbone for all subsequent experiments.

Table 4. **Benchmarking results of text-to-image generation capability on GenEval [8] benchmark**. '†' refers to the methods using the LLM rewriter. **Bold** indicates the best result, and underlined denotes the second best.

Method	Single	Two	Counting	Colors	Position	Color	Overall
Janus Pro [3]	0.99	0.89	0.59	0.90	**0.79**	0.66	0.80
BLIP3-o 8B [2]	-	-	-	-	-	-	0.84
UniWorld-V1 [13]	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1† [13]	0.98	0.93	0.81	0.89	0.74	0.71	0.84
OmniGen2 [32]	**1.00**	0.95	0.64	0.88	0.55	0.76	0.80
X-Omni† [6]	0.98	0.95	0.75	0.91	0.71	0.68	0.83
BAGEL [4]	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL† [4]	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
UAE	**1.00**	0.89	**0.84**	0.90	0.71	0.79	0.86
UAE†	**1.00**	**0.97**	0.82	**0.95**	0.73	**0.84**	**0.89**

Table 5. Comparisons of **challenging instruction following generation ability** with other unified multimodal models on Geneval++ [8]. **Bold** indicates the best result, and underlined denotes the second best.

Method	Color	Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
Janus-Pro [3]	0.450	0.300	0.125	0.300	0.075	0.350	0.125	0.246
T2I-R1 [10]	**0.675**	0.325	0.200	0.350	0.075	0.250	0.300	0.311
BLIP3-o 4B [2]	0.125	0.225	0.100	0.450	0.125	0.550	0.225	0.257
BLIP3-o 8B [2]	0.250	0.250	0.125	**0.600**	0.125	**0.575**	0.225	0.307
OmniGen2 [32]	0.550	0.425	0.200	0.275	0.125	0.250	**0.450**	0.325
Bagel [4]	0.325	0.600	0.250	0.325	0.250	0.475	0.375	0.371
UAE	0.550	0.525	**0.550**	0.550	**0.450**	0.400	0.400	**0.475**

Table 6. **Protocol-2 of Unified-Bench**: evaluating the quality of output caption of our trained understanding model (3B) against different opponents on Unified-Bench, evaluated by four judge models (using official commercial API). We use the metric of **Pairwise winning rate (%)** for evaluation. The **Avg** column reports the mean score across judges.

Opponent	# Param	Our Wining Rate (%)				
		Claude-4.1	GPT-4o	Grok-4	o4-mini	Avg
GPT-4o [16]	-	47.4	89.4	30.6	21.2	47.2
Bagel [4]	7B	57.7	92.9	58.3	48.2	64.3
OmniGen2 [32]	3B	67.9	97.6	63.5	56.5	71.4
Show-o [35]	1.3B	97.8	100.0	89.8	91.0	94.7
Qwen-2.5-VL-3B [1]	3B	76.3	99.0	67.0	63.0	76.3
Qwen-2.5-VL-7B [1]	7B	68.8	99.0	62.0	56.0	71.5

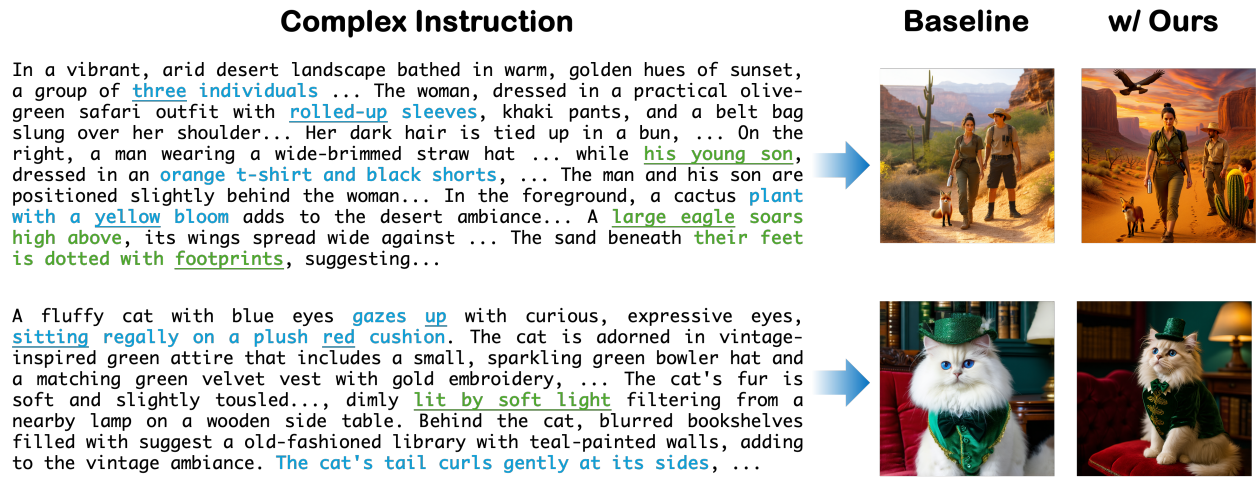

Figure 4. **Qualitative results on the complex and long-context generation.** Our method can recover very detailed semantics from the highly descriptive input caption over the baseline, demonstrating that improved understanding can notably benefit generation.

Figure 5. **Reconstruction results *vs.* RL training steps.** With the RL steps increasing, the understanding model (encoder) achieves better perception capability to produce an informative, detailed, yet accurate description to reconstruct the original image comprehensively; while the generation model (decoder) can take the rich description as input for recovering all semantics faithfully.

3.2. Unification Evaluation

We assess the unified degree with the proposed UnifiedBench. Tab. 2 shows that our **UAE** achieves the best **Overall** unified score (86.09), surpassing GPT-4o-Image (85.95). Specifically, UAE obtains the top results on CLIP (90.50), DINO-v2 (81.98), and DINO-v3 (77.54), and statistical parity on LongCLIP (94.35 vs. 94.37). These consistent gains across contrastive (CLIP-family) and self-supervised (DINO-family) features suggest that our **UAE** framework can preserve layout- and texture-level semantics that translate into more faithful reconstructions.

3.3. Text-to-Image Generation Evaluation

We evaluate **UAE** on two standard benchmarks: GenEval and its improved version GenEval++, which probe compositional understanding and instruction-following in increasingly challenging settings. More evaluations are in the Supplementary.

GenEval. As shown in Tab. 4, without considering LLM rewriting, our **UAE** attains the best *Overall* score among unified models (**0.86**). It leads on *Counting* (0.84) and *Color attribution* (0.79; **+16** points vs. Bagel's 0.63 and +3 vs. OmniGen2's 0.76), co-leads on *Colors* (0.90), is second-best on *Position* (0.71), and reaches 0.89 on *Two object* (below the strongest 0.94–0.95). When considering LLM rewriting, e.g., using the same rewritten prompts with Bagel, our UAE achieves an overall score of 0.89 on average, demonstrating the SOTA performance in image generation.

GenEval++ (harder compositional control). GenEval++ [42] extends GenEval to prompts with *three or more* objects, each bearing distinct attributes and spatial relations, demanding comprehensive, multi-constraint satisfaction. In Tab. 5, **UAE** achieves the best

Figure 6. **Qualitative examples on GenEval++.** Under the complex and anti-realistic cases, our method demonstrates a clear advantage in multi-attribute instruction-following generation over other methods.

Figure 7. **Qualitative examples showing how reconstruction-driven RL improves image-to-text understanding.** Compared to the baseline, our model better identifies subtle differences and performs accurate visual grounding, demonstrating that reconstruction-driven RL encourages richer and more precise semantic extraction in image-to-text understanding.

Table 7. **High-level meta-tasks evaluation results on the comprehensive multimodal understanding benchmark:** MMT-Bench [43]. Accuracy is the metric, and the Overall score is computed as the mean of all displayed subtasks.

Model	Overall	VR	Loc	Count	HLN	VC	VG	AR	PLP	I2IT	RR	Emo	VI	OCR	DU	IR	3D
Frequency Guess	32.3	30.0	28.2	28.2	43.4	28.2	29.1	30.0	29.4	30.8	33.5	30.1	52.1	30.4	37.6	29.9	26.5
Random Guess	27.9	27.1	28.1	25.0	41.6	25.0	24.8	26.6	21.2	33.4	10.5	25.4	50.8	27.2	30.3	24.3	25.5
InternVL-Chat-v1.2-34B	58.7	81.3	59.4	66.4	82.4	82.3	49.4	52.6	37.4	32.8	55.0	48.7	61.5	60.5	68.3	56.3	45.5
Qwen-VL-Plus	56.8	82.6	55.3	61.1	69.9	86.5	43.6	53.4	43.1	37.8	53.0	41.6	50.3	65.6	77.3	40.7	46.5
GPT-4V	54.1	85.3	55.6	51.6	69.6	80.3	25.0	47.7	48.2	31.8	52.5	45.1	47.9	68.0	69.8	44.9	42.0
GeminiProVision	56.2	84.7	43.6	56.4	65.9	80.1	33.0	57.4	40.3	31.5	58.5	55.2	47.5	59.5	71.6	68.4	45.2
DeepSeek-VL-7B	48.0	75.6	42.0	44.5	60.6	69.1	38.4	44.8	38.3	23.5	48.8	43.8	47.7	61.1	51.9	30.5	47.2
Claude3V-Haiku	47.4	74.3	44.8	51.1	63.6	67.6	26.9	46.2	35.5	22.8	50.0	35.2	42.9	54.4	69.8	34.6	38.2
ShareGPT4V-7B	47.8	74.2	36.0	50.9	62.4	71.6	35.4	46.2	39.2	21.8	59.8	44.3	54.5	47.8	47.9	27.8	45.2
LLaVA-v1.5-7B	46.1	72.8	34.3	47.5	61.6	68.1	34.0	46.6	36.0	22.2	58.0	42.5	57.6	45.0	40.8	26.1	44.8
Qwen-2.5-VL-3B	56.3	78.7	40.3	42.8	72.5	83.6	46.2	53.0	40.8	32.5	71.3	47.5	48.4	75.0	70.0	56.8	42.5
Ours (Qwen-3B)	56.5	80.1	47.3	44.7	72.8	84.1	47.1	53.5	46.6	32.7	71.3	48.3	57.6	68.8	58.4	50.6	40.0
vs. Baseline	**+0.2**	**+1.4**	**+7.0**	**+1.9**	**0.3**	**+0.5**	**+0.9**	**+0.5**	**+5.8**	**+0.2**	**+0.0**	**+0.8**	**+9.2**	**-6.2**	**-11.6**	**-6.2**	**-2.5**

Table 8. **Evaluation results on fine-grained visual perception oriented sub-tasks** on MMT-Bench [43]. Accuracy is the metric, and the Overall score is computed as the mean of all displayed subtasks. We show notable improvements across various fine-grained understanding tasks, highlighting the positive impact of generation on understanding.

Model	Overall	Fine-grained Visual Recognition					Color and Geometry Perception				
		Salient Obj. Detection RGBD	Transparent Object Det.	Small Object Detection	Rotated Object Detection	Person Re-ID	Color Constancy	Color Assimilation	Geometrical Relativity	Geometrical Perspective	Polygon Localization
InternVL-Chat-V1.2-34B	63.4	28.5	66.5	64.5	46.7	60.0	34.5	44.5	82.5	75.0	46.1
Qwen-VL-Plus	62.3	44.5	47.5	59.5	60.0	50.5	47.5	29.0	58.3	43.0	63.8
GPT-4V	62.0	42.0	56.5	52.0	79.0	49.0	65.0	24.7	43.3	35.7	66.0
GeminiProVision	61.6	45.0	38.5	43.0	50.0	72.5	38.9	53.5	46.0	43.3	36.0
DeepSeek-VL-7B	53.2	40.0	53.5	43.5	36.7	32.5	27.5	52.0	54.2	56.0	23.4
Claude3V-Haiku	52.2	43.0	19.5	44.0	46.7	35.0	38.5	58.5	55.8	56.5	66.7
ShareGPT4V-7B	51.5	40.5	39.0	37.5	27.8	24.0	52.8	26.5	60.0	65.8	32.0
LLaVA-v1.5-7B	49.5	37.5	40.0	31.5	30.0	23.0	56.9	28.0	64.0	70.0	34.0
Frequency	31.7	26.0	26.0	27.5	28.9	30.0	52.8	51.0	50.5	53.3	31.5
Random	28.5	28.5	29.0	27.0	24.4	26.0	48.6	50.0	50.5	51.7	27.5
Qwen-2.5-VL-3B	32.5	25.0	15.0	5.0	33.3	15.0	28.6	50.0	60.0	58.3	35.0
Ours (Qwen-3B)	56.9	45.0	45.0	45.0	55.6	75.0	42.9	60.0	65.0	75.0	60.0
vs. Baseline	**+24.4**	**+20**	**+30**	**+40**	**+22.3**	**+60**	**+14.3**	**+10**	**+5**	**+16.7**	**+25**

Overall score (**0.475**), leading on *Color/Count* (0.550) and *Pos/Count* (0.450), with runner-up performance on *Color/Pos* (0.550) and *Multi-Count* (0.400). Qualitative visualizations in Fig. 5 further show accurate attribute binding, disambiguation across multiple entities, and robust position–count consistency under long prompts. This highlights that our method can achieve notable improvement in complex instruction following.

3.4. Image-to-Text Understanding Evaluation

Here, we conduct experiments to verify that after our post-training, the encoder can achieve improved image-to-text, in terms of caption quality, and greater "generation-friendly".

Caption quality evaluation by commercial LLMs. As shown in Tab. 6, our understanding model (using Qwen-2.5-VL-3B as the baseline) attains high average win rates: **94.7** vs. Show-o, **71.4** vs. OmniGen2, **64.3** vs. Bagel, and **76.3/71.5** vs. Qwen-2.5-VL (3B/7B), while remaining competitive with GPT-4o (47.2). The cross-judge agreement suggests our captions improve along multiple axes, completeness, attribute binding, relational, and spatial fidelity.

Improving the understanding model as a better captioner suitable for generation. Under the Unified-Bench "caption→generate→compare" protocol, captions produced by our trained understanding model yield the highest reconstruction similarity across all four backbones (Tab. 3): **90.50** (CLIP), **94.35** (LongCLIP), **81.98** (DINO-v2), **77.54** (DINO-v3), with **86.09** Overall. These results indicate that the caption generated by our understanding model is more suitable for generation.

3.5. Evaluation on the Understanding Benchmark.

We evaluate on *MMT-Bench* [43], which comprises high-level meta-tasks[1] The overall score remains essentially unchanged with a marginal improvement over the baseline (+0.2%; Tab. 7). However, if we zoom in to observe *fine-grained visual recognition* suite (Tab. 8), the benefits of our generation-augmented training for perception become pronounced: we observe large absolute gains in Small Object Detection (+40.0%) and Person Re-ID (+60.0%), yielding a +24.4% increase in the fine-grained overall. These results indicate that generation does not harm understanding, but can instead **enhance fine-grained visual perception capability.**

[1]The tasks include VR (Visual Recognition), Loc (Spatial Localization), OCR (Text Reading), Count (Object Counting), HLN (Hallucination), IR (Image Retrieval), 3D, VC (Visual Caption), VG (Visual Grounding), DU (Document Understanding), AR (Action Recognition), PLP (Pixel-Level Perception), I2IT (Image-to-Image Translation), RR (Relation Reasoning), Emo (Emotion), and VI (Visual Illusion).

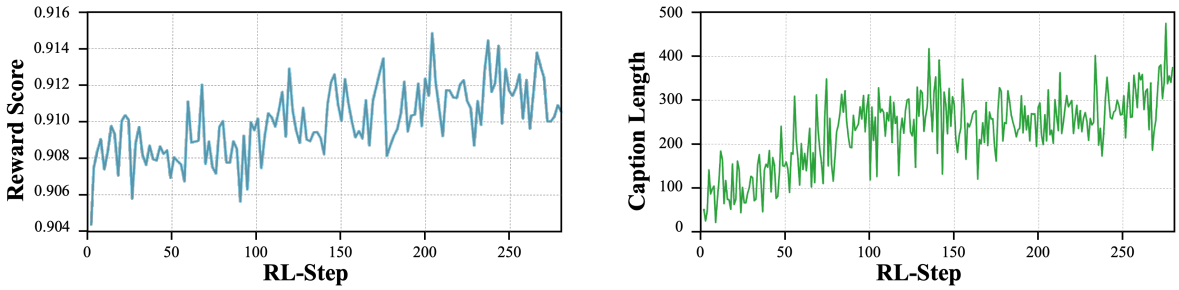

Figure 8. **Training dynamics of our reconstruction-oriented RL stage. Left:** The reward score steadily increases as the policy learns to generate captions that more faithfully preserve the visual information in the input image. **Right:** The caption length gradually grows throughout training, indicating that the model is producing richer and more detailed textual descriptions. Together, these trends show that the RL optimization encourages the model to encode progressively more complete image information into text, ensuring that the downstream decoder receives a maximally informative representation.

3.6. Case-Study in Fine-Grained Visual Perception

Spotting Subtle Differences. Fig. 7 (top) presents a challenging visual comparison task, where the baseline model fails to detect the fine difference between two images: one image shows *two* men standing near the car, while in the other image *only one* man is visible. The baseline incorrectly answers option A due to missing the subtle change. In contrast, our model, trained with Unified-GRPO, provides a detailed analysis of both images, accurately recognizing the presence and position of each person, the vehicle's location, the outdoor parking setting, and the contextual cues.

Visual Object Grounding. Fig. 7 (bottom) demonstrates another demanding task requiring precise grounding. The input instruction asks the model to caption the scene *and identify the corresponding region ID* for "a skier in yellow, blue, orange, and pink clothing." The baseline generates a generic caption and completely ignores the grounding instruction. After our training, the model not only follows the instruction faithfully but also grounds the described skier to the correct region ID by identifying the color composition of the outfit and matching it to the labeled bounding box.

4. Related Work

Recent advancements in multimodal AI have led to the development of Unified Multimodal Models (UMMs) [46]. The architectural designs of current UMMs can be broadly categorized into two paradigms: **(1) AR-based Approaches**: In this setup, all modalities, including images and text, are tokenized and processed sequentially using an autoregressive transformer. Systems like Chameleon and EMU generate image tokens akin to language modeling by predicting the next token in a sequence [3, 11, 20, 25, 31, 33]. An evolution of this idea is seen in Show-o [35], which enhances token prediction with a discrete diffusion mechanism, introducing a structured denoising process during generation. **(2) Hybrid AR-Diffusion Architectures:** Some models combine autoregressive modeling with diffusion-based image synthesis [41]. For instance, Transfusion and similar systems [4, 15, 23, 38, 47] extend a shared transformer backbone with a dedicated diffusion or flow-matching head for high-fidelity image generation. Alternatively, other approaches freeze a pre-trained MLLM and use learnable query modules or MLPs to extract and route intermediate representations to an external image generator [2, 13, 18]. A more recent direction integrates standard autoregressive language processing with masked-autoregressive reconstruction for visual data. MAR [12] enables image generation without relying on vector quantization, instead reconstructing patches in a flexible order. This approach has been adopted in models such as Harmon [5, 28, 34]. Meanwhile, some works [2, 6] use a discretized SigLIP [26] to convert images into tokens, training a single autoregressive model over these visual and language tokens, while employing a diffusion model for the final image decoding. Similar post-training works [30, 37] based on reconstruction demonstrate that using the dense image feature as the "rich text" condition for training diffusion models, which improves image generation. Additionally, RL-based frameworks have been proposed to enhance multimodal learning [14, 19, 39, 40].

5. Conclusion

We show that an auto-encoder can serve as a foundational architecture for unifying image-to-text understanding and text-to-image generation. This paradigm leverages text as a shared intermediate latent representation. By introducing Unified-GRPO, we jointly optimize both, creating a synergistic feedback loop, enabling the auto-encoder principle to benefit both understanding and generation tasks simultaneously. This simple yet powerful design yields stronger fine-grained visual perception, richer semantic encoding, and improved complex instruction-following capability. Our findings highlight the value of treating multimodal tasks not as isolated objectives but as mutually reinforcing components of a unified system, paving the way for more coherent and synergistic multimodal learning.

Acknowledgment. This work was supported in part by the Natural Science Foundation of China (No. 62332002, 62425101), and Shenzhen Science and Technology Program (KQTD20240729102051063)

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[2] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 2, 4, 8

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 3, 4, 8

[4] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 4, 8

[5] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025. 8

[6] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. 3, 4, 8

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 2

[8] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[9] Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. *arXiv preprint arXiv:2203.11931*, 2022. 2

[10] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 3, 4

[11] Hao Li, Yanhao Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. Freestyleret: retrieving images from style-diversified queries. In *European Conference on Computer Vision*, pages 258–274. Springer, 2024. 8

[12] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 8

[13] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2, 3, 4, 8

[14] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 8

[15] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. 8

[16] OpenAI. Gpt-4o. `https://openai.com/index/introducing-4o-image-generation`, 2025. 4

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[18] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2, 3, 8

[19] Caiyong Piao, Zhiyuan Yan, Haoming Xu, Yunzhen Zhao, Kaiqing Lin, Feiyang Xu, and Shuigeng Zhou. Towards policy-adaptive image guardrail: Benchmark and method. *arXiv preprint arXiv:2603.01228*, 2026. 8

[20] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 8

[21] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[23] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 8

[24] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4

[25] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 8

[26] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 8

[27] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024. 2

[28] Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*, 2025. 8

[29] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2

[30] XuDong Wang, Xingyi Zhou, Alireza Fathi, Trevor Darrell, and Cordelia Schmid. Visual lexicon: Rich image features in language space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19736–19747, 2025. 8

[31] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025. 2, 8

[32] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 4

[33] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024. 8

[34] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 8

[35] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 4, 8

[36] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation, 2024. 2

[37] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025. 8

[38] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 8

[39] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Wang, Weiyun Ye, Shihao Geng, Yiren Zhao, Jiaming Li, Cunjian Li, Hang Sun, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. 8

[40] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 8

[41] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025. 2, 8

[42] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 2, 5

[43] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024. 2, 7

[44] Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025. 3

[45] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 3

[46] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025. 8

[47] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2, 8