REASONUNI: REASONING-AUGMENTED RECONSTRUCTION ADVANCES UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-text (I2T) understanding and text-to-image (T2I) generation are dual tasks in unified multimodal models (UMMs). Reconstructive post-training couples these tasks, but existing reconstruction-based methods remain perception-centric and provide limited supervision for reasoning about structure, intent, and transformation. We propose **`ReasonUni`**, a reasoning-augmented reconstructive post-training framework that extends this coupling beyond perception. Given an original image and an editing instruction, the UMM first reasons about *what* and *how* to edit and describes the target state, then generates the corresponding target image conditioned on this understanding output. To jointly optimize both modules, we introduce **`TreeGRPO`**, a tree-structured reinforcement learning method that samples multiple reasoning paths and candidate images per path, then aggregates visual rewards to update both modules under a unified objective. Experiments across multiple benchmarks demonstrate consistent gains in both understanding and generation, showing that reasoning-augmented reconstruction yields genuine mutual gains for UMMs.

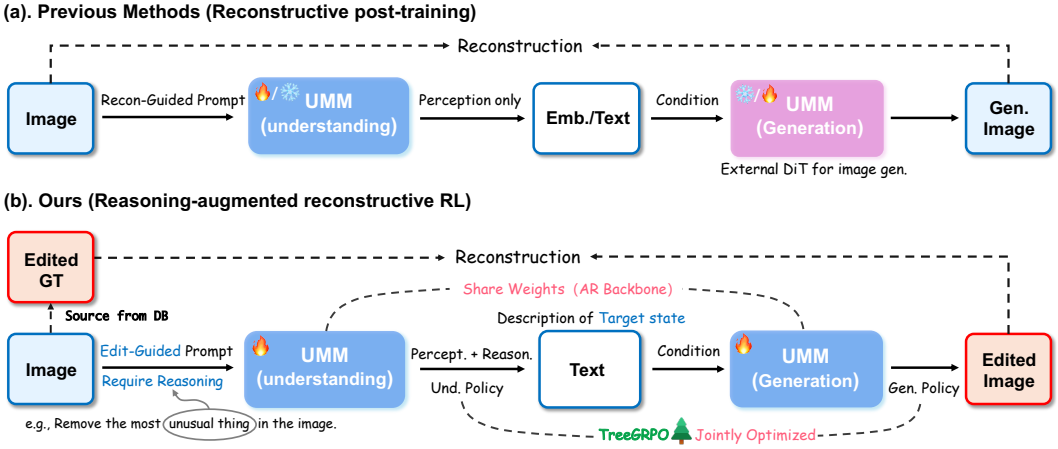


Figure 1: **From perception-only to reasoning-augmented reconstruction.** (a) Previous methods reconstruct the observed image. (b) ReasonUni infers an edited target state and uses TreeGRPO to jointly optimize understanding and generation through reconstruction feedback.

1 INTRODUCTION

Unified Multimodal Models (UMMs) Wang et al. (2024b); Chen et al. (2025c); Wu et al. (2025a); Xie et al. (2024); Pan et al. (2025); Tong et al. (2025b); Zhou et al. (2024); Yan et al. (2025b) have been proposed to both understand and generate across modalities–generating visual content and text within a single unified architecture. However, naively coupling understanding and generation often yields suboptimal results: prior UMM studies Pan et al. (2025); Chen et al. (2025a) report that optimizing diffusion-based generative objectives can impair understanding performance and learned representations (and vice versa).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

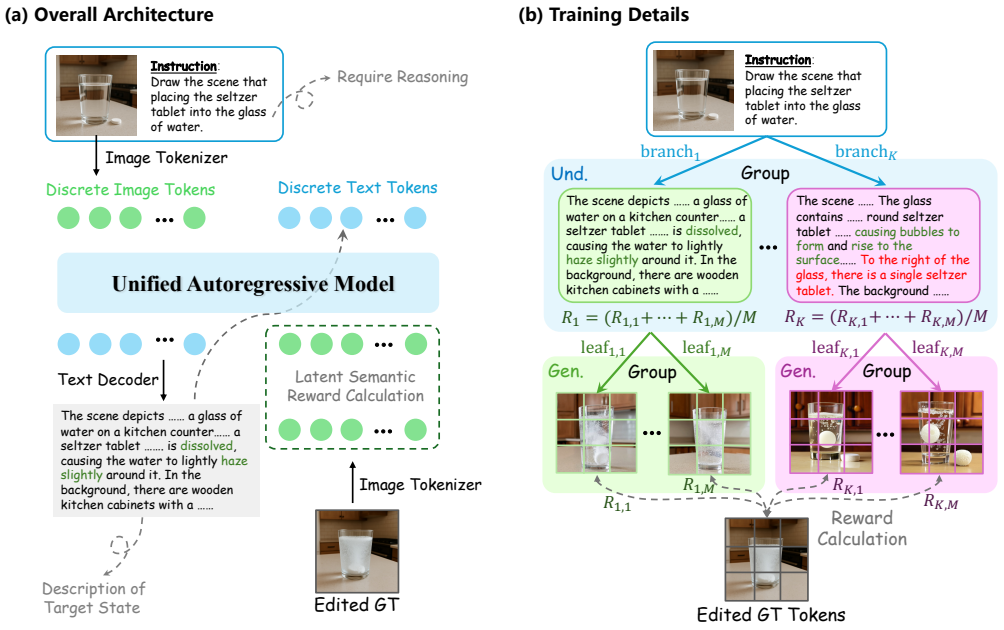


Figure 2: Overview of the ReasonUni architecture and method (left). Given an input image and reasoning-required editing instruction, the UMM infers and describes the target state, then generates candidate edited images conditioned on these descriptions. TreeGRPO formulates this process as a hierarchical branch–leaf structure: reasoning paths form branches, generated images form leaves, and visual rewards are aggregated and propagated to optimize understanding and generation jointly (right).

To address these conflicts, Janus Wu et al. (2025a) separates visual encoders for understanding and generation, while TokenFlow Qu et al. (2025) aligns semantic and pixel-level codebooks through shared indices. However, architectural unification alone does not establish mutual gains between understanding and generation, motivating training strategies that explicitly couple the two tasks.

Rather than redesigning the architecture, recent works have explored an alternative direction: dedicated *post-training* strategies to advance UMMs and explore potential mutual benefits Chen et al. (2025b); Su et al. (2026); Yan et al. (2025a); Xie et al. (2025a). One promising instantiation is to frame UMMs as auto-encoders and post-train both modules with a reconstruction objective Yan et al. (2025a); Xie et al. (2025a). To illustrate, RecA Xie et al. (2025a) uses embeddings from the visual understanding module as dense captions to condition text-to-image generation, yielding consistent gains across benchmarks. UAE Yan et al. (2025a) models the UMM as an ”Image–Text–Image” auto-encoder: the encoder captions visual details, and the decoder reconstructs the image from the caption. By optimizing reconstruction, UAE improves generation quality and also strengthens perceptual understanding of fine-grained details.

However, the gains in understanding are still limited at the ”perception” level Yan et al. (2025a); Wang et al. (2024a), as the prior reconstruction frameworks emphasize only recognizing what is present, without requiring the model to reason anything about structure, intent, or causal relations in the scene. As a result, their ability to generalize to complex, open-ended tasks that require higher-order visual reasoning remains limited.

To bridge this gap, we propose ReasonUni, a reasoning-augmented reconstructive post-training method that extends supervision beyond perception to visual reasoning. Given an input image and an editing instruction, the UMM first reasons about the required transformation and produces a detailed target description of the desired edits. The generation module then reconstructs the target image conditioned on the description.

To jointly optimize both modules, we propose TreeGRPO, a hierarchical tree-based framework that enables cleaner credit assignment between understanding and generation. Each sampled reasoning path forms a branch with multiple generated images as leaves. Aggregating their rewards provides a more reliable branch-level signal for understanding, while within-branch normalization guides

generation under a fixed target-state description. Applying Group Relative Policy Optimization (GRPO) Shao et al. (2024) at these two levels jointly optimizes both modules under a unified objective, rather than optimizing them in separate stages Yan et al. (2025a).

To make this joint optimization practical, we adopt a more unified version of the UMM series (i.e., X-Omni Geng et al. (2025) style) with a pure AR backbone that generates both image and text tokens, so image and text generation are optimized under one AR objective. In contrast, note that prior work like UAE mainly uses a UniWorld-style Lin et al. (2025) split design (AR for text with an external diffusion image generator), and only the AR component is directly optimized.

Experiments demonstrate consistent gains in both understanding and generation across multiple benchmarks. Compared with UAE, ReasonUni delivers larger gains in both capabilities, outperforming it by up to 5.98 points on DPGBench, 13.00 on WISE, and 16.78 on GenEval++. On MMBench, ReasonUni improves both perception and reasoning, with logical reasoning gaining 6.77 points over X-Omni versus 3.38 for UAE. These understanding gains extend across all 12 evaluated benchmarks, supporting the broad effectiveness of our approach.

2 METHOD

2.1 PRELIMINARIES: UNIFIED DISCRETE AUTOREGRESSIVE MODELING

We build on a unified discrete autoregressive (AR) multimodal model (UMM), instantiated with X-Omni Geng et al. (2025), with parameters θ. The key design is a shared *discrete* token space for both image and text. Specifically, an image is quantized into visual tokens by a visual tokenizer, and text is represented as language tokens. Importantly, visual inputs (for understanding) and visual outputs (for generation) use the same semantic codebook, so both directions are aligned at the representation level.

Due to this unified and simplified architecture design, the UMM can be trained with standard next-token AR likelihood. A separate decoder can map visual tokens back to RGB, but semantic layout and content are modeled in the latent token space by AR. Under this formulation, one backbone *naturally* supports two conditional policies: understanding $p_\theta(T \mid V)$ and generation $p_\theta(V \mid T)$. This dual-policy view is the foundation of our method: by chaining the two policies in sequence, we can construct training signals that flow from image-level feedback back to the reasoning process, enabling true end-to-end joint optimization.

2.2 REASONUNI: REASONING-AUGMENTED RECONSTRUCTIVE FRAMEWORK FOR UMMS

Conventional reconstructive post-training Yan et al. (2025a); Xie et al. (2025a) improves semantic completeness for UMMs, but often emphasizes perception more than visual reasoning. In practice, the model may learn to preserve or copy content, while receiving weak supervision on *what to change*, *how to change it*, and *what to keep unchanged*, leaving the model's higher-order visual reasoning abilities largely unsupervised.

To address this, we propose **ReasonUni**, which turns reconstruction into a reasoning-driven editing task: the model must interpret an editing instruction, reason about its implications, and produce a coherent target description before generating the edited image. For each training tuple (x_{src}, c, x_{gt}), where x_{src} is the source image, c is an edit instruction, and x_{gt} is the target image, we apply two coupled steps.

Step 1: Understanding. The understanding module first predicts a textual description y of the desired target state: $y \sim p_\theta(\cdot \mid x_{src}, c)$. To produce y, the module must infer the edit required by the instruction, including implicit targets and transformations, while perceiving the objects, attributes, and spatial relations in the source image that should remain unchanged. The resulting description therefore integrates the module's reasoning about the intended edit with its perception of the source scene. The model directly outputs a description of the edited scene rather than an explicit chain-of-thought trace; the prompting template is provided in Supplementary Section 5.

Step 2: Generation. The generation module then synthesizes an edited image conditioned on the target-state description: $\hat{x} \sim p_\theta(\cdot \mid y)$. Because generation receives only y (not the source image directly), faithful reconstruction requires the description to capture both the intended edit and the scene content to be preserved.

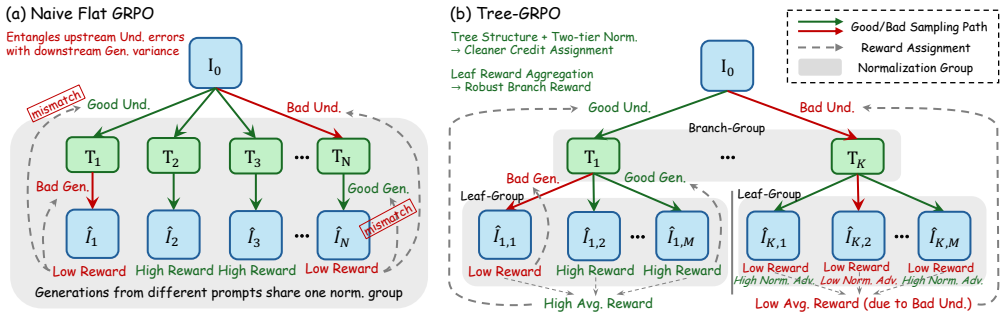


Figure 3: Flat vs. tree-structured sampling. **Left (Flat):** each reasoning path and its generated image form an independent sample. **Right (Tree):** each of K reasoning paths branches into M generated images, enabling group-relative comparisons across branch-level mean rewards and among image rewards within each branch. Even when a branch has a low mean reward, images scoring above that branch mean receive positive generation advantages.

This coupling is central: image-level feedback supervises not only generation quality but also the quality of the upstream reasoning signal, so both components receive learning signals accordingly.

2.3 TreeGRPO: Tree-based Joint Group Relative Policy Optimization

To jointly optimize understanding and generation, we introduce **TreeGRPO**. Standard GRPO Shao et al. (2024) samples multiple outputs for a single input and uses relative comparisons to compute advantages. However, our two-stage pipeline introduces variation in both the target-state description and the image generated from it. A flat sampling strategy would conflate these two sources, making credit assignment difficult.

As illustrated in Figure 3, TreeGRPO organizes candidate outputs into a tree, with reasoning paths as branches and generated images as leaves. Comparing mean rewards across branches supervises understanding, while comparing individual image rewards within each branch guides image generation under a fixed target-state description.

Hierarchical sampling. Given (x_{src}, c, x_{gt}), we sample K reasoning paths from the understanding module and M edited image candidates from the generation module for each path:

$$y_k \sim p_\theta(\cdot \mid x_{src}, c), \quad \hat{x}_{k,j} \sim p_\theta(\cdot \mid y_k), \quad k = 1, \ldots, K,\ j = 1, \ldots, M. \tag{1}$$

This produces a tree with K branches and $K \times M$ leaves. The two-level sampling separates uncertainty at two levels: *which reasoning path is best* and *how well each path can be rendered*, with multiple images per path providing a robust estimate of each path's quality.

Latent semantic reward. Unlike UAE Yan et al. (2025a), which computes reconstruction rewards using CLIP features of decoded images, we instead compute reconstruction rewards directly in X-Omni's semantically aligned visual token space, which effectively captures high-level semantic similarity while remaining lightweight and efficient.

Let $q(\cdot)$ denote the visual tokenizer and $e(v_i)$ the codebook embedding associated with visual token v_i. SigLIP-VQ (Geng et al., 2025) learns a visual codebook by quantizing contextualized patch features from a SigLIP encoder. Positional embeddings and self-attention incorporate spatial context into these features before quantization, allowing the learned token representations to encode spatial cues alongside semantic content. To compute the reward, we mean-pool the codebook embeddings of the generated tokens and the tokenized target image, then measure the cosine similarity between the two representations. The pooled representations can thus retain implicit spatial cues even after this aggregation. Operating directly in the model-native latent space also avoids decoding tokens into pixels and re-encoding them with an external vision encoder, thereby reducing information loss and representation mismatch. Let $\hat{\mathbf{v}}_{k,j}$ denote the generated visual token sequence corresponding to $\hat{x}_{k,j}$.

$$\begin{aligned} h(\mathbf{v}) &= \frac{1}{L}\sum_{i=1}^{L} e(v_i), \qquad \mathbf{v} = (v_1, \ldots, v_L), \\ R_{k,j} &= \mathrm{CosSim}(h(\hat{\mathbf{v}}_{k,j}), h(q(x_{gt}))) \,. \end{aligned} \tag{2}$$

Group-relative advantages. The key insight is that the two modules should be evaluated at different granularities: the understanding module is rewarded based on the average quality of all images from its reasoning path, while the generation module is compared against siblings under the same path. We aggregate leaf rewards into branch scores, then standardize at two levels:

$$S_k = \frac{1}{M}\sum_{j=1}^{M} R_{k,j}, \qquad A_k^{und} = \frac{S_k - \mu^{und}}{\sigma^{und} + \epsilon}, \qquad A_{k,j}^{gen} = \frac{R_{k,j} - \mu_k^{gen}}{\sigma_k^{gen} + \epsilon}. \tag{3}$$

Here $(\mu^{und}, \sigma^{und})$ are computed over branch scores $\{S_k\}_{k=1}^{K}$, while $(\mu_k^{gen}, \sigma_k^{gen})$ are computed over candidate rewards $\{R_{k,j}\}_{j=1}^{M}$ within branch k. This two-level normalization ensures each module is compared only against alternatives at its own level of the hierarchy.

Unified tree objective. Both modules are jointly optimized under a single objective, with each branch contributing an understanding term and a generation term:

$$\mathcal{J}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\left[\mathcal{J}_{und}(y_k; A_k^{und}) + \lambda_{gen}\frac{1}{M}\sum_{j=1}^{M}\mathcal{J}_{gen}(\hat{x}_{k,j} \mid y_k; A_{k,j}^{gen})\right], \tag{4}$$

where λ_{gen} controls the relative weight of generation updates (set to 1 in our experiments).

Both objectives use token-level GRPO with KL regularization to reference policy π_{ref}:

$$\begin{aligned}\mathcal{J}_{token}(\theta) = \mathbb{E}_t\Big[&\min\big(r_t A_t, \mathrm{clip}(r_t, 1-\epsilon_c, 1+\epsilon_c)A_t\big) \\ &- \beta D_{KL}\big(\pi_\theta(\cdot \mid o_{<t}) \,\|\, \pi_{ref}(\cdot \mid o_{<t})\big)\Big],\end{aligned} \tag{5}$$

where $r_t = \frac{\pi_\theta(o_t|o_{<t})}{\pi_{\theta_{old}}(o_t|o_{<t})}$. Here o_t is a token from the shared vocabulary, and A_t is assigned from A_k^{und} for text tokens in y_k or from $A_{k,j}^{gen}$ for visual tokens in $\hat{\mathbf{v}}_{k,j}$. The clipping mechanism and KL penalty jointly maintain training stability.

3 Experiments

We evaluate whether reasoning-augmented reconstruction improves understanding and generation jointly, rather than trading one capability for the other. Our experiments address three questions: (Q1) Can ReasonUni achieve simultaneous improvements in both understanding and generation across different initialization settings? (Q2) Are the gains robust across diverse benchmark families? (Q3) Which aspects of understanding and generation contribute to the gains?

We train on a mixture of existing editing datasets and synthetic reasoning-driven editing data. Visual understanding is evaluated using VLMEvalKit (Duan et al., 2024) with a decoding temperature of 0. Details of data synthesis, training configurations, and evaluation protocols are provided in Supplementary Sections 1–2.

3.1 Main Quantitative Results

To answer Q1, Table 1 compares ReasonUni with X-Omni and UAE Yan et al. (2025a) from the X-Omni-SFT and X-Omni-RL checkpoints, which we refer to as the SFT and RL settings, respectively. UAE is reimplemented on X-Omni for both settings, with implementation details and matched-data comparison settings provided in Supplementary Section 2.1. ReasonUni achieves simultaneous gains in understanding and generation across both initialization settings, outperforming X-Omni and UAE on all seven reported metrics. In the SFT setting, generation improves over X-Omni by 13.34 points on DPGBench, 13.00 on WISE, and 15.71 on GenEval++, alongside gains of 0.76–4.00 points on the four understanding benchmarks. The advantage also holds across all seven metrics in the matched-data control (UAE†), indicating that the gains are not simply due to training data. The same joint improvement holds in the RL setting, including gains over X-Omni of 1.65 on MMBench, 2.00 on MMVP, 1.75 on CV-Bench, and 6.79 on GenEval++. Together, these results establish consistent mutual gains across initialization settings, rather than improving one capability at the expense of the other.

Table 1: **Quantitative comparison on multimodal understanding and generation.** We benchmark our method against X-Omni and UAE Yan et al. (2025a) under the SFT and RL initialization settings. The gray † row uses the same training data as ReasonUni, while the unmarked UAE rows use UAE's original training data. **Bold** denotes the best results. Gain indicates the absolute improvement over the corresponding X-Omni baseline.

Method	Understanding				Generation		
	MMBench	MMVP	CV-Bench	MathVerse	DPGBench	WISE	GenEval++
X-Omni-SFT	74.40	48.67	67.98	17.01	65.25	24.00	7.50
+ UAE	74.40	50.67	67.06	17.51	72.61	24.00	6.43
+ UAE†	75.26	51.33	66.87	17.64	65.49	26.00	6.43
+ Ours	**75.60**	**52.67**	**68.85**	**17.77**	**78.59**	**37.00**	**23.21**
Gain (↑)	+1.20	+4.00	+0.87	+0.76	+13.34	+13.00	+15.71
X-Omni-RL	74.31	50.67	68.26	17.39	82.29	45.00	40.00
+ UAE	75.43	50.00	68.05	17.51	83.32	43.00	38.57
+ Ours	**75.96**	**52.67**	**70.01**	**18.40**	**83.58**	**48.00**	**46.79**
Gain (↑)	+1.65	+2.00	+1.75	+1.01	+1.29	+3.00	+6.79

Table 2: **Comparison with state-of-the-art unified models.** We conduct a comprehensive evaluation across six domains, including **General** (MMBench Liu et al. (2023), SEED-Bench Li et al. (2023a), MMStar Chen et al. (2024), MMT Ying et al. (2024)), **Vision-Centric** (VStar Wu & Xie (2024), MMVP Tong et al. (2024b), CV-Bench Tong et al. (2024a)), **Hallucination** (POPE Li et al. (2023b), HallusionBench Guan et al. (2024)), **OCR** (OCRBench Liu et al. (2024)), **Knowledge** (MMMU) Yue et al. (2024), and **Math** (MathVerse Zhang et al. (2024)). – denotes results not reported in original papers.

Method	General				Vision-Centric			Hallucination		OCR	Knowledge	Math
	MMB	SEED	MMStar	MMT	VStar	MMVP	CVB.	POPE	Hallu.	OCRB.	MMMU	MathV.
Show-o Xie et al. (2024)	47.9	–	–	–	–	50.0	–	73.8	–	–	25.1	–
Emu3 Wang et al. (2024b)	58.5	68.2	–	–	51.8	36.6	–	85.2	–	68.7	31.6	–
VILA-U Wu et al. (2024b)	66.6	57.1	–	–	38.7	22.0	–	85.8	–	–	32.2	–
Janus Pro Chen et al. (2025c)	79.2	72.1	–	–	–	–	–	87.4	–	59.5	41.0	–
MetaMorph Tong et al. (2025b)	75.2	71.8	–	–	44.0	48.3	–	–	–	–	41.8	–
Mogao Liao et al. (2025)	75.0	74.6	–	–	–	–	–	88.9	–	–	44.2	–
Bagel Deng et al. (2025)	79.2	–	–	–	–	54.7	–	–	–	–	43.2	–
X-Omni-RL	74.31	74.04	52.60	57.08	48.68	50.67	68.26	88.98	60.46	70.9	37.33	17.39
+Ours	75.96	74.14	53.27	57.40	50.26	52.67	70.01	89.13	61.30	71.6	40.00	18.40
Gain (↑)	+1.65	+0.10	+0.67	+0.32	+1.58	+2.00	+1.75	+0.15	+0.84	+0.70	+2.67	+1.01

Table 3: **Fine-grained analysis on MMBench.** We report detailed performance across **Perception** (Coarse-Perception, Fine-grained Cross-instance, Fine-grained Single-instance) and **Reasoning** (Attribute Reasoning, Logical Reasoning, Relation Reasoning) sub-tasks. Our method improves both perception and reasoning capabilities, and the gains in reasoning are more significant.

Model	Perception			Reasoning			Overall
	CoarseP.	FP-Cross	FP-Single	AttributeR.	LogicalR.	RelationR.	
X-Omni-RL	84.12	66.43	77.13	76.88	48.31	73.91	74.31
+UAE	84.46	67.13	**79.18**	77.39	51.69	73.91	75.43
+Ours	**84.46**	**67.83**	78.50	**78.39**	**55.08**	**74.78**	**75.96**
Gain(↑)	+0.34	+1.40	+1.37	+1.51	+6.77	+0.87	+1.65

Unless otherwise noted, subsequent experiments use the RL setting. To answer Q2, Table 2 extends the understanding evaluation to 12 metrics across six benchmark families. ReasonUni improves over X-Omni on every metric, with the largest gains on MMMU (+2.67), MMVP (+2.00), and CV-Bench (+1.75). These consistent gains across diverse benchmarks provide broader evidence that ReasonUni improves visual understanding.

Table 4: **Performance comparison on GenEval++.** This benchmark evaluates fine-grained image generation capabilities across compositional attributes (e.g., Color, Position, Count). Our method achieves the highest overall score among the compared models.

Model	Color	Pos/Cnt	Col/Pos	Col/Cnt	Mul-Cnt	Pos/Size	Count	Overall
Janus-Pro Chen et al. (2025c)	45.0	7.5	30.0	12.5	12.5	35.0	30.0	24.60
T2I-R1 Jiang et al. (2025a)	67.5	7.5	35.0	20.0	30.0	25.0	32.5	31.10
BLIP3o Chen et al. (2025a)	25.0	12.5	60.0	12.5	22.5	57.5	25.0	30.70
OmniGen2 Wu et al. (2025b)	55.0	12.5	27.5	20.0	45.0	25.0	42.5	32.50
Bagel Deng et al. (2025)	32.5	25.0	32.5	25.0	37.5	47.5	60.0	37.10
X-Omni-RL	25.0	32.5	47.5	15.0	45.0	55.0	60.0	40.00
+Ours	42.5	35.0	50.0	12.5	47.5	80.0	60.0	46.79

3.2 Fine-grained Quantitative Analysis

We next examine the understanding and generation subcategories behind these gains (Q3).

Understanding side. On MMBench, gains are strongest in reasoning-related dimensions rather than perception-only dimensions (Table 3). Compared with X-Omni, ReasonUni improves all six dimensions, with the largest gain on logical reasoning (LR: 55.08 vs. 48.31, +6.77). Compared with UAE, ReasonUni keeps perception comparable (CP tied at 84.46; FP-C +0.70) while providing consistent reasoning gains (AR +1.00, LR +3.39, RR +0.87), improving overall MMBench from 75.43 to 75.96. This suggests the main benefit is not merely better recognition, but stronger higher-order reasoning.

Generation side. On GenEval++ (Table 4), ReasonUni achieves the best overall score of 46.79, improving over X-Omni-RL by +6.79 and outperforming UAE baseline. The gains are most pronounced on Pos/Size (+25.0) and Color (+17.5), with consistent improvements on Pos/Count, Color/Position, and Multi-Count (+2.5 each). These results demonstrate stronger control over attributes, spatial relations, and multi-object composition.

3.3 Ablation Study

Reward design. Latent rewards improve training efficiency without sacrificing performance: Table 5(a) shows the highest Und. Avg (54.26) and Gen. Avg (47.40) among the compared rewards, together with a $\sim$60% reduction in training time (25h vs. 58–68h on 8$\times$H100). Thus, the efficiency gain accompanies stronger understanding and generation, rather than a quality–cost trade-off. We compute rewards directly in SigLIP-VQ token space, bypassing diffusion decoding and external re-encoding. This design avoids diffusion-induced noise and external-encoder mismatch, while using SigLIP-VQ codebook embeddings that can implicitly retain spatial cues. Figure 4 provides qualitative evidence that the latent reward captures task-relevant differences in reconstruction quality consistently with external visual encoders.

Sampling strategy. TreeGRPO makes more effective use of the image-sampling budget than flat GRPO (Table 5(b)). With the same 16 candidate images, our default $K{=}4, M{=}4$ configuration, highlighted in gray, improves Und. Avg from 53.90 to 54.26 and Gen. Avg from 45.40 to 47.40, while reducing training time from 30h to 25h. These joint gains demonstrate the value of tree-structured optimization at a fixed image-sampling budget. Sharing each description across four images requires only four target-state descriptions instead of sixteen, reducing text-generation overhead. Scaling to $K{=}8, M{=}8$ further increases Gen. Avg to 49.02, showing that generation benefits from a larger tree-sampling budget.

3.4 Qualitative Case Study

We visualize instruction-grounded editing and WISE knowledge-grounded generation. Supplementary Section 4.1 provides detailed WISE results and analysis, while Section 3 presents visualizations of optimization dynamics.

Instruction-grounded editing. In Figure 5, the instruction is to remove the largest text string on the vehicle's front. X-Omni retains "HARRISONS" but mistakenly removes the license-plate

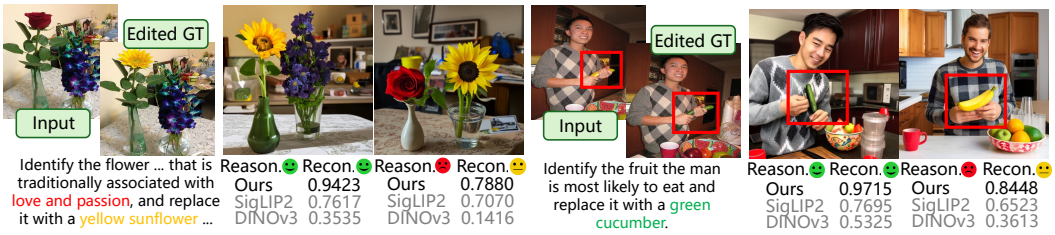


Figure 4: **Comparison of reconstruction rewards.** Across these examples, the latent reward produces the same preference ordering over candidate reconstructions as SigLIP2 and DINOv3, while operating directly on visual tokens without image decoding or external re-encoding.

Table 5: **(a)** Reward design: latent reward vs. alternative formulations. **(b)** Sampling strategy: K/M sweep and flat GRPO comparison under matched budget. MM-P/R denote MMBench Perception/Reasoning. Und. Avg averages MMBench, MMVP, CV-Bench, and MathVerse; Gen. Avg averages WISE and GenEval++. Both are equally weighted means of percentage scores.

(a) Reward design

	MM-P	MM-R	Und. Avg	Gen. Avg	Time
X-Omni-RL	75.89	66.36	52.66	42.50	—
Latent (Ours)	76.93	69.42	54.26	47.40	25 h
SigLIP-2	76.92	67.84	52.77	45.92	60 h
DINOv3	75.78	66.93	53.26	43.97	60 h
MSE	76.91	67.54	53.43	42.29	58 h
GPT-4.1	75.08	66.58	52.38	44.15	68 h

(b) Sampling strategy

	MM-P	MM-R	Und. Avg	Gen. Avg	Time
Flat 4	75.78	67.73	53.14	43.60	22 h
$K=2,\ M=2$	76.47	68.45	53.08	46.10	20 h
Flat 16	77.03	67.83	53.90	45.40	30 h
$K=4,\ M=4$	76.93	69.42	54.26	47.40	25 h
$K=8,\ M=8$	76.57	68.36	53.25	49.02	64 h

text, whereas ReasonUni removes the intended text while preserving the plate. This example illustrates that our training improves not only fine-grained perception but also visual reasoning, enabling the model to infer the correct editing target from the instruction's spatial and comparative constraints. Both models receive the same prompt and edit instruction, so the difference stems from the trained weights rather than additional textual guidance. Additional examples and a reconstruction-only ablation (Supplementary Sections 3–4) further support the contribution of reasoning-augmented reconstruction to visual understanding.

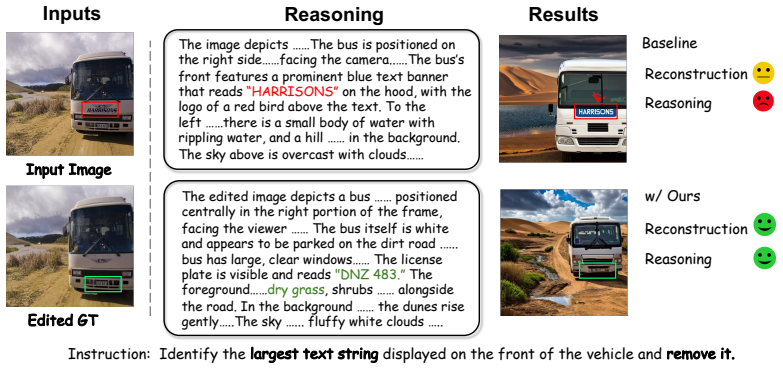


Figure 5: **Vehicle text-removal case.** We compare the intermediate target-state descriptions and final reconstructions under the same edit instruction. The baseline retains the target text, whereas ReasonUni removes the largest text string on the vehicle's front while preserving non-target details such as the license plate.

4 Related Work

Architecture Design of Unified Multimodal Models. Unified Multimodal Models (UMMs) integrate understanding and generation within a single trainable framework Zhang et al. (2025), with two dominant architectural paradigms. **(1) Hybrid AR–Diffusion Architectures.** These models combine autoregressive language modeling with diffusion-based visual generation. Show-o Xie et al. (2024) uses discrete diffusion, while other approaches employ continuous diffusion or flow matching. Transfusion Zhou et al. (2024) and related models Yan et al. (2025b); Deng et al. (2025); Ma et al. (2024); Shi et al. (2024); Xie et al. (2025b) share a backbone across modalities while



Figure 6: Qualitative results on WISE. For knowledge-focused prompts, ReasonUni generates more semantically accurate concepts than the baseline. The illustrated examples cover animal recognition, instrument categories, art-related knowledge, and historical technology.

using dedicated generative components, whereas another line connects a pre-trained MLLM to an external image generator through learnable adapters Pan et al. (2025); Chen et al. (2025a); Lin et al. (2025). **(2) Fully AR-based Approaches.** These models generate both text and discrete image tokens autoregressively, while differing in how visual inputs are represented Team (2024); Qu et al. (2025); Wu et al. (2025a); Chen et al. (2025c); Wu et al. (2024a). Among fully autoregressive models, X-Omni Geng et al. (2025) and Tar Han et al. (2025a) further employ semantically aligned visual tokens, unifying visual understanding and generation within the same discrete representation space.

Post-Training in Unified Multimodal Models. Recent post-training methods for UMMs can be broadly grouped into two paradigms. **(1) Post-training for Generation Quality.** These methods Chen et al. (2025b); Tong et al. (2025a); Jiang et al. (2025a); Tian et al. (2025); Li et al. (2025); Wei et al. (2025) adapt reinforcement learning and related post-training techniques to improve generation and editing quality. T2I-R1 Jiang et al. (2025a) further introduces reasoning over text prompts solely to improve image generation; in contrast, ReasonUni reasons over visual inputs and turns generation feedback into supervision for visual reasoning, jointly optimizing understanding and generation.

(2) Post-training for Understanding–Generation Alignment. This paradigm exploits the interaction between understanding and generation to improve both capabilities. Prior work explores self-generated supervision Han et al. (2026); Jin et al. (2025); Chen et al. (2026); Niu et al. (2025); Tang et al. (2026); Yang et al. (2025); Han et al. (2025b) or joint reinforcement learning Jiang et al. (2025b); Mao et al. (2025). More recently, reconstruction-based approaches Xie et al. (2025a); Yan et al. (2025a) explicitly connect understanding and generation by conditioning image reconstruction on extracted semantics. **However, their supervision largely remains at the *perception level*, providing limited explicit supervision for *visual reasoning* and thus limiting generalization to tasks that require reasoning beyond perception.**

5 Conclusion

`ReasonUni` extends reconstructive post-training to visual reasoning, with **`TreeGRPO`** jointly optimizing understanding and generation through tree-structured reward comparisons. Across both initialization settings, ReasonUni achieves substantial gains on generation benchmarks alongside consistent improvements in visual understanding. Gains on all 12 evaluated understanding benchmarks, including improved logical reasoning on MMBench, further support its benefits beyond fine-grained perception. Together, these findings support reasoning-augmented reconstruction as an effective approach to using generation feedback for mutual gains in unified multimodal models.

AI USE STATEMENT

We used Gemini-3-Pro and Nano Banana to construct synthetic editing data, as described in Supplementary Section 1, and GPT-4o for evaluation tasks requiring external model queries, as detailed in Supplementary Section 2.2. We also used AI tools to assist with coding and improve the clarity and presentation of the manuscript. The synthetic data were manually spot-checked, and all other AI-assisted work was reviewed and verified by the authors. The authors take full responsibility for the final content of this paper and the associated research artifacts.

REFERENCES

Jiahui Chen, Philippe Hansen-Estruch, Xiaochuang Han, Yushi Hu, Emily Dinan, Amita Kamath, Michal Drozdzal, Reyhane Askari-Hemmat, Luke Zettlemoyer, and Marjan Ghazvininejad. Unified text-image generation with weakness-targeted post-training. *arXiv preprint arXiv:2601.04339*, 2026.

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*, 2025a.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

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*, 2025c.

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.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025a.

Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, et al. Unicorn: Towards self-improving unified multimodal models through self-generated supervision. *arXiv preprint arXiv:2601.03193*, 2026.

Yujin Han, Hao Chen, Andi Han, Zhiheng Wang, Xinyu Liu, Yingya Zhang, Shiwei Zhang, and Difan Zou. Turning internal gap into self-improvement: Promoting the generation-understanding unification in mllms. *arXiv preprint arXiv:2507.16663*, 2025b.

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*, 2025a.

Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.17534*, 2025b.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv preprint arXiv:2510.12784*, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

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.

Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. *arXiv preprint arXiv:2505.23380*, 2025.

Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025.

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.

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*, pp. 2545–2555, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

Zhenchen Tang, Songlin Yang, Zichuan Wang, Bo Peng, Yang Li, Beibei Dong, and Jing Dong. Endogenous reprompting: Self-evolving cognitive alignment for unified multimodal models. *arXiv preprint arXiv:2601.20305*, 2026.

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

Rui Tian, Mingfei Gao, Haiming Gang, Jiasen Lu, Zhe Gan, Yinfei Yang, Zuxuan Wu, and Afshin Dehghan. Unigen-1.5: Enhancing image generation and editing through reward unification in reinforcement learning. *arXiv preprint arXiv:2511.14760*, 2025.

Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025a.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024b.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025b.

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

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*, 2024b.

Hongyang Wei, Baixin Xu, Hongbo Liu, Size Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025.

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*, pp. 12966–12977, 2025a.

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*, 2025b.

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*, 2024a.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.

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

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.

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

Zhiyuan Yan, Kaiqing Lin, Zongjian Li, Junyan Ye, Hui Han, Zhendong Wang, Hao Liu, Bin Lin, Hao Li, Xue Xu, et al. Unified multimodal model as auto-encoder. *arXiv preprint arXiv:2509.09666*, 2025a.

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*, 2025b.

Ling Yang, Xinchen Zhang, Ye Tian, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. Hermesflow: Seamlessly closing the gap in multimodal understanding and generation. *arXiv preprint arXiv:2502.12148*, 2025.

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. URL https://arxiv.org/abs/2404.16006.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

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.

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.