LongCat-Image Technical Report

Meituan LongCat Team
longcat-team@meituan.com

ABSTRACT

We introduce **LongCat-Image**, a pioneering open-source and bilingual (Chinese-English) foundation model for image generation, designed to address core challenges in multilingual text rendering, photorealism, deployment efficiency, and developer accessibility prevalent in current leading models. 1) We achieve this through rigorous data curation strategies across the pre-training, mid-training, and SFT stages, complemented by the coordinated use of curated reward models during the RL phase. This strategy establishes the model as a new state-of-the-art (SOTA), delivering superior text-rendering capabilities and remarkable photorealism, and significantly enhancing aesthetic quality. Extensive human evaluations rank our model's generation fidelity as superior to all existing models. 2) Notably, it sets a new industry standard for Chinese character rendering. By supporting even complex and rare characters, it outperforms both major open-source and commercial solutions in coverage, while also achieving superior accuracy. 3) The model achieves remarkable efficiency through its compact design. With a core diffusion model of only 6B parameters, it is significantly smaller than the nearly 20B or larger Mixture-of-Experts (MoE) architectures common in the field. This ensures minimal VRAM usage and rapid inference, significantly reducing deployment costs. Beyond generation, LongCat-Image also excels in image editing, achieving SOTA results on standard benchmarks with superior editing consistency compared to other open-source works. 4) To fully empower the community, we have established the most comprehensive open-source ecosystem to date. We are releasing not only multiple model versions for text-to-image and image editing, including checkpoints after mid-training and post-training stages, but also the entire toolchain of training procedure. We believe that the leading performance, high efficiency, and openness of LongCat-Image will provide robust support for developers and researchers, collectively pushing the frontiers of multilingual visual content creation.

LongCat Chat: https://longcat.ai
Hugging Face: https://huggingface.co/meituan-longcat/LongCat-Image
GitHub: https://github.com/meituan-longcat/LongCat-Image

Contents

Figure 1: **High-fidelity text-to-image generation results.**

Figure 2: **Showcase of versatile capabilities in general image editing.**

Figure 3: **Showcase on complex and comprehensive editing scenarios.** Beyond basic edits, LongCat-Image-Edit exhibits robust handling of intricate modifications and composite instructions.

1 Introduction

In recent years, significant advancements in Diffusion Models (DM) [Ho et al., 2020, Chen et al., 2024a, Labs, 2024, Wu et al., 2025a] have revolutionized the field of image generation, rapidly propelling the technology from academic research to widespread commercial applications, leading to the emergence of milestone products like Midjourney [Midjourney, 2025] and Seedream [Gong et al., 2025, Gao et al., 2025, Seedream et al., 2025]. As the technology has matured, the evaluation criteria for text-to-image (T2I) models have shifted from foundational metrics like instruction following and visual plausibility to more demanding benchmarks focusing on three core pillars: photorealism, aesthetics, and text rendering capabilities.

Concurrently, image editing has emerged as another critical domain, gaining prominence with the release of various open-source and commercial products (*i.e.*, Flux.1 Kontext [Batifol et al., 2025], Nano Banana (Gemini-2.5-flash-image)[1]). The primary challenges in image editing currently center on two key problems: executing editing instructions with high precision and maintaining strict visual consistency between the original and edited images [Wang et al., 2025a]. Although existing work [Batifol et al., 2025, Wang et al., 2025b, Wu et al., 2025a] has made important strides, a significant gap remains in achieving a seamless and reliable editing experience.

To address these challenges in both generation and editing, a prevailing trend has been the dramatic scaling of model parameters—from PixArt-α [Chen et al., 2024a] at 0.6B, to Stable Diffusion3.0 [Esser et al., 2024] at 8B, and further to Qwen-Image [Wu et al., 2025a] at 20B and even larger Mixture-of-Experts (MoE) architectures like Hunyuan-3.0 [Cao et al., 2025] with 80B full parameters. The expectation has been that, similar to Large Language Models (LLMs), diffusion models would experience a breakthrough in performance through brute-force scaling. However, our observations reveal a critical issue: unrestrained parameter growth has not delivered the anticipated qualitative leap. Instead, it has led to a host of problems, including soaring computational costs, higher deployment barriers, and increased inference latency. This not only hinders the democratization of the technology but also poses challenges for open academic research.

In this context, and guided by the LongCat team's consistent design philosophy of "Building efficient and powerful model", we introduce **LongCat-Image**—a novel, lightweight diffusion model for image generation and editing. We contend that a more optimal equilibrium must be struck between state-of-the-art performance and efficiency in training and inference. Through systematic experimentation, we determine that a parameter scale of 6B serves as the ideal foundation for balancing capability and efficiency without compromising generative quality. Specifically, the model's core diffusion architecture employs a hybrid MM-DiT and Single-DiT structure, consistent with Flux1.dev [Labs, 2024], while leveraging the Qwen2.5VL-7B [Bai et al., 2025] as its text encoder to provide a unified and powerful conditional space for both generation and editing tasks.

To further enhance photorealism, we implement a systematic overhaul of our data pipeline. We observe that even a small proportion of AIGC-contaminated data can cause the model to prematurely converge to a narrow local optimum during training. While this may accelerate initial convergence, it severely limits the model's potential to achieve higher levels of realism during subsequent fine-tuning. Consequently, we rigorously exclude all AIGC data during the pre-training and intermediate training stages. In the Supervised Fine-Tuning (SFT) phase, any high-quality synthetic data introduced was meticulously hand-selected. Finally, during the Reinforcement Learning (RL) phase, we innovatively incorporate an AIGC detection model as one of the reward models, using its adversarial signal to guide the model toward generating images with the texture and fidelity of the real physical world.

To overcome the industry-wide challenge of complex Chinese text rendering, we adopt a comprehensive strategy spanning data, architecture, and training. On the data front, we utilized the SynthDoG tool [Kim et al., 2022] to generate a large volume of text-in-image data, primarily with monotonous backgrounds, to minimize interference and improve the model's focus on learning character glyphs. Architecturally, we modify the text encoder to apply character-level encoding to the text designated for rendering in the prompt (identified by " "), which effectively reduces the memorization burden and enhances learning efficiency. During training, we use real-world text-in-image data in the SFT phase and introduce OCR and aesthetic reward models in the RL phase, significantly improving both the accuracy of text rendering and its natural integration with the background.

To solve the core challenge of visual consistency in image editing, we develop a meticulous training paradigm and a stringent data filtering strategy. We deliberately choose to initialize the editing model with weights from the mid-training stage of the T2I model, rather than from a highly optimized state after SFT or RL. The latter models exist in a narrowed state space, which is less conducive to learning and generalizing across diverse editing tasks. During the pre-training and SFT phases of the editing model, we employ multi-task joint training, combining editing tasks with T2I tasks. This approach effectively mitigates catastrophic forgetting of generative knowledge and enhances the model's comprehension

[1] https://aistudio.google.com/models/gemini-2-5-flash-image

of editing instructions. Data quality is paramount for ensuring visual consistency. We filter out samples with poor visual consistency during pre-training using a task-specific strategy and exclusively use high-consistency, human-annotated data during the SFT phase. Experiments confirm that this series of measures enables our model to achieve an exceptional standard in both instruction-following accuracy and visual consistency.

To empower the broader academic and industrial ecosystem, we are not only open-sourcing the final model but also releasing the mid-training checkpoint as a development model. Furthermore, we are providing the complete training and fine-tuning codebase, covering the entire workflow from pre-training to RL. Our goal is to foster a thoroughly open and accessible development ecosystem.

Our main contributions are five-fold:

- **Exceptional Efficiency and Performance:** With only 6B parameters, LongCat-Image surpasses numerous open-source models that are several times larger across multiple benchmarks, demonstrating the immense potential of efficient model design.
- **Remarkable Photorealism:** Through an innovative data strategy and training framework, our model achieves remarkable photorealism in generated images.
- **Powerful Chinese Text Rendering:** The model demonstrates superior accuracy and stability in rendering common Chinese characters compared to existing SOTA open-source models and achieves industry-leading coverage of the Chinese dictionary.
- **Superior Editing Performance**: The LongCat-Image editing model achieves state-of-the-art performance among open-source models, delivering a leading performance of instruction following and image quality, as well as superior visual consistency.
- **Comprehensive Open-Source Ecosystem:** We provide a complete toolchain, from intermediate checkpoints to the full training code, significantly lowering the barrier for further research and development within the community.

2 Data

The performance of generative models depends critically on the scale, diversity, and quality of the training corpus. Accordingly, we curate a massive dataset comprising **1.2 billion** samples. Fig. 4 provides a detailed statistical overview of the data composition.

2.1 Data Curation

As illustrated in Fig. 5, our data curation pipeline consists of four stages: filtering low-quality and duplicate samples, image metadata extraction, recaptioning, and data stratification for multi-stage training.

2.1.1 Filtering

Deduplication. To address data redundancy across diverse sources, we employ a two-tiered deduplication strategy: first, MD5 hashing is used to detect exact duplicates; second, SigLIP-based similarity assessment is applied to identify and eliminate near-duplicate entries.

Resolution & Aspect Ratio. Low-resolution images and extreme aspect ratios often correlate with poor visual quality. We exclude images with a shortest edge below 384 pixels. Furthermore, only images with aspect ratios between 0.25 and 4.0 are retained to ensure a uniform and structurally coherent dataset.

Watermark Detection. Watermarked images can introduce undesirable artifacts into the generated outputs. To mitigate this, we utilize a specialized watermark detector to identify and remove samples exhibiting visible watermark patterns.

Laion Aesthetics. To guarantee a baseline of visual quality, each image is evaluated using the LAION-Aesthetics predictor [Schuhmann et al., 2022]. We discard images with scores below 4.5. This threshold is empirically selected to filter out low-quality samples while preserving sufficient diversity for model training.

AIGC Detection. Our experiments indicate that a small fraction of AI-generated content (AIGC) in the training data can disrupt optimization, resulting in a "plastic" or "greasy" texture in generated images. Consequently, we develop an internal AIGC detector to purge synthetic data from the corpus.

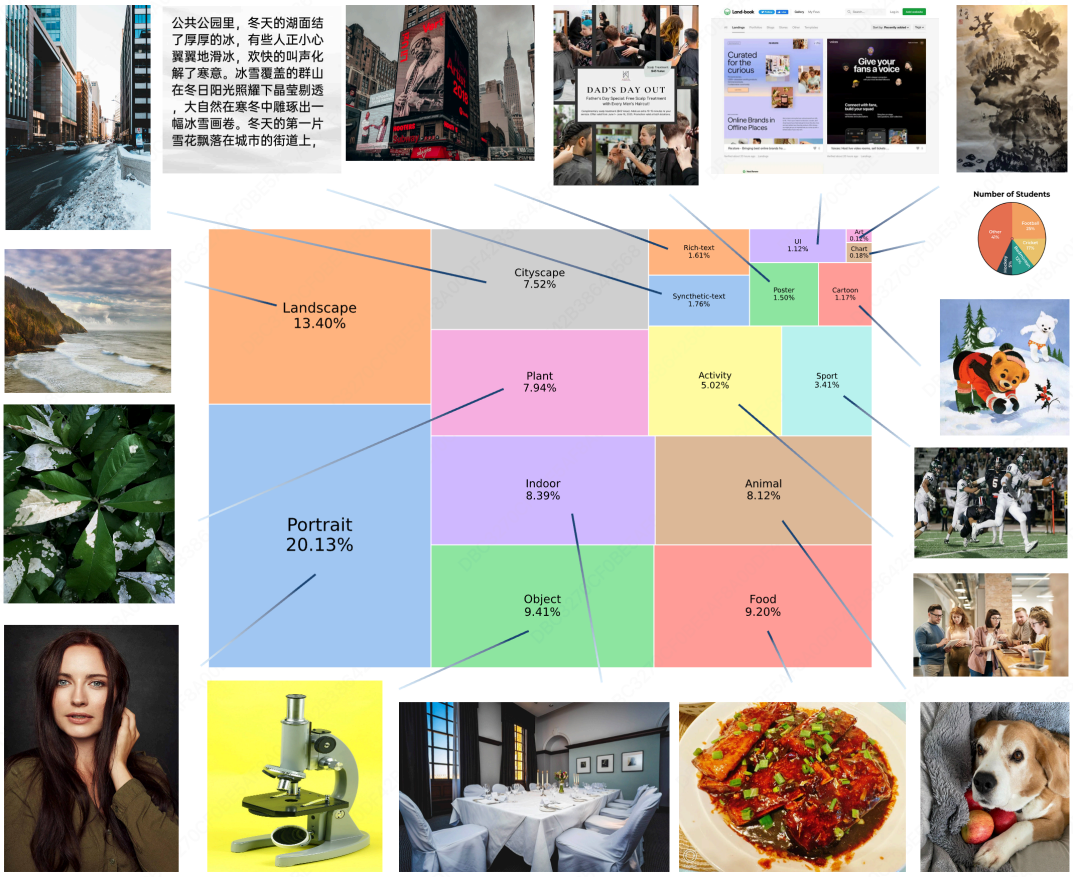

Figure 4: **Overview of training data**.

2.1.2 Meta Infomation Extraction

We delineate five key attributes essential for downstream processing: category, style, named entity, OCR, and aesthetics. In particular, category, style, and aesthetics are instrumental in achieving balanced data distribution and enabling hierarchical data structuring. Conversely, named entities and recognized text content play a pivotal role in enhancing the accuracy and informativeness of image captions.

Category. The category represents the semantic classification of an image according to its visual content. In our work, images are assigned to one of the following categories: *portrait*, *sport*, *activity*, *plant*, *animal*, *food*, *object*, *landscape*, *cityscape*, *indoor*, *UI*, *cartoon*, *chart*, *rich-text*, *poster*, and *synthetic text*. This categorization scheme provides a structured framework for organizing the dataset and supports subsequent processes such as content analysis and distribution balancing.

Style. Style serves as an indicator of the artistic characteristics inherent in an image. In our approach, we instruct the open-source VLM to produce a set of plausible style descriptions in the form of phrases, rather than constraining the output to a fixed set of predefined labels.

Named Entity. Named entities serve as indicators of the world knowledge encapsulated within visual content. In our work, we employ the available VLMs to identify potential celebrities, fictional characters, biological species, commercial brands, and intellectual properties depicted in the images. Subsequently, this extracted semantic information is incorporated into the subsequent image recaptioning process to enhance descriptive accuracy and contextual richness.

OCR Text. To enhance text rendering performance, an OCR model is employed to extract textual information from the images. The extracted text is subsequently processed and integrated into the corresponding image captions through a specialized handling procedure.

Comprehensive Aesthetics Evaluation. Quantifying aesthetics is inherently challenging due to subjectivity and the generalization limits of existing single-metric models. To address this, we decouple image evaluation into two

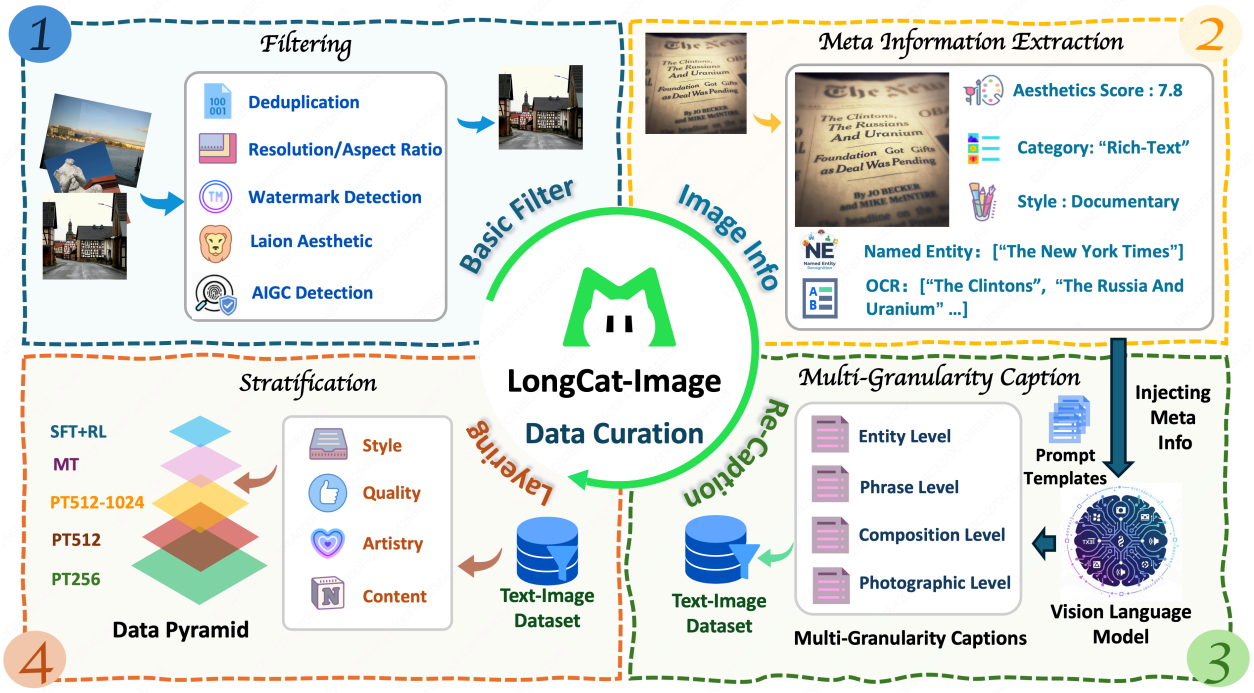

Figure 5: **Data curation pipeline.** The pipeline consists of four stages: (1) **Filtering**: Raw data undergoes deduplication and quality assessment, including watermark and AIGC detection. (2) **Meta Information Extraction**: We extract comprehensive metadata, such as aesthetic scores, named entities, and OCR text. (3) **Multi-Granularity Captioning**: Leveraging the extracted metadata and prompt templates, a VLM generates captions ranging from entity-level tags to detailed photographic descriptions. (4) **Stratification**: The dataset is stratified into a pyramid structure based on style, quality, and content to support progressive training stages.

orthogonal dimensions, namely *Quality* and *Artistry*, and employ an ensemble of six complementary assessment methodologies. The overall pipeline is illustrated in Fig. 6.

- *Quality*: This dimension measures technical fidelity. We combine low-level signal statistics, including saturation, contrast, and color richness in RGB and HSV spaces, with deep reference-free assessment metrics derived from MUSIQ [Ke et al., 2021] and Q-Align [Wu et al., 2023].
- *Artistry*: This dimension evaluates photographic merit and artistic expression. We leverage VLM-based analysis to assess high-level attributes such as composition, lighting, shadow, and color tonality, complemented by the Q-Align-Aesthetics [Wu et al., 2023] score.

2.1.3 Mutli-Granularity Captioning

Image captioning with advanced Vision-Language Models (VLMs) has recently emerged as a prominent paradigm. However, there exist three critical limitations: (1) insufficient integration of world knowledge embedded within generated captions; (2) restricted diversity in caption formats; and (3) low information density resulting from verbose captions.

The lack of world knowledge often leads to inaccurate or incomplete depictions of named entities, thereby undermining the semantic fidelity of the generated images. Furthermore, captions produced by the same VLM frequently conform to similar structural patterns and lengths, limiting robustness. Finally, verbose descriptions consume valuable token space, thereby reducing the efficiency of content representation within constrained caption lengths.

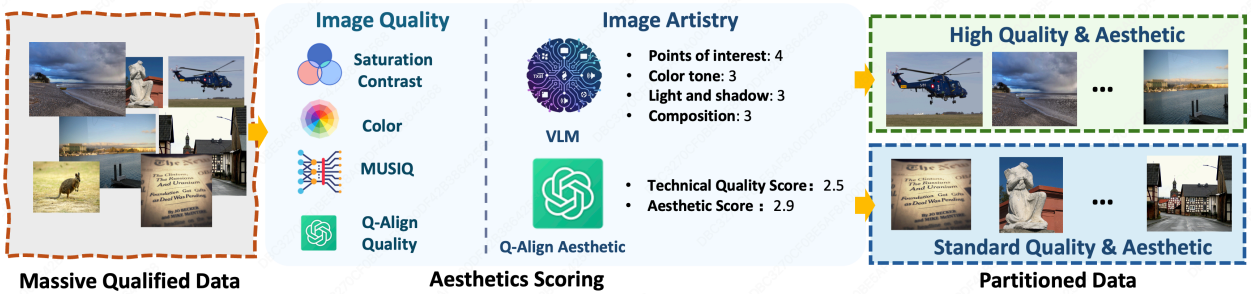

Figure 6: **Comprehensive aesthetic scoring**. The aesthetic scoring takes image quality and image artistry into account. In this context, quality denotes the signal-related attributes of the image, such as resolution, clarity, and noise levels, while artistry pertains to the perceptual appeal or visual attractiveness of the image as judged by humans or models.

To overcome the aforementioned limitations, we introduce a Multi-Granularity Captioning (MGC) framework that systematically organizes semantic abstraction into four hierarchical levels. Specifically, the *Entity Level Caption* aims to identify and describe the principal visual entities present in an image; the *Phrase Level Caption* encapsulates salient visual attributes using concise linguistic expressions; the *Composition Level Caption* provides an integrative interpretation that captures the overall semantic structure of the scene; and the *Photographic Level Caption* offers finest-grained visual depictions, incorporating both the specific content of the image and relevant world knowledge in a succinct yet informative manner.

At the Entity Level and Phrase Level, we employ Qwen2.5-VL [Bai et al., 2025] to concurrently extract the relevant semantic information from the input image. Subsequently, at the Composition Level, we integrate the image itself, the Entity Level descriptions, and the extracted image meta-information, and feed them into InternVL2.5 [Chen et al., 2024b] to generate a comprehensive, Composition Level caption. These example prompts are shown in Fig. 7.

At the Photographic Level, we develop a customized captioning model, called the Photographic Captioner, based on the Qwen2.5-VL backbone. Empirical analysis reveals that, while this open-source backbone can produce descriptions enriched with extensive world knowledge, the output format exhibits notable inconsistencies. To mitigate this issue, we apply LoRA to fine-tune the model, using meticulously annotated synthetic image–text pairs. This approach improves the informational density of the captions while retaining their embedded world knowledge. A qualitative comparison of caption outputs is provided in Fig. 8.

During training, we employ a weighted sampling strategy for these multi-granularity captions, prioritizing detailed descriptions to maximize information density. Specifically, the sampling probabilities for the four increasing levels of granularity are set to $[0.05, 0.1, 0.2, 0.65]$, respectively. This distribution enables the model to accommodate diverse prompt formats while robustly encoding complex world knowledge. Fig. 9 illustrates representative training samples across these granularity levels.

Entity & Phrase Level

```
# Task
You are an image content understanding expert, tasked with analyzing the given image and
understanding its information according to the following dimensions. The subjects of the image, the
number of subjects, the positional relationships between subjects, the color of the subjects, and the
content of each image's subject should be determined based on the image itself.

# Step
 1. Identify the most important part of the image, specify the color and quantity of the subject.
 2. Description of the background in the image (e.g., mountains, sea, urban, cityscape, indoor, café, etc.),
 3. Describe the subject information of the image as detailed as possible in the form of a set of phrases,

# Output Format
```
{
  "Entity": ["important parts of the image"],
  "Phrase": ["a set of keywords describing the subject information"],
}
```
```

Composition Level

```
# Task
Your task is to generate informative and concise summary for a given image.

# Image information
 The category is {category}.  The style is {style}. The entities in the images are {entity}. The named entities in the image are
{named_entity}. The recognized texts in the image are {OCR}

# Step
 1. Make a summary of the image information
 2. Supplement the summary with the artistic information like medium, techniques, light and shadow, atmosphere.
 3. Verify the summary with the given image and fix the inconsistency.

# Output Format
```
{
  "Explanation": "the explanation for verification step",
  "Summary": "the verified summary",
}
```
```

Figure 7: **Prompts for different level captions.** Entity Level and Phrase Level captions are generated concurrently using a single model, whereas composition-level captions are subsequently produced in a sequential stage utilizing a separate model.

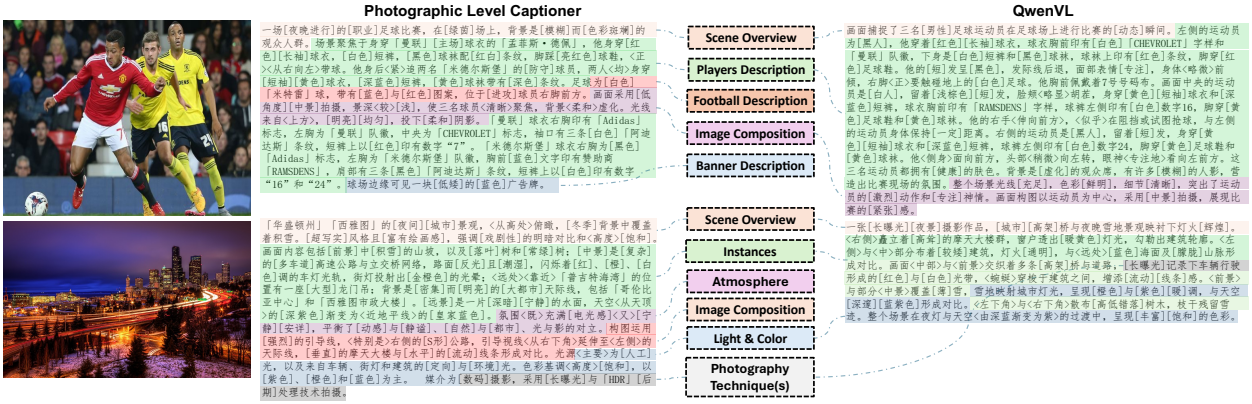

Figure 8: **Quality comparison of our Photographic Level Captioner**. This captioner produces more concise and information-dense captions compared to baseline. Different color blocks indicate different aspects of the captions.

Entity 一幅色彩斑斓的马赛克艺术作品

Phrase 图案，彩色，几何，形状，马赛克，对称设计，紫色，蓝色，圆形，三角形，红色，艺术作品，橙色

Composition 一幅色彩鲜艳的马赛克艺术作品，由各种几何形状和图案组成，包括三角形和圆形，并以对称的方式排列。作品的背景色彩丰富，融合了紫色、橙色、蓝色和红色等色调。

Photographic 一幅充满活力的马赛克艺术作品，由各种几何形状和图案组成，包括三角形和圆形，并以对称的方式排列。作品的背景色彩丰富，融合了紫色、橙色、蓝色和红色等色调。画面中包含多种抽象形状，营造出动感十足、引人入胜的视觉效果。照片采用近景镜头拍摄，突显了马赛克瓷砖的精细纹理。照片风格为抽象艺术，强调了鲜艳的色彩和复杂的图案。照片采用平视视角拍摄，完整展现了作品的全貌

Entity 家居用品推广海报

Phrase 极简家居陈设，黑色圆形茶几，浅灰布艺单人椅，白墙背景，金属镜面装饰，干枝插花摆件，商业广告排，黑白灰调

Composition 一张室内陈列拍摄图片，展示现代简约风格的客厅角落，配有文字说明环保无胶水墙贴功能。

Photographic 画面中央偏左为一张圆形黑色桌面茶几，桌面上放置一本打开的书籍和一个银色金属台灯，台灯位于左侧。茶几右侧是一把浅灰色布艺单人椅，椅背弧形，四腿细长。背景墙面为纯白色，墙上悬挂一束干枯植物枝条装饰，插在透明玻璃花瓶中，花瓶置于茶几后方矮柜上。矮柜为深灰色，表面光滑，横向延伸至画面右边缘。画面左上方有两行竖排黑色粗体汉字“不用胶水 环保无味 即贴即住”，下方有一行较小字体“精致纹理 拓展视觉感和空间”，再下方为一行英文“COLORFUL HOME BEAUTIFUL LIFE”，字体为无衬线小写，颜色为浅灰。所有文字位于墙面左上区域，未遮挡家具。整体图像清晰，布局平衡，文字与实物比例协调。

Entity A young man with dyed white hair

Phrase purple sweater, young man, holding a handkerchief, denim jacket, shadow, sunlight, black pants, gray concrete wall

Composition A street-style photograph depicts a young platinum-blonde man wearing a magenta hoodie standing in front of a concrete wall; the image is blurry and the atmosphere is cool and modern.

Photographic A male portraiture, presenting a modern, minimalist, and street-style photography aesthetic, with the overall picture appearing blurry and indistinct. A young man with platinum blonde hair stands facing the camera, wearing a maroon crew-neck sweatshirt and black pants, holding a light blue denim jacket in his left hand. The background is a gray concrete wall with horizontal line textures. Strong sunlight from the right casts a distinct shadow on the wall behind him, creating a cool and independent atmosphere.

Entity Hollywood Sign in L.A.

Phrase iconic, natural, surroundings, Hollywood Sign, modern house, cloudy sky, illuminated, Los Angeles

Composition A cinematic photograph showcasing the Hollywood sign, set against a gloomy, cloudy sky with a blurred house and vegetation in the foreground, creating a solemn and slightly melancholic atmosphere.

Photographic The famous landmark in Los Angeles, the "HOLLYWOOD" sign. The style is strongly cinematic and dramatic, with colors processed to have a low-saturation, dark tone. The content captures the giant white "HOLLYWOOD" sign situated on a hill, with a background of dense clouds. In the foreground, the roof of a house and some trees are visible but blurred by a shallow depth of field effect. The overall atmosphere of the image feels somewhat solemn, quiet, and even a bit melancholic due to the gloomy weather and vignette effect.

Figure 9: **Examples of Multi-Granularity Captioning.**

2.1.4 Stratification

We stratify the training corpus into distinct subsets tailored to specific training stages, utilizing the extracted meta-data—specifically image style, semantic content, and aesthetic scores.

Pre-training. Empirical evidence suggests that exposing the model to a high concentration of artistic data (e.g., illustrations, cartoons, and anime) during the early pre-training stage biases the model towards learning simplified visual patterns. This tendency can compromise the model's ability to generate high-fidelity photorealistic images, effectively causing a "collapse" in the realistic generation subspace. Consequently, we restrict artistic data to approximately 0.5% of the pre-training corpus, deferring the integration of broader stylistic data to the subsequent mid-training phase.

Mid-training. The mid-training phase focuses on two objectives: *Quality Enhancement* and *Artistic Style Injection*. For quality enhancement, we curate a subset of high-resolution images (exceeding 1,024 pixels) from the pre-training corpus,

selected for their superior sharpness, balanced composition, and high aesthetic scores. Simultaneously, for artistic style injection, we reintroduced the previously filtered artistic data to unlock the model's stylistic capabilities. We gradually increase the proportion of stylized data from 0.5% to 2.5% following a calibrated schedule. This progressive integration strategy effectively expands the model's stylistic repertoire while preserving its photorealistic foundation.

SFT. In the SFT phase, we employ a mix of real and synthetic data to align the model with human aesthetic preferences. For real data, human experts manually curate high-fidelity samples from the Mid-training dataset, evaluating dimensions such as composition, lighting, color tonality, and emotional expression, while ensuring a balanced categorical distribution. Complementing this, we incorporate model-synthesized images that have undergone rigorous manual filtering to eliminate structural distortions, visual unreality, and aesthetic flaws. The strong stylistic consistency of this synthetic data facilitates the model's rapid convergence towards the manifold of human preferences.

2.2 Data Synthesis

Synthetic Data Generation. To address the long-tail distribution inherent in real-world datasets, we construct a specialized synthetic corpus targeting rare concepts and corner cases. Specifically, we train multiple domain-specific LoRA adapters on limited samples to capture infrequent compositional patterns and distinctive artistic styles. These adapters generate high-quality synthetic images, which are integrated into the mid-training phase at a controlled low ratio. This strategy effectively boosts performance on tail categories without compromising the diversity of the generated output space.

Text Rendering. Empirical evidence suggests that mastering textual structures synergistically enhances a model's ability to generate other structured visual elements, thereby improving overall scene coherence. Motivated by this, we integrate synthetic text data into the pre-training corpus. As illustrated in Fig. 10, our pipeline renders text from classical literature onto diverse textures, utilizing varied color palettes and fonts.

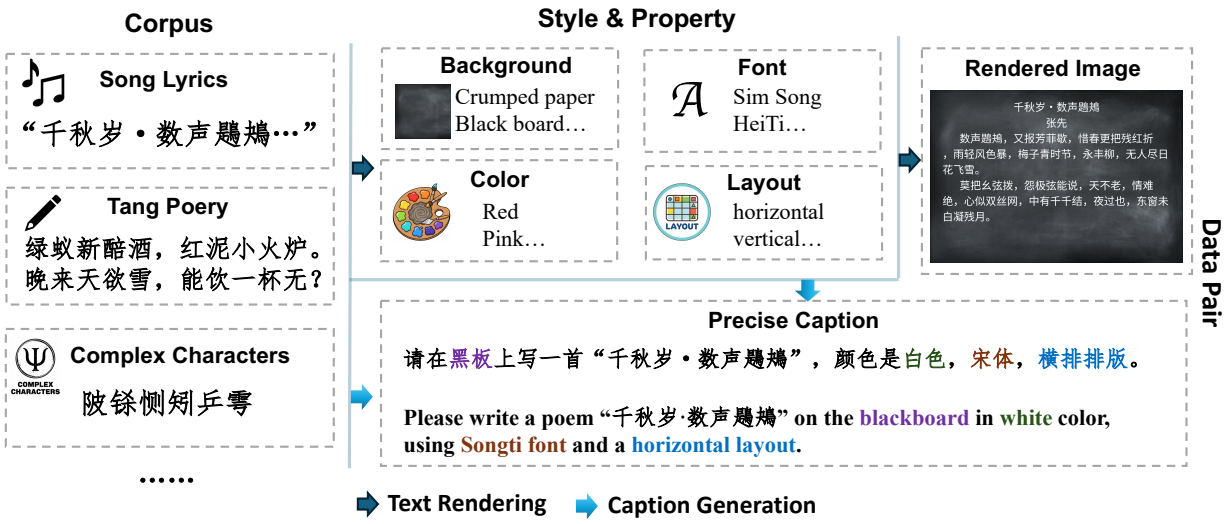

Figure 10: **Process for synthesizing text rendering data.**

3 Model Design

3.1 Diffusion Model

We adopt the transformer architecture of FLUX.1-dev [Labs, 2024], employing a double-stream attention mechanism in the initial layers and transitioning to a single-stream mechanism in the subsequent layers. To ensure parameter balance, the ratio of double-stream to single-stream blocks is maintained at approximately 1:2. The overall framework design is illustrated in Fig. 11.

For the VAE component, we utilize the implementation from FLUX.1-dev. Empirical evaluations demonstrate its superior reconstruction fidelity in challenging scenarios, such as fine typography and intricate textures. Specifically,

input images undergo $8\times$ spatial compression; the resulting latents are further processed via 2×2 token merging, yielding a final sequence length of $\frac{H\times W}{16\times 16}$ before entering the DiT module.

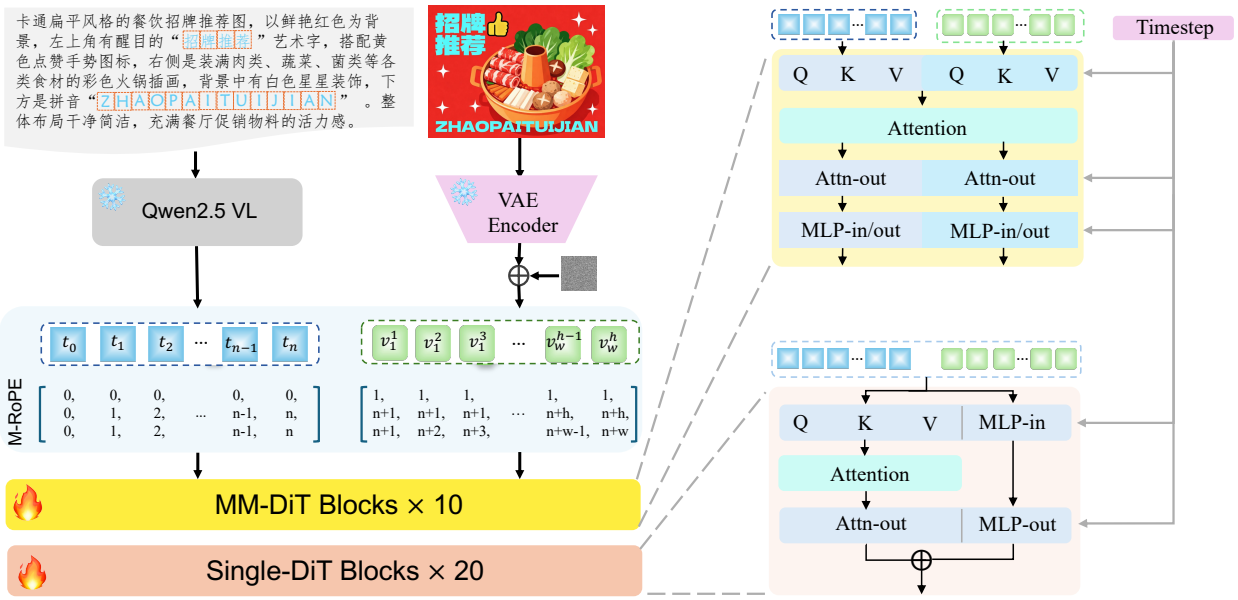

Figure 11: **Overview of model architecture.**

3.2 Text Encoder

The text encoder embeds user prompts into continuous representations to condition the DiT model. While prior works [Podell et al., 2023, Team, 2024, Li et al., 2024a, Esser et al., 2024, Labs, 2024] predominantly rely on CLIP [Radford et al., 2021] and T5 [Chung et al., 2024], recent studies [Gao et al., 2025, Wu et al., 2025a] have shifted toward LLMs or MLLMs to enhance multilingual compatibility, particularly for Chinese. Following this paradigm, we adopt Qwen2.5VL-7B [Bai et al., 2025] as our unified text encoder. This choice ensures robust English instruction following while significantly improving Chinese processing capabilities. Furthermore, we discard the conventional injection of text embeddings into timestep embeddings for adaLN [Peebles and Xie, 2022] modulation, as empirical evidence suggests negligible performance gains from this operation.

For visual text rendering, we employ a character-level tokenizer specifically for content demarcated by quotation marks. This strategy mitigates generation complexity without incurring the computational costs and memory footprint of specialized encoders (*e.g.*, GlyphByT5 [Liu et al., 2024]). Experiments demonstrate that this approach not only improves data efficiency but also accelerates convergence for text rendering tasks.

3.3 Positional Embedding

Positional embedding (PE) design is pivotal for handling variable aspect ratios and resolutions. While prior arts [Chen et al., 2024a, Li et al., 2024a, Gong et al., 2025] rely on intricate heuristics—such as coordinate centering, frequency scaling, or interpolation—to align spatial distributions, we adopt the vanilla Multimodal Rotary Position Embedding (MRoPE) [Su et al., 2024, Wang et al., 2024a] without modification. Our empirical observations indicate that the model possesses intrinsic adaptability to varying positional strides across different resolutions, rendering these explicit geometric constraints unnecessary. Consequently, MRoPE enables seamless generalization to unseen resolutions during pretraining without the computational or design overhead of complex adaptation strategies.

Specifically, we employ a 3D variant of MRoPE. The first dimension is designated for modality differentiation. In the text-to-image task, distinct values are assigned to distinguish tokens belonging to noise latents from those of text latents. For image editing tasks, this dimension further differentiates the latents of reference images from the aforementioned types. The remaining two dimensions encode the 2D spatial coordinates: for images, they correspond to the (x, y) positions, while for text, both coordinates are set to an identical value, analogous to the behavior of 1D-RoPE. This approach not only supports flexible image generation across arbitrary aspect ratios but also facilitates seamless interaction with other modalities, such as text and reference images.

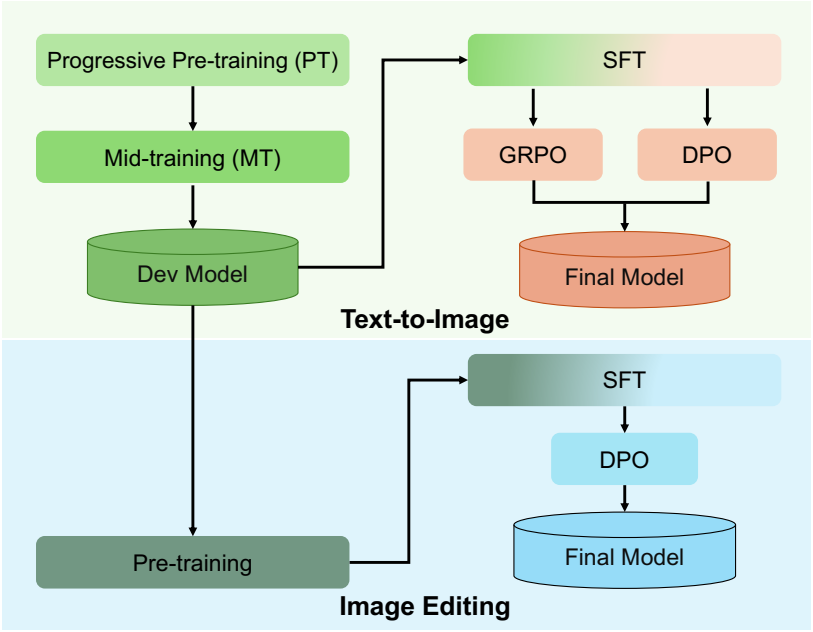

Figure 12: **Schematic overview of the multi-stage training pipeline.** The **upper panel** delineates the Text-to-Image training trajectory, progressing from progressive pre-training and mid-training to post-training alignment via SFT, GRPO, and DPO. The **lower panel** illustrates the Image Editing workflow, which initializes from the T2I development checkpoint.

3.4 Prompt Engineering

To bridge the gap between the dense captions used in training and the concise, often ambiguous queries provided by users, prompt refinement is essential. While external APIs or large language models (LLMs) may offer superior rewriting capabilities, their integration often introduces deployment constraints and latency issues. To address this, we provide a default **built-in** solution that efficiently repurposes the existing condition encoder, Qwen2.5-VL. This design ensures an out-of-the-box capability for generating high-fidelity images, eliminating dependencies on external services while maintaining ease of use.

4 Model Training

We establish a comprehensive multi-stage training pipeline, as illustrated in Fig. 12, structured into three distinct phases: Pre-training, Mid-training, and Post-training.

- **Pre-training**: This phase adopts a progressive multi-resolution strategy, facilitating the efficient acquisition of global semantic knowledge in early iterations while prioritizing high-frequency detail refinement in later stages.
- **Mid-training**: Serving as a crucial bridge between raw pre-training and alignment, this phase aims to elevate the model's baseline generation quality. We leverage a large-scale aesthetic assessment model alongside human curation to filter a high-fidelity dataset, ensuring the underlying model possesses robust aesthetic priors.
- **Post-training**: The final phase focuses on alignment and stylization, comprising SFT and Reinforcement Learning (RL). In the SFT stage, we target stylized data distributions and implement a model fusion strategy to synthesize diverse stylistic capabilities. Subsequently, the RL stage incorporates advanced alignment techniques—specifically DPO and GRPO—integrating ensemble reward models to significantly enhance instruction adherence and quality.

The detailed training hyperparameters for each phase are provided in Table 1.

4.1 Pre-training

Progressive Mixed-Resolution Training. We implement a progressive training curriculum commencing at 256_{px}. To optimize training efficiency and facilitate smooth resolution adaptation, we explicitly retain an intermediate 512_{px}

Table 1: Progressive training recipe for LongCat-Image.

	PT 256_{px}	**PT** 512_{px}	**PT** 512-1024_{px}	**MT**	**SFT**	**DPO**	**GRPO**
Learning rate	1e-4	5e-5	2e-5	1e-5	1e-5	1e-5	1e-5
LR scheduler	Constant	Constant	Constant	Constant	Consine	Consine	Consine
Warm-up steps	0	0	0	0	1000	1000	0
Training steps	900K	300K	200K	70K	20K	4K	300
Global batch size	4608	4608	3072	3072	128	64	32
Weight decay	0.01						
Gradient clip	1.0						
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)						

stage, avoiding the computational instability of transitioning directly from 256_{px} to the final phase. Subsequently, the training culminates in a dynamic stage covering a continuous resolution range between 512_{px} and 1024_{px}. Throughout these phases, we employ bucket sampling to accommodate variable aspect ratios.

Real-time Evaluation Protocol. To facilitate dynamic strategy adjustment and convergence analysis, we implement a comprehensive monitoring system tracking validation loss, image-text alignment, aesthetic scores, and OCR-based text rendering accuracy. Empirical observations indicate that while these metrics serve as pivotal indicators during the pre-training phase, their discriminative power and sensitivity notably diminish during the mid-training and post-training stages as performance saturates.

Dynamic Sampling of Synthetic Text Rendering Data. The Chinese character set exhibits a distinct long-tail distribution, comprising approximately 3,000 common characters and over 5,000 rare ones that appear sparsely in natural data. To address this sparsity, we employ SynthDoG to generate a large-scale dataset (over 10 million samples) rendered on simple textures—such as paper, glass, and blackboards—with high typographic diversity (see Fig. 10). While this synthetic data significantly enhances text rendering accuracy, it inevitably compromises the overall visual harmony. To mitigate this trade-off, we implement a dynamic sampling strategy based on real-time character-wise accuracy monitoring. Specifically, we increase the sampling probability for characters with high error rates, while gradually reducing synthetic exposure for well-learned characters in favor of real images. Furthermore, to prevent the model from overfitting to the simplistic synthetic domain, we completely phase out synthetic data in the final stage of pre-training.

4.2 Mid-training

While the pre-training phase successfully endows the model with robust global semantic priors and text-to-image mapping capabilities, the resulting visual outputs often lack high-fidelity textures and aesthetic coherence due to the inherent noise in large-scale pre-training data. The Mid-training stage is therefore designed to constrain the learned manifold, guiding the model's distribution toward a subspace characterized by superior aesthetic quality and visual realism. This refinement serves as a critical initialization for subsequent post-training optimization.

To achieve this, we implement a rigorous data curation protocol that is significantly more stringent than that of the pre-training phase. Our pipeline integrates a hierarchical assessment system comprising advanced aesthetic scoring models, image quality estimators, and domain-specific classifiers, culminating in a human-in-the-loop verification process. This systematic filtration yields a high-fidelity corpus of millions of samples, ensuring balanced representation across diverse domains. Continued training on this curated dataset significantly elevates the model's generation quality and visual fidelity.

We designate the model derived from this stage as a foundational Developer Version. Unlike fully aligned models subjected to extensive RL, this version retains high plasticity and adaptability, avoiding the potential mode collapse or rigidity often introduced by aggressive alignment. We release this model to the community to facilitate downstream fine-tuning and further research.

4.3 Post-training

4.3.1 SFT

The primary objective of the SFT phase is to elevate the model's visual aesthetics through a rigorous data-centric approach. This involves optimizing photorealistic attributes, such as compositional integrity, lighting dynamics, and photographic techniques, while simultaneously ensuring stylistic fidelity across various artistic domains.

High-Fidelity Data Curation. We employ a hybrid dataset comprising hundreds of thousands of samples that blends real-world imagery with high-quality synthetic data. To prevent the degradation of realism often associated with synthetic artifacts, we enforce a strict expert verification protocol. This process ensures that only data possessing superior textural quality and aesthetic value are retained, and it strictly filters out any samples that might compromise the model's generation fidelity.

Model Weight Averaging. Recognizing that diverse training subsets yield models with complementary strengths, we fine-tune multiple candidate models where each is specialized in distinct visual dimensions, including illumination, portraiture, and artistic style. To synthesize these capabilities, we adopt a model parameter averaging strategy. By merging the weights of these specialized models, we effectively balance performance across multiple attributes. This fusion process significantly enhances overall robustness and stability, effectively mitigating the specific biases or deficits inherent in individual single-domain models.

Optimization of Timestep Sampling. Unlike the pre-training phase, which prioritizes global structural formation, SFT focuses on refining high-frequency details that typically emerge during the later stages of the diffusion process. Consequently, we transition from the Logit-Normal sampling strategy to Uniform Sampling. This adjustment ensures balanced exposure across all timesteps, specifically increasing the training weight of high-frequency denoising steps to maximize the model's capacity for learning intricate textures and fine details.

4.3.2 RLHF

We develop fine-grained reward models (RMs), including distortion detection, AIGC detection, human preference assessment, and OCR accuracy, to comprehensively evaluate the model's detailed capabilities. Using these RMs, we employ three distinct RL strategies: Direct Preference Optimization (DPO) [Rafailov et al., 2023, Wallace et al., 2024], Group Relative Policy Optimization (GRPO) [Xue et al., 2025], and our proposed Monolithic Policy Optimization (MPO). DPO excels at offline preference modeling for flow-matching models with high computational efficiency, whereas GRPO and MPO perform on-policy sampling during training with reward model evaluation. MPO fundamentally improves upon GRPO by eliminating the group-relative paradigm and its associated synchronization bottlenecks, achieving superior training efficiency and stability. To leverage the scalability advantages of offline preference learning, we conduct relatively large-scale RL with DPO and reserve on-policy methods (MPO or GRPO) for small fine-grained RL refinement. Details of each algorithm are provided below.

(A) Direct Preference Optimization (DPO)

Data Construction 1) PromptSet: A category-balanced PromptSet is constructed from real user queries of public datasets, refined through clustering and data-cleaning techniques to ensure representativeness and diversity. 2) ImagePair: We utilize diverse random initialization seeds to generate 6 candidate images for each prompt sample. An annotation team then assigns subjective quality scores (1–5) to each sample. To ensure clarity and effectiveness of training, we discard neutral samples (score = 3), treating high-quality (4–5) samples as positive and low-quality (1–2) samples as negative, thereby forming win-lose pairs for DPO training. This strategy ensured high confidence and preference discernment in the training data.

Algorithm We employ the DPO algorithm to mitigate common structural deficiencies in the model. Based on the SFT model, we construct a preference dataset through diversified data sampling and manual curation, optimizing the model as follows:

$$L(\theta) = -\mathbb{E}_{(x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),x_t^w\sim q(x_t^w|x_0^w),x_t^l\sim q(x_t^l|x_0^l)} \log\sigma\Bigg(-\beta T\omega(\lambda_t)\Bigg(\|v^w - v_\theta(x_t^w,t)\|_2^2 - \|v^w - v_{\text{ref}}(x_t^w,t)\|_2^2 - \left(\|v^l - v_\theta(x_t^l,t)\|_2^2 - \|v^l - v_{\text{ref}}(x_t^l,t)\|_2^2\right)\Bigg)\Bigg). \tag{1}$$

We optimize the model according to Equation 2. During training, we further explore strategies including multi-round DPO iterations, gradient norm-based dirty data skipping, and KL constraints, conducting comparative analyses of different approaches' impacts on model performance. Ultimately, DPO significantly reduce the model's bad case rate and enhance the robustness of image generation.

(B) Group Relative Policy Optimization (GRPO)

Algorithm After training with DPO, we perform further fine-grained training using GRPO following the Dance-GRPO [Ma et al., 2024] framework. Given text hidden state h, the flow model predicts a group of G images $\{x_0^i\}_{i=1}^G$ and the corresponding trajectory $\{x_T^i, x_{T-1}^i, ..., x_0^i\}_{i=1}^G$. Within each group, the advantage function is formulated as:

$$A_i = \frac{R(x_0^i,h) - \text{mean}(\{R(x_0^i,h)\}_{i=1}^G)}{\text{std}(\{R(x_0^i,h)\}_{i=1}^G)}. \tag{2}$$

The training objective of GRPO is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{h\sim\mathcal{D},\{x_T^i,...,x_0^i\}_{i=1}^G\sim\pi_\theta}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\left(\min\left(r_t^i(\theta)A_i, \text{clip}(r_t^i(\theta),1-\epsilon,1+\epsilon)A_i\right)\right)\right], \tag{3}$$

where $r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i,h)}{p_{\theta_{\text{old}}}(x_t^i|x_t^i,h)}$.

During trajectory sampling, we reformulate the deterministic flow-matching ODE as an SDE for effective exploration:

$$dx_t = \left(v_t + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_t)\right)dt + \sigma_t dw, \tag{4}$$

with Euler-Maruyama discretization:

$$x_{t+\Delta t} = x_t + \left[v_\theta(x_t,t,h) + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta(x_t,t,h))\right]\Delta t + \sigma_t\sqrt{\Delta t}\epsilon. \tag{5}$$

(C) Monolithic Policy Optimization (MPO)

Algorithm MPO employs a single-stream on-policy optimization paradigm that eliminates group-relative synchronization bottlenecks inherent in GRPO. For each prompt, MPO generates a single trajectory using a Stochastic Differential Equation (SDE) sampler and performs one gradient update, achieving superior computational efficiency.

The generative process is governed by the SDE:

$$d\mathbf{z} = \mathbf{v}_\theta(\mathbf{z}_t,c,t)dt + g(t)d\mathbf{w}, \tag{6}$$

where $g(t)d\mathbf{w}$ is the diffusion term enabling exploration within a single trajectory. Using Euler-Maruyama discretization:

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \mathbf{v}_\theta(\mathbf{z}_t,c,t)\Delta t + g_t\sqrt{\Delta t}\epsilon_t, \quad \epsilon_t\sim\mathcal{N}(0,\mathbf{I}). \tag{7}$$

MPO incorporates three synergistic components for stable variance control:

1. Gaussian Value Tracker with KL-Adaptive Forgetting: A persistent Bayesian tracker maintains per-prompt reward estimates $\mathcal{V}(c) = \mathcal{N}(\mu_c,\sigma_c^2)$. The mean μ_c serves as a stable baseline, while the variance σ_c^2 quantifies epistemic uncertainty. Updates follow Kalman filter principles:

$$\begin{aligned} K_t &= \frac{\sigma_{c,t-1}^2}{\sigma_{c,t-1}^2 + \sigma_{\text{obs}}^2}, \\ \mu_{c,t} &\leftarrow \mu_{c,t-1} + K_t(r - \mu_{c,t-1}), \\ \sigma_{c,t}^2 &\leftarrow (1-K_t)\sigma_{c,t-1}^2 + Q_t, \end{aligned} \tag{8}$$

where $Q_t = \alpha \cdot D_{\mathrm{KL}}(\pi_{\theta'} \| \pi_\theta)$ adaptively scales process noise with policy drift.

2. Global Advantage Normalization: Raw advantages $A = r - \mu_c$ are normalized using exponential moving averages:

$$\tilde{A} = \frac{A - \mu_A}{\sqrt{\sigma_A^2 + \epsilon}}. \tag{9}$$

3. Uncertainty-Powered Curriculum: Prompt sampling probability follows $p(c) \propto \sigma_c + \eta/\sqrt{n_c + 1}$, prioritizing high-uncertainty prompts.

The policy update uses advantage-weighted regression:

$$\mathcal{L}_{\mathrm{MPO}}(\theta) = \mathbb{E}_{t, \mathbf{z}_t \sim \tau}\left[\mathrm{stop\_grad}(w_c \cdot \tilde{A}) \cdot \left\|\mathbf{v}_\theta(\mathbf{z}_t, c, t) - \mathbf{u}(\mathbf{z}_t, \mathbf{z}_0)\right\|^2\right], \tag{10}$$

where $\mathbf{u}(\cdot)$ is the target flow-matching vector field and $w_c = 1 + \gamma \cdot |r - \mu_c|/(\sigma_c + \epsilon)$.

Training Strategy and Implementation Details. GRPO and MPO experiments are initialized from the same DPO base model. We optimize using the AdamW optimizer [Loshchilov and Hutter, 2017] with a constant learning rate of 5×10^{-6} and a global batch size of 64. For MPO and GRPO, we employ a 12-step SDE sampler with Euler-Maruyama discretization. The diffusion coefficient g_t is linearly annealed from 0.1 to 0 during training. For MPO-specific hyperparameters, we set the EMA decay for advantage normalization to $\lambda = 0.99$, the curriculum balance coefficient to $\eta = 1.0$, the adaptive scaling factor $\alpha = 1.0$, and the surprise reweighting factor to $\gamma = 0.5$.

5 Model Performance

5.1 Benchmarks

We conduct a comprehensive evaluation using a suite of established public benchmarks, including GenEval [Ghosh et al., 2023], DPG-Bench [Hu et al., 2024], and WISE [Niu et al., 2025] for general generative capabilities, as well as GlyphDraw2 [Ma et al., 2025a] and CVTG-2K [Du et al., 2025] for text rendering proficiency. Furthermore, to rigorously assess Chinese character coverage, we construct a complete dictionary-based benchmark, *ChineseWord*, following the protocol of Qwen-Image. Finally, to validate performance in production environments, we introduce a proprietary dataset focusing on business-critical scenarios such as poster design and natural scenes with text.

5.1.1 Text-Image Alignment

GenEval evaluates the fine-grained controllability of generative models, specifically targeting attribute binding, quantitative relations, and spatial composition. As shown in Table 2, LongCat-Image exhibits superior performance on GenEval, demonstrating robust capabilities in handling complex compositional constraints and entity attributes.

DPG-Bench comprises 1,065 dense and structurally complex prompts designed to challenge the semantic alignment of text-to-image models. As presented in Table 3, LongCat-Image achieves competitive alignment performance, ranking closely behind leading models such as Qwen-Image and Seed4.0, thereby validating its proficiency in interpreting verbose captions.

WISE comprises 1,000 curated prompts aimed at rigorously testing semantic comprehension and world knowledge. During evaluation, we leverage the off-the-shelf text encoder for intrinsic prompt enhancement. Results indicate that LongCat-Image achieves state-of-the-art (SOTA) scores among open-source diffusion models, underscoring its robust reasoning capabilities and responsiveness to semantic nuances. Table 4 details these findings.

5.1.2 Text Rendering

GlyphDraw2 assesses text generation across two distinct subsets. The *Poster-Set* (200 prompts) evaluates bilingual generation in design contexts, while the *Complex-Set* challenges models with random combinations drawn from a pool of 2,000 frequent Chinese characters to test coverage. As illustrated in Table 5, LongCat-Image excels particularly in the Complex-Set, highlighting its robustness in rendering intricate character structures.

CVTG-2K focuses on English text rendering across diverse real-world scenarios, including street views, advertisements, and memes. Each prompt features multi-region layouts (2 to 5 regions) to test spatial text placement. LongCat-Image attains SOTA performance on this benchmark (see Table 5), demonstrating exceptional effectiveness in complex, multi-turn English text rendering tasks.

Table 2: Quantitative Evaluation results on GenEval.

Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall↑
Show-o [Xie et al., 2024]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3 [Wang et al., 2024b]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
PixArt-α [Chen et al., 2024a]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD-3-Medium [Esser et al., 2024]	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1-dev [Labs, 2024]	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD-3.5-large [Stabilityai, 2024]	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow [Ma et al., 2025b]	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Lumina-Image 2.0 [Qin et al., 2025]	-	0.87	0.67	-	-	0.62	0.73
Janus-Pro-7B [Chen et al., 2025]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream-I1-Full [Cai et al., 2025]	1.00	0.98	0.79	0.91	0.60	0.72	0.83
GPT Image 1 [High] [OpenAI, 2025]	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 3.0 [Gao et al., 2025]	0.99	0.96	0.91	0.93	0.47	0.80	0.84
Seedream 4.0 [Gao et al., 2025]	0.99	0.92	0.72	0.91	0.76	0.74	0.84
Qwen-Image [Wu et al., 2025a]	0.99	0.92	0.89	0.88	0.76	0.77	**0.87**
HunyuanImage-3.0 [Cao et al., 2025]	1.00	0.92	0.48	0.82	0.42	0.63	0.72
LongCat-Image	0.99	0.98	0.86	0.86	0.75	0.73	**0.87**

Table 3: Quantitative evaluation results on DPG.

Model	Global	Entity	Attribute	Relation	Other	Overall↑
PixArt-α [Chen et al., 2024a]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [Zhuo et al., 2024]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [Podell et al., 2023]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [Li et al., 2024b]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [Li et al., 2024a]	84.59	80.59	88.01	74.36	86.41	78.87
Janus [Wu et al., 2025b]	82.33	87.38	87.70	85.46	86.41	79.68
PixArt-Σ [Chen et al., 2024c]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [Wang et al., 2024b]	85.21	86.68	86.84	90.22	83.15	80.60
Janus-Pro-1B [Chen et al., 2025]	87.58	88.63	88.17	88.98	88.30	82.63
DALL-E 3 [OpenAI, 2023]	90.97	89.61	88.39	90.58	89.83	83.50
FLUX.1-dev [Labs, 2024]	74.35	90.00	88.96	90.87	88.33	83.84
SD-3-Medium [Esser et al., 2024]	87.90	91.01	88.83	80.70	88.68	84.08
Janus-Pro-7B [Ma et al., 2025b]	86.90	88.90	89.40	89.32	89.48	84.19
HiDream-I1-Full [Cai et al., 2025]	76.44	90.22	89.48	93.74	91.83	85.89
Lumina-Image 2.0 [Qin et al., 2025]	-	91.97	90.20	**94.85**	-	87.20
Seedream 3.0 [Gao et al., 2025]	**94.31**	**92.65**	91.36	92.78	88.24	88.27
GPT Image 1 [High] [OpenAI, 2025]	88.89	88.94	89.84	92.63	90.96	85.15
Seedream 4.0 [Seedream et al., 2025]	94.10	92.28	92.75	93.67	92.77	88.25
Qwen-Image Wu et al. [2025a]	91.32	91.56	**92.02**	94.31	**92.73**	**88.32**
HunyuanImage-3.0 [Cao et al., 2025]	92.12	92.53	89.13	92.13	91.92	86.10
LongCat-Image	89.10	92.54	92.00	93.28	87.50	86.80

Table 4: Quantitative evaluation results of world knowledge reasoning on WISE.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	**Overall**
Unified Models							
GPT4o [OpenAI, 2025]	**0.81**	**0.71**	**0.89**	**0.83**	**0.79**	**0.74**	**0.80**
MetaQuery-XL [Pan et al., 2025]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
Liquid [Wu et al., 2025c]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Emu3 [Wang et al., 2024b]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Janus-1.3B [Wu et al., 2025b]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
JanusFlow [Ma et al., 2025b]	0.13	0.26	0.28	0.20	0.19	0.11	0.18
Janus-Pro-1B [Chen et al., 2025]	0.20	0.28	0.45	0.24	0.32	0.16	0.26
Janus-Pro-7B [Chen et al., 2025]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Orthus-7B-instruct [Kou et al., 2024]	0.23	0.31	0.38	0.28	0.31	0.20	0.27
Show-o-512 [Xie et al., 2024]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Text-to-Image Models							
FLUX.1-dev [Labs, 2024]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
FLUX.1-schnell [Labs, 2024]	0.39	0.44	0.50	0.31	0.44	0.26	0.40
PixArt-α [Chen et al., 2024a]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Playground-v2.5 [Li et al., 2024b]	0.49	0.58	0.55	0.43	0.48	0.33	0.49
SD-3-medium [Esser et al., 2024]	0.42	0.44	0.48	0.39	0.47	0.29	0.42
SD-3.5-medium [Stabilityai, 2024]	0.43	0.50	0.52	0.41	0.53	0.33	0.45
SD-3.5-large [Stabilityai, 2024]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
Seedream 4.0 [Seedream et al., 2025]	0.78	0.73	0.85	0.79	0.84	0.67	**0.78**
Qwen-Image [Wu et al., 2025a]	0.62	0.63	0.77	0.57	0.75	0.40	0.62
HunyuanImage-3.0 [Cao et al., 2025]	0.58	0.57	0.70	0.56	0.63	0.31	0.57
LongCat-Image	0.66	0.61	0.72	0.66	0.72	0.49	<u>0.65</u>

Table 5: Quantitative evaluation results of GlyphDraw2.

Model	Complex-en	Complex-zh	Poster-en	Poster-zh	Avg
Seedream 4.0 [Seedream et al., 2025]	0.99	0.91	0.99	0.99	**0.97**
Qwen-Image [Wu et al., 2025a]	0.90	0.87	0.95	0.98	0.93
HunyuanImage-3.0 [Cao et al., 2025]	0.47	0.85	0.90	0.90	0.78
LongCat-Image	0.94	0.92	0.95	0.99	<u>0.95</u>

Table 6: Quantitative evaluation results of CVTG-2K.

Model	**Word Accuracy↑**					**NED↑**	**CLIPScore↑**
	2 regions	3 regions	4 regions	5 regions	average		
Seedream 4.0 [Seedream et al., 2025]	0.8898	0.9147	0.8991	0.8873	**0.8917**	**0.9507**	0.7853
Qwen-Image [Wu et al., 2025a]	0.8370	0.8364	0.8313	0.8158	0.8288	0.9297	<u>0.8059</u>
HunyuanImage-3.0 [Cao et al., 2025]	0.8300	0.7635	0.7384	0.7279	0.7650	0.8765	**0.8121**
LongCat-Image	0.9129	0.8737	0.8557	0.8310	<u>0.8658</u>	<u>0.9361</u>	0.7859

ChineseWord To evaluate the full spectrum of Chinese character rendering, especially for long-tail characters, we constructed a comprehensive benchmark comprising 8,105 prompts based on the *General Standard Chinese Characters*

Table 7: Quantitative evaluation results of ChineseWord.

Model	L1	L2	L3	Overall
Seedream 4.0 [Seedream et al., 2025]	94.8	41.2	2.3	58.5
Qwen-Image [Wu et al., 2025a]	92.5	37.1	6.1	56.6
HunyuanImage-3.0 [Cao et al., 2025]	83.5	31.3	4.1	49.3
LongCat-Image	98.7	90.8	70.3	**90.7**

Table 8: Quantitative evaluation results of internal poster and scene text scenarios.

Model	Poster	Real Scene	Avg
Seedream 4.0 [Seedream et al., 2025]	93.2	90.0	**91.6**
Qwen-Image [Wu et al., 2025a]	88.7	89.6	89.2
HunyuanImage-3.0 [Cao et al., 2025]	89.0	85.1	87.1
LongCat-Image	92.0	91.0	91.5

Table[2], aligning with the protocol of Qwen-Image. Each character is embedded in a standardized template (e.g., "On the blackboard, the word '华' is written in purple Song font."). We employ PPOCRv5 [Cui et al., 2025] for objective accuracy quantification, as existing MLLMs often struggle to recognize rare characters. Results in Table 7 demonstrate that LongCat-Image outperforms all existing models by a significant margin. However, we acknowledge that while exhibiting dominance in single-character rendering, the model experiences a noticeable decline in stability when generating multi-character sequences, primarily due to the insufficient scale of real-world textual training data. In future work, we aim to address this by rigorously expanding our text-rich dataset collection to enhance robustness in complex multi-character generation tasks.

Poster&SceneBench To bridge the gap between academic benchmarks and industrial applications, we curate a dataset of 500 prompts covering both poster typography and natural scene text. Unlike flat poster layouts, the latter specifically evaluates the model's capability to seamlessly integrate text into real-world environments (*e.g.*, signage on textured surfaces or shop fronts with complex lighting). As indicated in Table 8, LongCat-Image delivers SOTA-level performance, proving its reliability and effectiveness in these practical operational contexts.

5.2 Human Evaluation

To assess perceptual quality, we adopt the Mean Opinion Score (MOS) protocol, focusing on four distinct dimensions: text-image alignment, visual plausibility, visual realism, and aesthetics.

- **Text-Image Alignment** measures the semantic fidelity of the generated image to the input prompt. It evaluates the accurate depiction of key elements, including entities, attributes, spatial relationships, and stylistic constraints.
- **Plausibility** examines the image's adherence to physical coherence and logical consistency. This metric penalizes anatomical distortions, unnatural proportions, and spatial anomalies that violate real-world physics.
- **Realism** assesses the degree of photorealism and texture fidelity. It serves as a proxy for the "Turing test" of image generation, determining how indistinguishable the synthesized output is from authentic photography, free from "AI-generated" artifacts.
- **Aesthetics** evaluates the artistic quality and perceptual appeal, considering factors such as composition, lighting, color harmony, and overall visual impact.

Evaluation dataset. To ensure a rigorous and unbiased assessment, we constructed a diverse dataset of 400 prompts tailored for black-box evaluation. This corpus spans a broad spectrum of difficulty, ranging from fundamental entity depiction to intricate scene synthesis. It comprehensively covers challenging generative dimensions, including multi-entity interactions, spatial layouts, dynamic actions, artistic creativity, text rendering, and world knowledge reasoning.

[2] http://www.moe.gov.cn/jyb_sjzl/ziliao/A19/201306/t20130601_186002.html

Result Analysis. As illustrated in Fig. 13, LongCat-Image demonstrates comprehensive superiority over HunyuanImage 3.0 across all metrics. Compared to Qwen-Image, our model achieves parity in both Text-Image Alignment and Visual Plausibility. Notably, LongCat-Image excels in Visual Realism, outperforming Qwen-Image and even exhibiting a slight advantage over the commercial baseline, Seedream 4.0. While there remains a marginal gap in Visual Aesthetics compared to Qwen-Image, the overall human evaluation indicates that LongCat-Image delivers performance comparable to SOTA open-source models.

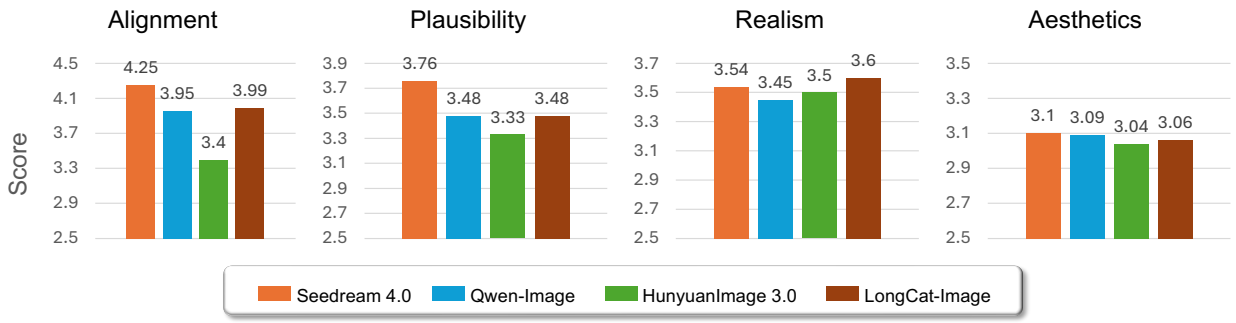

Figure 13: **Comparison of human evaluation MOS.**

5.3 Qualitative Results

In Fig. 14, 15, 16, 17, we present qualitative comparisons between our LongCat-Image and leading SOTA baselines. Visual inspection confirms that our model exhibits robust performance across critical dimensions, including text-image alignment, visual plausibility, realism, and aesthetic quality. Furthermore, it demonstrates exceptional proficiency in text rendering tasks.

6 Image Editing

Extending a Text-to-Image foundation model for image editing is a well-established paradigm, effectively validated by prior works [Batifol et al., 2025, Wang et al., 2025b, Wu et al., 2025a]. In this section, we detail the adaptation of LongCat-Image into **LongCat-Image-Edit**, which achieves SOTA performance among open-source models.

6.1 Data Curation

Distinct from standard text-to-image pre-training, image editing necessitates source-target image pairs. We curate a comprehensive training set from diverse sources, including open-source datasets, synthetic pipelines, video sequences, and interleaved web corpora. To enhance the model's instruction-following capabilities, which range from simple descriptions to complex reasoning, we apply extensive instruction rewriting strategies to maximize linguistic diversity. Fig. 18 illustrates the proportional distribution of editing tasks within our training set.

6.1.1 Open-Source Datasets

Given the high cost of annotating editing pairs, we prioritize leveraging high-quality open-source repositories, specifically OmniEdit [Wei et al., 2024], OmniGen2 [Wu et al., 2025d], and NHREdit [Kuprashevich et al., 2025]. We implement a rigorous data cleaning pipeline on these datasets and rewrite the original instructions. This process efficiently yields high-fidelity training pairs tailored for diverse editing tasks.

6.1.2 Synthesized Data

Specialized expert models demonstrate exceptional performance in specific common editing tasks. Leveraging their capabilities, we synthesize high-quality training pairs for tasks such as object manipulation, style transfer, background alteration, and reference-based generation. For each task, we establish a dedicated pipeline where MLLMs craft editing instructions, and the corresponding expert models generate the target images. Complementarily, we employ traditional image processing algorithms to construct data for low-level adjustments like filter transformations and lighting changes. Furthermore, we incorporate human-in-the-loop verification for critical samples to ensure semantic alignment and visual fidelity. This approach allows us to accumulate a substantial volume of high-quality data.

Seedream 4.0 Qwen-Image Hunyuan 3.0 LongCat-Image

一对年轻情侣坐在毛绒桌布前，男性穿西装，女性身穿白色婚纱，佩戴珍珠项链和头纱，双手戴白色长手套。她的姿态优雅地支撑在桌上，右手手指微张。桌上摆放着一只奶油草莓蛋糕，周围散落几颗草莓。左侧有金色烛台，右侧有白色蜡烛，光线柔和，增添了一丝浪漫氛围。背景为浅灰色墙壁，被柔和的焦点光环衬托。全图光照柔和，整体氛围温馨且梦幻。

画面前景是一个荷花池，池中漂浮着大量绿色的荷叶，其间点缀着粉色和白色的荷花，荷花花瓣饱满，花蕊清晰可见。荷花池的水面映照出建筑的倒影。荷花池后方是一座宏伟的中国古典建筑，建筑主体为两层，屋顶覆盖着金黄色的琉璃瓦，层层叠叠，飞檐翘角。建筑的墙体呈红色，柱子为朱红色，雕梁画栋，门窗装饰有精美的金色图案。建筑前方有白色的石质台阶和栏杆，通向建筑入口。建筑两侧各有一棵古老的树木，树干粗壮，枝繁叶茂，树叶呈绿色，部分被夕阳染成金黄色，与建筑的金色屋顶相得益彰。建筑后方可见红色的围墙，围墙上方有绿色的树木。天空呈现出淡黄色到浅蓝色的渐变，光线柔和，表明拍摄时间为傍晚。呈现出温暖的金色调，色彩丰富，以红色、金色、绿色和粉白色为主，饱和度较高，营造出一种华丽而宁静的氛围。画面质感细腻，展现了古典建筑的庄重与荷花的柔美。

主体为一个透明杯，杯中装有分层的粉色茶饮，顶部撒有细腻的桂花，杯身带有“TEA”标识。茶饮的分层色彩从粉色渐变至白色，展现出细腻的过渡效果。玻璃杯垂直放置在圆形木质托盘的中心，呈现出自然的木质棕色。托盘上散落着两个新鲜的山楂，色泽鲜红，旁边点缀着几朵桂花，增添了画面的细腻感。一根中式枝条在图片右上角伸出。背景为浅橙色渐变，营造出温暖的养生氛围，左上角射光光影映射。拍摄采用中心构图，平视角度，清晰展现茶饮的分层过渡、桂花的质感、山楂的色泽以及托盘的纹理。红色器皿中盛放着山楂，呈现出规整的中式自然摆放形态，整体风格参考养生茶饮品牌宣传图，追求中式元素与产品的完美融合。

Figure 14: **Comparison of overall capability in image generation.**

Seedream 4.0 | Qwen-Image | Hunyuan 3.0 | LongCat-Image

At dusk on the city streets, on an old brick wall, there is a long graffiti written in black spray paint: "我有一个梦想。-科技不再是权力的象征，而是人人都可使用的魔法;我有一个梦想，每一个平凡人都能借助AI书写不平凡的故事.." The street atmosphere carries hope and reflection, with the sunlight casting its last rays on the wall. The graffiti's messy font is full of power, presenting a realistic and vivid scene. Shadows pass by but do not disturb the center of the picture.。

主体为一个复古黑板风格的菜单，黑板表面呈现出自然的木质纹理。黑板顶部用白色粉笔书写的艺术字“云朵咖啡”，字体优雅且略带弯曲。饮品栏详细列出几款饮品，分别是“美式 | Americano ￥26”、“拿铁 | Latte ￥28”、“桂花特调 | Osmanthus Special ￥32”、“冰滴咖啡 | Cold Brew Coffee ￥38”。黑板底部用白色粉笔书写标语“每日新鲜烘焙 | Freshly Baked Daily”，字体略大且醒目。背景为饮品店的门口，门口装饰有绿色植物和木质装饰，阳光透过玻璃窗洒在黑板上，营造出温馨且自然的氛围。

From a bird's-eye view, a primary school student holds a crayon, facing a drawing assignment on the desk. The paper reads: “1. Draw a happy sun. 2. Color the apple red. 3. Draw a tall tree. 4. Color the sky blue. 5. Draw a fluffy cloud.”. The text is printed in a large, playful font, centered. The student has already drawn a bright yellow sun with a smiling face for the first question. On the student's left-hand palm, a small 'star' sticker is attached.

Figure 15: **Comparison of text rendering capability in image generation.**

Seedream 4.0 | Qwen-Image | Hunyuan 3.0 | LongCat-Image

Style poster design, main title "和自然做邻居", with a childlike handwritten texture font, exaggerated and eye-catching. The small text reads "2025 年4月7日，观山湖环湖", encircling the title with handwritten texture font. The background features a comic dot-style natural scenery, with dot effects and paper tearing texture, adorned with neon-colored expression decorations.

图片主体为一盘精致的三文鱼牛油果寿司卷，寿司卷呈现出鲜艳的橙色和绿色，三文鱼切片光泽鲜亮，牛油果柔滑细腻。寿司米拌日式米醋，酸甜适中，米粒饱满且有光泽。外层裹着海苔脆。背景为一个木质餐桌，桌面上散落着几片新鲜的柠檬和薄荷叶，营造出清新自然的氛围。图片上方有醒目的标题“（鲜渔寿司）”，字体为现代风格，颜色为深蓝色。下方是宣传语“新鲜直达，一口尝到深海鲜味”，字体为白色，搭配浅蓝色背景框。右下角标注规格和价格：“6 块 / 份，¥38”，字体清晰，背景为浅灰色。整体构图简洁，色彩搭配和谐，突出寿司的卖点：挪威冰鲜三文鱼和墨西哥牛油果。

字体海报设计，极简风格，渐变浅灰背景，海报中间有一个布制沙发座椅，现代感十足的排版层次，白底黑字清晰锐利，几何构图简洁。主标题区域（左上）：粗体黑色无衬线字体“LIVING IN A DREAM HOME”，视觉冲击力强，现代工业风字体粗细。中文副标题：粗体黑色字体“住进梦想家)”在英文正下方。时间信息（左中）：粗体红色字体“TIME 08.05 - 08.07”，结构化格式。产品亮点：黑色字体“超软座椅 超大舒适度”位于右侧，简洁的无衬线字体。底部文字元素：左侧的“超软座椅 木质底座”，右侧为英文大写字母“¥399.00 纤维扶手椅 FIBER ARMCHAIR”，其中“399.00”使用红色粗体描绘。下方一排小字：细体黑色字体底部“无论在公共场所还是在家中都能提供最大的舒适度”，右上角侧为垂直排列文字：“WOOD BASE”。

Figure 16: **Comparison of text rendering capability in image generation.**

A museum exhibition poster features a rustic deep green and black background, creating a sophisticated historical atmosphere. Central to the design is a bronze vessel with ornate handles, showcasing ancient craftsmanship. Above it, gold calligraphic lettering reads “饕餮纹青铜簋”. Below, smaller text provides essential details: “所属年代：西周早期” and “展览时间：10/10-10/30”. The balanced composition highlights the vessel's significance, conveying timelessness and cultural heritage.

一座古色古香的传统中式门楼，上方悬挂着一块写有“吉祥如意”的黑色牌匾，字体庄重古朴。门楼两侧的柱子上分别竖立着“犇犇骏马迎新岁”和“翔翔凤凰贺大年”的竖排黑色文字。画面采用写实摄影风格，细节清晰。门楼位于画面中央，以白色和灰色为主色调，搭配深色木质结构，纹理清晰可见。背景是蓝天白云。正午阳光从上方垂直照射下来，光线明亮均匀，突显建筑的立体感和历史沧桑感，营造出古朴典雅且具有浓厚历史感的氛围。

An authentic photographic style captures a classroom setting focused on a blackboard. The blackboard prominently displays “有机化学” in large chalk font and “醇、酚、醚、醛、酮的区别”. The background includes wooden desks and chairs, adding context, with natural lighting from side windows casting soft shadows.

Figure 17: **Comparison of text rendering capability in image generation.**

6.1.3 Video Frames

Synthetic methods often struggle with complex structural changes, such as human pose and perspective, frequently introducing artifacts. To bridge this gap, we leverage video sequences that naturally capture realistic temporal transitions. Our pipeline employs multimodal models to identify target objects and optical flow estimation to quantify changes between frames. We extract keyframe pairs that exhibit significant yet coherent variations and automatically annotate them with editing instructions. A subset of these pairs undergoes manual verification to guarantee accuracy.

6.1.4 Interleaved Corpus

While the aforementioned data sources effectively cover standard editing categories, they often lack coverage for long-tail scenarios. To bridge this gap and significantly enrich the diversity of editing instructions, we exploit web-scale interleaved corpora. By mining large-scale image-text sequences with inherent semantic correlations, we extract implicit editing signals from naturally occurring data. These raw pairs undergo rigorous filtering and multimodal-assisted instruction rewriting to ensure their suitability for training. However, mining valid training samples from such massive unstructured corpora is an extremely resource-intensive endeavor. Consequently, the scale of data we have curated to date remains limited. We firmly believe that this represents a critical direction for long-term data engineering and encourage broader community participation to further explore this valuable frontier.

6.1.5 Instruction Rewriting

Since synthetic instructions often diverge from real-world user prompts and complex reasoning benchmarks, we employ GPT-4o [Hurst et al., 2024] to enhance instruction diversity. We implement a one-to-many strategy, associating each editing pair with multiple rewritten variants, including natural language paraphrases and compound commands. This approach aligns training data with diverse inference scenarios, a benefit subsequently validated by our experimental results.

Figure 18: **Overview of image editing pre-training data.**

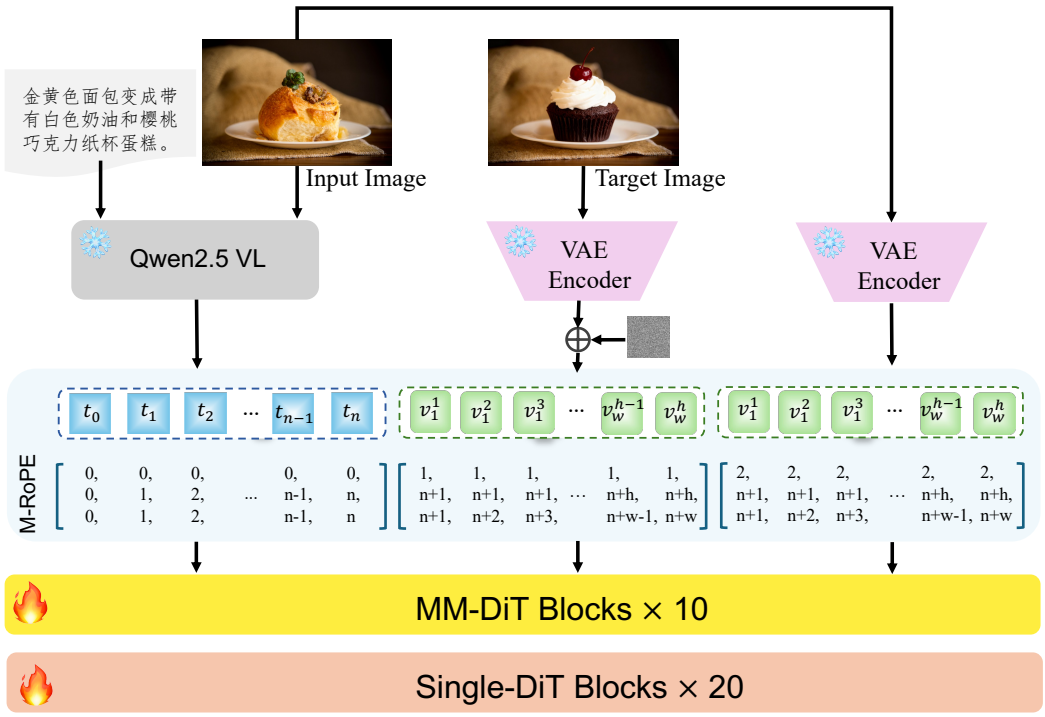

Figure 19: **Overview of LongCat-Image-Edit model architecture.**

6.2 Model Design

Building upon the base architecture and drawing inspiration from prior works [Wang et al., 2025b, Wu et al., 2025a], we introduce an image conditioning branch via modifications to VAE features, 3D RoPE embeddings, and token sequencing. Specifically, reference images are encoded into VAE latents and distinguished from noised latents by manipulating the first dimension of the 3D RoPE embeddings, while preserving spatial alignment in the remaining dimensions. These reference tokens are then concatenated with noised latents along the sequence dimension to serve as input for the diffusion visual stream. Furthermore, we feed both the source image and instructions into the multimodal encoder (*i.e.*, Qwen2.5-VL). To differentiate editing tasks from standard text-to-image generation, we implement a distinct system prompt during this feature extraction process. The overall schematic of the model architecture is illustrated in Fig. 19.

6.3 Model Training

As illustrated in Fig. 12, the training framework comprises three progressive stages: Pre-training, SFT, and DPO. This multi-stage curriculum is designed to systematically enhance the resolution and visual fidelity of the generated images.

6.3.1 Pre-training

We initialize the model using a mid-training T2I checkpoint, as its unconstrained parameter space offers superior plasticity compared to post-trained models. Our training follows a multi-scale strategy: we begin at 512×512 resolution with massive, noisy datasets for rapid convergence, then progress to 1024×1024 with high-quality data to refine details. Simultaneously, we adopt a joint training strategy, mixing editing data with T2I mid-training data at a balanced batch ratio. Since experiments confirm this approach improves both semantic understanding and image quality, we retain it for the following SFT stage. Furthermore, to enhance instruction generalization, we associate each sample with 3-5 candidate prompts (in Chinese and English) and randomly select one during each training iteration.

6.3.2 SFT

During the SFT stage, we curate a high-fidelity dataset comprising hundreds of thousands of samples from real photographs, professional manual retouches, and synthetic sources. To guarantee generation stability, we implement a rigorous human-in-the-loop filtering protocol, specifically targeting the structural alignment between source and

edited images. Our experiments reveal a high sensitivity to data quality: even a marginal relaxation of these alignment standards leads to a precipitous drop in the model's ability to maintain consistency. Consequently, we enforce the strictest criteria to ensure precise content preservation. Furthermore, by jointly training this strictly filtered corpus with high-quality T2I SFT data, we achieve significant improvements in both instruction adherence and aesthetic quality.

6.3.3 DPO

To further align the model with human aesthetic standards and mitigate persistent structural artifacts, we employ Direct Preference Optimization (DPO) following the SFT stage. We construct a high-quality preference dataset via diverse sampling and rigorous manual annotation, optimizing the diffusion DPO objective defined as:

$$
\begin{aligned}
L(\theta) = &- \mathbb{E}_{(I^w_{\text{src}},P^w,I^l_{\text{src}},P^l)\sim\mathcal{D},\ t\sim\mathcal{U}(0,T),\ x^w_t\sim q(x^w_t|I^w_{\text{src}},P^w),\ x^l_t\sim q(x^l_t|I^l_{\text{src}},P^l)} \\
&\log\sigma\Bigg(-\beta T\omega(\lambda_t)\Bigg(\|v^w - v_\theta(x^w_t, I^w_{\text{src}}, P^w, t)\|_2^2 - \|v^w - v_{\text{ref}}(x^w_t, I^w_{\text{src}}, P^w, t)\|_2^2 \\
&- \left(\|v^l - v_\theta(x^l_t, I^l_{\text{src}}, P^l, t)\|_2^2 - \|v^l - v_{\text{ref}}(x^l_t, I^l_{\text{src}}, P^l, t)\|_2^2\right)\Bigg)\Bigg).
\end{aligned} \tag{11}
$$

Data Construction. The preference dataset is curated in two steps. (1) Prompt Curation: We synthesize category-balanced image-instruction pairs leveraging both Multimodal LLMs and real-world user queries. These inputs undergo clustering and filtration to ensure semantic diversity and representativeness. (2) Preference Annotation: For each unique prompt, we generate five candidate outputs using distinct random seeds. A professional annotation team evaluates these candidates to identify the most successful edit (winner) and the failed instances (losers). This rigorous selection process ensures high-confidence preference signals, forming robust win-lose pairs for training.

Training Strategy. We optimize the model using Eq. (11). To stabilize training and prevent reward hacking, we incorporate advanced strategies such as gradient-based outlier rejection (to skip noisy data) and KL divergence constraints. Comparative analysis confirms that these techniques are crucial for performance gains. Ultimately, DPO significantly reduces the failure rate (*e.g.*, structural artifacts) and enhances the overall robustness of the image generation.

6.4 Discussion

Initially, we aim to unify T2I and image editing into a single model to leverage potential task synergies and minimize deployment costs. However, experiments reveal a critical data quality mismatch: the heavy reliance on synthetic data during editing pre-training noticeably degrades the photorealism of T2I generation compared to models trained solely on real data. Consequently, we decided to separate the models to ensure optimal performance for each task. We emphasize that this is a data-driven issue, not an architectural flaw. We believe that by substituting synthetic datasets with large-scale interleaved corpora, future iterations can successfully merge these capabilities into a unified model without sacrificing generation quality.

6.5 Model Performance

In this section, we comprehensively evaluate our model across three quantitative benchmarks: CEdit-Bench (Ours), GEdit-Bench [Liu et al., 2025], and ImgEdit-Bench [Ye et al., 2025]. Furthermore, we conduct a qualitative comparison against leading models on complex editing tasks to demonstrate practical utility.

6.5.1 Benchmarks

CEdit. While numerous benchmarks exist for image editing, they often exhibit specific limitations in task coverage and granularity. For instance, GEdit-Bench, despite its popularity, lacks tasks involving reference image generation, structural modification, and viewpoint transformation. Similarly, ImgEdit-Bench offers a limited scope, KontextBench [Batifol et al., 2025] suffers from coarse task granularity and low instruction diversity, and Reason-Edit [Huang et al., 2024] prioritizes reasoning over conventional editing capabilities.

To address these gaps, we introduce CEdit-Bench, a comprehensive evaluation suite derived from the integration and expansion of these existing benchmarks. We curate new data to enrich task diversity, resulting in a robust dataset

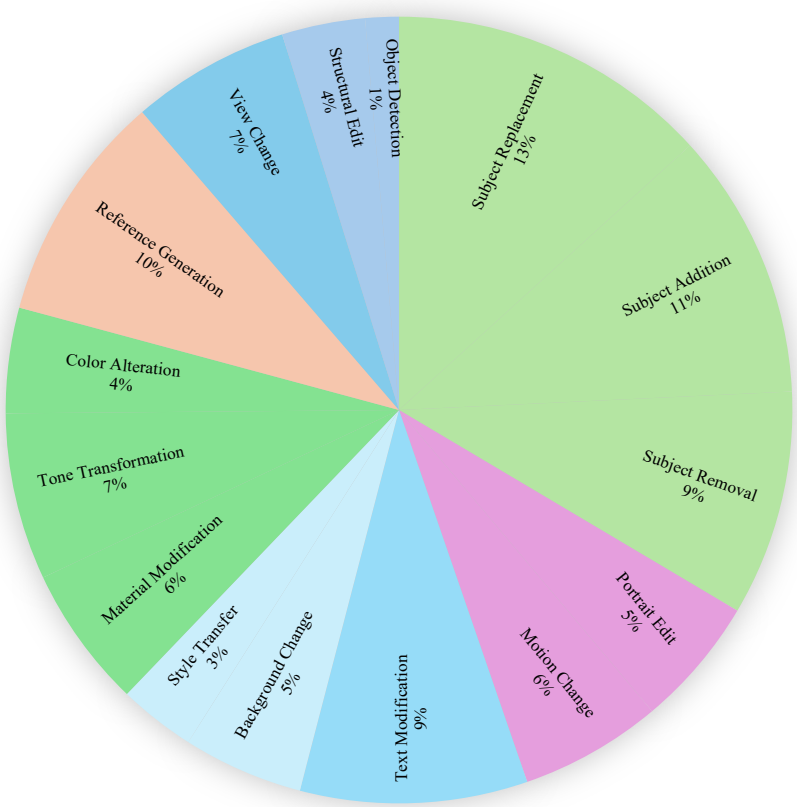

Figure 20: **Task category distribution of CEdit-Bench.**

comprising 1,464 bilingual (Chinese and English) editing pairs across 15 fine-grained task categories, as illustrated in Fig. 20. This establishes CEdit-Bench as a more holistic and rigorous evaluation standard.

We benchmark our model against FLUX.1 Kontext [Batifol et al., 2025], Step1X-Edit [Liu et al., 2025], Qwen-Image-Edit [Wu et al., 2025a], Seedream 4.0 [Seedream et al., 2025], and Nano Banana (Gemini-2.5-flash-image)[3] on CEdit-Bench. Following standard evaluation protocols, we employ Semantic Consistency (SQ), Perceptual Quality (PQ), and an Overall Score (O) as metrics, utilizing GPT-4o for automated evaluation. As shown in Table 9, our model achieves SOTA performance among open-source models.

GEdit. To benchmark against established standards, we evaluate our model on GEdit-Bench, comparing it with both popular open-source and proprietary products. As reported in Table 10, our model achieves top-tier performance, demonstrating superior instruction-following capabilities in both Chinese and English.

ImgEdit. Serving as a complement to GEdit, ImgEdit-Bench evaluates models with a focus on instruction adherence, editing quality, and detail preservation. We compare our model against competitive baselines using the official metrics provided by the benchmark. The results in Table 11 indicate that our model outperforms all competitors, further validating its comprehensive capabilities across diverse evaluation dimensions.

6.5.2 Human Evaluation

To benchmark our model against SOTA open-source models and leading commercial products, we conduct a Side-by-Side (SBS) human evaluation. This assessment focuses on two primary dimensions: comprehensive quality and consistency.

[3] https://aistudio.google.com/models/gemini-2-5-flash-image

Table 9: Performance comparison on CEdit-Bench.

Model	CEdit-Bench-EN↑			CEdit-Bench-CN↑		
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
FLUX.1 Kontext [Pro] [Batifol et al., 2025]	6.79	7.80	6.53	1.15	8.07	1.43
GPT Image 1 [High] [OpenAI, 2025]	**8.64**	**8.26**	**8.17**	**8.67**	**8.26**	**8.21**
Nano Banana [Google, 2025]	7.51	8.17	7.20	7.67	8.21	7.36
Seedream 4.0 [Seedream et al., 2025]	8.12	7.95	7.58	8.14	7.95	7.57
FLUX.1 Kontext [Dev] [Batifol et al., 2025]	6.31	7.56	5.93	1.25	7.66	1.51
Step1X-Edit [Liu et al., 2025]	6.68	7.36	6.25	6.88	7.28	6.35
Qwen-Image-Edit [Wu et al., 2025a]	8.07	7.84	7.52	8.03	7.78	7.46
Qwen-Image-Edit [2509] [Wu et al., 2025a]	8.04	7.79	7.48	7.93	7.71	7.37
LongCat-Image-Edit	**8.27**	**7.88**	**7.67**	**8.25**	**7.85**	**7.65**

Table 10: Performance comparison on GEdit-Bench.

Model	GEdit-Bench-EN↑			GEdit-Bench-CN↑		
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Gemini 2.0 [DeepMind, 2025]	6.73	6.61	6.32	5.43	6.78	5.36
FLUX.1 Kontext [Pro] [Batifol et al., 2025]	7.02	7.60	6.56	1.11	7.36	1.23
GPT Image 1 [High] [OpenAI, 2025]	7.85	7.62	7.53	7.67	7.56	7.30
Nano Banana [Google, 2025]	7.86	**8.33**	7.54	7.51	**8.31**	7.25
Seedream 4.0 [Seedream et al., 2025]	**8.24**	8.08	**7.68**	**8.19**	8.14	**7.71**
InstructPix2Pix [Brooks et al., 2023]	3.58	5.49	3.60	-	-	-
AnyEdit [Yu et al., 2025]	3.18	5.82	3.21	-	-	-
MagicBrush [Zhang et al., 2023]	4.68	5.66	4.52	-	-	-
UniWorld-v1 [Lin et al., 2025]	4.93	7.43	4.85	-	-	-
OmniGen [Xiao et al., 2025]	5.96	5.89	5.06	-	-	-
OmniGen2 [Wu et al., 2025d]	7.16	6.77	6.41	-	-	-
FLUX.1 Kontext [Dev] [Batifol et al., 2025]	6.52	7.38	6.00	-	-	-
BAGEL [Deng et al., 2025]	7.36	6.83	6.52	7.34	6.85	6.50
Step1X-Edit [Liu et al., 2025]	7.66	7.35	6.97	7.20	6.87	6.86
Qwen-Image-Edit [Wu et al., 2025a]	8.00	7.86	7.56	7.82	7.79	7.52
Qwen-Image-Edit [2509] [Wu et al., 2025a]	8.15	7.86	7.54	8.05	7.88	7.49
LongCat-Image-Edit	**8.18**	**8.00**	**7.64**	**8.08**	**7.99**	**7.60**

- **Comprehensive Quality.** This metric evaluates the overall performance of image editing across multiple aspects, including instruction adherence, visual plausibility, aesthetics, and the consistency between original and edited images. Annotators provide a holistic judgment by categorizing the result as a *Win*, *Tie*, or *Loss*.
- **Consistency.** We conduct a dedicated evaluation for this dimension, distinct from the comprehensive score, to emphasize its critical role in multi-turn editing. This metric specifically scrutinizes whether attributes in non-edited regions, such as layout, texture, color tone, and subject identity, remain invariant unless targeted by the instruction.

Evaluation Dataset. We curate a diverse dataset comprising approximately 400 samples tailored for black-box evaluation. The dataset covers a broad spectrum of difficulty levels and includes various editing tasks, such as global editing, local editing, text modification, and reference-guided editing.

Table 11: Performance comparison on ImgEdit-Bench.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
FLUX.1 Kontext [Pro] [Batifol et al., 2025]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT Image 1 [High] [OpenAI, 2025]	**4.61**	4.33	2.90	4.35	3.66	**4.57**	**4.93**	3.96	**4.89**	4.20
Nano Banana [Google, 2025]	4.50	**4.47**	**3.75**	**4.64**	**4.51**	4.44	4.14	**4.03**	4.65	**4.35**
Seedream4.0 [Seedream et al., 2025]	4.52	4.41	2.93	4.56	4.44	4.30	4.76	3.33	4.36	4.18
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
InstructPix2Pix [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 [Yu et al., 2025]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [Zhao et al., 2024]	3.44	2.81	2.13	2.96	1.45	2.83	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
ICEdIt [Zhang et al., 2025]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
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
BAGEL [Deng et al., 2025]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
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
OmniGen2 [Wu et al., 2025d]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
FLUX.1 Kontext [Dev] [Batifol et al., 2025]	4.12	3.80	2.04	4.22	3.09	3.97	4.51	3.35	4.25	3.71
Qwen-Image-Edit [Wu et al., 2025a]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Qwen-Image-Edit [2509] [Wu et al., 2025a]	4.32	4.36	**4.04**	4.64	4.52	4.37	4.84	3.39	4.71	4.35
LongCat-Image-Edit	**4.51**	**4.57**	3.93	**4.76**	**4.60**	**4.49**	**4.85**	**4.01**	**4.74**	**4.50**

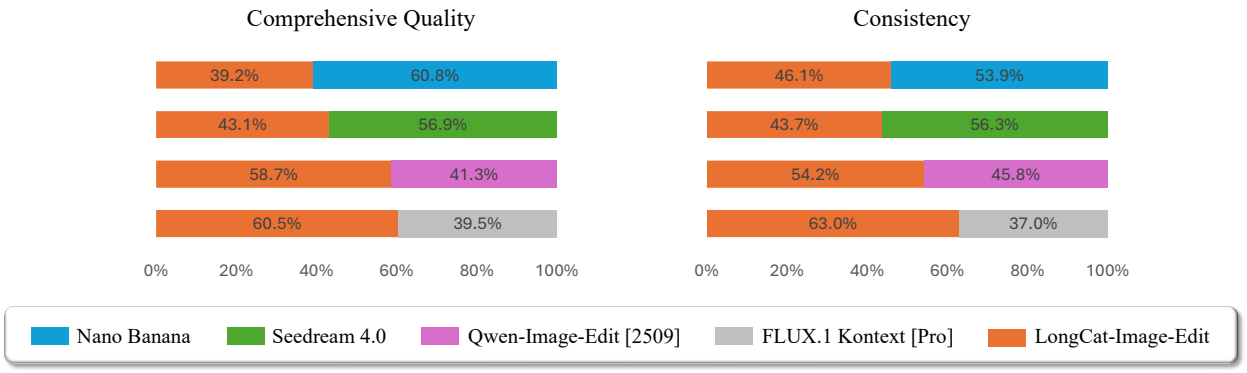

Figure 21: **Comparison of human evaluation win rates between LongCat-Image-Edit and competing methods.**

Result Analysis. We calculate the win rate using the formula: $(\#\text{Win} + 0.5 \times \#\text{Tie})/\#\text{Total}$. As illustrated in Fig. 21, LongCat-Image-Edit outperforms Qwen-Image-Edit [2509] and FLUX.1 Kontext [Pro] in both comprehensive quality and consistency. However, a performance gap remains when compared to commercial systems such as Nano Banana and Seedream 4.0.

6.5.3 Qualitative Results

To comprehensively evaluate our model's versatility, we conduct qualitative comparisons against leading instruction-based image editing baselines. We begin by highlighting the model's performance in two distinct, high-demand real-world scenarios: **Multi-turn Editing** and **Portrait and Human-Centric Editing**. These tasks are selected for their prevalence in practical applications and the rigorous requirements they impose on editing precision.

Multi-turn Editing. Sequential editing imposes stringent demands on a model's ability to preserve visual consistency across iterative steps. We evaluate our model on representative editing chains and further challenge it with *compound instructions*—where multiple edits are requested in a single prompt. As illustrated in Fig. 22, our model maintains exceptional semantic and structural consistency throughout the entire editing sequence. Remarkably, even when processing a complex prompt containing six distinct operations, the model executes all directives accurately. The result aligns closely with the sequential output, underscoring its robust capability in handling both fine-grained iterative updates and complex composite tasks.

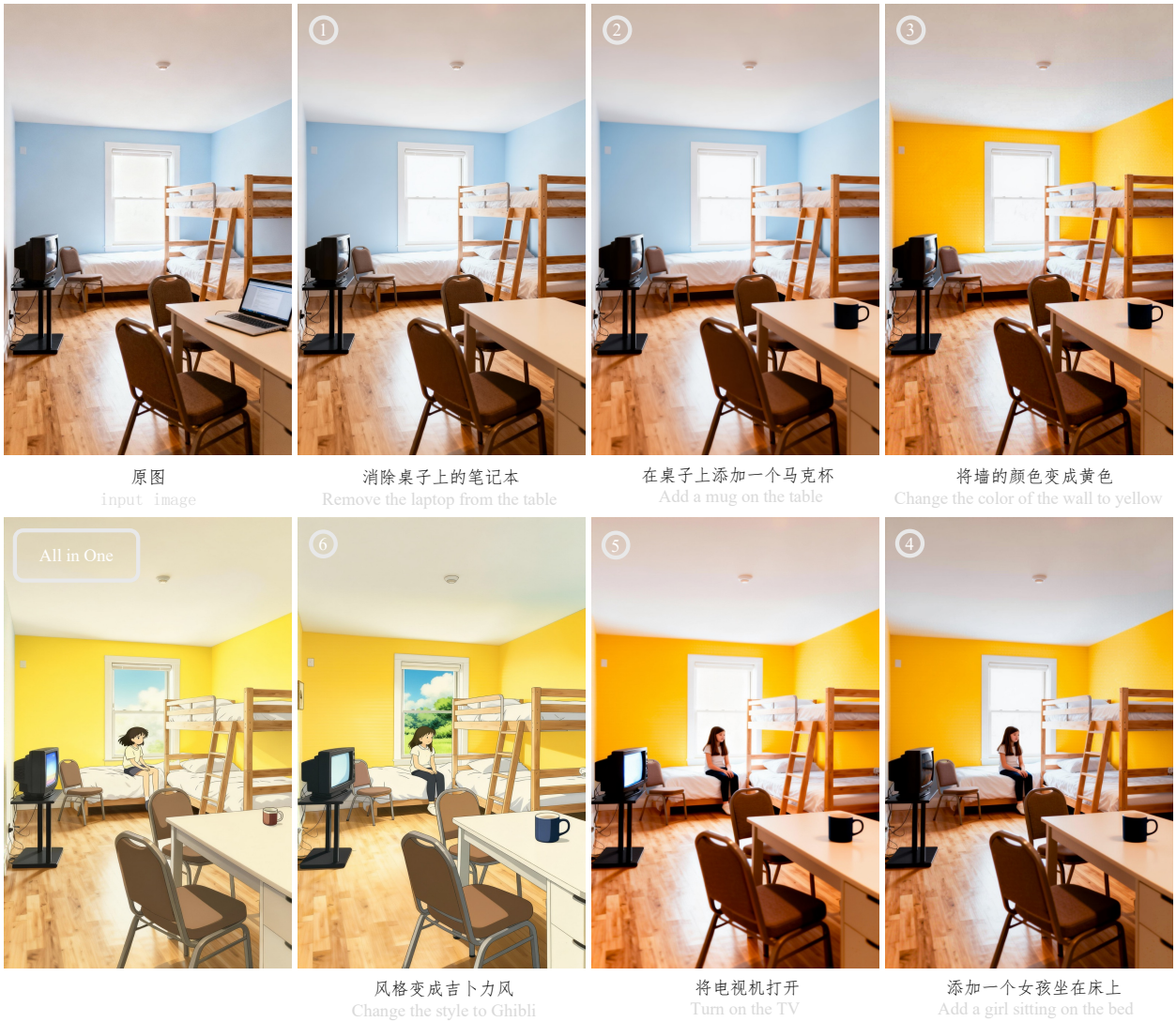

Figure 22: **Visual comparison of multi-turn editing versus one-shot composite editing.** The numbered sequence (①–⑥) illustrates the progressive results of multi-turn editing. In contrast, the "All in One" image (bottom-left) demonstrates the outcome of a single complex instruction containing all six operations: *Remove the laptop, add a mug, change the wall to yellow, add a girl sitting on the bed, turn on the TV, and change the style to Ghibli.*

Portrait and Human-Centric Editing. Fig. 23 validates the model's precision in fine-grained portrait editing, confirming its ability to accurately execute diverse facial attribute modifications while preserving identity. Expanding beyond facial details, Fig. 24 demonstrates robust performance in structural body editing. The model successfully handles complex challenges ranging from viewpoint transformation to multi-person interaction synthesis. Collectively, these results highlight the model's significant practical utility, indicating its potential for integration into mobile photography pipelines for high-quality post-processing.

General Image Editing Capabilities. We conduct a comprehensive qualitative evaluation against SeedDream 4.0, Nano Banana, Qwen-Image-Edit [2509], and FLUX.1 Kontext [Pro]. The assessment covers a broad spectrum of editing dimensions: object manipulation (addition, removal, extraction), attribute modification, viewpoint transformation, scene text editing, and reference-guided generation. As illustrated in Fig. 25, 26, 27, 28, our model consistently outperforms Qwen-Image-Edit [2509] and FLUX.1 Kontext [Pro] across all dimensions. Notably, in certain challenging scenarios, it also achieves superior results compared to the commercial counterparts.

Figure 23: **Demonstration of fine-grained portrait editing.** From left to right: The original input image, followed by results for blemish removal, hairstyle modification, lighting adjustment, eyelash addition, face slimming, and ID photo generation. The results highlight the model's precision in manipulating specific facial attributes while preserving the subject's identity.

Figure 24: **Qualitative results on Human-centric Editing.** The figure displays pairs of input (left) and edited (right) images across three dimensions: **Pose & Interaction** (top row), involving complex interaction synthesis (*e.g.*, hugging) and large-scale body pose alteration; **Viewpoint** (bottom-left), transforming a subject from side view to front view; and **Lighting** (bottom-right), simulating directional illumination. Note the preservation of background details and subject identity despite significant structural changes.

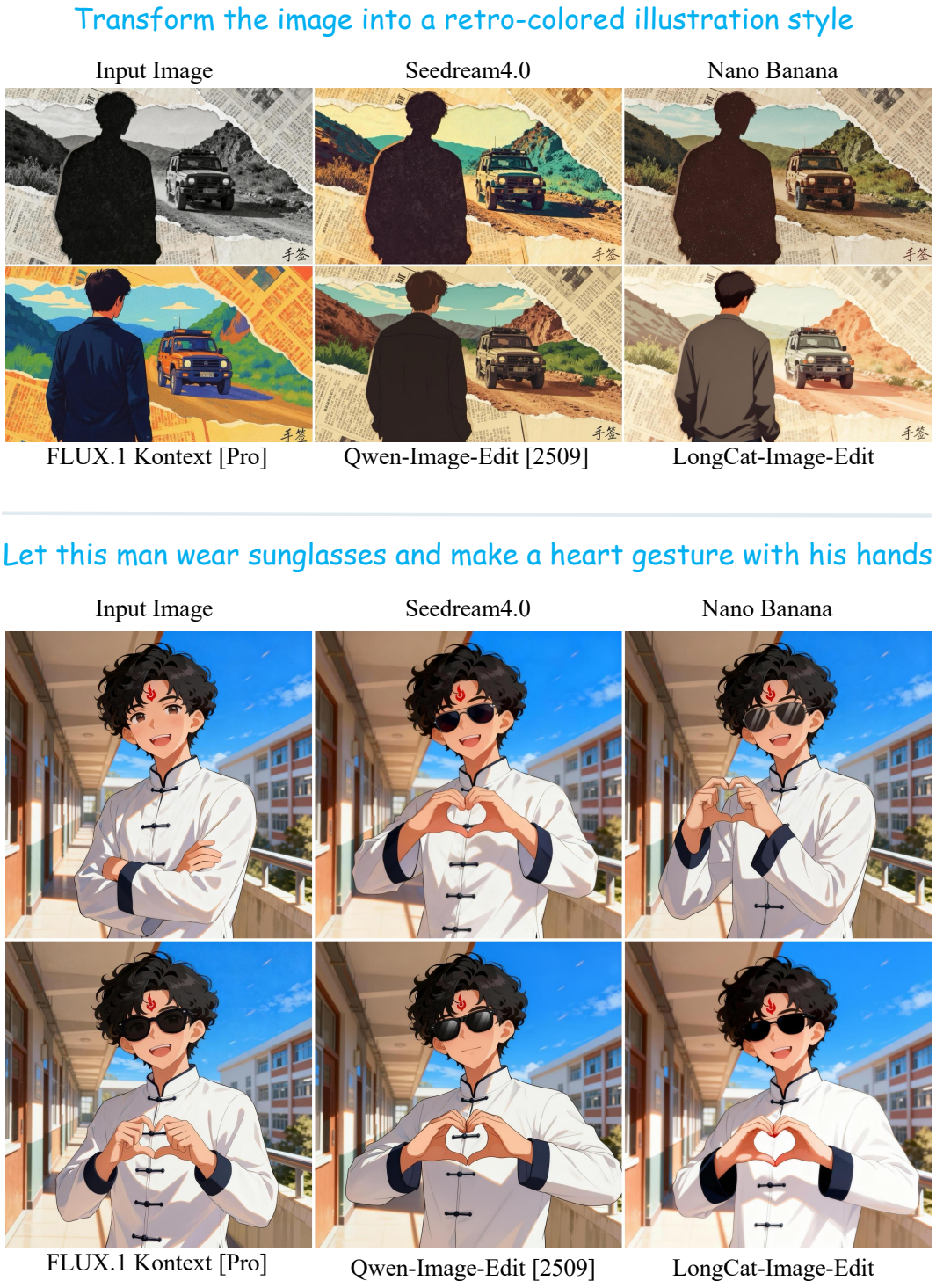

Figure 25: **Qualitative comparison on Style Transfer and Attribute Editing.** The upper panel demonstrates the transformation of a photorealistic scene into a retro-colored illustration style. The lower panel illustrates a complex instruction involving both accessory addition (sunglasses) and hand pose modification (heart gesture), highlighting our model's ability to preserve facial identity while executing significant structural changes.

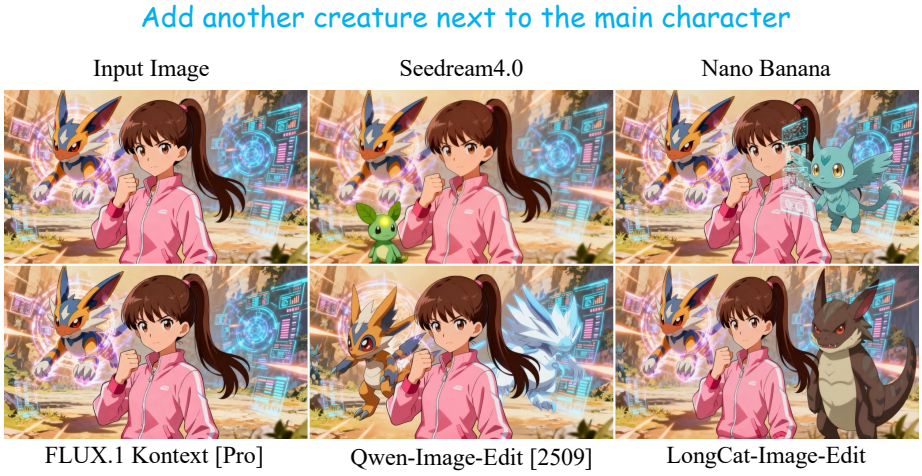

Figure 26: **Qualitative comparison on Object-centric Editing.** We evaluate the performance across three distinct scenarios: **Object Insertion** (top), where an additional creature is added while maintaining scene consistency; **Subject Extraction** (middle), isolating the foreground subject (the cat) from a complex background; **Object-Preserved Generation** (bottom), where the reference object (the device) is seamlessly integrated into a new context (held by a robot).

Figure 27: **Qualitative comparison on Scene Text Editing and Region-Controlled Editting.**

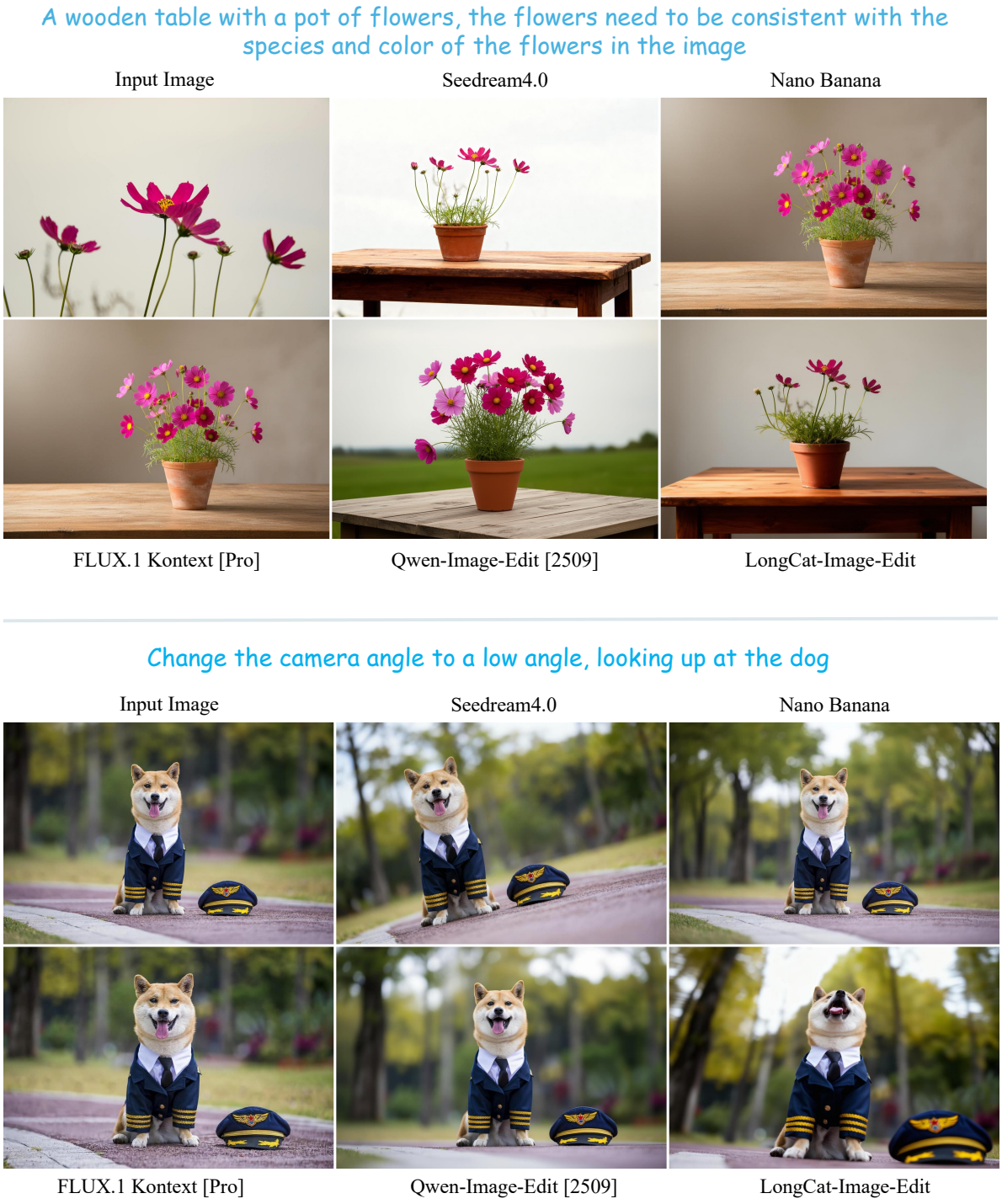

Figure 28: **Qualitative comparison on Camera Control and Viewpoint Transformation.** The upper panel shows **camera distance adjustment** (zooming out), where the view is expanded to place the flowers into a pot on a table. The lower panel displays **camera angle modification**, transitioning the view to a low-angle perspective looking up at the dog.

7 Conclusion

In this work, we present **LongCat-Image**, a 6B-parameter diffusion framework that challenges the prevailing reliance on brute-force scaling by demonstrating that exceptional performance can be achieved through efficient architectural design and refined training methodologies. By integrating a hybrid MM-DiT architecture with a unified multimodal context encoder, our model establishes an optimal equilibrium between high-fidelity generation and inference efficiency, effectively surpassing the generation quality of numerous open-source models with significantly larger parameters. Across specific domains, LongCat-Image delivers exceptional results. In text-to-image generation, our strategic data curation enables photorealism and Chinese text rendering capabilities that compete with top-tier proprietary systems. In the realm of image editing, our model sets a new benchmark for the open-source community. Supported by rigorous data filtering and a robust training paradigm, it achieves state-of-the-art performance, exhibiting both precise instruction following and superior visual consistency that significantly outperform existing alternatives. Finally, we distinguish our contribution by democratizing the entire research lifecycle. By open-sourcing not only the final model but also intermediate checkpoints and the complete training codebase, we aim to lower the barriers to entry and foster a more transparent, accessible, and collaborative ecosystem for future research.

8 Contributions and Acknowledgments

Contributors are defined as individuals who undertook primary responsibilities in data curation, model design, model training, and relative infrastructures throughout the LongCat-Image1.0 complete development cycle. Acknowledgment include those who are working part-time on tasks such as data collection, annotation, model evaluation, and technical discussions. All people are cataloged **alphabetically by first name**. Names with a dagger (†) are the project leader and sponsors, and names with an asterisk (*) are former team members.

Contributors:

Hanghang Ma*	Jie Hu†	Lishuai Gao	Xiaoqi Ma*
Haoxian Tan	Junqiang Wu	Songlin Xiao	Xunliang Cai†
Jiale Huang	Jun-Yan He	Xiaoming Wei†	Yayong Guan

Acknowledgments:

Bingcan Wang	Jia Wang	Shengxi Li	Yanbing Zeng
Cong Wei	Jiajun Liu	Tianye Dai	Yingsen Zeng*
Dengsheng Chen*	Kaiwen Wang*	Tiezhu Yue	Yuchen Tang
Fei Peng	Lingfeng Tan*	Wei Wang	Zizhe Zhao
Fengjiao Chen	Liya Ma	Xiaopeng Sun*	
Hao Lu	Man Gao	Xiaoyu Li	

References

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024a.

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

Midjourney. Midjourney, 2025. URL https://www.midjourney.com. [Text-to-image model].

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

Jia Wang, Jie Hu, Xiaoqi Ma, Hanghang Ma, Xiaoming Wei, and Enhua Wu. Image editing with diffusion models: A survey. *arXiv preprint arXiv:2504.13226*, 2025a.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025b.

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 *ICML*, 2024.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

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.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024a.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 25(70):1–53, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *ECCV*, pages 361–377. Springer, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pages 8228–8238, 2024.

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

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *CVPR*, pages 15762–15772, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 36:52132–52152, 2023.

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.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5955–5963, 2025a.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 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.

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

Stabilityai. stable-diffusion-3.5-large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7739–7751, June 2025b.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.

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

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

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

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. *NeurIPS*, 37: 131278–131315, 2024.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024b.

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 *CVPR*, pages 12966–12977, 2025b.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, pages 74–91. Springer, 2024c.

OpenAI. DALL·E 3. https://openai.com/research/dall-e-3, September 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.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *IJCV*, 2025c.

Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

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.

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*, 2025d.

Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. *arXiv preprint arXiv:2507.14119*, 2025.

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.

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.

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.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024.

Google. Gemini 2.5 flash & 2.5 flash image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf, 2025.

Google DeepMind. Gemini 2.0, 2025. URL https://gemini.google.com/.

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*, pages 18392–18402, 2023.

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*, pages 26125–26135, 2025.

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.

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.

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*, pages 13294–13304, 2025.

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.

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.

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*, 2025.