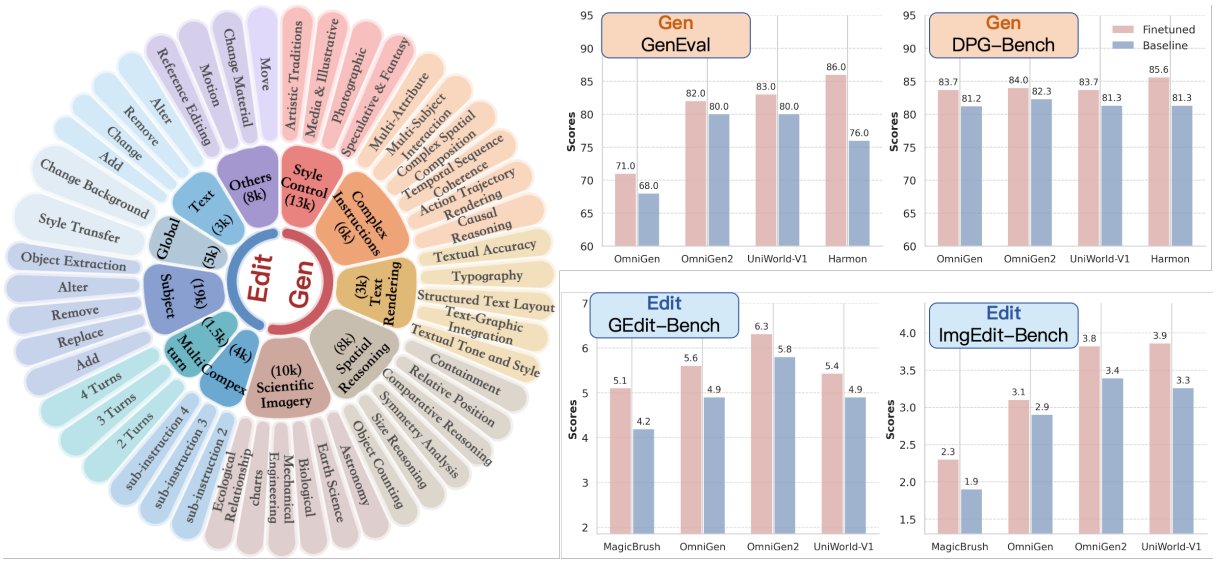
arXiv:2509.24900v2 [cs.CV] 17 Aug 2026

OpenGPT-4o-Image: A Comprehensive Dataset for Advanced Image Generation and Editing

Zhihong Chen[1,*], Xuehai Bai[3,*], Yang Shi[2,4,*,§], Chaoyou Fu[5],
Huanyu Zhang[6], Haotian Wang[7], Xiaoyan Sun[1], Zhang Zhang[6], Liang Wang[6],
Yuanxing Zhang[4,†], Pengfei Wan[4], Yi-Fan Zhang[6,♠,†]
[1] USTC [2] PKU [3] HDU [4] Kling Team [5] NJU [6] CASIA [7] THU
[*] Equal Contribution [♠] Project Leader [†] Corresponding Author

https://huggingface.co/datasets/WINDop/OpenGPT-4o-Image
https://github.com/NROwind/OpenGPT-4o-Image



Abstract

The performance of unified multimodal models for image generation and editing is fundamentally constrained by the quality and comprehensiveness of their training data. While existing datasets have covered basic tasks like style transfer and simple object manipulation, they often lack the systematic structure and challenging scenarios required for real-world applications. To address this bottleneck, we introduce **OpenGPT-4o-Image**, a large-scale dataset constructed using a novel methodology that combines hierarchical task taxonomy with automated data generation. Our taxonomy not only includes fundamental capabilities such as text rendering and style control but also introduces highly practical yet challenging categories like **scientific imagery** for chemistry illustrations and **complex instruction editing** requiring simultaneous execution of multiple operations. Through an automated pipeline leveraging structured resource pools and GPT-4o, we generate 80k high-quality instruction-image pairs with controlled diversity, covering 11 major domains and 51 subtasks. Extensive experiments show that fine-tuning leading models on our dataset achieves significant performance gains across multiple benchmarks, with improvements of up to 18% on editing tasks (UniWorld-V1 (Lin et al., 2025) on ImgEdit-Bench (Ye et al., 2025)) and 13% on generation tasks (Harmon (Wu et al., 2025d) on GenEval (Ghosh et al., 2024)). Our work demonstrates that systematic data construction is key to advancing multimodal AI capabilities.

§ Work done during an internship at Kling Team.

Contents

1 Introduction **3**

2 Related Work **5**
2.1 Unified Multimodal Large Language Models 5
2.2 Datasets for Image Generation 5
2.3 Datasets for Image Editing 5

3 Method **6**
3.1 Generation Type Definition 6
3.2 Edit Type Definition 8
3.3 Automatic Dataset Pipeline for Image Generation 8
3.4 Automatic Dataset Pipeline for Image Editing 9

4 Experimental Analysis **9**
4.1 Experimental Setting 10
4.2 Data Scaling Experiments 10
4.3 Compressive Evaluations 10

5 Conclusion **12**

A Dataset Distribution **15**
A.1 Distribution of Image Editing Data 15
A.2 Distribution of Image Generation Data 15

B Extended Experiments **16**
B.1 Supplementary Quantitative Experiments 16
B.2 Supplementary Qualitative Experiments 17
B.3 Supplementary Quantitative Experiments on Unified dataset 17

C Discussion **17**
C.1 Data Curation and Quality Control Strategy 17

Figure 1: **Illustrative examples of image generation from OpenGPT-4o-Image.** We comprehensively categorize image generation tasks into five groups based on the core capabilities they target: (a) Style Control, focusing on rendering diverse artistic and aesthetic styles; (b) Complex Instruction Following, which tests the model's ability to adhere to intricate compositional and logical constraints; (c) In-Image Text Rendering, involving the accurate generation and placement of text within images; (d) Spatial Reasoning, which demands geometric precision in tasks like object counting and relative positioning; and (e) Scientific Imagery, extending the model's application to specialized domains such as science, engineering, and data visualization.

1 Introduction

The field of AI-powered content creation is being transformed by unified multimodal models capable of both generating and editing images from natural language instructions (Deng et al., 2025; Lin et al., 2025; Liu et al., 2025),. Despite these achievements, creating training data that comprehensively addresses the full spectrum of real-world applications remains challenging. Existing datasets have made valuable contributions in areas such as style transfer, basic object manipulation, and increasingly, **text rendering** (Wu et al., 2025a)—a capability that has rightfully gained attention for its practical importance. However, certain complex scenarios that require specialized knowledge or sophisticated reasoning still present difficulties for current models. For instance, tasks involving technical illustrations for scientific education, or editing operations that require executing multiple instructions simultaneously, often reveal limitations in model capabilities. These challenges suggest opportunities for more structured approaches to data collection and taxonomy design.

To help address these opportunities, we introduce **OpenGPT-4o-Image**, a dataset designed to support the development of more capable and robust multimodal systems. Our work builds upon previous research while expanding coverage to include additional challenging scenarios. The dataset is organized around a hierarchical taxonomy that systematically addresses both established challenges and less explored areas. As shown in Figure 1 and 2, this includes not only refining existing focus areas like **style control** and **text rendering**, but also introducing categories such as: *1. Scientific Imagery*: Supporting technical illustration needs in fields like physics, chemistry, and biology, where visualizations play a critical role in education. *2. Complex Instruction Editing*: Addressing scenarios where users naturally provide multiple editing instructions that should be executed in concert. *3. Spatial Reasoning and Causal Inference*: Expanding

Figure 2: **Illustrative examples of image editing from OpenGPT-4o-Image.** We comprehensively categorize image editing tasks into six groups: (a) Subject Manipulation, focusing on region-based editing; (b) Text Editing, involving modifications to textual content embedded in images; (c) Complex Editing, involving the combination of multiple simple editing instructions; (d) Multi-Turn Editing, consisting of iterative, multi-round editing interactions; (e) Other Challenging Editings, covering additional difficult editing scenarios; and (f) Global Editing, which targets holistic modifications across the entire image.

beyond basic object recognition to include more sophisticated relational understanding.

To ensure scalability and consistency, we develop an automated pipeline that generates high-quality instruction-image pairs. This approach allows us to create 80,000 samples spanning 11 major domains and 51 subtasks with controlled diversity and difficulty levels. Our quantitative and qualitative experimental results demonstrate the utility of this approach. When fine-tuned on OpenGPT-4o-Image, several leading models show consistent improvements across multiple benchmarks. For example, UniWorld-V1 (Lin et al., 2025) achieves a 18% relative improvement on ImgEdit-Bench (Ye et al., 2025), while Harmon (Wu et al., 2025d) shows a 13% gain on GenEval (Ghosh et al., 2024). These improvements suggest that our structured approach to data construction can help models better handle complex and specialized tasks. In summary, our primary contributions are:

- ◇ A hierarchical taxonomy for image generation and editing that systematically decomposes complex tasks into 51 fine-grained sub-capabilities across 11 major domains. For image generation, this includes five core modules: *Style Control*, *Complex Instruction Following*, *In-Image Text Rendering*, *Spatial Reasoning*, and *Scientific Imagery*. For image editing, we define six categories with 21 subtasks, including *Subject Manipulation*, *Text Editing*, *Complex Instruction Editing*, *Multi-Turn Editing*, *Global Editing*, and other challenging forms.
- ◇ An automated, scalable pipeline for generating high-quality training data that GPT-4o to produce 80k instruction-image pairs with controlled diversity and difficulty levels. Our pipeline ensures comprehensive coverage of both fundamental capabilities and challenging scenarios through systematic task definition and template-based generation.
- ◇ Our experimental results demonstrate the substantial utility of our dataset and methodology. We employ four leading models spanning different architectural paradigms—including UniWorld-V1, Harmon, OmniGen2, and MagicBrush—to ensure the generalizability of our findings. The models

are evaluated on four standardized benchmarks: GEdit-Bench and ImgEdit-Bench for image editing capabilities, and GenEval and DPG-Bench for text-to-image generation quality. The evaluation reveals consistent and significant improvements across all tested configurations.

We hope that OpenGPT-4o-Image will contribute to future research by providing a resource that addresses a broader range of real-world applications, and that our methodology inspires further work on systematic data construction for multimodal AI.

2 Related Work

2.1 Unified Multimodal Large Language Models

The development of Unified Multimodal Large Language Models (Wu et al., 2025d; Deng et al., 2025; Lin et al., 2025) that integrate both multimodal understanding and generation capabilities has become a key research direction. This unification is motivated by the inherent synergy between understanding and generation—deep semantic comprehension enables controllable, high-quality synthesis (Wu et al., 2025a; Deng et al., 2025), while generative capability enhances complex reasoning through mechanisms like "thinking with generated images" Zhang et al. (2025b;a). Recent advances in Multimodal Large Language Models for perception (Bai et al., 2025) and diffusion models (Labs, 2024) for synthesis have made such unified approaches increasingly feasible. Training data represents another critical challenge in UFM development. For multimodal understanding tasks, the research community has established numerous large-scale, diverse datasets covering various capabilities such as visual question answering, image captioning, and visual reasoning. In contrast, high-quality datasets for generation and editing tasks remain significantly more limited. Existing generation datasets primarily focus on basic capabilities like style transfer and simple object manipulation, while complex scenarios requiring specialized knowledge, multi-instruction execution, or sophisticated reasoning remain underexplored. This data imbalance between understanding and generation capabilities motivates the construction of more comprehensive and challenging datasets to support the development of truly capable unified models.

2.2 Datasets for Image Generation

The advancement of text-to-image generation has been driven by several large-scale datasets (see Table 1 for representative examples). For instance, LAION-Aesthetics-UMAP (dclure, 2022), focus on curating images with high aesthetic scores, while DenseFusion-1M (Li et al., 2024a) provides dense descriptions rich in visual detail. Others, such as Megalith (madebyollin, 2024) and Public Domain 12M (Meyer et al., 2024), contribute vast image resources from real-world or public-domain sources. However, often constrained by the capabilities of the models available at the time of their creation, these earlier datasets exhibit limitations in their ability to support complex semantic understanding and precise instruction following. Recently, ShareGPT-4o-Image (Chen et al., 2025b) has made significant strides in data quality by leveraging the advanced generation capabilities of GPT-4o. While this dataset effectively distills the model's generative potential, its taxonomy is relatively coarse-grained and lacks a systematic, fine-grained structure. This somewhat limits its utility for targeted evaluation and in-depth analysis of specific model capabilities. To address this gap, we introduce OpenGPT-4o-Image, a text-to-image dataset built upon a clear, hierarchical taxonomy. It is designed to provide a resource that combines diversity, practicality, and evaluative depth, enabling more precise training and analysis of a model's multi-dimensional instruction-following abilities.

2.3 Datasets for Image Editing

Table 1 compares representative instruction-driven image editing datasets (Brooks et al., 2023; Zhang et al., 2023; Ge et al., 2024; Zhao et al., 2024; Hui et al., 2024; Yu et al., 2025; Wei et al., 2024; Ye et al., 2025). Most of these datasets (Brooks et al., 2023) rely primarily on open-source models to generate edited images. Although MagicBrush (Zhang et al., 2023) and SEED-Data-Edit (Ge et al., 2024) incorporate varying degrees of human quality control, their limited ability to fully interpret editing instructions often leads to suboptimal results. Specifically, InstructPix2Pix (IP2P) employs GPT-3 (Floridi &Chiriatti, 2020) to generate editing instructions and P2P (Hertz et al., 2022) for image editing; MagicBrush leverages extensive human annotations to improve data fidelity; HQ-Edit (Hui et al., 2024) uses DALL·E to produce paired images, but the outputs lack realism; AnyEdit (Yu et al., 2025) and OmniEdit (Wei et al., 2024) design alternative pipelines to support diverse editing tasks; and UltraEdit(Zhao et al., 2024) introduces a large-scale region-based editing dataset. In addition, ShareGPT-4o-Image (Chen et al., 2025b) constructs 46K instruction-driven image editing pairs using GPT-4o (Hurst et al., 2024), but it fails to adequately cover a broad spectrum of editing types, such as reference image editing. In contrast, we introduce

Table 1: **Comparison of Existing Datasets and OpenGPT-4o-Image. MTS** indicates support for multi-turn image editing, **CS** for complex instruction-based image editing, **TES** for visual text-based image editing, **ITR** for In-Image Text Rendering, and **SIG** for Scientific Image Generation.

Dataset	Size	Types	MTS	CS	TES	ITR	SIG
Image Editing							
Magicbrush (Zhang et al., 2023)	10K	5	✓	✗	✗	-	-
Seed-Data-Edit (Ge et al., 2024)	3.7M	6	✓	✗	✗	-	-
HQ-Edit (Hui et al., 2024)	197k	6	✗	✗	✗	-	-
AnyEdit (Yu et al., 2025)	2.5M	25	✗	✗	✗	-	-
IP2P (Brooks et al., 2023)	313K	4	✗	✗	✗	-	-
UltraEdit (Zhao et al., 2024)	4M	9	✗	✗	✗	-	-
OmniEdit (Wei et al., 2024)	1.2M	7	✗	✗	✗	-	-
ImgEdit (Ye et al., 2025)	1.2M	13	✓	✓	✗	-	-
Image Generation							
text-to-image-2M (jackyhate, 2024)	2M	-	-	-	-	✗	✗
Laion-aesthetics-umap (dclure, 2022)	12M	-	-	-	-	✗	✗
Densefusion-1m (Li et al., 2024a)	1M	-	-	-	-	✗	✗
Journeydb (Pan et al., 2023)	4M	-	-	-	-	✗	✗
Public Domain 12M (Meyer et al., 2024)	12M	-	-	-	-	✗	✗
Megalith (madebyollin, 2024)	10M	-	-	-	-	✗	✗
Text-Render-2M (Chen et al., 2025c)	2M	-	-	-	-	✓	✗
Unify							
ShareGPT-4o-Image (Chen et al., 2025b)	91K	29	✗	✗	✗	✗	✗
OpenGPT-4o-Image	80K	**51**	✓	✓	✓	✓	✓

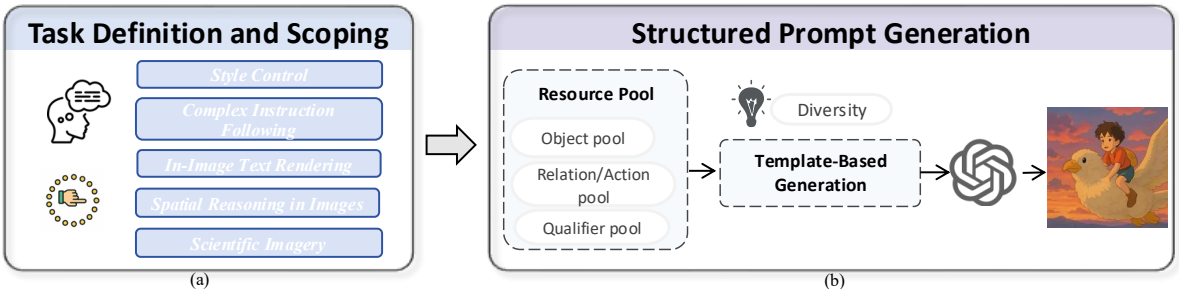


Figure 3: **Image Generation Data Construction Pipeline.** (a) illustrates the Task Definition and Scoping phase, where target capabilities are precisely defined, decomposed into hierarchical categories, and assigned difficulty grades. (b) shows the Structured Prompt Generation phase, where instructions are generated at scale by populating diverse syntactic templates with components from structured resource pools, which are then used to produce the final image.

OpenGPT-4o-Image, which selects high-quality source images, designs a comprehensive taxonomy of editing categories, and leverages GPT-4o to generate diverse instructions and high-quality edited images.

3 Method

OpenGPT-4o-Image encompasses a broader spectrum of categories, more precise and comprehensive instructions, and a substantial collection of practical as well as challenging generation and editing tasks. Most importantly, the quality of both image generation and editing achieved are remarkable. Section 3.1 and Section 3.2 enumerate the categories of image generation and editing, while Section 3.3 and Section 3.4 provide a detailed account of the data creation pipeline. More detailed information on the dataset distribution can be found in the Appendix A.

3.1 Generation Type Definition

Style Control. This module is designed to enhance the model's ability to render a diverse range of visual styles. Comprising a substantial 13k samples, it is organized into a comprehensive collection of

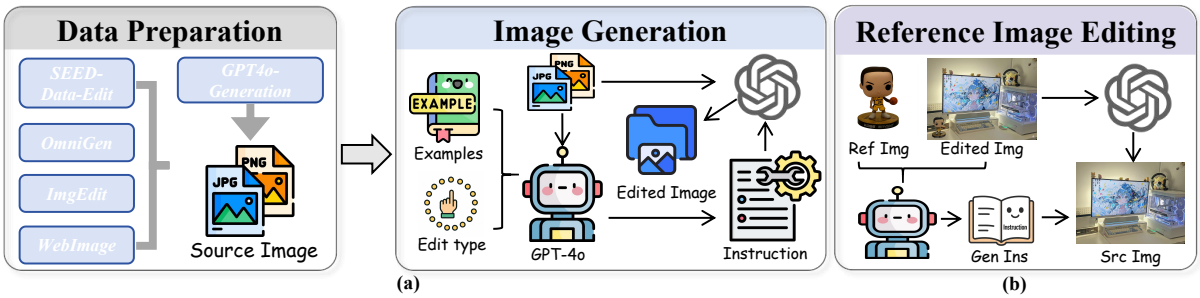


Figure 4: **Image Editing Data Construction Pipeline.** (a) illustrates the source image acquisition and demonstrates the use of GPT-4o for generating instructions and producing the edited image. (b) shows the process of reference image editing, where Subject-Driven Image Generation is used to create reference and edited images, with GPT-4o applied for inpainting the source image.

four distinct categories. The first, Artistic Traditions, covers historical and cultural art forms, including Western movements (Impressionism, Post-Impressionism, Cubism), Eastern traditions (Japanese Ukiyo-e, Ink Wash Painting, Traditional Chinese Painting), and contemporary forms like Graffiti Art. The second category, Media and Illustration, focuses on aesthetics from popular media and illustration, spanning animation (Ghibli Animation Style, Pixar Animation Style, Classic Animation Style), comics (Japanese Manga Style), digital-native styles (Pixel Art, Block-Based Art, Mosaic Style), and foundational techniques like Sketch and Line Art. The third, Photographic Styles, is dedicated to the visual language of photography, encompassing technical styles (Analog Film Aesthetic, High Dynamic Range (HDR)), artistic choices (Monochrome Photography), and mood-based aesthetics (Ethereal and Dreamlike, Vintage and Retro Aesthetics). Finally, Speculative and Fantasy Styles explores fictional genres such as Cyberpunk, Steampunk, Cosmic and Space Opera, and Digital Futurism.

Complex Instruction Following. This module, comprising a focused set of 6k samples, addresses the model's capacity for precise instruction adherence, particularly for prompts involving multiple constraints and complex logical relationships. We decompose this capability into several specific sub-tasks. These include the composition of static scenes through Multi-Attribute Combination and Multi-Subject Interaction and Action, as well as arranging elements according to a Complex Spatial Composition. Furthermore, the module assesses the understanding of dynamic and sequential concepts via Temporal Sequence Coherence and Action Trajectory Rendering. Finally, it probes the model's abstract reasoning abilities with prompts requiring Causal Reasoning.

In-Image Text Rendering. This module addresses in-image text rendering, a capability of high practical value that has been largely unaddressed in prior instruction-following datasets. It is designed around a core of 3k samples to systematically improve the model's ability to create accurate and aesthetic text-graphic compositions. The module covers foundational aspects such as Textual Accuracy (verbatim content rendering) and Typography (font control). It then progresses to more complex structural and relational challenges, including Structured Text Layout for multi-line arrangements and Text-Graphic Integration for the coherent placement of text within the image. Finally, the scope is expanded to include Multilingual Support for non-English scripts and the nuanced task of aligning Textual Tone and Style with the image's overall aesthetic.

Spatial Reasoning. In contrast to the prior modules, which primarily emphasize semantic and aesthetic interpretation, the Spatial Reasoning component focuses on the model's understanding of fundamental spatial and logical relationships. Consequently, tasks in this section demand geometric precision and logical fidelity. To assess this capability, we utilize a dedicated dataset of 8k samples divided into several categories, including basic topological relations and discrete 2D Relative Position. The model's numerical capacity is tested via Object Counting, while its understanding of geometric properties and comparisons is evaluated through Size Reasoning, Symmetry Analysis, and Comparative Reasoning.

Scientific Imagery. The Scientific Imagery module extends the applicability of text-to-image models to specialized professional domains, addressing the growing demand for image generation in scientific research and education where relevant training data remains scarce. To address this existing gap, we provide 10k samples covering a breadth of disciplines, including Mathematics, Physics, and Mechanical Engineering; natural sciences such as Astronomy and Earth Science; life sciences like Biological studies and Ecology; and topics from Culture and History. The introduction of this module provides a data foundation for exploring the model's potential in professional knowledge visualization and its applications in scientific communication and education.

3.2 Edit Type Definition

We define six categories of editing tasks: Subject Manipulation, Text Editing, Complex Editing, Multi-Turn Editing, Global Editing, and other challenging forms of editing. More specifically, these categories are further divided into 21 subtasks, which collectively encompass a wide range of practical instruction-based image editing scenarios.

Subject Manipulation. Building on classical instruction-based image editing, we define Subject Manipulation as the precise local modification of specific target objects within an image, while preserving the integrity and consistency of the background and other semantically irrelevant content. Subject Manipulation encompasses five operations: Add, Remove, Replace, Alter, and Object Extraction, comprising 19k samples. Specifically, Alter refers to modifications that change only the attributes of an object, whereas Replace involves a comprehensive transformation of the object itself. Additionally, Add and Remove denote the insertion or deletion of an object, respectively, whereas Object Extraction refers to isolating and extracting a specific object from the image.

Text Editing. Given the substantial gap between GPT-4o (Hurst et al., 2024) and existing open-source unified models in terms of text editing capabilities, we define Text Editing as the task of manipulating and modifying text elements embedded within images. In contrast to conventional image editing, which often focuses on coarse-grained objects, text editing requires a deeper comprehension of the rich semantics embedded in editing instructions and finer-grained editing capabilities. Drawing inspiration from Subject Manipulation, we further define four types: Text Add, Replace, Alter, and Remove, including 3k samples.

Complex Instruction Editing. With the increasing demand for image editing, users often issue multiple instructions that they expect to be executed simultaneously. However, under such scenarios, existing models frequently suffer from limited instruction-following capabilities and reduce image generation quality. To address this limitation, we introduce the task of Complex Instruction Image Editing, comprising a total of 4k samples, designed to enhance the ability of unified models to handle complex instructions. Furthermore, we categorize complex instructions based on their level of complexity. Each complex instruction consists of two to four distinct sub-editing operations, which are drawn from Text Editing and Subject Manipulation except Object Extraction.

Multi-turn Editing. Building on the insights from interactive text modifications in multimodal large language models, we extend the concept of multi-turn interactions to the domain of image editing. In this context, we introduce Multi-Turn Image Editing, including 1,500 samples, where iterative user interactions progressively guide the modification of image content. Specifically, we categorize the task by the number of interaction rounds, distinguishing between two-round, three-round, and four-round editing scenarios. This task is expected to strengthen unified models in multi-turn image editing and to achieve controllable editing aligned with user satisfaction.

Global Editing. In addition to local image editing, we also introduce global editing, comprising a total of 4k samples, which encompasses Background Replacement and Style Transfer. For style transfer, we design 11 distinct styles, including Cyberpunk Style, Ghibli Style, Ink Painting Style, Disney Animation Style, Hand-drawn Style, Monai Style, and others. In contrast, background replacement operations focus on substituting the surrounding environment while maintaining the integrity of foreground objects. This editing type maintains the subject elements and their spatial relationships, ensuring seamless integration into entirely different environments.

Other Challenging Editing. Beyond the aforementioned categories, we define a set of challenging editing tasks, including reference image editing, motion modification, material transformation, and object movement. Reference image editing involves the seamless incorporation of specified subjects into the original image. Motion modification emphasizes altering and adjusting the expressions and movements of objects, thereby enabling flexible control over their dynamic characteristics. Material transformation focuses on modifying the texture or composition of clothing and other materials. Finally, we introduce a novel task, Object movement, which entails repositioning an object from one location to another within the image while preserving spatial coherence and visual realism.

3.3 Automatic Dataset Pipeline for Image Generation

To systematically construct the large-scale, hierarchical training dataset for OpenGPT-4o-gen, we formulated and adhered to a general data construction pipeline, illustrated in Figure 3. This pipeline is designed to ensure that the data for each submodule has clear training objectives, controlled diversity, a reasonable difficulty distribution, and reliable quality. The process transforms abstract capability targets into concrete, trainable data and consists of two primary phases.

Task Definition and Scoping. This initial phase marks the starting point of the construction process,

where we clearly delineate the target capability for each submodule. We begin with Capability Definition and Boundary Setting, precisely defining the core skill each module aims to train. To ensure the purity of the training signal, we establish strict boundaries, specifying its scope and exclusions. For example, in the "Relative Position" module, we focused on planar relations like left of and above, while intentionally excluding other spatial concepts like 3D perspective or topological containment. For more complex capabilities, we perform Hierarchical Categorization, decomposing them into logically interconnected sub-categories. In the "Causal Reasoning" module, for instance, we divided the task into levels such as "Explicit Causality," "Implicit Inference," and "Complex Causal Chains" to enable more granular training and evaluation. Finally, we implement Difficulty Grading for each submodule, assessing difficulty based on factors like instruction complexity, required background knowledge, and the length of the reasoning chain. This allows us to generate instruction sets with a well-distributed range of difficulty.

Structured Prompt Generation. Once the task definitions are established, this phase focuses on the efficient and high-quality generation of instruction. We first implement Resource Pool Design, constructing a series of structured resource pools to serve as foundational components for prompt generation. These pools typically include an Object Pool (various entities), a Relation/Action Pool (core verbs, prepositions, and their synonyms), and a Qualifier Pool (scenes, materials, adjectives) to increase the naturalness of the prompts. We then employ Template-Based Generation, designing multiple templates with diverse syntactic structures and populating them with randomly sampled components from the resource pools. This method allows us to generate instructions at scale that possess both structural consistency and linguistic variety. To further ensure data quality, we apply a Diversity Strategy, which includes varying the presentation format of instructions, controlling the combinatorial logic of content elements, and judiciously injecting content-aligned stylistic elements into a subset of prompts to enhance their realism.

3.4 Automatic Dataset Pipeline for Image Editing

Data Preparation. We construct the dataset corpus by integrating multiple high-quality sources, including SEED-Data-Edit (Ge et al., 2024), ImgEdit (Ye et al., 2025), GPT-4o (Hurst et al., 2024) generated images, OmniEdit (Wei et al., 2024), and a curated collection of high-resolution images as the initial source corpus. Compared with existing datasets, our corpus contains source images of substantially higher quality, a broader spectrum of editing categories, more diverse instruction types, and correspondingly higher-quality edited results. To ensure dataset quality, we select distinct source corpora for each editing type. For text editing, GPT-4o is used to generate images with embedded textual elements, serving as the source corpus for text-related tasks. Reference image editing utilizes reference-target image pairs from the subject-driven generation component of OmniEdit, which are adopted as target and reference images in our corpus. Motion and material modifications are based on original images from ImgEdit. For object removal tasks, we include a subset of carefully curated high-resolution images, each with a resolution exceeding 1024 pixels. Finally, for multi-turn and other editing tasks, we leverage the multi-turn corpus from SEED-Data-Edit as it contains high-quality real-world images.

Instruction Generation. We provide the original image, the editing type, and a set of in-context examples to facilitate prompt generation. For style transfer, we define 10 distinct styles and generate the associated instructions. Additionally, we provide sub-reference editing types for complex instruction-based images, including Subject Manipulation and Text Editing, and supply carefully designed examples to guide the generation of diverse instructions. For reference image editing, we draw inspiration from subject-driven image generation. As illustrated in Figure 4(b), we provide the reference image, the result generated from subject-driven generation, and several in-context examples, which are then used with GPT-4o to generate instructions. For multi-turn image editing, we define 10 subcategories of tasks, including Text Editing, Global Editing, and Subject Manipulation. We further provide in-context examples and generate interactive multi-turn editing instructions.

Image Generation. To construct the OpenGPT-4o-Image dataset, the core of our data generation process relied on the gpt-image-1 API, which was employed to regenerate or augment all image–instruction pairs. Furthermore, as illustrated in Figure 4(b), for subject-driven image editing, we utilize inpainting instructions together with the edited outputs to generate the corresponding original source images. Finally, multi-turn image editing generates the results of each round progressively, ensuring the overall quality of the editing process.

4 Experimental Analysis

In this section, we conduct a comprehensive evaluation of our dataset. Section 4.1 provides a detailed specification of the baseline models, the evaluation benchmarks, and the experimental setup. Section 4.2 presents the results of data scaling experiments, where subsets of different sizes were sampled to assess

Table 2: Comparison of fine-tuning results of different models on our dataset on ImgEdit-Bench. ‡ indicates results from our own tests without fine-tuning. † denotes results without fine-tuning.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
Open-source Models										
IP2P (Brooks et al., 2023)	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit (Jiang et al., 2025b)	3.18	2.95	1.88	2.47	2.23	2.23	2.85	1.56	2.65	2.45
UltraEdit (Zhao et al., 2024)	3.44	2.81	2.13	2.96	1.45	2.86	3.76	1.91	2.98	2.70
OmniGen (Xiao et al., 2025)	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
Step1X-Edit (Liu et al., 2025)	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
ICEdit (Zhang et al., 2025c)	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
BAGEL (de Jong et al., 2006)	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
OmniGen2 (Wu et al., 2025c)	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	**4.68**	3.44
Ovis-U1 (Wang et al., 2025)	4.13	3.62	**2.98**	**4.45**	**4.06**	**4.22**	**4.69**	**3.45**	4.61	**4.00**
FluxKontext dev (Labs et al., 2025)	3.76	3.45	2.15	3.98	2.94	3.78	4.38	2.96	4.26	3.52
Proprietary Models										
GPT-4o	4.61	4.33	2.9	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Finetuning										
MagicBrush† (Zhang et al., 2023)	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
MagicBrush	2.66	2.12	1.45	2.04	2.03	2.18	2.04	2.19	1.60	2.30
OmniGen‡ (Xiao et al., 2025)	3.27	3.05	1.90	2.82	2.43	3.00	4.11	1.67	3.27	2.90
OmniGen	3.75	3.48	2.17	2.68	2.03	3.23	4.17	2.87	3.81	3.10
OmniGen2‡ (Wu et al., 2025c)	3.60	3.44	1.94	3.81	2.54	3.79	4.57	2.71	4.49	3.39
OmniGen2	4.15	3.69	2.53	4.11	3.64	4.10	**4.69**	2.99	**4.68**	3.82
UniWorld-V1† (Lin et al., 2025)	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
UniWorld-V1	**4.34**	**4.28**	2.66	3.92	3.30	4.15	4.62	3.43	3.97	3.86

the impact of scaling on performance trends. Section 4.3 provides both qualitative and quantitative analyses of the experimental outcomes.

4.1 Experimental Setting

Baselines. For the data scaling experiments, we employ UniWorld-V1 (Lin et al., 2025), which integrates Qwen2.5-VL (Bai et al., 2025) as the comprehension model and FLUX-dev (Labs et al., 2025) as the vision generation model. For the comprehensive evaluations, we conducted extensive comparisons across different modeling paradigms, including diffusion-based and autoregressive frameworks. The evaluated models comprise UniWorld-V1, Harmon (Wu et al., 2025d), OmniGen2 (Wu et al., 2025c), MagicBrush (Zhang et al., 2023), and OmniGen (Xiao et al., 2025).

Benchmarks. Our evaluation methodology encompasses a thorough quality assessment for both image generation and editing datasets. For the image generation dataset, we comprehensively evaluate its quality using GenEval and DPG-Bench; GenEval specifically assesses compositionality, while DPG-Bench focuses on semantic alignment. Similarly, our image editing dataset is comprehensively evaluated across two widely used benchmarks: GEdit-Bench and ImgEdit-Bench. GEdit-Bench covers 11 distinct evaluation dimensions and assesses visual text and portrait editing, whereas ImgEdit-Bench further provides fine-grained evaluation of complex instruction-based image editing.

4.2 Data Scaling Experiments

To validate the effectiveness of our dataset, as shown in Table 6, we perform uniform sampling on subsets of the dataset at 20K, 30K, and 40K sizes. Since relying on a single benchmark can introduce bias and lead to incomplete evaluations, we employ two benchmarks for a comprehensive assessment. We also compute the overall average across both benchmarks and observe a consistent upward trend in the average performance as the dataset size increases. Additionally, given the relatively small increase in performance between 30K and 40K, we select the 40K dataset size.

4.3 Compressive Evaluations

Performance Improvement on Image Editing Model. As shown in Table 2 and Table 5, we finetune on MagicBrush (Zhang et al., 2023), OmniGen (Xiao et al., 2025), UniWorld-V1 (Lin et al., 2025), and OmniGen2 (Wu et al., 2025c). MagicBrush achieves improvements of 21.1% and 21.7% on ImgEdit-Bench and GEdit-Bench, respectively. OmniGen gains of 6.9% and 14.0% on the two benchmarks. UniWorld-V1 improves by 18.4% on ImgEdit-Bench and 12.0% on GEdit-Bench, while OmniGen2 achieved 12.7% and 8.8% improvements, respectively. The results indicate that leveraging more advanced comprehension and editing models, together with fine-tuning on our high-quality dataset, enables UniWorld-V1 and OmniGen2 to achieve notable improvements across multiple dimensions, such as Adjust, and Compose.

Table 3: Comparison of fine-tuning results of different models on our dataset on GenEval (Ghosh et al., 2024) benchmark. Results of GPT-4o are tested by (Yan et al., 2025). ‡ indicates results from our own tests without fine-tuning. † denotes results without fine-tuning.

Method	Single object	Two object	Counting	Colors	Position	Color attribution	Overall
Open-source Models							
SDv2.1 (Rombach et al., 2022)	0.98	0.5	0.44	0.85	0.07	0.17	0.50
SDXL (Podell et al., 2023)	0.98	0.74	0.39	0.85	0.15	0.23	0.55
IF-XL	0.97	0.74	0.66	0.81	0.13	0.35	0.61
LUMINA-Next (Zhuo et al., 2024)	0.92	0.46	0.48	0.70	0.09	0.13	0.46
SD3-medium (AI, 2024)	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (Labs, 2024)	0.99	0.81	0.79	0.74	0.20	0.47	0.67
OmniGen (Xiao et al., 2025)	0.98	0.84	0.66	0.74	0.40	0.43	0.68
TokenFlow-XL (Qu et al., 2025)	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus (Wu et al., 2025b)	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus Pro (Chen et al., 2025d)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Emu3-Gen (Wang et al., 2024)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
Show-o (Xie et al., 2024)	0.98	0.80	0.66	0.84	0.31	0.50	0.68
MetaQuery-XL (Pan et al., 2025)	-	-	-	-	-	-	0.80
BLIP3-o 8B (Chen et al., 2025a)	-	-	-	-	-	-	0.84
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82
Proprietary Models							
GPT-4o	0.99	0.92	**0.85**	**0.92**	0.75	0.61	0.84
Finetuning							
OmniGen† (Xiao et al., 2025)	0.98	0.84	0.66	0.74	0.40	0.43	0.68
OmniGen	0.99	0.88	0.68	0.79	0.45	0.44	0.71
OmniGen2† (Wu et al., 2025c)	0.99	0.92	0.75	0.87	0.60	0.69	0.80
OmniGen2	0.99	0.93	0.76	0.88	0.66	0.72	0.82
UniWorld-V1† (Lin et al., 2025)	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1	0.99	0.96	0.82	0.88	0.60	**0.73**	0.83
Harmon† (Wu et al., 2025d)	0.99	0.86	0.66	0.85	0.74	0.48	0.76
Harmon	**0.99**	**0.96**	0.79	0.88	**0.85**	0.69	**0.86**

Performance Improvement on Image Generation Model. We fine-tune four representative models: OmniGen, OmniGen2, UniWorld-V1, and Harmon. As detailed in Table 3 and Table 4, all models exhibit significant performance improvements after this process. For instance, Harmon's performance surges by 13.2% on Geneval and 5.3% on DPG-Bench. Even Omnigen2, which shows the most modest gains, still achieves robust improvements of 2.5% and 1.9%. We hypothesize that Harmon's markedly superior performance compared to the other models stems from its efficient architecture and smaller 1.5B parameter size. Notably, these substantial results are achieved using our dataset of only 40k samples. This strongly demonstrates that our dataset's clear taxonomy and high-quality instructions enhance a model's text-to-image capabilities, particularly in its precision for following complex instructions.

Performance Improvement on Unified Dataset. As shown in Table 7, we finetune UniWorld-V1 using both ShareGPT-4o and our image editing dataset. Our dataset achieves significant improvements, surpassing ShareGPT-4o by 3.2% on ImgEdit-Bench, 1.7% on GEdit-Bench, 1.2% on Geneval and 1.1% on DPG-Bench. Under the same training configuration, our dataset enhances the model's performance due to its more comprehensive categorization, richer set of editing instructions.

Quantitative Comparison. Due to the limitations of quantitative metrics in evaluating editing tasks, we further conduct qualitative evaluations to assess the effectiveness of our dataset, as illustrated in Figure. 6. Before finetuning, UniWorld-V1 exhibits suboptimal performance in following complex instructions, particularly in tasks involving object replacement and action modification. After efficient fine-tuning with the OpenGPT-4o-Image dataset, the model demonstrates substantial qualitative improvements. As shown in Figure 6, the fine-tuned UniWorld-V1 successfully replaces the hat with a teapot and raises the person's right arm, whereas the original model fails to execute these editing instructions effectively. Next, we turn to the Harmon model's improvements on text-to-image generation tasks. Here, the fine-tuned model demonstrates comprehensive performance gains, particularly in its ability to follow complex instructions. Specifically, we observe significant enhancements in handling various fine-grained capabilities, including in-image text rendering, multi-entity composition, temporal and spatial reasoning, and relative size comparison. These improvements collectively indicate that our dataset effectively strengthens the model's integrated understanding of multi-dimensional, complex semantics and its generation fidelity.

We conduct comprehensive quantitative and data scaling experiments, as detailed in Appendix B. These experiments demonstrate that fine-tuning with our image generation and editing dataset yields consistent improvements across a wide range of benchmarks, while also exhibiting qualitative superiority. Notably, our approach achieves substantial improvements over concurrent work such as ShareGPT-4o-Image. The detailed characterization of our dataset distribution and our data curation and quality control strategy

are presented in Appendix A and Appendix C, respectively.

5 Conclusion

This paper introduces OpenGPT-4o-Image, a comprehensive dataset designed to advance multimodal AI capabilities in image generation and editing through systematic task decomposition and automated data construction. The work addresses significant gaps in existing datasets by providing 80,000 high-quality instruction-image pairs across 11 major domains and 51 subtasks, with particular emphasis on previously underexplored areas such as scientific imagery, complex instruction following, and multi-turn editing. The key contributions include a hierarchical taxonomy that systematically categorizes image generation into five core modules (Style Control, Complex Instruction Following, In-Image Text Rendering, Spatial Reasoning, and Scientific Imagery) and image editing into six categories with 21 subtasks. The automated pipeline leverages GPT-4o to ensure scalable, consistent data generation while maintaining controlled diversity and difficulty distribution. Experimental validation demonstrates the dataset's effectiveness across multiple model architectures and benchmarks. However, the work has limitations. The reliance on GPT-4o for data generation may introduce biases inherent to that model, and the evaluation is primarily conducted on existing benchmarks which may not fully capture real-world application scenarios.

References

Stability AI. Sd3-medium. https://stability.ai/news/stable-diffusion-3-medium, 2024.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

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.

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. *arXiv preprint arXiv:2506.18095*, 2025b.

SiXiang Chen, Jianyu Lai, Jialin Gao, Tian Ye, Haoyu Chen, Hengyu Shi, Shitong Shao, Yunlong Lin, Song Fei, Zhaohu Xing, et al. Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework. *arXiv preprint arXiv:2506.10741*, 2025c.

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*, 2025d.

dclure. Laion-aesthetics-umap. https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap, 2022.

Anne de Jong, Sacha AFT van Hijum, Jetta JE Bijlsma, Jan Kok, and Oscar P Kuipers. Bagel: a web-based bacteriocin genome mining tool. *Nucleic acids research*, 34(suppl_2):W273–W279, 2006.

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.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 30(4):681–694, 2020.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

jackyhate. text-to-image-2m. https://huggingface.co/datasets/jackyhate/text-to-image-2M, 2024.

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.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025b.

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

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

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.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024a.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024b.

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.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

madebyollin. Megalith-huggingface. https://huggingface.co/datasets/madebyollin/megalith-10m, 2024.

Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv preprint arXiv:2410.23144*, 2024.

AI Open. Introducing 4o image generation, 2025.

OpenAI. Dall·e 3. https://openai.com/index/dall-e-3, 2024.

Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

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

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

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, 2025b.

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*, 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.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

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.

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

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. *arXiv preprint arXiv:2505.20275*, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yifan Zhang, Haochen Tian, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, et al. Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes. *arXiv preprint arXiv:2504.15037*, 2025a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025b.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025c.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.

A Dataset Distribution

A.1 Distribution of Image Editing Data

In this section, we provide a more detailed characterization of the distribution of the image editing dataset. As summarized in Table 9, the corpus is organized into six major categories: Subject Manipulation (19k), Text Editing (3k), Complex Instruction Editing (4k), Multi-turn Editing (1.5k), Global Editing (5k), and Other Challenging Editing (8k). Because Subject Manipulation and Global Editing represent practical and canonical editing tasks, we allocate comparatively larger quantities to these categories to strengthen performance in these areas. For Text Editing, Multi-turn Editing, and the sub-tasks under Other Challenging Editing, we employ templated generation to curate a relatively smaller but targeted set of instances, enabling an examination of how high-quality fine-tuning translates into improvements for these capabilities. Finally, we include a moderate amount of Complex Instruction Editing to further enhance instruction following; this capability shows substantial gains after fine-tuning, as reflected by marked improvements on ImgEdit-Bench.

Table 4: Comparison of fine-tuning results of different models on our dataset on DPG-Bench (Hu et al., 2024). ‡ indicates results from our own tests without fine-tuning. † denotes results without fine-tuning.

Method	Global	Entity	Attribute	Relation	Other	Overall
Open-source Models						
SDXL (Podell et al., 2023)	83.27	82.43	80.91	86.76	80.41	74.65
Hunyuan-DiT (Li et al., 2024b)	84.59	80.59	88.01	74.36	86.41	78.87
DALLE3 (OpenAI, 2024)	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium (AI, 2024)	87.90	**91.01**	88.83	80.70	88.68	84.08
FLUX.1-dev (Labs, 2024)	82.1	89.5	88.7	91.1	89.4	84.0
Show-o (Xie et al., 2024)	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 (Wang et al., 2024)	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL (Qu et al., 2025)	78.72	79.22	81.29	85.22	71.20	73.38
Janus Pro (Chen et al., 2025d)	86.90	88.90	89.40	89.32	89.48	84.19
T2I-R1 (Jiang et al., 2025a)	91.79	90.23	89.05	90.13	89.48	84.76
BLIP3-o 4B (Chen et al., 2025a)	-	-	-	-	-	79.36
BLIP3-o 8B (Chen et al., 2025a)	-	-	-	-	-	81.60
BAGEL (Deng et al., 2025)	88.94	90.37	**91.29**	90.82	88.67	85.07
Finetuning						
OmniGen† (Xiao et al., 2025)	87.90	88.97	88.47	87.95	83.56	81.16
OmniGen	86.69	88.08	90.40	91.39	90.77	83.76
OmniGen2‡ (Wu et al., 2025c)	85.69	88.98	86.88	**91.54**	**91.38**	82.37
OmniGen2	88.75	89.31	89.55	90.67	91.08	84.01
UniWorld-V1† (Lin et al., 2025)	83.64	88.39	88.44	89.27	87.22	81.38
UniWorld-V1	**92.81**	89.56	89.41	88.93	88.52	83.66
Harmon† (Wu et al., 2025d)	83.53	85.45	86.31	87.93	89.03	81.27
Harmon	92.30	90.35	90.45	91.47	91.19	**85.60**

A.2 Distribution of Image Generation Data

Similarly, the generation dataset is strategically structured to build a comprehensive range of creative and technical capabilities, with its distribution detailed in Table 10. The corpus is organized into

five major categories: Style Control (13k), Scientific Imagery (10k), Spatial Reasoning (8k), Complex Instruction Following (6k), and In-Image Text Rendering (3k). The largest allocations are given to Style Control and Scientific Imagery to, respectively, master a foundational pillar of creative expression and establish a robust data foundation for professional domains where high-quality training data is typically scarce. Substantial resources are also devoted to Spatial Reasoning and Complex Instruction Following to systematically enhance the model's grasp of logical relationships, geometric precision, and compositional directives. Finally, a more targeted collection for In-Image Text Rendering (3k) is included to specifically address persistent challenges in textual accuracy and typography, aiming to improve the model's reliability in this critical area.

B Extended Experiments

Table 5: Comparison of fine-tuning results of different models on our dataset on GEdit-Bench. ‡ indicates results from our own tests without fine-tuning. † denotes results without fine-tuning.

Model	Background	Color	Material	Motion	Portrait	Style	Add	Remove	Replace	Text	Tone	Avg
Open-source Models												
AnyEdit (Jiang et al., 2025b)	4.31	4.25	2.64	0.67	1.9	1.95	3.72	3.75	3.23	0.77	4.21	2.85
IP2P (Brooks et al., 2023)	3.94	5.4	3.52	1.27	2.62	4.39	3.07	1.5	3.48	1.13	5.1	3.22
OmniGen (Xiao et al., 2025)	5.23	5.93	5.44	3.12	3.17	4.88	6.33	6.35	5.34	4.31	4.96	5.01
Step1X-Edit (Liu et al., 2025)	7.03	6.26	6.46	3.66	5.23	7.24	7.17	6.42	7.39	**7.40**	6.62	6.44
Bagel (de Jong et al., 2006)	7.44	6.99	6.26	5.09	4.82	6.04	**7.94**	7.37	7.31	7.16	6.17	6.60
Bagel-thinking	7.22	7.24	6.69	7.12	6.03	6.17	7.93	**7.44**	**7.45**	3.61	6.36	6.66
Ovis-U1 (Wang et al., 2025)	**7.49**	6.88	6.21	4.79	**5.98**	6.46	7.49	7.25	7.27	4.48	6.31	6.42
OmniGen2 (Wu et al., 2025c)	-	-	-	-	-	-	-	-	-	-	-	6.42
Step1X-Edit (v1.1)	7.45	**7.38**	**6.95**	**4.73**	4.70	7.11	8.2	7.59	7.8	7.91	6.85	**6.97**
FluxKontext dev (Labs et al., 2025)	7.06	7.03	5.52	5.62	4.68	5.55	6.95	6.76	6.13	6.10	7.48	6.26
Proprietary Models												
Doubao (Shi et al., 2024)	8.07	7.36	7.20	5.38	6.28	7.2	8.05	7.71	7.87	4.01	7.67	6.98
Gemini (Kampf &Brichtova, 2025)	7.11	7.14	6.47	5.67	3.99	4.95	8.12	6.89	7.41	6.85	7.01	6.51
GPT-4o (Open, 2025)	6.96	6.85	7.10	5.41	6.74	7.44	7.51	8.73	8.55	8.45	8.69	7.49
Finetuning												
MagicBrush† (Zhang et al., 2023)	6.17	5.41	4.75	1.55	2.9	4.1	5.53	4.13	5.1	1.33	5.07	4.19
MagicBrush	5.84	6.07	5.08	3.41	4.17	5.94	5.84	5.95	5.01	2.41	6.36	5.10
OmniGen‡ (Xiao et al., 2025)	4.87	5.57	4.75	2.57	4.09	5.84	6.04	4.77	5.42	4.41	5.21	4.87
OmniGen	5.83	6.79	5.25	4.82	4.64	5.83	6.08	5.66	6.02	3.90	6.24	5.55
OmniGen2‡ (Wu et al., 2025c)	7.04	6.32	6.21	3.56	2.94	6.74	6.42	6.14	6.93	4.86	6.62	5.80
OmniGen2	7.23	6.46	6.73	4.65	4.81	7.07	6.69	6.42	7.03	5.43	6.83	6.31
UniWorld-V1† (Lin et al., 2025)	4.92	6.37	4.79	1.85	4.03	5.64	7.23	6.17	5.70	1.15	5.54	4.85
UniWorld-V1	5.38	**7.38**	5.22	3.52	4.03	**6.88**	7.07	5.23	5.50	2.13	**7.39**	5.43

B.1 Supplementary Quantitative Experiments

To provide a fuller account of the evaluation, we partition the additional experiments into three facets: scaling the training data, assessing benchmark outcomes, and comparing unified dataset. Table 6 reports results obtained by fine-tuning UniWorld-V1 on incrementally larger samples drawn from our dataset; the measurements consistently indicate performance growth with increasing data volume, reinforcing the importance of dataset size in enabling stronger generalization.

In parallel, Table 5 presents a comprehensive evaluation on GEdit-Bench, which demonstrates that UniWorld-V1 attains state-of-the-art results among closed-source systems on representative editing tasks, including color change and tone transfer. Similarly, on the DPG-Bench, the fine-tuned Harmon model

Table 6: Data scaling results on GEdit-Bench and ImgEdit-Bench, obtained by randomly sampling datasets of different sizes. Two Avg represents the average performance on GEdit-Bench together with the overall average on ImgEdit-Bench.

Size	Background	Color	Material	Motion	Portrait	Style	Add	Remove	Replace	Text	Tone	Avg
GEdit-Bench												
20K	5.60	7.30	6.46	3.82	3.77	6.57	7.73	7.73	5.60	1.90	6.83	5.50
30K	6.80	7.64	6.23	4.98	3.19	6.84	7.66	5.52	5.45	1.13	6.87	5.79
40K	5.38	7.38	5.22	3.52	4.03	6.88	7.07	5.23	5.50	2.13	7.39	5.43

Size	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall	Two Avg
ImgEdit-Bench											
20K	4.09	4.28	2.46	3.85	2.93	3.99	4.58	3.19	4.07	3.72	6.46
30K	4.16	4.02	2.53	3.68	2.75	4.06	4.49	2.98	3.81	3.64	6.53
40K	4.34	4.28	2.66	3.92	3.30	4.15	4.62	3.43	3.97	3.86	6.58

Table 7: Comparison of fine-tuning results on UniWorld-V1 with ShareGPT-4o-Image. Two Avg represents the average performance on GEdit-Bench together with the overall average on ImgEdit-Bench.

Dataset	Background	Color	Material	Motion	Portrait	Style	Add	Remove	Replace	Text	Tone	Avg
GEdit-Bench												
ShareGPT-4o-Image	4.87	7.70	5.59	2.20	4.04	7.29	6.94	5.03	5.13	2.12	6.93	5.26
Ours	5.38	7.38	5.22	3.52	4.03	6.88	7.07	5.23	5.50	2.13	7.39	5.43

Dataset	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall	Two Avg
ImgEdit-Bench											
ShareGPT-4o-Image	4.16	4.24	2.45	3.85	2.9	3.98	4.65	2.99	3.77	3.7	6.33
Ours	4.34	4.28	2.66	3.92	3.3	4.15	4.62	3.43	3.97	3.86	6.58

Dataset	Single object	Two object	Counting	Colors	Position	Color attribution	Overall
GenEval-Bench							
ShareGPT-4o-Image	0.99	0.94	0.82	0.87	0.56	0.71	0.82
Ours	0.99	0.96	0.82	0.88	0.60	0.73	0.83

Size	Global	Entity	Attribute	Relation	Other	Overall
DPG-Bench						
ShareGPT-4o-Image	87.46	88.75	88.97	90.36	88.76	82.71
Ours	92.81	89.56	89.41	88.93	88.52	83.66

achieves the best performance, as shown in Table4. Moreover, the same fine-tuning protocol yields marked improvements for the remaining models considered, suggesting that the gains are not confined to specific models or benchmarks but instead reflect stronger generalization capacity. In terms of the unified dataset, we conducted a comparison between our dataset and ShareGPT4o. Specifically, we employed UniWorld-V1 distributions to finetune both the generation and editing components. As summarized in Table 7, our dataset consistently outperforms ShareGPT4o in both generation and editing tasks. This advantage can be attributed to our meticulous categorization, the diversity of editing instructions, and the high-quality editing outcomes.

B.2 Supplementary Qualitative Experiments

Qualitative results for generation are shown in Figure 5, with analysis in Section 4.3. For the editing task, beyond the qualitative example in Figure 6 (which shows the model simultaneously removing a laptop and adding a light blue sofa), we also present a comprehensive quantitative study. This study, covering multiple editing types like Add Subject and Change Background, indicates that fine-tuning improves not only editing performance but also image fidelity. Moreover, for hybrid instructions, our dataset enhances the model's ability to follow complex, compositional directives.

B.3 Supplementary Quantitative Experiments on Unified dataset

We further compare unified training with separate training for generation and editing. As shown in Table 8, for image editing, separate fine-tuning surpasses unified training on ImgEdit-Bench, whereas unified training outperforms separate fine-tuning on GEdit-Bench. For image generation, separate fine-tuning performs on par with unified training on GenEval-Bench, but surpasses it on DPG-Bench. These results suggest a benchmark-dependent trade-off between unified and task-specific training, indicating potential task interference and underscoring the importance of tailored optimization for different evaluation regimes.

C Discussion

C.1 Data Curation and Quality Control Strategy

A primary challenge in curating our dataset is ensuring high fidelity to complex, compositional instructions. While frontier models like GPT-4o consistently produce aesthetically pleasing images, their instruction-following capabilities often degrade when faced with highly complex prompts demanding adherence to numerous fine-grained details. This limitation is also prevalent in existing open-source models. Consequently, a simple post-hoc filtering strategy proved ineffective, as high aesthetic quality is often a poor proxy for semantic correctness and instruction-following accuracy.

To address this, we adopted a proactive quality control strategy centered on meticulous, fine-grained data curation prior to generation. Our methodology is guided by two core principles:

Table 8: Comparison of fine-tuning performance on UniWorld-V1 across different training strategies. The table evaluates model performance when fine-tuned separately on generation-only and editing-only datasets versus a unified dataset combining both.

Dataset	Background	Color	Material	Motion	Portrait	Style	Add	Remove	Replace	Text	Tone	Avg
GEdit-Bench												
Edit-Only	5.38	7.38	5.22	3.52	4.03	6.88	7.07	5.23	5.50	2.13	7.39	5.43
Unify	6.01	7.47	6.97	2.37	4.54	7.05	7.11	5.99	5.36	2.01	6.95	5.62

Dataset	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
ImgEdit-Bench										
Edit-Only	4.34	4.28	2.66	3.92	3.30	4.15	4.62	3.43	3.97	3.86
Unify	4.13	3.86	2.45	3.52	2.81	3.86	4.65	3.32	3.78	3.60

Dataset	Single object	Two object	Counting	Colors	Position	Color attribution	Overall
GenEval-Bench							
Gen-Only	0.99	0.96	0.82	0.88	0.60	0.73	0.83
Unify	0.99	0.96	0.80	0.87	0.60	0.72	0.83

Dataset	Global	Entity	Attribute	Relation	Other	Overall
DPG-Bench						
Gen-Only	92.81	89.56	89.41	88.93	88.52	83.66
Unify	89.62	89.39	88.80	87.66	88.39	82.45

Hierarchical Categorization. We first establish a clear hierarchy by defining distinct modules (e.g., Style Control, Spatial Reasoning) and then partitioning them into more granular, well-defined sub-classes. This ensures thematic coherence and targeted data collection.

Difficulty Calibration. Within each category, we carefully calibrate the task difficulty to occupy a specific "sweet spot". The instructions are designed to be challenging for current open-source models yet demonstrably solvable by a state-of-the-art model like GPT-4o.

This principled approach ensures that our dataset is not only of high visual quality but is also semantically accurate, posing a meaningful and well-defined challenge for advancing model capabilities.

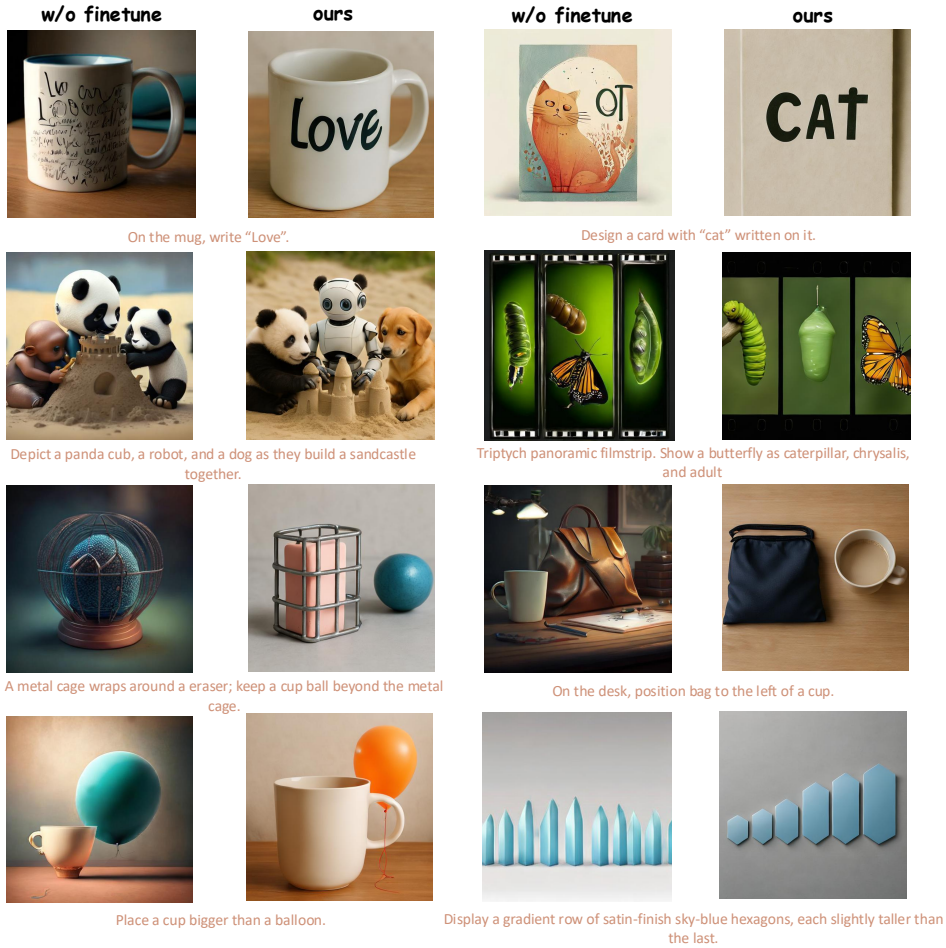


Figure 5: Qualitative comparison of Harmon before and after fine-tuning on our dataset.

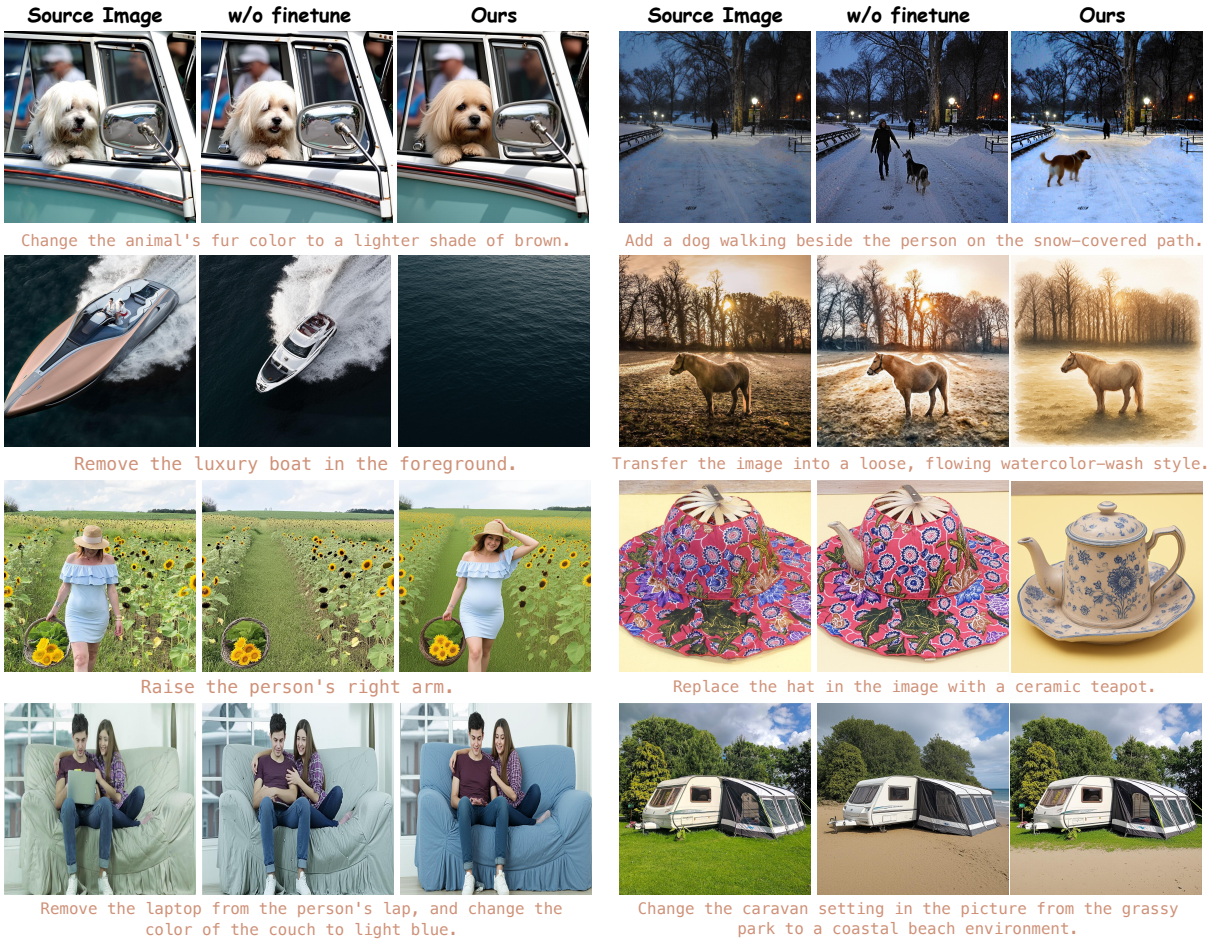


Figure 6: Qualitative comparison of UniWorld-V1 before and after fine-tuning on our dataset.

Table 9: Overview of the image editing dataset, including task categories, number of instances, and corresponding definitions.

Task	Number	Definition
		Subject Manipulation (19k)
Add	4k	Add introduces a new element into the source image.
Replace	4.2k	Replace substitutes an object in the image with a different object.
Alter	1.8k	Alter refers to modifying an existing object's attributes.
Remove	7.2k	Remove refers to eliminating an existing object from the image.
Obj Extraction	1.8k	Object extraction isolates and extracts a specific object from an image.
		Text Editing (3k)
Text Add	750	Text Add inserts textual elements into an image.
Text Replace	750	Text Replace substitutes a textual element in the image with a new one.
Text Alter	750	Text Alter modifies the attributes of an textual element (e.g., color).
Text Remove	750	Text Remove eliminates an existing textual element from an image.
		Complex Instruction Editing (4k)
Sub-Ins 2	1k	A complex instruction consists of two simple editing operations.
Sub-Ins 3	2k	A complex instruction consists of three simple editing operations.
Sub-Ins 4	1k	A complex instruction consists of four simple editing operations.
		Multi-turn Editing (1.5k)
2 Turns	500	A multi-turn editing operation consists of two simple editing rounds.
3 Turns	500	A multi-turn editing operation consists of three simple editing rounds.
4 Turns	500	A multi-turn editing operation consists of four simple editing rounds.
		Global Editing (5k)
Change BG	3.2k	Background replacement refers to substituting the surrounding environment of the subject.
Style Transfer	1.8k	Style transfer refers to the process of modifying the style of an image according to given instructions.
		Other Challenging Editing (8k)
Ref Image	3.5k	Reference image editing adds specified subjects into the source image.
Change Motion	2k	Motion modification alters the expressions and movements of objects.
Change Material	2k	Material transformation modifies the texture of clothing.
Obj Movement	500	Object movement refers to moving an object from one location to another within the image.

Table 10: Overview of the image generation dataset, including task categories, number of instances, and corresponding definitions.

Task	Number	Definition
Style Control (13k)		
Artistic Traditions	3.5k	Renders historical and cultural art styles.
Media and Illustration	4.5k	Renders aesthetics from media and illustration.
Photographic Styles	3k	Emulates various photographic techniques and moods.
Speculative and Fantasy Styles	2k	Creates speculative and fantasy genre aesthetics.
Complex Instruction Following (6k)		
Multi-Attribute Combination	500	Applies multiple attributes to subjects.
Multi-Subject Interaction and Action	500	Depicts interactions between multiple subjects.
Complex Spatial Composition	750	Arranges elements in complex spatial layouts.
Temporal Sequence Coherence	500	Generates a logical sequence of events.
Action Trajectory Rendering	750	Renders the trajectory of moving objects.
Causal Reasoning	3k	Depicts cause-and-effect relationships.
In-Image Text Rendering (3k)		
Textual Accuracy	500	Renders text verbatim from the prompt.
Typography	500	Controls text font, style, and appearance.
Structured Text Layout	500	Arranges text in structured layouts (e.g., multi-line).
Text-Graphic Integration	500	Integrates text coherently with image graphics.
Multilingual Support	500	Renders text in non-English languages.
Textual Tone and Style	500	Aligns text style with the image's aesthetic.
Spatial Reasoning in Images (8k)		
Containment	2k	Depicts one object inside another.
Relative Position	2k	Places objects in specified relative positions.
Comparative Reasoning	2k	Compares object attributes like size or color.
Symmetry Analysis	500	Generates symmetrical object arrangements.
Size Reasoning	500	Renders objects with correct relative sizes.
Object Counting	1k	Generates a specific number of objects.
Scientific Imagery (10k)		
Mathematics	1k	Visualizes mathematical concepts.
Ecology	1k	Creates imagery of ecosystems and species.
Astronomy	1k	Generates images of celestial bodies and phenomena.
Biological	1.2k	Illustrates biological structures and organisms.
Culture and History	2.2k	Depicts historical events and cultural artifacts.
Earth Science	1.4k	Visualizes geological and weather phenomena.
Mechanical Engineering	1.2k	Renders mechanical engineering diagrams and systems.
Physics	1.2k	Illustrates physical laws and concepts.