A UNIFIED REASONING FRAMEWORK FOR WORLD KNOWLEDGE-ALIGNED IMAGE GENERATION AND EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models often struggle with complex synthesis tasks that demand deep reasoning with world knowledge, and typically treat text-to-image generation and image editing as isolated capabilities rather than interconnected reasoning steps. To address this, we propose **UniReason**, a unified framework that harmonizes these two tasks through a dual reasoning paradigm. We formulate generation as *world knowledge-enhanced planning* to inject implicit constraints, and leverage editing capabilities for *fine-grained visual refinement* to further correct visual errors via self-reflection. This approach unifies generation and editing within a shared representation, mirroring the human cognitive process of planning followed by refinement. We support this framework by systematically constructing a large-scale reasoning-refinement dataset (∼300K samples) covering five major knowledge domains (*e.g.*, cultural commonsense, physics, etc.) for planning, alongside an agent-generated corpus for visual self-correction. Extensive experiments demonstrate that **UniReason** achieves advanced performance on reasoning-intensive benchmarks such as WISE and KrisBench, while maintaining superior general synthesis capabilities on GenEval and ImgEdit.

1 INTRODUCTION

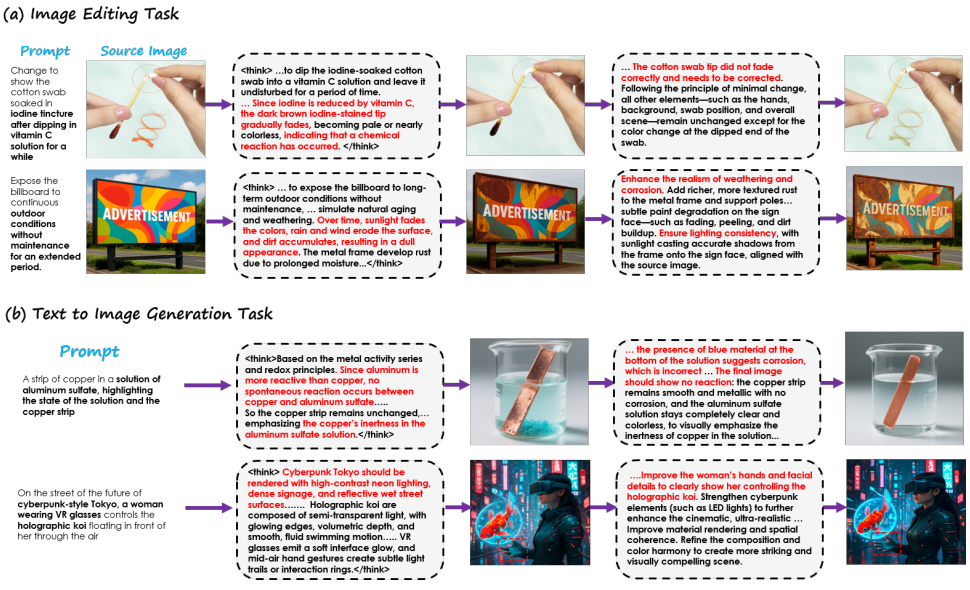

Figure 1: Illustrative cases of UniReason on Image Editing and T2I Generation tasks. Given an instruction, the model first performs world knowledge-enhanced textual reasoning to generate grounded, fine-grained guidance for image synthesis. It then applies fine-grained editing-like visual refinement, correcting errors arising during the initial generation and improving the synthesis quality.

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.

Unified multimodal models have emerged as a promising paradigm for jointly handling visual understanding and generation tasks (Xie et al., 2025; Chen et al., 2025c; Wu et al., 2025d; 2024; Pan et al., 2025; Deng et al., 2025; Wang et al., 2026). By integrating perception and synthesis within a shared architecture, these models enable seamless interplay between comprehending visual content and producing new images conditioned on multimodal inputs. Among various capabilities, *text-to-image (T2I) generation* and *image editing* stand out as particularly challenging yet impactful applications. However, current unified models still struggle with complex scenarios, where tasks demand not only precise instruction following, but also *world knowledge* that is not explicitly stated in the instructions, such as commonsense, physical laws, and spatiotemporal logic. Such challenges fundamentally demand *reasoning* capabilities to bridge the gap between abstract user intent and faithful visual output.

To enhance reasoning capabilities, a prominent line of work focuses on *prompt enhancement* or *reprompting* strategies (Jiang et al., 2025; Fang et al., 2025; Xiao et al., 2025). These methods employ chain-of-thought (CoT) reasoning to expand abstract user prompts into explicit semantic and spatial guidance before generation. While effective in improving instruction alignment, these "reason-then-generate" approaches are inherently limited as reasoning occurs only *before* generation without access to visual feedback, preventing reflection on and correction of output errors. More recently, *interleaved reasoning* mechanisms (Huang et al., 2025; Qin et al., 2025) alternate between textual reasoning and visual generation. By first generating an initial image, then performing textual reflection based on visual feedback, and finally refining the output, these approaches enable post-generation correction that was previously infeasible.

Despite this progress, existing methods still exhibit two key limitations. (1) Reasoning in these methods largely remains at the level of semantic reorganization, decomposing instructions into finer-grained descriptions or spatial layouts (Xiao et al., 2025; Jiang et al., 2025; Qin et al., 2025; Ghosh et al., 2023). This addresses only the explicit component of user intent, whereas faithful synthesis in practice demands *world knowledge* that is implicitly assumed rather than explicitly stated. Such knowledge must be *inferred*, not merely *parsed* from instructions. This leaves a fundamental knowledge gap that semantic reorganization alone cannot address. (2) Existing methods typically address text-to-image generation and image editing as separate tasks (Huang et al., 2025), leaving their inherent synergies within a unified interleaved framework untapped. We argue that these two tasks share substantial reasoning overlap and can mutually reinforce each other. Specifically, post-generation critique and refinement in interleaved reasoning is structurally analogous to editing. Isolating them therefore forgoes such synergy and leads to redundant learning.

To address these challenges, we propose **UniReason**, a unified reasoning framework that harmonizes text-to-image generation and image editing within a shared architecture, with illustrative cases shown in Fig. 1. Our framework supports two complementary reasoning paradigms. ***(1) World Knowledge-Enhanced Textual Reasoning*** aims to bridge the knowledge gap prior to synthesis. Given an underspecified instruction, the model performs textual reasoning to infer implicit world knowledge and produces grounded guidance that specifies fine-grained details for the subsequent image synthesis. To support this, we construct training data across five knowledge categories: cultural commonsense, natural science, spatial, temporal, and logical reasoning. We use large language models to generate reasoning traces and apply multi-dimensional filtering to ensure high-quality supervision. ***(2) Fine-grained Editing-like Visual Refinement*** aims to improve synthesis quality after initial generation. Given the initial image and prior reasoning, the model performs self-reflection to identify discrepancies or missing details, then applies targeted corrections to produce a refined image. Observing that this process is structurally analogous to image editing, we jointly learn T2I generation and editing for mutual benefit. We design an agent pipeline that iterates through generation, verification, refinement, and comparison to construct high-quality training data. The two paradigms apply flexibly across diverse synthesis scenarios, including tasks that require world modeling and simulation.

We adopt a two-stage training strategy: the first stage strengthens foundational generation capability, and the second stage enables interleaved reasoning by jointly training the understanding and generation branches. Through this unified framework, we achieve comprehensive world knowledge-grounded reasoning capabilities for both T2I generation and image editing. Extensive experiments demonstrate advanced performance on multiple benchmarks, including GenEval, WISE for T2I generation, and UniREditBench, KrisBench for image editing.

2 Related Work

Image Generation Image generation includes T2I generation and image editing, depending on whether the conditional signals are textual descriptions or reference images. Recently, Diffusion Transformers (Vaswani et al., 2017; Peebles & Xie, 2023) (DiTs) have served as the backbone of state-of-the-art generation frameworks, with flow-matching (Esser et al., 2024; Ma et al., 2024) adopted as the prevailing training scheme. Together with data and model scaling, these advances have enabled photo-realistic synthesis and substantially improved instruction following in T2I generation (Esser et al., 2024; Wu et al., 2025a; Black Forest Labs, 2024). Building upon these powerful generators, recent image editing systems (Labs et al., 2025; Wu et al., 2025a) achieve precise content manipulation while preserving overall visual consistency. However, despite their generative prowess, these specialized models lack the intrinsic capacity for world comprehension and self-reflection, motivating the integration of reasoning and generation within a coherent unified framework.

Unified Multimodal Models Unified multimodal models (Wu et al., 2025d; 2024; Pan et al., 2025; Chen et al., 2025a; Deng et al., 2025; Chen et al., 2025c) aim to jointly support image understanding and generation within a single framework. Broadly, existing approaches can be grouped into two paradigms. A first, more modular paradigm aligns pretrained LMMs and DiTs via LLM hidden states (Wu et al., 2025d; Lin et al., 2025; Wu et al., 2025b) or learnable queries (Chen et al., 2025a; Pan et al., 2025; Wu et al., 2025c). Another line of work (Xie et al., 2025; Chen et al., 2025c; Deng et al., 2025) adopts a shared LLM architecture for perception and synthesis, encouraging a tight coupling between the two tasks. In our study, we focus on the second paradigm, since a shared backbone naturally supports interleaved reasoning between language and image generation in a unified inference process.

Reasoning in Unified Multimodal Models The structural convergence of understanding and generation within unified models unlocks the potential for grounding high-fidelity image synthesis in complex multimodal reasoning. Initial efforts primarily involve the adaptation of textual Chain-of-Thought (CoT) to image generation (Jiang et al., 2025; Fang et al., 2025; Xiao et al., 2025), following a 'reason-then-generate' paradigm that expands user instructions into detailed descriptions prior to synthesis. More recently, interleaved reasoning mechanisms (Huang et al., 2025; Qin et al., 2025) extend the process into iterative "reason-generate-reflect" cycles to incorporate visual feedback. Despite these advancements, existing methods are often confined to prompt reorganization and rigidly separate generation and editing tasks. In this work, we address these limitations by inferring implicit world knowledge rather than merely parsing instructions. Furthermore, we exploit the inherent synergies between T2I generation and image editing within a unified reasoning framework.

3 Preliminary

Architecture We build upon BAGEL (Deng et al., 2025) to develop a unified and interleaved reasoning framework for both T2I generation and image editing. Bagel adopts a Mixture-of-Transformers (MoT) architecture with a ViT encoder (Tschannen et al., 2025) to process multimodal inputs and enables unified image understanding and generation within a single foundation model.

Specifically, multimodal understanding is formulated as generating context-aware textual outputs via standard next-token prediction through a language modeling head. This process is conditioned on multimodal context inputs and handled by the understanding expert. Formally, the training objective minimizes the negative log-likelihood:

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}, C)\,, \tag{1}$$

where x_t denotes the target text token, $x_{<t}$ is the preceding tokens and C is the multimodal context.

Multimodal generation focuses on producing high-quality and semantically aligned images via a rectified flow process (Liu et al., 2022) within a VAE decoder (Black Forest Labs, 2024), conditioned on multimodal inputs and handled by the generation expert. The training objective is to minimize the

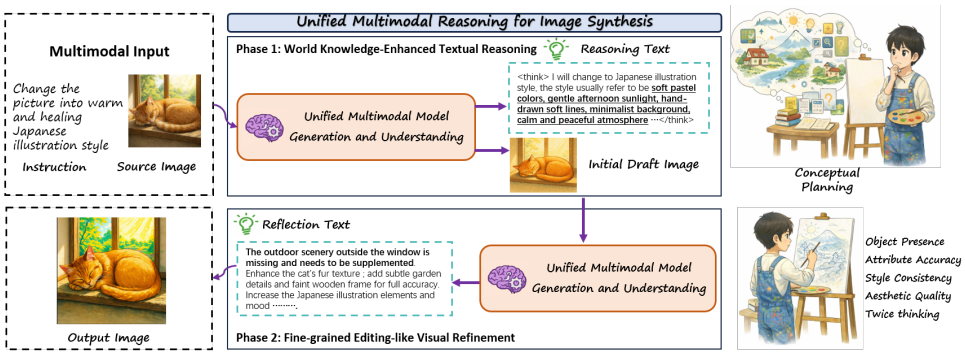

Figure 2: Overview of our proposed UniReason framework for two complementary reasoning paradigms in image synthesis.

the latent flow-matching loss:

$$\mathcal{L}_{\text{img}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \left\| u_\theta(z_t, t\,;\, C) - u^\star(z_t, t) \right\|_2^2 , \tag{2}$$

where $u^\star$ denotes the target velocity, u_θ is the learned time-conditioned velocity field in latent space.

Reasoning Paradigms The unified architecture supports interleaving textual reasoning and visual synthesis in both T2I generation and image editing tasks. Specifically, T2I generation takes a textual instruction as input and outputs a sequence of intermediate reasoning tokens together with a synthesized image. Image editing takes an existing image and a textual instruction as input and outputs refined reasoning text and a correspondingly edited image.

Most prior work focuses on single-turn reasoning-augmented image synthesis and lacks iterative refinement of intermediate results. We formulate interleaved reasoning as an iterative process: $(I^{k+1}, T^{k+1}) = \mathcal{F}\big(I^{\leq k}, T^{\leq k}, C\big)$ where I^k and T^k denote the image and reasoning text at iteration k, C denotes the multimodal context, and $\mathcal{F}$ is the unified model ($k = 1$ in our implementation). Under this formulation, each refinement step is an image editing operation conditioned on updated reasoning. We therefore jointly learn T2I generation and image editing in a unified interleaved reasoning framework, allowing refinement to benefit from editing supervision and improve interleaved reasoning for both generation and editing.

4 Method

In this section, we present our unified multimodal reasoning framework for both T2I generation and image editing, iteratively interleaving textual reasoning and visual synthesis, as illustrated in Fig 2. In practice, the framework operates in two phases, (1) *World Knowledge-Enhanced Textual Reasoning* for initial synthesis; (2) *Fine-grained Editing-like Visual Refinement* for iterative improvement. We introduce each phase along with its corresponding data creation pipeline in Section 4.1 and 4.2, respectively shown in Fig 4, followed by the training strategy in Section 4.3.

4.1 World Knowledge-Enhanced Textual Reasoning

Prior work (Fang et al., 2025; Xiao et al., 2025) primarily focuses on semantic reorganization in textual reasoning that enriches details explicitly parsed from instructions. In contrast, our core objective is to equip the unified multimodal model to infer implicit world knowledge underlying abstract instructions, and integrate world knowledge inference and semantic reorganization into textual reasoning. This process provides explicit and structured guidance for synthesizing an initial visual output, mirroring human conceptual planning prior to rendering. Building on this capability, the model supports initial synthesis for both T2I generation and image editing guided by textual reasoning, which resolves ambiguity and specifies fine-grained details such as object attributes, interactions, spatial relationships, and stylistic preferences, while explicitly incorporating underlying world knowledge from various aspects to ensure that generated images are both instruction-aligned and knowledge-consistent.

Data Preparation To enable world knowledge-enhanced textual reasoning for initial visual synthesis, we construct challenging input instructions for both T2I generation and image editing tasks that require inferring implicit world knowledge, along with their associated reasoning processes. Specifically, we cover five major categories of world knowledge and adopt post-generation filtering to ensure high-quality supervision.

- **Cultural Commonsense** instructions require using shared cultural knowledge, such as historical events, iconic figures, social customs, and idiomatic expressions, to resolve unnamed or underspecified entities into explicit, contextually meaningful visual content, ensuring generated images aligned with real-world cultural understanding.
- **Natural Science** instructions requires incorporating principles from physics, biology, medicine, or chemistry to ensure that generated images remain consistent with scientific laws, and reflect plausible real-world observations.
- **Spatial** reasoning focuses on understanding correct spatial relationships among entities, including relative position, orientation, viewpoint, and camera transformations. Such instructions requires deriving precise spatial configurations from abstract descriptions to generate visuals consistent with real-world geometric logic.
- **Temporal** reasoning models time-dependent relationships, such as event sequences, state transitions, and causal ordering. This type of instructions require inferring the temporal progression of events and ensuring that visual outputs reflect coherent and plausible temporal dynamics aligned with natural chronological flow.
- **Logical** reasoning emphasizes causal coherence and logical consistency during image generation, such as in maze-solving or constraint satisfaction problems, by adhering to explicit or implicit logical structures. These instructions require applying deductive principles to translate abstract logical constraints into visually valid solutions.

For T2I generation in each category, we manually construct seed prompts based on Wikipedia, together with explicit category definitions, and use Gemini-2.5 Pro (Google, 2025) to expand them into a larger prompt set and then employed to generate textual CoT reasoning for each prompt. All prompts with their corresponding CoTs are subsequently fed into Qwen-Image (Wu et al., 2025a) for image rendering to form paired training samples. For image editing, we utilize data triples (original image, editing instruction, desired outcome) from UniREdit-Data-100K (Han et al., 2025) covering diverse knowledge dimensions, combined with Gemini-2.5 Pro-generated textual reasoning process.

Gemini-2.5 Pro serves as a comprehensive evaluator to assess generated images across three dimensions: instruction alignment, visual fidelity, and reasoning correctness, verifying faithful instruction adherence and knowledge-consistent reasoning without hallucinations. Only verified samples are retained to construct a high-quality training set for training visual synthesis with textual reasoning.

4.2 Fine-grained Editing-like Visual Refinement

After initial generation or editing, the draft typically captures the essential content and aligns with the instruction and world knowledge, yet still contains imperfections requiring refinement. We therefore introduce a second-stage visual refinement process. Building on prior textual reasoning, the model reassesses the synthesized image, reflectively identifies and verbalizes inconsistencies or missing details, and performs a second round of reasoning to refine semantic attributes, aesthetics, stylistic coherence, and instruction fidelity. Since this reflection-guided refinement is structurally analogous to image editing, we jointly learn T2I generation and image editing within a unified interleaved framework that alternates knowledge-enhanced textual reasoning with editing-like visual refinement, forming a synergistic loop in which reasoning and refinement mutually improve each other.

Data Preparation We design an agent pipeline to construct high-quality supervision data for training interleaved reasoning framework across both T2I generation and image editing tasks. The pipeline consists of (i) an initial generator (the base model) that produces a draft image with its textual reasoning from the input; (ii) a verifier (Gemini-2.5 Pro) that diagnoses caption–image mismatches and outputs structured, actionable edit directives across five dimensions: object presence, attribute accuracy, style consistency, realism, and aesthetic quality; (iii) a refinement teacher, Qwen-Image-Edit (Wu et al., 2025a), which applies the feedback and textual reasoning via instruction-guided image editing to obtain an improved image; and (iv) a final judge (Gemini-2.5 Pro) that performs

Table 1: Evaluation of world knowledge-intensive Text-to-Image Generation on the WISE (Niu et al., 2025) benchmark. "*" denotes generation with textual reasoning only, "†" denotes generation with both reasoning and refinement. The first block reports the performance of closed-source models. Bold entries represent the best performance among open-source models.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
GPT-4o	0.81	0.71	0.89	0.83	0.79	0.74	0.80
Seedream 4.0	0.78	0.73	0.85	0.79	0.84	0.67	0.78
Unified Understanding and Generation w/o Reasoning.							
Harmon	0.38	0.48	0.52	0.37	0.44	0.29	0.41
Show-o	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Janus Pro	0.30	0.37	0.49	0.36	0.42	0.26	0.35
MetaQuery-XL	0.56	0.55	0.62	0.49	0.63	0.41	0.55
BLIP3-o	–	–	–	–	–	–	0.62
UniWorld-V1	0.53	0.55	0.73	0.45	0.59	0.41	0.55
OmniGen2	0.42	0.52	0.64	0.43	0.50	0.34	0.47
Hunyuan-Image 3.0	0.58	0.57	0.70	0.56	0.63	0.31	0.57
Qwen-Image	0.62	0.63	0.77	0.57	0.75	0.40	0.62
Unified Understanding and Generation w/ Reasoning.							
T2I-R1*	0.56	0.55	0.63	0.54	0.55	0.30	0.54
MindOmni*	0.75	0.70	0.76	0.76	0.72	0.52	0.71
IRG†	0.78	**0.72**	0.76	**0.81**	0.82	0.78	0.77
BAGEL*	0.76	0.69	0.75	0.65	0.75	0.58	0.70
UniCoT†	0.76	0.70	0.76	0.73	0.81	0.73	0.75
Ours†	**0.80**	0.68	**0.79**	0.77	**0.83**	**0.81**	**0.78**

comparative evaluation between the initial and refined images, retaining refined images only if they exhibit measurable improvements over the initial generation and faithfully reflects the verifier's suggested modifications.

Specifically, we sample long-form captions from ShareGPT-4o-Image dataset (Chen et al., 2025b) and short-form captions from midjourney prompts[1] for T2I generation, and image-instruction pairs from UniREdit-Data-100K (Han et al., 2025) for image editing. These inputs are fed to the initial generator for reasoning-augmented initial synthesis. The resulting caption–image pairs then undergo the full verification, refinement and comparison cycle. To enable the construction of high-quality and robust training data for image synthesis with multimodal interleaved reasoning, we apply strict multi-stage filtering and validation during dataset construction detailed in Appendix A.2.

4.3 Two-stage Training Strategy

We adopt a simple yet effective two-stage supervised fine-tuning (SFT) strategy to first strengthen the foundational generation capability of the unified multimodal model, then train interleaved knowledge-enhanced reasoning and refinement capabilities across diverse image synthesis queries.

Stage 1: Foundational Generation Strengthening In the first stage, we freeze the multimodal understanding branch of the base model and train only the generation branch. This stage focuses exclusively on image synthesis supervision using existing T2I generation and image editing datasets without textual reasoning, aiming to enhance the instruction-following ability and foundational image synthesis quality.

Stage 2: Interleaved Reasoning Tuning In the second stage, we unfreeze all model parameters and jointly train the understanding and generation branches using our prepared interleaved reasoning data, including single-turn knowledge-enhanced reasoning samples and iterative visual refinement samples. This enables the model to perform world knowledge-enhanced reasoning and to iteratively evaluate and refine visual content based on multimodal feedback. Specifically, for single-turn reasoning data, we supervise both the textual reasoning traces and the image synthesis outputs, and for visual refinement data, we supervise textual reflections and refined images while leaving the initial reasoning text and visual draft unsupervised. The overall objective is formulated as

$$\mathcal{L} = \lambda_{\text{text}}\,\mathcal{L}_{\text{text}} + \lambda_{\text{img}}\,\mathcal{L}_{\text{img}}. \tag{3}$$

Table 2: Evaluation of knowledge-intensive Image Editing on KrisBench (Wu et al., 2025e) and UniREditBench (Han et al., 2025) benchmarks. "*" denotes textual reasoning only for editing, "†" denotes interleaved reasoning with both reasoning and refinement. Bold entries represent the best performance among open-source models.

Model	KrisBench				UniREditBench		
	Factual	Conceptual	Procedural	Overall	Real World	Game World	Overall
GPT-4o	79.80	81.37	78.32	80.09	81.01	62.07	73.39
Gemini 2.0	65.26	59.65	62.90	62.41	–	–	–
Seedream 4.0	–	–	–	–	66.22	45.38	55.77
Unified Understanding and Generation w/o Reasoning.							
OmniGen2	57.36	44.20	47.79	49.71	53.69	33.14	43.41
Uniworld-V1	47.71	44.80	47.92	50.27	–	–	–
Lumina-DiMOO	–	–	–	–	51.44	45.61	48.54
LightFusion-World	66.69	63.50	52.38	61.85	–	–	–
Qwen-Image-Edit	–	–	–	–	70.95	41.92	56.52
Unified Understanding and Generation w/ Reasoning.							
BAGEL*	66.18	61.92	49.02	60.18	56.80	45.10	50.96
UniCoT†	**71.85**	67.16	**63.68**	68.00	61.82	40.54	51.18
Ours†	70.67	**72.38**	56.89	**68.23**	**74.82**	**65.30**	**70.06**

Table 3: Comparison of different models across general image generation and editing benchmarks. Bold entries represent the best performance among open-source models and underlined entries indicate the best performance among unified models with reasoning.

Type	Model	General T2I Generation		General Image Editing	
		GenEval	DPGBench	ImgEdit	GEdit-EN
Closed-source	GPT-4o	0.84	85.15	4.20	7.53
	Gemini 2.0	–	–	–	6.32
	Seedream 4.0	0.84	88.25	4.18	7.68
Unified Understanding and Generation w/o Reasoning.					
Open-source	TokenFlow-XL	0.55	73.38	–	–
	Harmon	0.76	–	–	–
	Show-o	0.53	67.48	–	–
	Janus Pro	0.80	84.19	–	–
	MetaQuery-XL	0.80	82.05	–	–
	BLIP3-o	0.84	81.60	–	–
	UniWorld-V1	0.80	81.38	3.26	4.85
	Mogao	0.89	84.33	–	–
	OmniGen2	0.80	83.57	3.43	6.41
	MMaDA	0.63	69.97	–	–
	Lumina-DiMOO	0.88	86.04	–	–
	LightFusion-World	–	–	3.85	6.58
	Hunyuan-Image 3.0	0.72	86.10	–	–
	Qwen-Image	0.87	**88.32**	–	–
	Qwen-Image-Edit	–	–	**4.27**	**7.56**
Unified Understanding and Generation w Reasoning.					
Open-source	T2I-R1	–	–	–	–
	GoT	0.64	–	–	–
	MindOmni	0.83	82.50	–	–
	IRG	0.85	–	–	–
	BAGEL	0.88	85.07	3.20	6.52
	UniCoT	0.83	66.72	3.26	6.74
	Ours	**0.90**	<u>86.21</u>	<u>4.06</u>	<u>6.76</u>

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training Details In the first stage, the training corpus comprises nearly 7 million T2I samples and 500k image editing samples from open-source datasets including BLIP3-o (Chen et al., 2025a), ShareGPT-4o-Image (Chen et al., 2025b), Echo-4o-Image (Ye et al., 2025a), OpenGPT4o-Image (Chen et al., 2025d), Nano-banana-consist (nan, 2024), and Pico-banana (Qian et al., 2025). We train the model's generation branch for 30,000 iterations using the Adam optimizer with a cosine learning rate schedule, including 3,000 warm-up steps, a maximum learning rate of 5×10^{-5} and a minimum learning rate of 1×10^{-5}.

In the second stage, the training corpus includes 150k self-constructed single-turn knowledge-enhanced reasoning samples for T2I generation, 100k image editing reasoning samples (Han et al.,

[1] https://huggingface.co/datasets/vivym/midjourney-prompts

2025), and 36k and 10k self-constructed interleaved reasoning samples for T2I generation and image editing, respectively. We fine-tune all parameters for 10,000 iterations with 1,000 warm-up steps, a maximum learning rate of 2×10^{-5} and a minimum learning rate of 1×10^{-6}. Loss weights are set to $\lambda_{\text{text}} = 2$ and $\lambda_{\text{img}} = 1$, with a packed sequence length of 50,000.

Evaluation Setup We evaluate world knowledge reasoning and fine-grained semantic alignment for T2I generation using the WISE (Niu et al., 2025) benchmark which comprises 1,000 world knowledge-informed prompts across culture, natural science, spatial and temporal comprehension. For image editing with world knowledge, we use UniREditBench (Han et al., 2025) with 2,700 meticulously curated samples covering both real- and game-world scenarios, and KrisBench (Wu et al., 2025e) with 1,267 samples across factual, conceptual, and procedural knowledge to assess world knowledge reasoning and refinement capabilities. Additionally, we evaluate general abilities using GenEval (Ghosh et al., 2023) and DPGBench (Hu et al., 2024) for T2I generation, and ImgEdit (Ye et al., 2025b) and GEdit-EN (Liu et al., 2025) for image editing.

5.2 Main Results

We present a comprehensive comparison of our model against existing state-of-the-art unified multimodal models support both generation and understanding in Table 1 and Table 2, for world knowledge-intensive T2I generation and image editing tasks, respectively. Detailed descriptions of the compared models are provided in Appendix A.1.

Our model achieves the best overall performance among open-source unified multimodal models, with or without explicit reasoning mechanisms, across knowledge-intensive image generation and editing tasks. Besides, it demonstrates comparable results to closed-source models, including Seedream 4.0 (ByteDance, 2025) and GPT4o (OpenAI, 2025) on T2I generation, and even surpasses Gemini 2.0 (Kampf & Brichtova, 2025) on KrisBench (Wu et al., 2025e) and outperforms Seedream 4.0 (ByteDance, 2025) on UniREditBench (Han et al., 2025). These results highlight the effectiveness of our unified reasoning framework.

Moreover, as shown in the fine-grained breakdown of performance across different knowledge domains in Tables 1 and 2, our model exhibits broad and consistent world-knowledge coverage. Notably, it achieves the highest performance in *Cultural-Commonsense*, *Spatial Reasoning*, *Natural Science* including *Physics* and *Chemistry*. For image editing tasks, it also demonstrates strong performance across diverse knowledge categories in both KrisBench and UniREditBench. Overall, our model's knowledge-enhanced reasoning capabilities cover a wide range of tasks and domains.

5.3 General Ability Retention

Beyond knowledge-intensive tasks, our model remains highly competitive on general image generation and editing benchmarks while improving knowledge-enhanced reasoning, demonstrating strong all-round capability. As shown in Table 3, on GenEval (Ghosh et al., 2023), our model surpasses leading systems, including Qwen-Image (Wu et al., 2025a), GPT-4o (Chen et al., 2025d), and Seedream 4.0 (ByteDance, 2025), without external LLM-based rewriting. On DPGBench (Hu et al., 2024), it achieves the best performance among models with reasoning during generation, demonstrating strong long-horizon instruction following. We further evaluate instruction-following image editing on ImgEdit (Ye et al., 2025b) and GEdit-EN (Liu et al., 2025), which are essential for practical refinement. Our model achieves the strongest results among reasoning-based models while remaining competitive with existing approaches. These results show that our model excels not only in reasoning-centric settings but also in general generation and editing, providing a robust unified foundation. Detailed results are provided in Appendix A.3, with case studies in Appendix A.4.

5.4 Ablation Study

We further investigate the contributions of the two-stage training strategy and the reasoning and refinement mechanisms for image synthesis. On three knowledge-intensive generation and editing benchmarks, we compare three progressive settings built upon the BAGEL base model: (i) *Two-Stage Training Base*, which performs direct image generation without reasoning paradigms after two-stage fine-tuning; (ii) + *Only Reasoning*, which elicits only textual reasoning prior to image synthesis; (iii) + *Only Refinement*, which introduces an explicit reflection step to produce final refined output ;and (iv) + *UniReason*, which utilizes UniReason framework for two complementary paradigms in image

Table 4: Ablation study of UniReason. The base model is BAGEL (Deng et al., 2025). "Two-Stage Training Base" refers to fine-tuning the base model using the two-stage training recipe, as described in Section. 4.3.

Method	WISE	KrisBench	UniREditBench
Base Model	0.52	56.21	50.96
+ UniReason (zero-shot)	0.72(+0.20)	63.94(+7.73)	52.31(+1.35)
Two-Stage Training Base	0.58(+0.06)	61.53(+5.32)	63.37(+12.41)
+ Only Reasoning	0.73(+0.21)	64.12(+7.91)	67.30(+16.34)
+ Only Refinement	0.64(+0.12)	63.57(+7.36)	65.13(+14.17)
+ UniReason	**0.78** (+0.26)	**68.23**(+12.02)	**70.06**(+19.10)

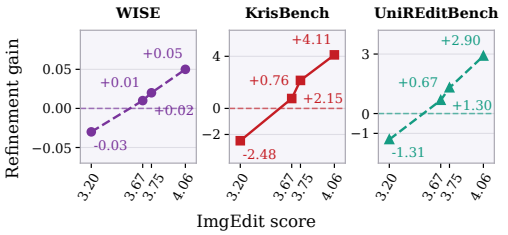

Figure 3: Correlation between image editing capability (ImgEdit score) and performance gains from refinement across three benchmarks. Higher editing proficiency leads to monotonically increasing refinement effectiveness.

synthesis(reasoning and refinement). (v)+ *UniReason(zero-shot)*, which utilizes the Base model applied directly to the proposed UniReason framework without additional training.

Table 4 shows consistent improvement across all benchmarks as each component is added. The two-stage training alone effectively improves the base model's instruction-following and synthesis capabilities. Then, introducing world knowledge-enhanced textual reasoning yields significant gains. while the visual refinement stage brings additional gains across all benchmarks. Notably, the full framework achieves substantially stronger results than either component alone, indicating that textual reasoning and visual refinement are complementary rather than redundant. These results suggest that the two-stage training strategy injects both knowledge-enhanced reasoning and fine-grained refinement capabilities into the unified multimodal model, rather than merely enhancing surface-level visual composition. We also visualize the test time scaling curve of UniReason in Appendix A.5

Moreover, we evaluate the direct application of the proposed UniReason framework to the base model, without any additional training. This setting also yields consistent performance improvements, suggesting that UniReason framework can activate and better leverage the model's existing reasoning capabilities, thereby producing measurable gains. When UniReason is combined with our proposed training strategy and data, the model achieves further improvements overall performance and reasoning abilities, demonstrating the effectiveness of the proposed training paradigm and data.

5.5 Correlation of Editing and Refinement

To show how image editing capability affects refinement effectiveness, we analyze performance gains with and without the refinement mechanism across models with varying editing capabilities. Specifically, we select different checkpoints during stage-1 training, each exhibiting different levels of editing proficiency, and apply identical stage-2 reasoning training to all checkpoints. We then evaluate performance on three knowledge-intensive benchmarks, measuring the gain achieved through refinement after initial textual reasoning. Fig 3 plots these performance gains against the editing performance of each checkpoint on ImgEdit.

The results reveal that performance gains from refinement increase monotonically with higher ImgEdit scores. This trend highlights the importance of jointly training image editing and T2I generation within a unified interleaved reasoning framework that integrates both textual reasoning and visual refinement. Since visual refinement relies on fine-grained and controllable editing, insufficient editing capacity can limit the effectiveness of reasoning-guided refinement.

6 Conclusion

In this paper, we introduce UniReason, a unified reasoning framework that harmonizes the text-to-image generation and image editing by exploiting their inherent structural synergies. Specifically, we proposed two complementary components: World Knowledge-Enhanced Textual Reasoning that infers implicit common sense and physical laws, and Fine-grained Editing-like Visual Refinement that enables iterative reflection and correction. By constructing high-quality datasets across five knowledge categories and employing a two-stage training strategy, UniReason demonstrates superior instruction following and visual fidelity. Extensive experiments on multiple benchmarks demonstrate that our unified reasoning approach achieves advanced performance across both T2I and editing tasks. Beyond these benchmarks, the framework's reasoning over world knowledge also exhibits latent potential for broader directions such as world modeling and simulation.

AI USE STATEMENT

AI assistance was used solely for language polishing. The authors are fully responsible for verifying the scientific content and the final submitted version of the manuscript.

REFERENCES

Nano-banana-150k. https://github.com/yejy53/Nano-banana-150k, 2024. GitHub repository.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

ByteDance. Seedream 4.0, 2025. URL https://seed.bytedance.com/en/seedream4_0. Accessed: 2025-08.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, Yang Li, Zhenxi Li, Zhimin Li, Jiaxin Lin, Linus, Lucaz Liu, Shu Liu, Songtao Liu, Yu Liu, Yuhong Liu, Yanxin Long, Fanbin Lu, Qinglin Lu, Yuyang Peng, Yuanbo Peng, Xiangwei Shen, Yixuan Shi, Jiale Tao, Yangyu Tao, Qi Tian, Pengfei Wan, Chunyu Wang, Kai Wang, Lei Wang, Linqing Wang, Lucas Wang, Qixun Wang, Weiyan Wang, Hao Wen, Bing Wu, Jianbing Wu, Yue Wu, Senhao Xie, Fang Yang, Miles Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Zheng Yuan, Chao Zhang, Jian-Wei Zhang, Peizhen Zhang, Shi-Xue Zhang, Tao Zhang, Weigang Zhang, Yepeng Zhang, Yingfang Zhang, Zihao Zhang, Zijian Zhang, Penghao Zhao, Zhiyuan Zhao, Xuefei Zhe, Jianchen Zhu, and Zhao Zhong. Hunyuanimage 3.0 technical report, 2025. URL https://arxiv.org/abs/2509.23951.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025a. URL https://arxiv.org/abs/2505.09568.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation, 2025b. URL https://arxiv.org/abs/2506.18095.

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, 2025c. URL https://arxiv.org/abs/2501.17811.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, Pengfei Wan, and Yi-Fan Zhang. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing, 2025d. URL https://arxiv.org/abs/2509.24900.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing, 2025. URL https://arxiv.org/abs/2503.10639.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Google. Gemini 2.5 pro. https://deepmind.google/models/gemini/pro/, 2025.

Chenhui Gou, Zilong Chen, Zeyu Wang, Feng Li, Deyao Zhu, Zicheng Duan, Kunchang Li, Chaorui Deng, Hongyi Yuan, Haoqi Fan, Cihang Xie, Jianfei Cai, and Hamid Rezatofighi. Vq-va world: Towards high-quality visual question-visual answering, 2025. URL https://arxiv.org/abs/2511.20573.

Feng Han, Yibin Wang, Chenglin Li, Zheming Liang, Dianyi Wang, Yang Jiao, Zhipeng Wei, Chao Gong, Cheng Jin, Jingjing Chen, and Jiaqi Wang. Unireditbench: A unified reasoning-based image editing benchmark, 2025. URL https://arxiv.org/abs/2511.01295.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. URL https://arxiv.org/abs/2403.05135.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, Junbo Qiao, Yue Guo, Yao Hu, Zhenfei Yin, Philip Torr, Yu Cheng, Wanli Ouyang, and Shaohui Lin. Interleaving reasoning for better text-to-image generation, 2025. URL https://arxiv.org/abs/2509.06945.

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, 2025. URL https://arxiv.org/abs/2505.00703.

Kat Kampf and Nicole Brichtova. Experiment with gemini 2.0 flash native image generation. Accessed: 05-08, 2025, March 2025. URL https://developers.googleblog.com/en/experiment-withgemini-20-flash-native-image-generation/.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 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, 2025. URL https://arxiv.org/abs/2505.05472.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. URL https://arxiv.org/abs/2506.03147.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. URL https://arxiv.org/abs/2504.17761.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models wfith scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation, 2025. URL https://arxiv.org/abs/2503.07265.

OpenAI. Gpt-image-1, 2025. URL https://openai.com/index/introducing-4o-image-generation/. Accessed: 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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing, 2025. URL https://arxiv.org/abs/2510.19808.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision, 2025. URL https://arxiv.org/abs/2508.05606.

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, 2025. URL https://arxiv.org/abs/2412.03069.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dianyi Wang, Ruihang Li, Feng Han, Chaofan Ma, Wei Song, Siyuan Wang, Yibin Wang, Yi Xin, Hongjian Liu, Zhixiong Zhang, et al. Deepgen 1.0: A lightweight unified multimodal model for advancing image generation and editing. *arXiv preprint arXiv:2602.12205*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025b. URL https://arxiv.org/abs/2506.18871.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025c.

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*, 2025d.

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*, 2024.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models, 2025e. URL https://arxiv.org/abs/2505.16707.

Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo, 2025. URL https://arxiv.org/abs/2505.13031.

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, 2025. URL https://arxiv.org/abs/2408.12528.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, Jinbin Bai, Qian Yu, Dengyang Jiang, Yuandong Pu, Haoxing Chen, Le Zhuo, Junjun He, Gen Luo, Tianbin Li, Ming Hu, Jin Ye, Shenglong Ye, Bo Zhang, Chang Xu, Wenhai Wang, Hongsheng Li, Guangtao Zhai, Tianfan Xue, Bin Fu, Xiaohong Liu, Yu Qiao, and Yihao Liu. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding, 2025. URL https://arxiv.org/abs/2510.06308.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. URL https://arxiv.org/abs/2505.15809.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025a. URL https://arxiv.org/abs/2508.09987.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark, 2025b. URL https://arxiv.org/abs/2505.20275.

A Appendix

A.1 Compared Baselines

We compared closed-source models including: GPT-4o (OpenAI, 2025), Gemini-2.0 (Google, 2025), Seedream4.0 (ByteDance, 2025),as well as open-source advanced unified multimodal models models which support both multimodal understanding and high quality image generation including autoregressive unified models, such as Harmon (Wu et al., 2025d), TokenFlow-XL (Qu et al., 2025), and Janus-Pro (Chen et al., 2025c). Discrete diffusion-based approaches, including Lumina-DiMOO (Xin et al., 2025), MMaDA (Yang et al., 2025), and Show-o (Xie et al., 2025). Another line of work connects VLMs and diffusion transformers via explicit connectors, exemplified by BLIP3-o (Chen et al., 2025a), UniWorld-V1 (Lin et al., 2025), OmniGen2 (Wu et al., 2025b), and the Qwen-Image series (Wu et al., 2025a). In contrast, deep fusion methods tightly integrate VLMs and DiTs within a unified architecture, such as Mogao (Liao et al., 2025), Hunyuan Image 3.0 (Cao et al., 2025) and LightFusion-World (Gou et al., 2025), the latter further enhanced with knowledge-centric fine-tuning.

Among open-source unified multimodal models that support naive reasoning, T2I-R1 (Jiang et al., 2025), MindOmni (Xiao et al., 2025), and BAGEL (Deng et al., 2025) primarily rely on textual reasoning to decompose abstract instructions into explicit semantic components that guide image generation. In contrast, GoT (Fang et al., 2025) introduces coordinate-based representations to provide explicit spatial guidance during synthesis. Another line of work, including IRG (Huang et al., 2025) and UniCoT (Qin et al., 2025), adopts interleaved reasoning mechanisms to reorganize semantics across modalities, progressively decomposing instructions into finer-grained and more structured descriptions for generation and refinement.

A.2 The Data Preparation Details

To construct high-quality supervision for training UniReason across both text-to-image (T2I) generation and image editing tasks, we design a two-phase data construction pipeline that integrates *world knowledge–enhanced textual reasoning* with *editing-like visual refinement*.

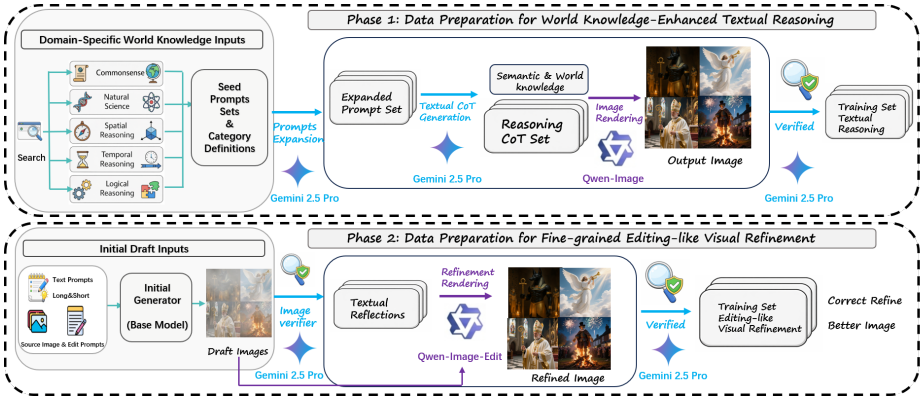

Figure 4: Overview of our data preparation framework.

Phase I: World Knowledge–Enhanced Reasoning Data Construction. We first build challenging instructions that require inferring implicit world knowledge, covering five categories of world knowledge: (i) Cultural Commonsense, which resolves culturally grounded but underspecified entities using shared knowledge of history, customs, and symbols; (ii) Natural Science, which enforces consistency with physical, biological, medical, or chemical laws; (iii) **Spatial Reasoning**, which derives correct relative positions, orientations, viewpoints, and camera transformations; (iv) Temporal Reasoning, which models time-dependent state transitions and causal event sequences; and (v) Logical Reasoning, which translates explicit or implicit logical constraints into visually valid solutions. For T2I generation, we manually curate seed prompts grounded in Wikipedia and category definitions, then use Gemini-2.5 Pro (Google, 2025) to expand them and generate corresponding textual CoT reasoning. Each prompt–reasoning pair is rendered into images using Qwen-Image (Wu

et al., 2025a), forming reasoning-grounded training samples. For image editing, we adopt triplets from UniREdit-Data-100K (Han et al., 2025), augmented with Gemini-2.5 Pro–generated reasoning processes with category definitions. All samples are filtered by Gemini-2.5 Pro to ensure instruction alignment, visual fidelity, and knowledge-consistent reasoning, retaining only verified high-quality data.

Phase II: Editing-like Visual Refinement Data Construction. To further train interleaved reasoning and refinement capabilities, we design an agent-based pipeline to generate iterative refinement supervision. Given an input instruction, an initial generator produces a draft image along with textual reasoning. A verifier (Gemini-2.5 Pro) then diagnoses caption–image mismatches and outputs structured, actionable feedback along five dimensions: object presence, attribute accuracy, style consistency, realism, and aesthetic quality. A refinement teacher, Qwen-Image-Edit (Wu et al., 2025a), applies this feedback and textual reasoning via instruction-guided image editing to produce a refined image. Finally, a judge (Gemini-2.5 Pro) performs comparative evaluation between the initial and refined images, retaining refined results only if they demonstrate measurable improvements and faithfully reflect the suggested modifications. Concretely, we sample long-form captions from ShareGPT-4o-Image (Chen et al., 2025b) and short-form captions from Midjourney prompts[2] for T2I generation, and image–instruction pairs from UniREdit-Data-100K for image editing. These inputs undergo the full generation–verification–refinement–selection cycle, yielding a high-quality training set that jointly supports world knowledge–enhanced reasoning and fine-grained visual refinement.

Keep the dataset high-quality and robust we apply strict multi-stage filtering and validation during dataset construction. Specifically, we ensure that (i) textual reasoning is consistent with world knowledge, (ii) deficiencies in the initial images are correctly identified, and (iii) refinement steps strictly follow the given instructions. We further use Gemini 2.5 Pro to automatically detect failure cases, identifying about one-third of the initial samples as problematic and revising them accordingly. This process yields a substantially cleaner final dataset. To assess data quality, we manually inspect a random 10% subset of the final dataset. Only $\tilde{3}$% of samples appear improvable, indicating that the supervision signal is high quality with limited residual noise. Finally, the consistent and significant performance gains observed after training on this dataset further support its quality and robustness.

A.3 Detailed Evaluation Results

We show the Detailed Evaluation Results on general tasks include GenEval (Ghosh et al., 2023) shown in table 5 and DPGBench (Hu et al., 2024) in table 6 for T2I generation, and ImgEdit (Ye et al., 2025b) and GEdit-EN (Liu et al., 2025) in table 7 for image editing. The results show our model delivers the strongest results among models with reasoning while remaining competitive with a broad range of existing approaches. These results indi- cate that our model is not only strong in reasoning-centric settings but also excels in general generation and editing, providing a robust and versatile unified foundation.

A.4 Case Study

We present additional UniReason results on both T2I generation and image editing tasks in Fig. 6. The results demonstrate that, while maintaining high-quality T2I generation and image editing performance, UniReason exhibits strong reasoning capabilities, enabling it to handle complex scenarios such as maze navigation, temporal evolution, and spatial camera viewpoint transformations. Moreover, UniReason shows robust refinement ability, effectively correcting fine-grained details such as faces, text, and hand gestures, thereby improving the quality of the initial images and rectifying errors introduced during the initial generation.

A.5 Test time scaling of UniReason

We report the average per-sample inference overhead of the reasoning-plus-refinement UniReason framework relative to the underlying model without reasoning ans shown in Fig 5. We observe a clear test-time scaling trend: as inference time increases with the introduction of reasoning, performance

[2]https://huggingface.co/datasets/vivym/midjourney-prompts

Table 5: Evaluation of General Text-to-Image Generation Capabilities on GenEval (Ghosh et al., 2023) Benchmark.

Model	Single object	Two object	Counting	Colors	Position	Attribution	Overall
GPT-4o	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 4.0	0.99	0.92	0.72	0.91	0.76	0.74	0.84
Unified Understanding and Generation w/o Reasoning.							
TokenFlow-XL	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Harmon	0.99	0.86	0.66	0.85	0.74	0.48	0.76
Show-o	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Janus Pro	0.99	0.89	0.59	0.90	0.79	0.66	0.80
MetaQuery-XL	–	–	–	–	–	–	0.80
BLIP3-o	–	–	–	–	–	–	0.84
UniWorld-V1	0.99	0.93	0.79	0.89	0.49	0.70	0.80
Mogao	1.00	0.97	0.83	0.93	0.84	0.80	0.89
OmniGen2	1.00	0.95	0.64	0.88	0.55	0.76	0.80
MMaDA	0.99	0.76	0.61	0.84	0.20	0.37	0.63
Lumina-DiMOO	1.00	0.94	0.85	0.89	0.85	0.76	0.88
Hunyuan-Image 3.0	1.00	0.92	0.48	0.82	0.42	0.63	0.72
Qwen-Image	0.99	0.92	0.89	0.88	0.76	0.77	0.87
Unified Understanding and Generation w Reasoning.							
GoT	–	–	–	–	–	–	0.64
MindOmni	0.99	0.94	0.71	0.90	0.71	0.71	0.83
IRG	0.98	0.94	0.83	0.86	0.74	0.73	0.85
BAGEL	0.98	0.95	0.84	0.95	0.78	0.77	0.88
UniCoT	0.99	0.96	0.84	0.92	0.57	0.71	0.83
Ours	1.00	0.96	0.82	0.90	0.88	0.82	0.90

Table 6: Evaluation of General Text-to-Image Generation Capabilities on DPG (Hu et al., 2024) Benchmark.

Model	Global	Entity	Attribute	Relation	Other	Overall
GPT-4o	88.89	88.94	89.84	92.63	90.96	85.15
Seedream 4.0	94.10	92.28	92.75	93.67	92.77	88.25
Unified Understanding and Generation w/o Reasoning.						
TokenFlow-XL	78.72	79.22	81.29	85.22	71.20	73.38
Show-o	–	–	–	–	–	67.48
Janus Pro	86.90	88.90	89.40	89.32	89.02	84.19
MetaQuery-XL	–	–	–	–	–	82.05
BLIP3-o	–	–	–	–	–	81.60
UniWorld-V1	83.64	88.39	88.44	89.27	87.22	81.38
Mogao	82.37	90.03	88.26	93.18	85.40	84.33
OmniGen2	88.81	88.83	90.18	89.37	90.27	83.57
MMaDA	77.81	78.48	81.74	84.79	63.20	69.97
Lumina-DiMOO	81.46	92.08	88.98	94.31	82.00	86.04
Hunyuan-Image 3.0	92.12	92.53	89.13	92.13	91.92	86.10
Qwen-Image	91.32	91.56	92.02	94.31	92.73	88.32
Unified Understanding and Generation w Reasoning.						
MindOmni	89.10	–	–	–	89.20	82.50
BAGEL	88.94	90.37	91.29	90.82	88.67	85.07
Ours	91.78	91.23	90.76	91.12	92.27	86.21

on knowledge-intensive benchmarks consistently improves for both generation and editing, with larger gains at higher test-time budgets.

A.6 Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Table 7: Evaluation of General Image Editing Capabilities on ImgEdit (Ye et al., 2025b) and GEdit-EN (Liu et al., 2025) Benchmarks.

Model	ImgEdit										GEdit-EN		
	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall	G_SC	G_PQ	G_O
GPT-4o	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20	7.85	7.62	7.53
Gemini 2.0	–	–	–	–	–	–	–	–	–	–	6.73	6.61	6.32
Seedream 4.0	4.52	4.41	2.93	4.56	4.44	4.30	4.76	3.33	4.36	4.18	8.24	8.08	7.68
Unified Understanding and Generation w/o Reasoning.													
Janus 4o	3.60	–	2.28	3.27	2.28	3.32	4.47	2.74	4.13	3.26	–	–	–
UniWorld-V1	3.82	3.66	2.31	3.45	3.02	2.99	4.71	2.96	2.74	3.26	4.93	7.43	4.85
OmniGen2	3.74	3.54	1.77	3.21	2.77	3.57	4.81	2.30	4.14	3.43	7.16	6.77	6.41
LightFusion-World	4.33	3.37	1.25	4.63	3.74	4.24	4.69	3.91	4.45	3.85	7.00	7.29	6.58
Qwen-Image-Edit	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27	8.00	7.86	7.56
Unified Understanding and Generation w Reasoning.													
BAGEL	3.56	3.31	1.88	2.62	2.88	3.44	4.49	2.38	4.17	3.20	7.36	6.83	6.52
UniCoT	3.49	4.25	1.46	3.52	2.88	3.49	4.25	2.71	4.56	3.26	7.91	6.24	6.74
Ours	4.14	4.06	2.49	4.42	4.31	4.23	4.65	2.58	4.68	4.06	7.30	7.40	6.76

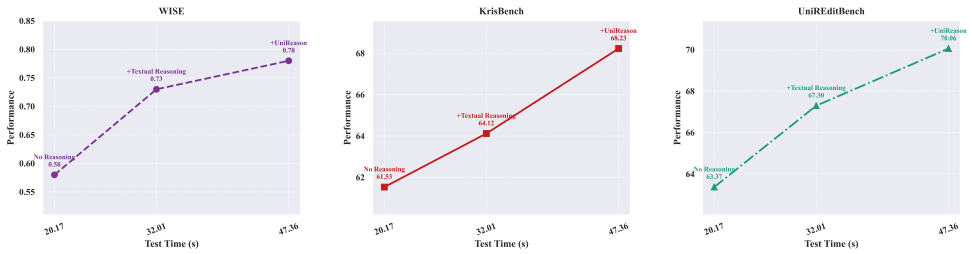

Figure 5: Test-time scaling of UniReason.

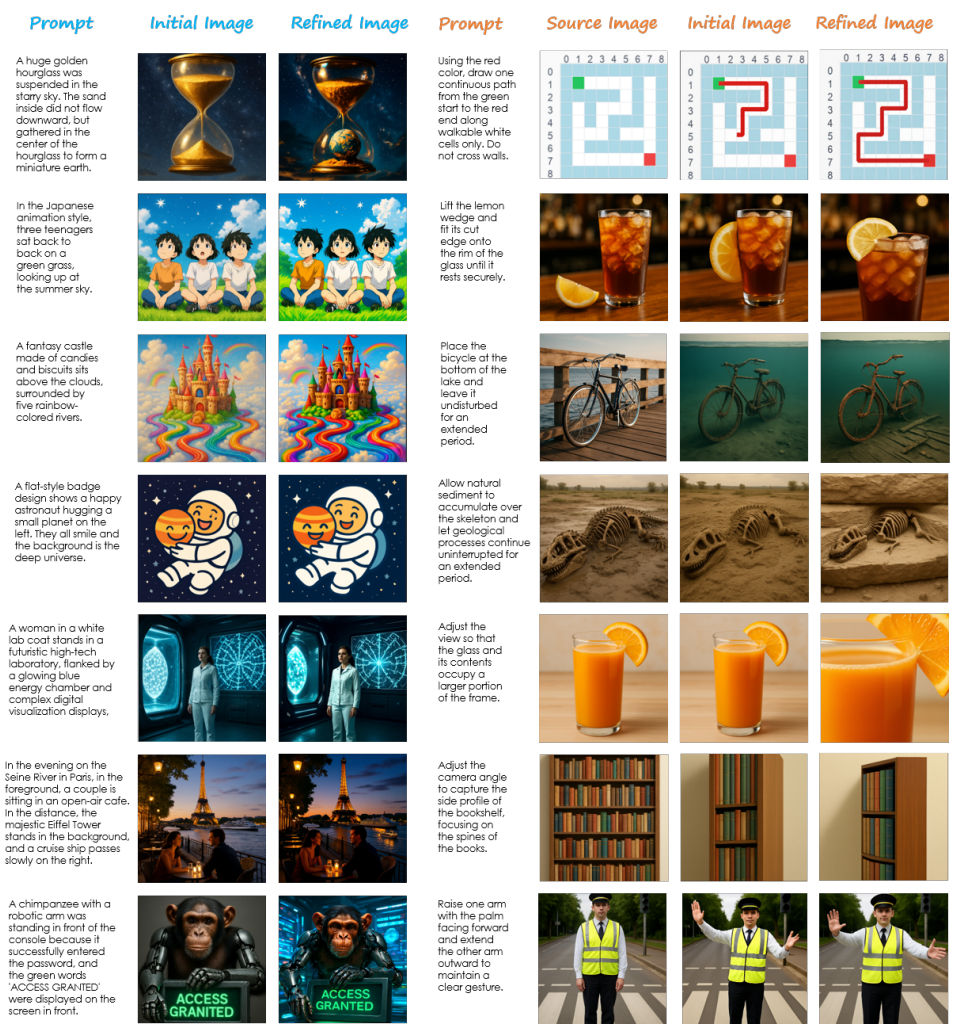

Figure 6: More case of UniReason on both T2I Generation(blue column) and Image Editing task(orange column).