

Qwen-Image 技术报告

Qwen-Image Technical Report

Qwen 团队

 <https://qwen.ai>
 <https://huggingface.co/Qwen/Qwen-Image>
 <https://modelscope.cn/models/Qwen/Qwen-Image>
 <https://github.com/QwenLM/Qwen-Image>

摘要

We present Qwen-Image, an image generation foundation model in the Qwen series that achieves significant advances in complex text rendering and precise image editing. To address the challenges of complex text rendering, we design a comprehensive data pipeline that includes large-scale data collection, filtering, annotation, synthesis, and balancing. Moreover, we adopt a progressive training strategy that starts with non-text-to-text rendering, evolves from simple to complex textual inputs, and gradually scales up to paragraph-level descriptions. This curriculum learning approach substantially enhances the model's native text rendering capabilities. As a result, Qwen-Image not only performs exceptionally well in alphabetic languages such as English, but also achieves remarkable progress on more challenging logographic languages like Chinese.

我们提出 Qwen-Image，这是 Qwen 系列中的图像生成基础模型，在复杂文本渲染和精确图像编辑上取得了显著进展。为应对复杂文本渲染的挑战，我们设计了一套完整的数据流水线，涵盖大规模数据收集、过滤、标注、合成与平衡。此外，我们采用渐进式训练策略：从非文本渲染起步，由简单文本输入逐步过渡到复杂文本输入，最终扩展到段落级描述。这种课程学习方法大幅提升了模型原生的文本渲染能力。因此，Qwen-Image 不仅在英文等字母文字语言上表现出色，在中文这类更具挑战性的表意文字语言上也取得了显著进步。

To enhance image editing consistency, we introduce an improved multi-task training paradigm that incorporates not only traditional text-to-image (T2I) and text-image-to-image (TI2I) tasks but also image-to-image (I2I) reconstruction, effectively aligning the latent representations between Qwen2.5-VL and MMDiT. Furthermore, we separately feed the original image into Qwen2.5-VL and the VAE encoder to obtain semantic and reconstructive representations, respectively. This dual-encoding mechanism enables the editing module to strike a balance between preserving semantic consistency and maintaining visual fidelity. We present a comprehensive evaluation of Qwen-Image across multiple public benchmarks, including GenEval, DPG, and OneIG-Bench for general image generation, as well as GEdit, ImgEdit, and GSO for image editing. Qwen-Image achieves state-of-the-art performance, demonstrating its strong capabilities in both image generation and editing. Furthermore, results on LongText-Bench, ChineseWord, and CVTG-2K show that it excels in text rendering—particularly in Chinese text generation—outperforming existing state-of-the-art models by a significant margin. This highlights Qwen-Image's unique position as a leading image generation model that combines broad general capability with exceptional text rendering precision.

为提升图像编辑的一致性，我们引入了一种改进的多任务训练范式：除传统的文本生成图像（T2I）和图文生成图像（TI2I）任务外，还纳入图像到图像（I2I）重建，从而有效对齐 Qwen2.5-VL 与 MMDiT 之间的潜在表示。此外，我们将原始图像分别送入 Qwen2.5-VL 和 VAE 编码器，各自得到语义表示与重建表示。这种双重编码机制使编辑模块能在保持语义一致性与维持视觉保真度之间取得平衡。我们在多个公开基准上全面

评估了 Qwen-Image，其中通用图像生成用 GenEval、DPG 和 OneIG-Bench，图像编辑用 GEdit、ImgEdit 和 GSO。Qwen-Image 取得了最先进的性能，在图像生成与编辑两方面都展现出强大能力。此外，在 LongText-Bench、ChineseWord 和 CVTG-2K 上的结果表明，它在文本渲染、尤其是中文文本生成上表现优异，大幅领先现有最先进模型。这凸显了 Qwen-Image 作为一流图像生成模型的独特地位——兼具广泛的通用能力与卓越的文本渲染精度。

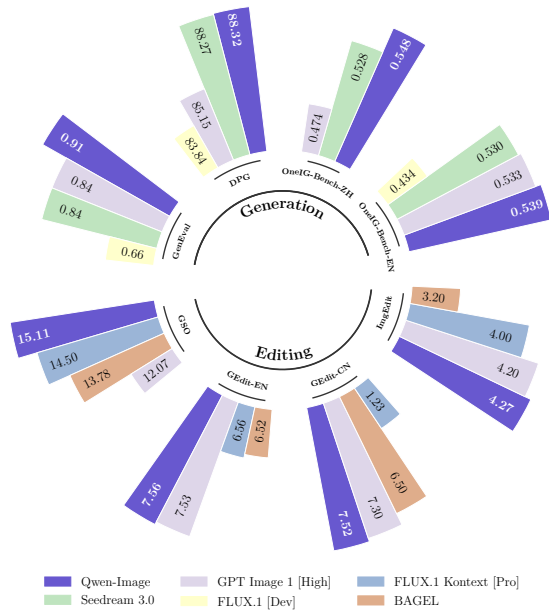


Figure 1: Qwen-Image exhibits strong general capabilities in both image generation and editing, while demonstrating exceptional capability in text rendering, especially Chinese.



图 2: Showcase of Qwen-Image in complex text rendering, including multi-line layouts, paragraph-level semantics, and fine-grained details. Qwen-Image supports both alphabetic languages (e.g., English) and logographic languages (e.g., Chinese) with high fidelity. Qwen-Image 在复杂文本渲染上的示例, 涵盖多行排版、段落级语义和细粒度细节。Qwen-Image 能高保真地支持字母文字语言 (如英文) 和表意文字语言 (如中文)。



图 3: Showcase of Qwen-Image in general image generation, supporting diverse artistic styles.
Qwen-Image 在通用图像生成上的示例，支持多样的艺术风格。



图 4: Showcase of Qwen-Image in general image editing, including style transfer, text editing, background change, object addition, removal, and replacement, pose manipulation, and more.
Qwen-Image 在通用图像编辑上的示例，包括风格迁移、文本编辑、背景更换、物体的添加、移除与替换、姿态操控等。

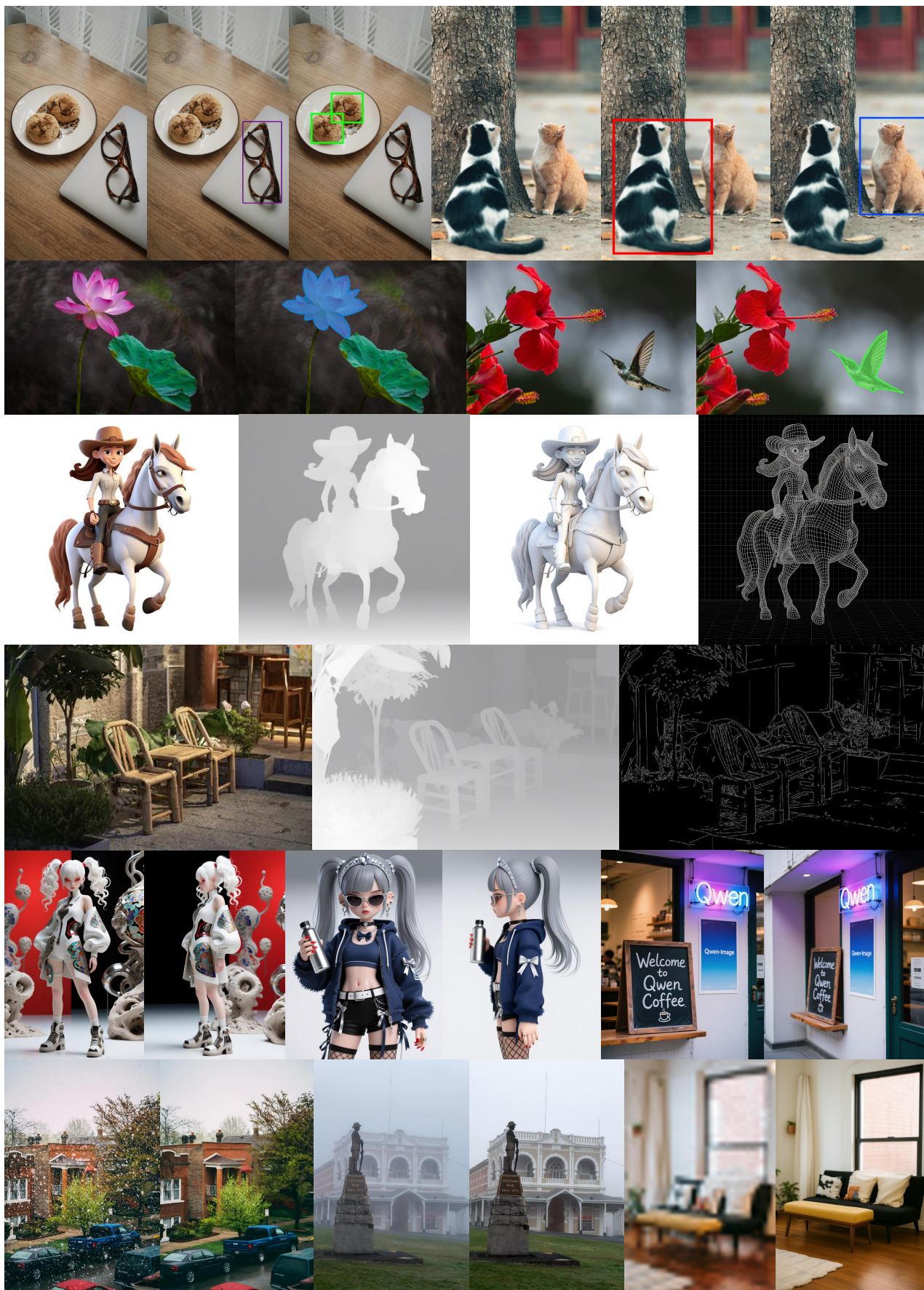


图 5: Showcase of Qwen-Image in general image understanding tasks, including detection, segmentation, depth/canny estimation, novel view synthesis, and super resolution—tasks that can all be viewed as specialized forms of image editing.

Qwen-Image 在通用图像理解任务上的示例，包括检测、分割、深度/canny 估计、新视角合成和超分辨率——这些任务都可视作图像编辑的特殊形式。

1 引言

Introduction

Image generation models—encompassing both text-to-image generation (T2I) (Rombach 等, 2021; Wu 等, 2022; Liang 等, 2022; OpenAI, 2023; Podell 等, 2023; Chen 等, 2024c; Li 等, 2024b; Esser 等, 2024; BlackForest, 2024; Gao 等, 2025; Gong 等, 2025; Cai 等, 2025) and image editing (TI2I) (Brooks 等, 2023; Zhang 等, 2023; Wang 等, 2025; Deng 等, 2025; Labs 等, 2025; Wu 等, 2025b; Liu 等, 2025b; Cai 等, 2025; OpenAI, 2025)—have emerged as a fundamental component of modern artificial intelligence, enabling machines to synthesize or modify visually compelling and semantically coherent content from text prompts. Over the past few years, remarkable progress has been achieved in this domain, particularly with the advent of diffusion-based architectures (Ho 等, 2020; Liu 等, 2022) that enable high-resolution image generation while capturing fine-grained semantic details.

图像生成模型——包括文本生成图像 (T2I) (Rombach 等, 2021; Wu 等, 2022; Liang 等, 2022; OpenAI, 2023; Podell 等, 2023; Chen 等, 2024c; Li 等, 2024b; Esser 等, 2024; BlackForest, 2024; Gao 等, 2025; Gong 等, 2025; Cai 等, 2025) 和图像编辑 (TI2I) (Brooks 等, 2023; Zhang 等, 2023; Wang 等, 2025; Deng 等, 2025; Labs 等, 2025; Wu 等, 2025b; Liu 等, 2025b; Cai 等, 2025; OpenAI, 2025)——已成为现代人工智能的基础组件, 使机器能够从文本提示词合成或编辑视觉效果出众且语义连贯的内容。近年来, 该领域取得了显著进展, 特别是基于扩散的架构 (Ho 等, 2020; Liu 等, 2022) 的出现, 使高分辨率图像生成成为可能, 同时捕捉细粒度语义细节。

Despite these advances, two critical challenges persist. First, for text-to-image generation, aligning model outputs with complex, multifaceted prompts remains a significant hurdle. Our evaluation reveals that even state-of-the-art commercial models such as GPT Image 1 (OpenAI, 2025) and Seedream 3.0 (Gao 等, 2025) struggle when faced with tasks requiring multi-line text rendering, non-alphabetic languages rendering (e.g., Chinese), localized text insertions, or seamless integration of text and visual elements. Second, for image editing, achieving precise alignment between the edited output and the original image poses dual challenges: (i) visual consistency, where only targeted regions should be modified while preserving all other visual details (e.g., changing hair color without altering facial details) and (ii) semantic coherence, where global semantics must be preserved during structural changes (e.g., modifying a person’s pose while maintaining identity and scene coherence).

尽管取得了这些进展, 两个关键挑战仍然存在。首先, 对于文本生成图像 (T2I) 任务, 将模型输出与复杂多样的提示词对齐仍然是重大难题。我们的评估表明, 即使是最先进的商业模型如 GPT Image 1 (OpenAI, 2025) 和 Seedream 3.0 (Gao 等, 2025), 在处理多行文本渲染、非字母文字渲染 (如中文)、局部文本插入或文本与视觉元素无缝融合等任务时也难以胜任。其次, 对于图像编辑 (TI2I) 任务, 实现编辑输出与原始图像间的精确对齐面临双重挑战: (i) 视觉一致性——仅修改目标区域, 同时保留所有其他视觉细节 (如改变头发颜色而不改变脸部细节); (ii) 语义连贯性——在结构变化期间保留全局语义 (如改变人物姿态同时保持身份和场景连贯性)。

In this work, we introduce Qwen-Image, a novel image generation model within the Qwen series, designed to overcome these challenges through comprehensive data engineering, progressive learning strategies, enhanced multi-task training paradigms, and scalable infrastructure optimization.

本工作提出 Qwen-Image, 这是 Qwen 系列中的新颖图像生成模型, 通过全面的数据工程、渐进学习策略、增强的多任务训练范式和可扩展基础设施优化来克服这些挑战。

To address the challenge of complex prompt alignment, we develop a robust data pipeline encompassing large-scale collection, annotation, filtering, synthetic augmentation, and class balancing. We further adopt a curriculum learning strategy, starting from basic text rendering tasks and progressively advancing to paragraph-level and layout-sensitive descriptions. This approach significantly enhances the model’s ability to follow diverse languages, especially logographic languages like Chinese.

为解决复杂提示词对齐的挑战, 我们建立了一条鲁棒的数据流水线, 涵盖大规模收集、标注、过滤、合成增强和类别平衡。进一步, 我们采用课程学习策略, 从基础文本渲染任务开始, 逐步进阶到段落级和布局敏感的描述。这种方法显著增强了模型对多语言的理解和生成能力, 特别是对表意文字语言 (如

中文) 的处理。

To address the challenge of image alignment, we propose an enhanced multi-task learning framework that seamlessly integrates T2I, I2I and TI2I objectives within a shared latent space. Specifically, the input image is encoded into two distinct yet complementary feature representations: semantic features are extracted via Qwen-VL (Bai 等, 2025), capturing high-level scene understanding and contextual meaning, while reconstructive features are obtained through the VAE encoder, preserving low-level visual details. Both sets of features are then jointly fed into the MMDiT architecture (Esser 等, 2024) as conditioning signals. This dual-conditioning design enables the model to simultaneously maintain semantic coherence and visual consistency.

为解决图像对齐的挑战，我们提出了一个增强的多任务学习框架，在共享潜在空间中无缝集成 T2I、I2I 和 TI2I 目标。具体而言，输入图像被编码为两个不同但互补的特征表示：语义特征通过 Qwen-VL (Bai 等, 2025) 提取，捕捉高级场景理解和上下文含义；重建特征通过 VAE 编码器获得，保留低级视觉细节。两组特征随后共同送入 MMDiT 架构 (Esser 等, 2024) 作为条件信号。这种双重条件设计使模型能够同时保持语义连贯性和视觉一致性。

To ensure training efficiency and stability at scale, we design a Producer-Consumer framework leveraging TensorPipe for distributed data loading and preprocessing. The Producer handles preprocessing tasks such as VAE encoding and data I/O, while the Consumer focuses on distributed model training using the Megatron (Shoeybi 等, 2019) framework. We also implement extensive monitoring tools to ensure reliable convergence and debugging capabilities throughout large-scale training.

为保证大规模训练的效率和稳定性，我们设计了生产者-消费者框架，利用 TensorPipe 进行分布式数据加载和预处理。生产者负责 VAE 编码和数据 I/O 等预处理任务，消费者则使用 Megatron (Shoeybi 等, 2019) 框架进行分布式模型训练。我们还实现了广泛的监控工具，确保大规模训练过程中的可靠收敛和调试能力。

Qwen-Image demonstrates significant advances in both generating high-quality images from complex textual prompts and performing accurate, context-aware image editing. The model is capable of interpreting intricate linguistic structures and producing visually compelling outputs that align with both semantic intent and visual constraints. To validate its effectiveness, we evaluate Qwen-Image across a diverse set of tasks, including text-to-image generation and image editing.

Qwen-Image 在从复杂文本提示词生成高质量图像和执行准确的、上下文感知的图像编辑两方面均取得了显著进展。该模型能够解释复杂的语言结构，生成视觉上引人注目的输出，同时符合语义意图和视觉约束。为验证其有效性，我们在多项任务上评估了 Qwen-Image，包括文本生成图像和图像编辑。

The key contributions of Qwen-Image can be summarized as follows:

- **Superior Text Rendering:** Qwen-Image excels at complex text rendering, including multi-line layouts, paragraph-level semantics, and fine-grained details. It supports both alphabetic languages (e.g., English) and logographic languages (e.g., Chinese) with high fidelity.
- **Consistent Image Editing:** Through our enhanced multi-task training paradigm, Qwen-Image achieves exceptional performance in preserving both semantic meaning and visual realism during editing operations.
- **Strong Cross-Benchmark Performance:** Evaluated on multiple benchmarks, Qwen-Image consistently outperforms existing models across diverse generation and editing tasks, establishing a strong foundation model for image generation.

Qwen-Image 的主要贡献可总结如下：

- **卓越的文本渲染能力：**Qwen-Image 擅长复杂文本渲染，包括多行布局、段落级语义和细粒度细节。它支持字母文字语言（如英文）和表意文字语言（如中文）的高保真渲染。
- **一致的图像编辑性能：**通过我们增强的多任务训练范式，Qwen-Image 在编辑操作中卓越地保持语义含义和视觉真实性。

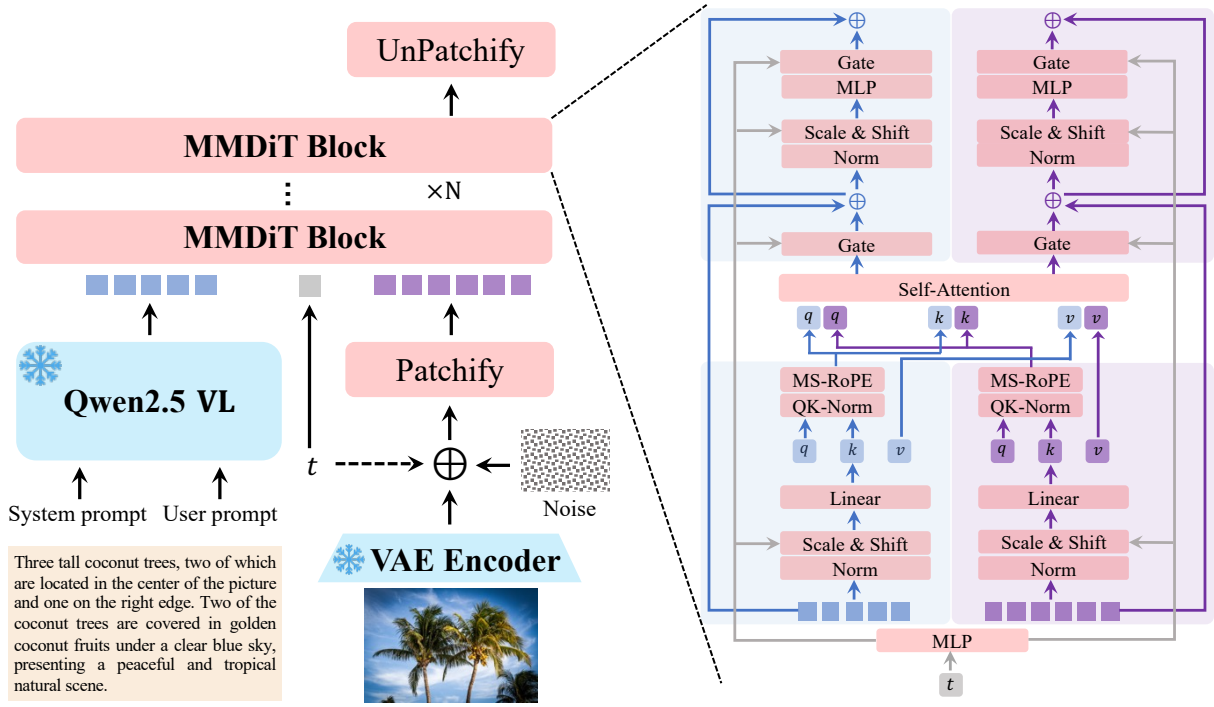


图 6: Overview of the Qwen-Image architecture. It adopts a standard double-stream MMDiT architecture. The input representations are provided by a frozen Qwen2.5-VL and a VAE encoder. The model employs RMSNorm (Zhang & Sennrich, 2019) for QK-Norm, while all other normalization layers use LayerNorm. Additionally, we design a new positional encoding scheme, MSRoPE (Multimodal Scalable RoPE), to jointly encode positional information for both image and text modalities.

Qwen-Image 架构概览。采用标准双流 MMDiT 架构。输入表示来自冻结的 Qwen2.5-VL 和 VAE 编码器。QK-Norm 采用 RMSNorm (Zhang & Sennrich, 2019)，其他规范化层采用 LayerNorm。新设计的位置编码方案 MSRoPE（多模态可扩展 RoPE）联合编码图像和文本模态的位置信息。

- **强大的跨基准性能：**在多个基准上评估，Qwen-Image 在生成和编辑任务上持续超越现有模型，建立了图像生成领域的强大基础模型。

2 模型

Model

本节介绍 Qwen-Image 模型的架构设计、训练数据和训练细节。

2.1 模型架构

Model Architecture

As shown in Figure 6, the Qwen-Image architecture is built upon three core components that work in concert to enable high-fidelity text-to-image generation. First, a **Multimodal Large Language Model (MLLM)** serves as the condition encoder, responsible extracting feature from textual inputs. Second, a **Variational AutoEncoder (VAE)** acts as the image tokenizer, compressing input images into compact latent representations and decoding them back during inference. Third, a **Multimodal Diffusion Transformer (MMDiT)** functions as the backbone diffusion model, modeling the complex joint distribution between noise and image latents under text guidance. While this section outlines their general roles, the specific model choices and architectural details are elaborated in the following sections.

如图 6 所示，Qwen-Image 架构基于三个核心组件，协同工作实现高保真的文本到图像生成。首先，**多模态大语言模型 (MLLM)** 充当条件编码器，负责从文本输入提取特征。其次，**变分自编码器 (VAE)** 充当图像 tokenizer，将输入图像压缩为紧凑的潜在表示，在推理时解码恢复。第三，**多模态扩散**

System Prompt for T2I task

```
<|im_start|>system
Describe the image by detailing the color, quantity, text, shape, size, texture, spatial relationships of
the objects and background: <|im_end|>
<|im_start|>user
<|user_text|><|im_end|>
<|im_start|>assistant
```

图 7: System prompt for Text-to-Image generation task, where `<|user_text|>` is the user input prompt. 文本到图像生成任务的系统提示词，其中 `<|user_text|>` 是用户输入提示词。

Transformer (MMDiT) 充当主干扩散模型，在文本指导下建模噪声与图像潜在表示的复杂联合分布。本节阐述它们的总体角色，具体的模型选择和架构细节见后续章节。

2.2 多模态大语言模型

Multimodal Large Language Model

Qwen-Image 采用 Qwen2.5-VL 模型 (Bai 等, 2025) 作为文本输入的特征提取模块，有三个主要原因：(1) Qwen2.5-VL 的语言和视觉空间已经对齐，相比 Qwen3 (Yang 等, 2025) 等纯语言模型，更适合文本到图像任务；(2) Qwen2.5-VL 保留了强大的语言建模能力，与纯语言模型相比没有显著下降；(3) Qwen2.5-VL 支持多模态输入，使 Qwen-Image 能够解锁更广泛的功能，如图像编辑 (Labs 等, 2025)。设 x 和 y 分别表示图像和文本输入。给定用户输入（如提示词和图像），我们采用 Qwen2.5-VL 模型提取特征。为了更好地指导模型生成精细化的潜在表示，同时适应不同任务的多种输入模态，我们分别为**纯文本输入**和**文本和图像输入**设计了不同的系统提示词。系统提示模板分别如图 7 和图 15 所示。最后，我们采用 Qwen2.5-VL 语言模型主干最后一层隐状态的潜在表示作为用户输入的代表。

2.3 变分自编码器

Variational AutoEncoder

强大的 VAE 表示是构建图像基础模型的关键。现有图像基础模型通常在大规模图像数据集上用 2D 卷积训练 VAE，得到高质量的图像表示。我们的工作目标是开发兼容图像和视频的更通用视觉表示。然而，现有联合图像视频 VAE（如 Wan-2.1-VAE (Wan 等, 2025)）常面临性能权衡，导致图像重建能力下降。为此，我们采用单编码器双解码器架构。该设计由一个兼容图像和视频的共享编码器和为各模态配置的专用解码器组成，使图像基础模型能为未来视频模型奠定基础。具体地，我们采用 Wan-2.1-VAE 的架构，冻结编码器，仅对图像解码器进行微调。

To enhance reconstruction fidelity, particularly for small texts and fine-grained details, we train the decoder on an in-house corpus of text-rich images. The dataset consists of real-world documents (PDFs, PowerPoint slides, posters) alongside synthetic paragraphs, covering both alphabetic (e.g., English) and logographic (e.g., Chinese) languages. During training, we observe that: (1) Balancing reconstruction loss with perceptual loss effectively reduces grid artifacts, which are often observed in repetitive textures like bushes. (2) As reconstruction quality increases, adversarial loss becomes ineffective because the discriminator is unable to provide effective guidance. Based on these observations, we use only reconstruction and perceptual losses, dynamically adjusting their ratio during fine-tuning. Interestingly, we find that finetuning only the decoder can effectively enhance details and improve the rendering of small text, thereby building a solid foundation for Qwen-Image’s text rendering abilities. Quantitative and qualitative results are presented in Section 5.2.1.

为了增强重建保真度，尤其是对小文字和细粒度细节的处理，我们在内部富文本图像语料库上训练解码器。该数据集包含真实文档（PDF、PowerPoint 幻灯片、海报）和合成段落，覆盖字母文字（如英文）和表意文字（如中文）。训练中我们发现：(1) 平衡重建损失和感知损失能有效减少网格伪影（常见于灌木丛等重复纹理）；(2) 随重建质量提高，对抗损失失效，因判别器无法提供有效指导。基于这些观

表 1: Configuration of Qwen-Image architecture.
Qwen-Image 架构配置。

配置	VLM		VAE		MMDiT
	ViT	LLM	Enc	Dec	
层数	32	28	11	15	60
头数 (Q / KV)	16 / 16	28 / 4	-	-	24 / 24
头维度	80	128	-	-	128
中间维度	3,456	18,944	-	-	12,288
Patch / 缩放因子	14	-	8x8	8x8	2
通道维度	-	-	16	16	-
参数量	7B		54M	73M	20B

察, 我们仅使用重建损失和感知损失, 在微调中动态调整其比例。值得注意的是, 仅微调解码器就能增强细节、改进小文字渲染, 提升 Qwen-Image 的文本渲染能力。量化和定性结果见第 5.2.1 节。

2.4 多模态扩散 Transformer

Multimodal Diffusion Transformer

Qwen-Image 采用多模态扩散 Transformer (MMDiT) (Esser 等, 2024) 联合建模文本和图像。该方法已被证明在 FLUX (BlackForest, 2024; Labs 等, 2025) 系列和 Seedream (Gong 等, 2025; Gao 等, 2025) 系列等众多工作中有效。

Within each block, we introduce a novel positional encoding method: Multimodal Scalable RoPE (MSRoPE). As illustrated in Figure 8, we compare various text-image joint positional encoding strategies. In the traditional MMDiT block, text tokens are directly concatenated after the flattened image positional embeddings. Furthermore, Seedream 3.0 (Gao 等, 2025) introduces Scaling RoPE, where the image positional encoding is shifted to the central region of the image, and text tokens are considered as 2D tokens with a shape of $[1, L]$. Then, 2D RoPE (Heo 等, 2024) is used for image-text joint positional encoding. Although this adjustment facilitates resolution scale training, certain rows of positional encodings for text and image, e.g., the 0-th middle row in Figure 8 (B), become isomorphic, making it harder for the model to distinguish between text tokens and the image latent tokens in the 0-th middle row. Yet, it is also non-trivial to determine an appropriate image row to concatenate the text tokens. To address the aforementioned challenges, we introduce Multimodal Scalable RoPE (MSRoPE). In this approach, text inputs are treated as 2D tensors with identical position IDs applied across both dimensions. As depicted in Figure 8 (C), the text is conceptualized as being concatenated along the diagonal of the image. This design allows MSRoPE to leverage resolution scaling advantages on the image side while maintaining functional equivalence to 1D-RoPE on the text side, thereby obviating the need to determine the optimal positional encoding for text. We show the architecture and configuration of Qwen-Image in Table 1.

在每个块中, 我们提出了一种新颖的位置编码方法: 多模态可扩展 RoPE (MSRoPE)。如图 8 所示, 我们对比了各种文本-图像联合位置编码策略。传统 MMDiT 块中, 文本 token 直接连接在展平的图像位置嵌入之后。Seedream 3.0 (Gao 等, 2025) 引入了缩放 RoPE, 将图像位置编码移至图像中心区域, 文本 token 视为形状为 $[1, L]$ 的 2D token。随后用 2D RoPE (Heo 等, 2024) 进行图像-文本联合位置编码。这种调整虽然便于分辨率缩放训练, 但文本和图像的位置编码某些行 (如图 8 (B) 的第 0 中间行) 变成同构, 使模型难以区分文本 token 与该行的图像潜在 token。然而, 确定适当的图像行来连接文本 token 也非易事。为了解决这些问题, 我们提出了多模态可扩展 RoPE (MSRoPE)。该方法将文本输入视为 2D 张量, 两个维度采用相同的位置 ID。如图 8 (C) 所示, 文本沿图像对角线连接。MSRoPE 既能利用图像端的分辨率缩放优势, 又能在文本端保持 1D-RoPE 的功能等价性, 无需单独确定文本的最优位置编码。Qwen-Image 的架构和配置见表 1。

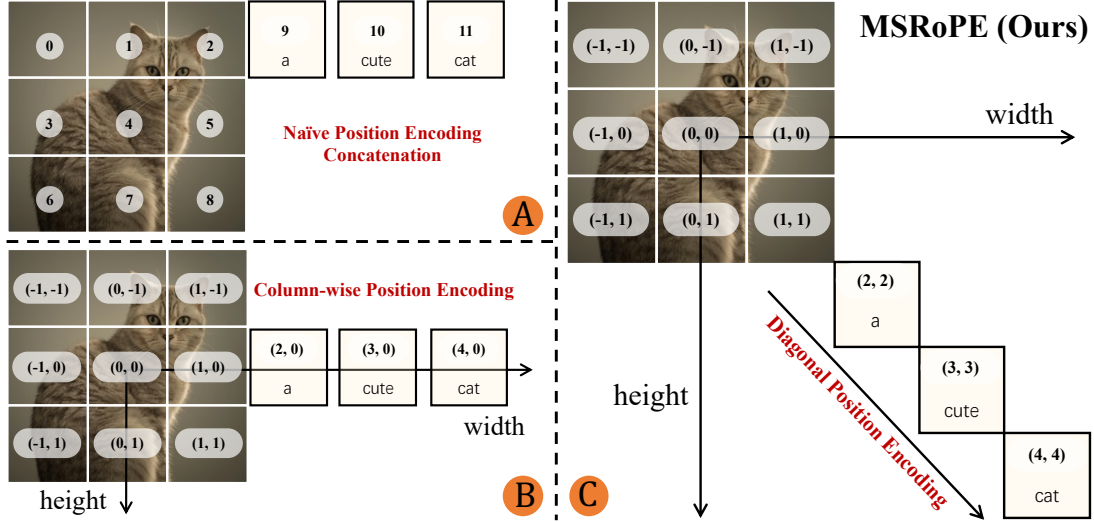


图 8: Comparison of different image-text joint positional encoding strategies. We design the Multimodal Scalable RoPE (MSRoPE) strategy, which starts encoding from the image center and positions textual encodings along the diagonal of the grid, enabling better image resolution scaling and improved text-image alignment.

不同图像-文本联合位置编码策略的比较。我们提出的多模态可扩展 RoPE（MSRoPE）策略从图像中心开始编码，将文本编码沿网格对角线定位，实现更好的图像分辨率缩放和改进的文本-图像对齐。

3 数据

Data

3.1 数据收集

Data Collection

We systematically collected and annotated billions of image-text pairs to support the training of our image generation model. Rather than focusing solely on the scale of raw dataset, we prioritize data quality and balanced data distribution, aiming to construct a well-balanced and representative dataset that closely mirrors real-world scenarios. As illustrated in Figure 9, the dataset is organized into four primary domains: Nature, Design, People and Synthetic Data.

我们系统地收集和标注了数十亿个图文对，支持图像生成模型的训练。我们不仅关注数据集规模，更优先考虑数据质量和均衡的数据分布，旨在构建一个均衡且代表性的数据集，紧密反映真实场景。如图 9 所示，数据集分为四个主要领域：Nature、Design、People 和 Synthetic Data。

Among these categories, Nature category constitutes the largest proportion, accounting for approximately 55% of the dataset. This category includes diverse subcategories such as Objects, Landscape, Cityscape, Plants, Animals, Indoor, and Food categories. Additionally, content that does not clearly belong to the People or Design categories is also classified under Nature category. This broad, general-purpose category serves as a crucial foundation for improving the model’s ability to generate realistic and diverse natural images.

其中，Nature 类别占比最大，约占数据集的 55%。该类别包括 Objects、Landscape、Cityscape、Plants、Animals、Indoor 和 Food 等多个子类别。此外，不明确属于 People 或 Design 类别的内容也归为 Nature 类别。这一宽泛的通用类别为提升模型生成真实和多样化自然图像的能力奠定了关键基础。

The second largest category is the Design category, comprising around 27% of the dataset. It primarily includes structured visual content such as Posters, User Interfaces, and Presentation Slides, as well as various forms of art including paintings, sculptures, art crafts, and digital arts. These types of data often contain rich textual elements, complex layouts, and design-specific visual semantics. This category is particularly important for enhancing the model’s capabilities in following intricate prompts about artistic styles, text rendering, and

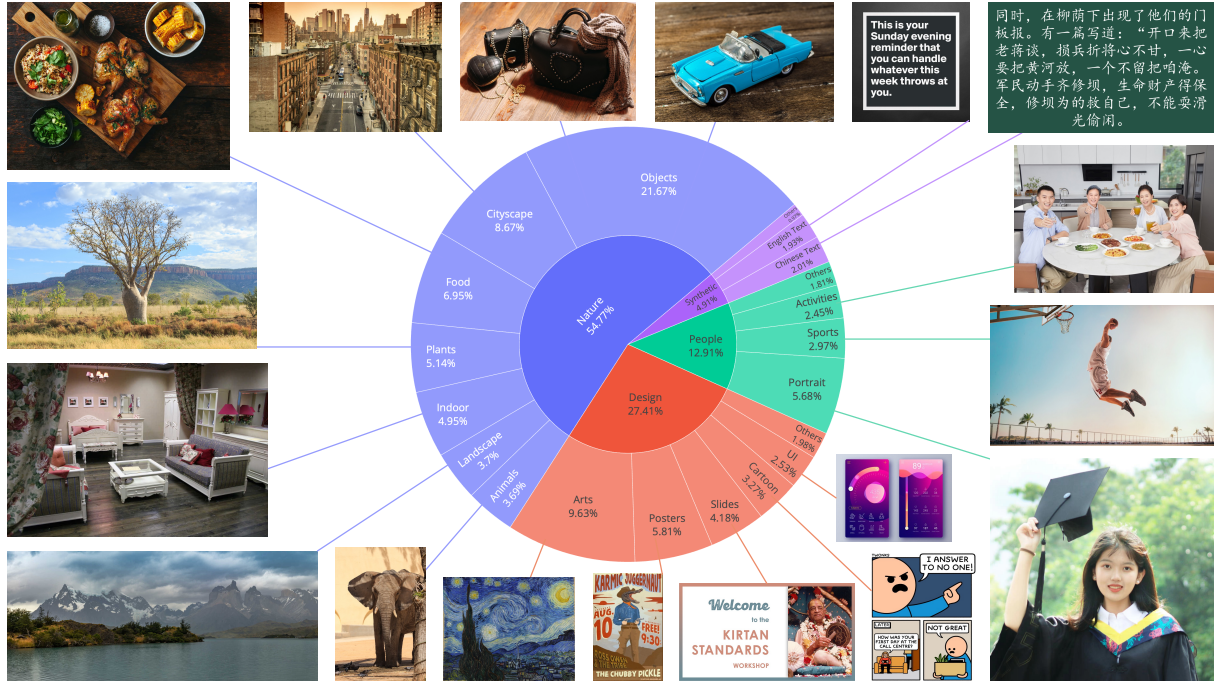


图 9: Overview of Data Collection. The dataset includes four main categories: Nature (general-purpose generation), People (human-centric generation), Design (artistic styles, text rendering, and complex layouts), and Synthetic Data (text rendering enhancement). Our collection strategy balances diversity and quality while training, ensuring broad coverage and high-fidelity annotations to support robust model training.

数据收集概览。数据集包含四个主要类别：Nature（通用生成）、People（以人为中心的生成）、Design（艺术风格、文本渲染和复杂布局）和 Synthetic Data（文本渲染增强）。收集策略在训练过程中平衡多样性和质量，确保覆盖范围广泛和标注高保真，为稳健的模型训练提供支撑。

layout design.

第二大类别是 Design 类别，约占数据集的 27%。主要包括 Posters、User Interfaces 和 Presentation Slides 等结构化视觉内容，以及绘画、雕塑、工艺品和数字艺术等多种艺术形式。这些数据通常包含丰富的文本元素、复杂的布局和设计特有的视觉语义。该类别对增强模型在艺术风格、文本渲染和布局设计等复杂提示词上的能力至关重要。

Next, the People category makes up about 13% of the dataset, encompassing subcategories such as Portrait, Sports, and Human Activities. It comprises a wide range of human-related images, including Portrait, Sports, Activities, and so on. This category is essential for improving the model’s ability to generate realistic and diverse human images, to ensure satisfactory user experiences and practical applicability.

其次，People 类别约占数据集的 13%，包括 Portrait、Sports 和 Human Activities 等子类别。包含了大量人类相关的图像，如 Portrait、Sports、Activities 等。该类别对提升模型生成真实且多样化人物图像的能力至关重要，确保用户体验满意度和实际应用价值。

Finally, the Synthetic Data category accounts for approximately 5% of the dataset. It is important to clarify that the synthetic data discussed here does not include images generated by other AI models, but rather data synthesized through controlled text rendering techniques (described in §§ 3.4). This excludes images synthesized by other AI models, which often introduce significant risks such as visual artifacts, text distortions, biases, and hallucinations. We adopt a conservative stance toward such data, as training on low-fidelity or misleading images may weaken the model’s generalization capabilities and undermine its reliability.

最后，Synthetic Data 类别约占数据集的 5%。需要澄清的是，此处讨论的合成数据不包括由其他 AI 模型生成的图像，而是通过受控的文本渲染技术合成的数据（详见 §§ 3.4）。这排除了由其他 AI 模型合成的图像，它们往往引入视觉伪影、文本失真、偏差和幻觉等重大风险。我们对这类数据采取谨慎态度，因为在低保真或误导性图像上训练会削弱模型的泛化能力和可靠性。

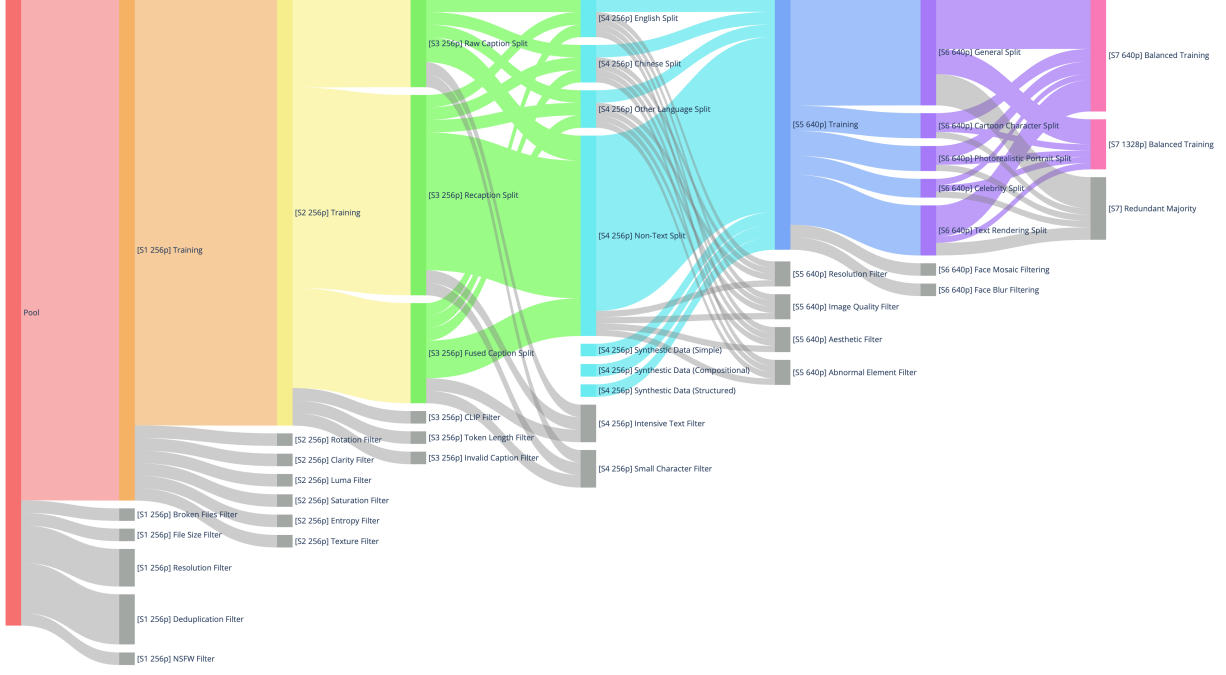


图 10: Overview of the multi-stage data filtering pipeline. Our filtering strategy consists of seven sequential stages (S1–S7), each targeting specific aspects of data quality, alignment, and diversity. From initial pre-training data curation to high-resolution refinement and multi-scale training, the pipeline progressively improves dataset quality while maintaining semantic richness and distributional stability.

多阶段数据过滤流水线概览。过滤策略包括七个顺序阶段（S1-S7），每个阶段针对数据质量、对齐和多样性的特定方面。从初始预训练数据整理到高分辨率精细化和多尺度训练，流水线逐步提升数据集质量，同时保持语义丰富性和分布稳定性。

3.2 数据过滤

Data Filtering

To curate high-quality training data throughout the iterative development of our image generation model, we propose a multi-stage filtering pipeline comprising seven sequential stages, as depicted in Figure 10. These stages are progressively applied throughout the training process, with data distributions continuously refined over time. Notably, synthetic data are introduced from Stage 4, when the foundational model has reached a certain level of stability. The following section presents a comprehensive description of each stage.

为在图像生成模型的迭代开发全程中整理高质量训练数据，我们提出了包含七个顺序阶段的多阶段过滤流水线，如 Figure 10 所示。这些阶段在训练过程中逐步应用，数据分布随时间不断精细化。值得注意的是，从阶段 4 开始引入合成数据，此时基础模型已达到一定的稳定程度。以下章节详细阐述每个阶段。

阶段 1：初始预训练数据整理 (Stage 1: Initial Pre-training Data Curation) At this early stage, the model is trained on images resized to 256p (256×256 pixels with various aspect ratios including 1:1, 2:3, 3:2, 3:4, 4:3, 9:16, 16:9, 1:3 and 3:1). To improve data quality, a series of filters are applied to remove low-quality or irrelevant images. The Broken Files Filter identifies and discards corrupted or partially damaged files (e.g., truncated images), many of which are also associated with abnormally small file sizes, as detected by the File Size Filter. The Resolution Filter removes images with original resolutions below 256p. The Deduplication Filter eliminates duplicate or near-duplicate image-text pairs. Furthermore, the NSFW Filter is applied to exclude content containing sexual, violent, or other offensive material.

在这一早期阶段，模型在调整大小至 256p 的图像上训练（256×256 像素，宽高比包括 1:1、2:3、3:2、3:4、4:3、9:16、16:9、1:3 和 3:1）。为提升数据质量，应用一系列过滤器移除低质量或无关图像。Broken Files Filter 识别并丢弃损坏或部分损坏的文件（如截断图像），其中许多文件还存在异常小的文件大小，

由 File Size Filter 检测。Resolution Filter 移除原始分辨率低于 256p 的图像。Deduplication Filter 消除重复或近似重复的图文对。此外，应用 NSFW Filter 以排除包含色情、暴力或其他冒犯性内容的图像。

阶段 2：图像质量增强 (Stage 2: Image Quality Enhancement) In this stage, we focus on systematically improving the image quality of the dataset. The Rotation Filter removes images with significant rotation or flipping, as indicated by the EXIF metadata. The Clarity Filter discards blurry or out-of-focus images, ensuring that only sharp and clear images are retained. The Luma Filter excludes images that are excessively bright or dark, while the Saturation Filter eliminates images with unnaturally high color saturation, which often suggests artificial rendering or unrealistic digital manipulations. Furthermore, the Entropy Filter identifies and removes images with significantly low entropy, typically characterized by large uniform regions or minimal visual content. Finally, the Texture Filter is employed to discard images with overly complex textures, which are often associated with noise or non-semantic patterns. Figure 11 illustrates some examples of the filtering operators used in this stage.

该阶段重点是系统地改进数据集的图像质量。Rotation Filter 移除具有明显旋转或翻转的图像，由 EXIF 元数据指示。Clarity Filter 丢弃模糊或失焦的图像，确保仅保留清晰锐利的图像。Luma Filter 排除过度明亮或过暗的图像，而 Saturation Filter 消除颜色饱和度不自然过高的图像，这通常表明人工渲染或不真实的数字处理。此外，Entropy Filter 识别并移除熵值明显过低的图像，通常表现为大范围均匀区域或最少视觉内容。最后，应用 Texture Filter 丢弃纹理过于复杂的图像，这些图像通常与噪声或非语义模式相关。Figure 11 展示了该阶段使用的某些过滤算子的示例。

阶段 3：图文对齐改进 (Stage 3: Image-Text Alignment Improvement) This stage focuses on improving the alignment between textual descriptions and visual content. To balance the training data distribution, the dataset is divided into three splits based on the source of captions: Raw Caption Split, Recaption Split, and Fused Caption Split. Raw Caption Split includes captions provided by websites as well as metadata such as titles or tags originally associated with the images. Although raw captions may introduce noise, they contribute to the model’s robustness in handling short text inputs and serve as a vital source of real-world knowledge (e.g., plant names, cartoon IPs) often absent in datasets with synthesized captions. Recaption Split consists of captions generated by the most advanced Qwen-VL Captioner (Bai 等, 2025), which provides more descriptive and structured annotations. Due to model limitations, not all IPs can be accurately identified. Fused Caption Split combines both raw captions and synthesized captions, offering a blend of general knowledge and detailed descriptions. To further improve alignment, we applied both the Chinese CLIP (Yang 等, 2022) Filter and the SigLIP 2 (Tschannen 等, 2025) Filter to remove mismatched image-text pairs from the Raw Caption Split. Additionally, a Token Length Filter was employed to eliminate excessively long captions, and the Invalid Caption Filter discards captions with abnormal content, such as “Sorry, I cannot provide a caption for this image.”.

该阶段重点是改进文本描述与视觉内容的对齐。为平衡训练数据分布，根据描述文本来源将数据集分为三个 split：Raw Caption Split、Recaption Split 和 Fused Caption Split。Raw Caption Split 包括网站提供的描述文本以及原始与图像关联的元数据（如标题或标签）。虽然原始描述文本可能引入噪声，但它有助于提升模型处理短文本输入的鲁棒性，并提供现实世界知识（如植物名称、动画 IP）的重要来源，这类信息在合成描述文本的数据集中往往缺失。Recaption Split 由最先进的 Qwen-VL Captioner (Bai 等, 2025) 生成的描述文本组成，提供更具描述性和结构化的标注。由于模型限制，并非所有 IP 都能准确识别。Fused Caption Split 结合原始描述文本和合成描述文本，提供通用知识和详细描述混合。为进一步改进对齐，我们应用了 Chinese CLIP (Yang 等, 2022) Filter 和 SigLIP 2 (Tschannen 等, 2025) Filter 移除 Raw Caption Split 中不匹配的图文对。此外，采用 Token Length Filter 消除过长的描述文本，Invalid Caption Filter 丢弃包含异常内容的描述文本，如“Sorry, I cannot provide a caption for this image.”。

阶段 4：文本渲染增强 (Stage 4: Text Rendering Enhancement) In this stage, we focus on improving the model’s capability in rendering text within images, which is crucial for generating images with high textual fidelity. To this end, we categorize the dataset based on the presence and language of text within images. Specifically, the dataset from Stage 3 is divided into four splits: English Split, Chinese Split, Other Language Split, and Non-Text

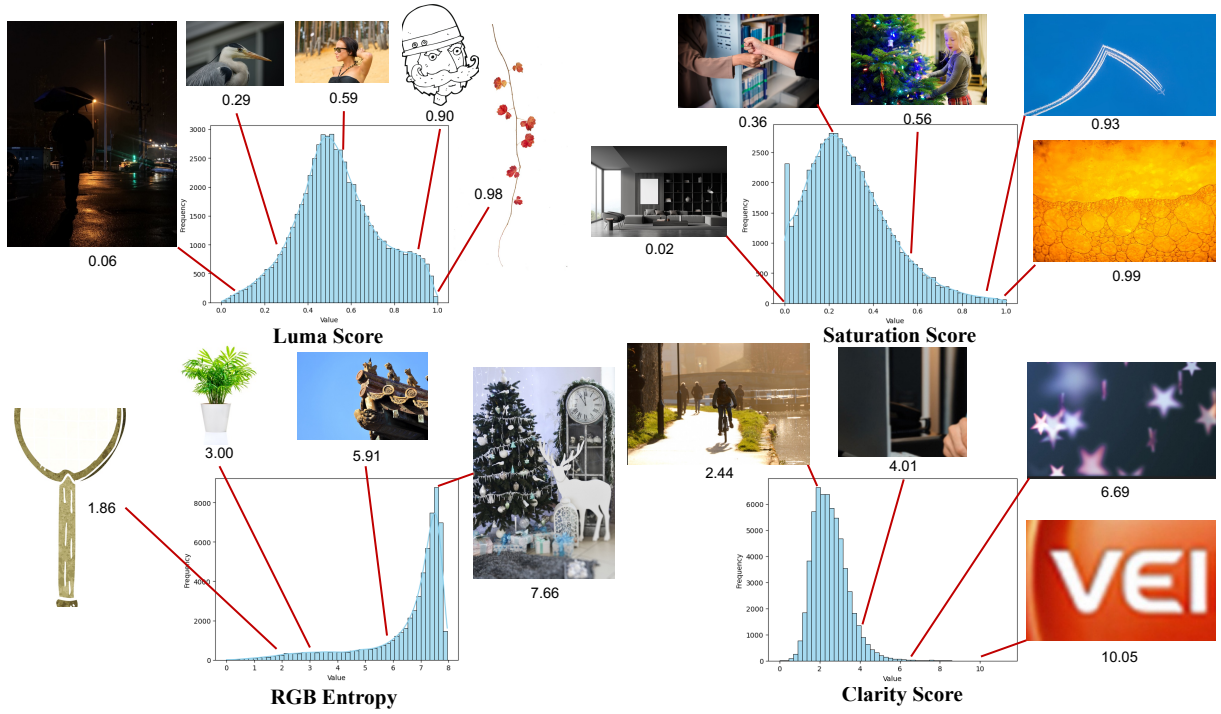


图 11: Examples of some filtering operators used in our pipeline. Extreme values in these operators often indicate atypical or low-quality images. For instance, images with excessively high Luma score typically contain large areas of white or overexposed regions.
流水线中使用的某些过滤算子的示例。这些算子中的极值通常表明非典型或低质量图像。例如，Luma 分数过高的图像通常包含大范围白色或曝光过度的区域。

Split, to ensure balanced training across different linguistic contexts. To address challenges such as low-frequency characters, mixed-language scenarios, and font diversity, we incorporate synthetic text rendering data, which are generated using the strategies described in §§ 3.4. Moreover, the Intensive Text Filter and the Small Character Filter are applied to remove images with overly dense or excessively small text, as such cases are challenging to annotate accurately and difficult to render legibly.

该阶段重点是改进模型在图像中渲染文本的能力，这对生成高文本保真度的图像至关重要。为此，我们根据图像内文本的存在和语言对数据集进行分类。具体地，将阶段 3 的数据集分为四个 split：English Split、Chinese Split、Other Language Split 和 Non-Text Split，以确保在不同语言背景下的均衡训练。为应对低频字符、混合语言场景和字体多样性等挑战，我们采用合成文本渲染数据，这些数据使用 §§ 3.4 中描述的策略生成。此外，应用 Intensive Text Filter 和 Small Character Filter 移除具有过度密集或过小文本的图像，因为这类情况难以准确标注且难以清晰渲染。

阶段 5：高分辨率精细化 (Stage 5: High-Resolution Refinement) In this stage, the model transitions to training with images at 640p resolution, accompanied by further dataset refinement to ensure both high quality and aesthetic appeal. The Image Quality Filter is applied to eliminate images with quality defects such as overexposure, underexposure, blur, or compression artifacts. The Resolution Filter ensures that all images meet the minimum resolution requirements. The Aesthetic Filter is employed to exclude images with poor composition or visual appeal. Finally, the Abnormal Element Filter removes images containing watermarks, QR codes, barcodes, or other elements that can interfere with viewing.

该阶段中，模型过渡到以 640p 分辨率训练图像，伴随进一步的数据集精细化以确保高质量和美观度。应用 Image Quality Filter 消除具有曝光过度、曝光不足、模糊或压缩伪影等质量缺陷的图像。Resolution Filter 确保所有图像符合最低分辨率要求。应用 Aesthetic Filter 排除构图或视觉吸引力较差的图像。最后，Abnormal Element Filter 移除包含水印、二维码、条码或其他妨碍观看的元素的图像。

阶段 6：类别均衡与人物增强 (Stage 6: Category Balance and Portrait Augmentation) After identifying underperformed categories through careful error analysis, this stage involves recategorizing the dataset into three primary categories: General, Portrait, Text Rendering, to facilitate category-based rebalancing during training. Keyword-based retrieval and image retrieval techniques are employed to augment the dataset with targeted patches, to enhance coverage of underrepresented categories. To further improve the model’s ability to generate high-quality portraits, we first retrieve photorealistic portraits, cartoon characters, and celebrity images from the People category. Synthesized captions are then generated to emphasize character-specific details, such as facial features, expressions, and clothing, as well as contextual elements such as background, lighting, and mood. This approach aims to enhance both the quality of generated images and the model’s instruction-following capability. Additional filters are applied to remove images with face mosaics or blurs, to avoid potential privacy issues and ensure the model’s robustness in handling human subjects.

通过仔细的错误分析确定表现不足的类别后，该阶段涉及将数据集重新分类为三个主要类别：General、Portrait、Text Rendering，以促进训练中基于类别的重新均衡。采用基于关键词的检索和图像检索技术对数据集进行有针对性的增强，提高代表性不足类别的覆盖率。为进一步提升模型生成高质量人物图像的能力，我们首先从 People 类别中检索真实感人物、卡通角色和名人图像。然后生成合成描述文本以强调人物特定细节（如面部特征、表情和服装）以及背景、光线和情绪等背景元素。这一方法旨在提升生成图像的质量和模型的指令跟随能力。应用额外过滤器移除含有面部马赛克或模糊的图像，避免潜在隐私问题并确保模型在处理人物主体时的鲁棒性。

阶段 7：均衡多尺度训练 (Stage 7: Balanced Multi-Scale Training) In the final stage, the model is trained jointly on images with resolutions of 640p and 1328p. Imposing a strict resolution threshold of 1328p would lead to significant data loss and distort the underlying data distribution. To improve training efficiency and ensure balanced data distribution, we design a hierarchical taxonomy system for image categorization, inspired by the design principles of WordNet (Miller, 1995). All images in Stage 6 are classified according to this hierarchical taxonomy. Within each category, we retain only images with the highest quality and aesthetic appeal. Furthermore, a specialized resampling strategy is employed to balance data containing text rendering, addressing the long-tail distribution of token frequencies. Such balanced multi-scale training allows the model to retain previously learned general knowledge and ensure stable convergence while adapting to higher-resolution inputs, thereby improving detail generation without sacrificing robustness.

在最后阶段，模型在 640p 和 1328p 分辨率的图像上联合训练。严格设置 1328p 分辨率阈值会导致大量数据损失并扭曲数据分布。为提高训练效率并确保均衡的数据分布，我们设计了层级化的图像分类体系，受 WordNet (Miller, 1995) 设计原则启发。阶段 6 的所有图像根据这一层级化分类体系进行分类。每个类别内，我们仅保留质量和美观度最高的图像。此外，采用专门的重采样策略平衡包含文本渲染的数据，解决令牌频率的长尾分布。这种均衡的多尺度训练使模型能够保持已学得的知识，在适应更高分辨率输入的同时确保稳定收敛，从而提升细节生成能力而不牺牲鲁棒性。

3.3 数据标注

Data Annotation

In our data annotation pipeline, we utilize a capable image captioner (e.g., Qwen2.5-VL) to generate not only comprehensive image descriptions, but also structured metadata that captures essential image properties and quality attributes.

数据标注流水线采用强大的描述生成器（如 Qwen2.5-VL）生成全面的图像描述，并同步捕获关键图像属性和质量特征的结构化元数据。

Instead of treating captioning and metadata extraction as independent tasks, we designed an annotation framework in which the captioner concurrently describes visual content and generates detailed information in a structured format, such as JSON. Critical details such as object attributes, spatial relationships, environmental context, and verbatim transcriptions of visible text are captured in the caption, while key image properties like type, style, presence of watermarks, and abnormal elements (e.g., QR codes or facial mosaics) are reported

Qwen-Image Annotation Prompt

```
# Image Annotator
You are a professional image annotator. Please complete the following tasks based on the input image.

## Step 1: Create Image Caption
1. Write the caption using natural, descriptive text without structured formats or rich text.
2. Enrich caption details by including: object attributes, vision relations between objects, and environmental details.
3. Identify the text visible in the image, without translation or explanation, and highlight it in the caption with quotation marks.
3. Maintain authenticity and accuracy, avoid generalizations.

## Step 2: Image Quality Assessment
1. Image Type Identification: Return the image type based on its source and usage.
2. Image Style Identification: Return the image style based on its overall visual characteristics.
3. Watermark Detection: Detect watermarks in the image. Return the detected watermarks in a list format.
4. Abnormal Element Detection: Check if any elements affect viewing, such as QR codes or mosaics.

## Sample Output Format
```json
{
 "Caption": "...",
 "Image Type": "...",
 "Image Style": "...",
 "Watermark List": [],
 "Abnormal Element": "yes/no",
}
```

图 12: Example of the annotation prompt used in Qwen-Image.  
Qwen-Image 中使用的标注提示词示例。

in a structured format. With the help of advanced captioner, this methodology transcends traditional image captioning and generate both detailed image descriptions and structured metadata in a single pass, as shown in Figure 12. The annotation pipeline is designed to be efficient and scalable, allowing us to process large-scale datasets without relying on additional models or post-processing steps. In practice, we further refine the above pipeline by integrating initial annotations with expert rules and lightweight classification models for critical tasks, such as watermark verification and content filtering.

我们设计的标注框架将描述生成和元数据提取统一进行，描述生成器同时描述视觉内容并以结构化格式（如 JSON）生成详细信息。关键细节如对象属性、空间关系、环境背景和可见文本的逐字记录纳入描述文本，而类型、风格、水印存在、异常元素（如二维码或面部马赛克）等关键图像属性则以结构化格式输出。这一方法超越了传统图像标注，一次性生成详细的图像描述和结构化元数据（如 Figure 12 所示）。标注流水线设计为高效且可扩展，使我们能够处理大规模数据集而无需依赖额外模型或后处理步骤。实际应用中，我们通过将初始标注与专家规则和轻量级分类模型集成来进一步精细化流水线，用于水印验证和内容过滤等关键任务。

Overall, our pipeline not only provides deep insights into image content, but also supports advanced data curation, providing a solid foundation for training robust and reliable image generation models.

总体而言，我们的流水线不仅提供对图像内容的深入见解，还支持高级数据整理，为训练稳健可靠的图像生成模型奠定坚实基础。

### 3.4 数据合成

#### Data Synthesis

Given the long-tail distribution of textual content in real-world images, particularly for non-Latin languages such as Chinese, where numerous characters exhibit extremely low frequency, relying solely on naturally occurring text is insufficient to ensure adequate exposure to these rare characters during model training. To address this

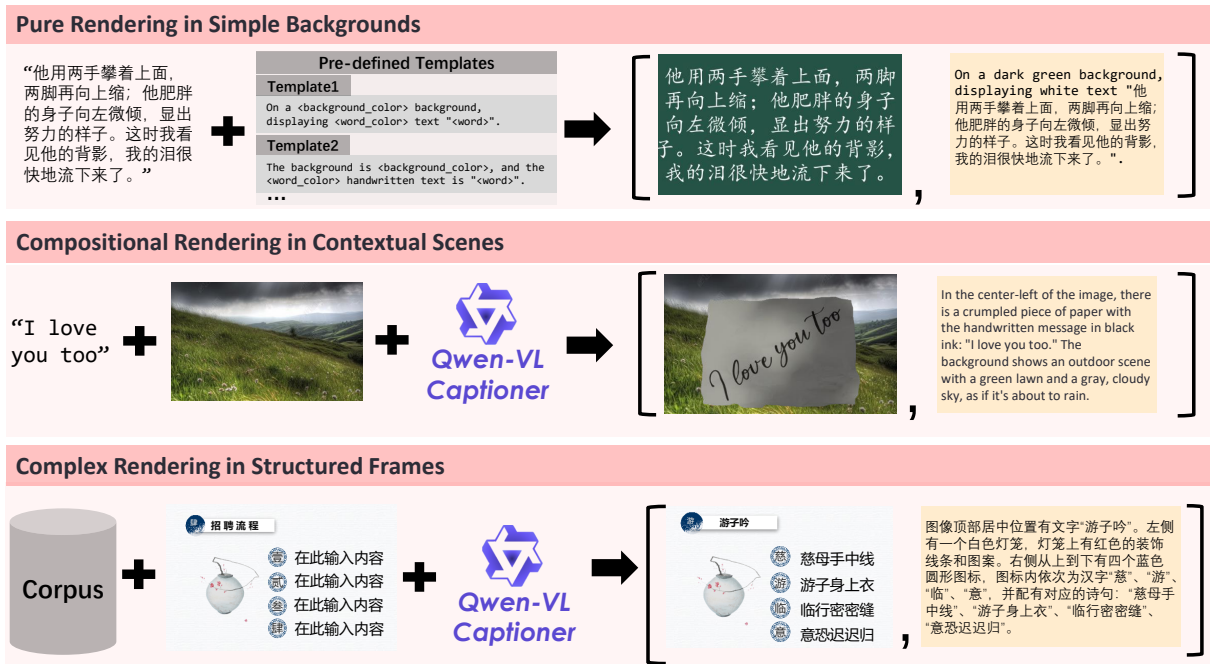


图 13: Overview of Data Synthesis. We designed three rendering strategies—Pure Rendering, Compositional Rendering, and Complex Rendering—to generate text-only data, text-in-context data, and complex-layout data, respectively.

数据合成概览。我们设计了三种渲染策略——纯渲染（Pure Rendering）、组合渲染（Compositional Rendering）和复杂渲染（Complex Rendering）——分别用于生成纯文本数据、文本上下文数据和复杂布局数据。

challenge and improve the robustness of text rendering across diverse contexts, we propose a multi-stage text-aware image synthesis pipeline. This pipeline integrates three complementary strategies: Pure Rendering, Compositional Rendering, and Complex Rendering. The details of each strategy are elaborated below.

在真实世界图像中，文本内容呈现长尾分布，尤其对于中文等非拉丁语言，众多字符出现频率极低，仅依靠自然出现的文本不足以确保模型训练时对这些稀有字符有充分的接触。为应对这一挑战并提升文本渲染在多样化背景下的鲁棒性，我们提出了多阶段文本感知的图像合成流水线。该流水线集成了三种互补策略：纯渲染（Pure Rendering）、组合渲染（Compositional Rendering）和复杂渲染（Complex Rendering）。各策略的详细内容如下阐述。

**简单背景下的纯渲染（Pure Rendering in Simple Backgrounds）** This strategy represents the most straightforward and effective method for training the model to recognize and generate characters (e.g., English and Chinese characters). Text paragraphs are extracted from large-scale high-quality corpora and rendered onto clean backgrounds using dynamic layout algorithms that adapt font size and spacing based on canvas size. To ensure high-quality synthesized samples, a rigorous quality control mechanism is employed: if any character within a paragraph cannot be rendered due to limitations (e.g., font unavailability or rendering errors), the entire paragraph is discarded. This strict filtering guarantees that only fully valid and legible samples are included in the training dataset, thereby maintaining high fidelity in character-level text rendering.

这一策略是训练模型识别和生成字符（如英文和中文字符）最直接有效的方法。文本段落从大规模高质量语料库中提取，使用动态布局算法渲染到干净背景上，该算法根据画布大小自适应调整字体大小和间距。为确保高质量合成样本，采用严格的质量控制机制：如果段落中任何字符因限制（如字体不可用或渲染错误）无法渲染，整个段落被丢弃。这一严格的过滤确保仅包含完全有效和清晰易读的样本被纳入训练数据集，从而在字符级文本渲染中保持高保真。

**背景场景下的组合渲染（Compositional Rendering in Contextual Scenes）** This strategy focuses on embedding synthetic text into realistic visual contexts, mimicking its appearance in everyday environments. Text

is simulated as being written or printed onto various physical media, such as paper or wooden boards, and then seamlessly composited into diverse background images to create visually coherent scenes. We employ the Qwen-VL Captioner to generate descriptive captions for each synthesized image, capturing contextual relationships between the text and its surrounding visual elements. This approach significantly improves the model’s ability to comprehend and generate text within real-world scenarios.

这一策略重点是将合成文本嵌入真实视觉背景，模拟其在日常环境中的外观。文本被模拟为写在或印在纸张或木板等多种物理媒介上，然后无缝合成到多样化的背景图像中以创建视觉上连贯的场景。我们采用 Qwen-VL Captioner 为每个合成图像生成描述文本，捕获文本与其周围视觉元素之间的背景关系。这一方法大幅提升了模型在真实场景中理解和生成文本的能力。

**结构化模板下的复杂渲染 (Complex Rendering in Structured Templates)** To improve the model’s capacity to follow complex, structured prompts involving layout-sensitive content, we propose a synthesis strategy based on programmatic editing of pre-defined templates, such as PowerPoint slides or User Interface Mockups. A comprehensive rule-based system is designed to automate the substitution of placeholder text while maintaining the integrity of layout structure, alignment, and formatting. These synthetic examples are crucial for helping model understand and execute detailed instructions that involve multi-line text rendering, precise spatial layouts, and control over text font and color.

为提升模型执行涉及布局敏感内容的复杂结构化提示词的能力，我们提出了基于预定义模板（如 PowerPoint 幻灯片或用户界面原型）程序化编辑的合成策略。设计了全面的基于规则的系统来自动替换占位符文本，同时保持布局结构、对齐和格式的完整性。这些合成示例对于帮助模型理解和执行涉及多行文本渲染、精确空间布局和文本字体与颜色控制的详细指令至关重要。

In conclusion, our data synthesis framework systematically addresses the challenges associated with the scarcity and imbalance of textual content in natural image datasets. By integrating multiple rendering strategies that span simplicity, realism, and structural complexity, the framework synthesizes comprehensive and diverse training data. Robustness across various text rendering tasks is achieved, thereby enhancing the model’s ability to generate high-quality images that accurately follow complex user prompts about text rendering.

总之，我们的数据合成框架系统地应对自然图像数据集中文本内容稀缺和不均衡的挑战。通过整合三种互补的渲染策略，框架生成全面且多样化的训练数据，增强模型在各类文本渲染任务中的鲁棒性，从而提升生成高质量图像并准确执行文本渲染提示词的能力。

## 4 训练

### Training

#### 4.1 预训练

##### Pre-training

We adopt a flow matching training objective to pre-train Qwen-Image, which facilitates stable learning dynamics via ordinary differential equations (ODEs) while preserving equivalence to the maximum likelihood objective. Formally, let  $x_0$  denote the latent of the input image. The latent representation  $z$  is obtained by encoding  $x_0$  through the variational autoencoder (VAE) encoder  $\mathcal{E}$ , i.e.,  $z = \mathcal{E}(x)$ , where  $\mathcal{E} : x \mapsto z$ . Next, a random noise vector  $x_1$  is sampled from the standard multivariate normal distribution, i.e.,  $x_1 \sim \mathcal{N}(0, \mathbf{I})$ . For a user input  $S$ , which may comprise text or prompt combined with image, the guidance latent  $h$  is obtained from an MLLM  $\phi$ , i.e.,  $h = \phi(S)$ , where  $\phi : S \mapsto h$ . In addition, a diffusion timestep  $t$  is sampled from a logit-normal distribution with  $t \in [0, 1]$ . According to Rectified Flow (Liu 等, 2022; Esser 等, 2024), the intermediate latent variable at timestep  $t$  and its corresponding velocity  $v_t$  can be calculated as:

$$\begin{cases} x_t = tx_0 + (1-t)x_1 \\ v_t = \frac{dx_t}{dt} = x_0 - x_1 \end{cases} \quad (1)$$



我们采用流匹配训练目标预训练 Qwen-Image，通过常微分方程（ODE）保证学习稳定性，同时等价于最大似然目标。形式上，设  $x_0$  为输入图像的潜在编码。潜在表示  $z$  通过变分自编码器（VAE）编码器  $\mathcal{E}$  对  $x_0$  进行编码得到，即  $z = \mathcal{E}(x)$ ，其中  $\mathcal{E} : x \mapsto z$ 。接着，从标准多元正态分布中采样随机噪声向量  $x_1$ ，即  $x_1 \sim \mathcal{N}(0, \mathbf{I})$ 。对于用户输入  $\mathbf{S}$ （可能包含文本或文本提示与图像的组合），引导潜在表示  $h$  从 MLLM  $\phi$  获得，即  $h = \phi(\mathbf{S})$ ，其中  $\phi : \mathcal{S} \mapsto h$ 。此外，从 logit-正态分布中采样扩散时间步  $t$ ，满足  $t \in [0, 1]$ 。根据修正流（Rectified Flow）（Liu 等, 2022; Esser 等, 2024），时间步  $t$  处的中间潜在变量及其对应速度  $v_t$  可计算为：

$$\begin{cases} x_t = tx_0 + (1-t)x_1 \\ v_t = \frac{dx_t}{dt} = x_0 - x_1 \end{cases} \quad (2)$$

Then, the model is trained to predict the target velocity, and the loss function is defined as the mean squared error (MSE) between the predicted output  $f_\theta(x_t, t)$  and the ground truth velocity  $v_t$ :

$$\mathcal{L} = \mathbb{E}_{(x_0, h) \sim \mathcal{D}, x_1, t} \|v_\theta(x_t, t, h) - v_t\|^2, \quad (3)$$

where  $v_\theta(x_t, t, h)$  is velocity predicted by the model and  $\mathcal{D}$  denotes the training dataset.

然后，模型被训练来预测目标速度，损失函数定义为预测输出  $f_\theta(x_t, t)$  与真值速度  $v_t$  之间的均方误差（MSE）：

$$\mathcal{L} = \mathbb{E}_{(x_0, h) \sim \mathcal{D}, x_1, t} \|v_\theta(x_t, t, h) - v_t\|^2, \quad (4)$$

其中  $v_\theta(x_t, t, h)$  是模型预测的速度， $\mathcal{D}$  表示训练数据集。

#### 4.1.1 生产者-消费者框架

##### Producer-Consumer Framework

为确保大规模 GPU 集群上的高吞吐量和训练稳定性，我们采用受 Ray (Moritz 等, 2018) 启发的生产者-消费者框架，将数据预处理与模型训练解耦。此设计支持两个阶段异步高效运行，并可动态更新数据管道而不中断训练。在生产者端，原始图像-标题对首先根据预定义的标准（如图像分辨率和检测算子）进行过滤。选中的数据随后使用 MLLM 模型（如 Qwen2.5 VL）和 VAE 编码成潜在表示。处理后的图像按分辨率分组存储在快速访问缓存桶中，并保存在共享的位置感知存储中，使消费者能够立即获取而无需等待。生产者与消费者之间的连接通过采用特定的 HTTP 传输层实现，该层原生支持两端之间异步、零拷贝调度所需的 RPC 语义。消费者部署在 GPU 密集集群上，专门用于模型训练。通过将所有数据处理卸载到生产者，消费者节点可将其全部计算资源用于训练 MMDiT 模型。MMDiT 参数按 4 路张量并行布局分布在这些节点上，每个数据并行组异步直接从生产者拉取预处理好的批次。消费者端的进一步优化细节将在下一节介绍。

#### 4.1.2 分布式训练优化

##### Distributed Training Optimization

鉴于 Qwen-Image 模型的参数规模很大，单独使用 FSDP (Zhao 等, 2023) 不足以将模型装入每个 GPU。因此，我们利用 Megatron-LM (Shoeybi 等, 2019; Korthikanti 等, 2023) 进行训练并应用以下优化：

**混合并行策略（Hybrid Parallelism Strategy）** We adopted a hybrid parallelism strategy, combining data parallelism and tensor parallelism, to efficiently scale training across large GPU clusters. Specifically, to implement tensor parallelism, we built the MMDiT model using the Transformer-Engine library<sup>1</sup>, which allows seamless and automatic switching between different degrees of tensor parallelism. Furthermore, for the multi-head self-attention blocks, we employ the head-wise parallelism to reduce the synchronization and communication overhead compared to tensor parallelism along the head dimension.

<sup>1</sup><https://github.com/NVIDIA/TransformerEngine>

---

我们采用混合并行策略，结合数据并行和张量并行，以高效地扩展大 GPU 集群上的训练。具体地，为实现张量并行，我们使用 Transformer-Engine 库<sup>2</sup>构建 MMDiT 模型，它允许在不同张量并行度之间进行无缝和自动切换。此外，针对多头自注意力块，采用按头并行可降低同步和通信开销。

**分布式优化器和激活检查点 (Distributed Optimizer and Activation Checkpointing)** To alleviate GPU memory pressure with minimal recomputation overhead during backpropagation, we experimented with both distributed optimizers and activation checkpointing. However, activation checkpointing introduces substantial computational overhead in the backward pass, which can significantly degrade training speed. Through empirical comparison with the 256 multi-resolution image training setup, we observed that enabling activation checkpointing reduced per-GPU memory consumption by 11.3% (from 71GB to 63GB per GPU), but at the cost of increasing per-iteration time by  $3.75\times$  (from 2s to 7.5s per iteration). Based on this trade-off, we ultimately opted to disable activation checkpointing and rely solely on distributed optimizers. During training, all-gather operations are performed in bfloat16, while gradient reduce-scatter operations utilize float32, ensuring both computational efficiency and improved numerical stability.

为缓解 GPU 内存压力，我们同时尝试了分布式优化器和激活检查点。然而激活检查点在反向传播时引入大量计算开销，反而降低训练速度。实证对比表明：在 256 多分辨率图像设置下，激活检查点减少 11.3% 内存（从 71GB 降至 63GB/GPU），但每迭代耗时增加 3.75 倍（从 2s 增至 7.5s）。综合考虑，我们禁用激活检查点，仅用分布式优化器。训练时，all-gather 采用 bfloat16，梯度 reduce-scatter 采用 float32，兼顾效率和数值稳定性。

#### 4.1.3 训练策略

##### Training Strategy

我们采用多阶段预训练策略，目标是逐步提升数据质量、图像分辨率和模型性能。在不同的训练阶段，我们融合各种训练策略以优化学习过程。这些训练策略列举如下：

##### 分辨率提升：从低分辨率到高分辨率 (Enhancing Resolution: From Low Resolution to High Resolution)

This strategy progressively upscales the multi-resolution, multi-aspect ratio inputs, starting from an initial resolution of  $256\times 256$  pixels (with various aspect ratios including 1:1, 2:3, 3:2, 3:4, 4:3, 9:16, 16:9, 1:3 and 3:1), then increasing to  $640\times 640$  pixels, and finally reaching  $1328\times 1328$  pixels. By enhancing image resolution, the model can capture more detailed features, leading to better performance. Richer feature spaces facilitate improved generalization to unseen data. For example, transitioning from low-resolution to high-resolution flower images allows the model to discern finer details such as petal textures and color gradients.

此策略逐步上采样多分辨率、多宽高比输入，从初始分辨率  $256\times 256$  像素（包括多个宽高比如 1:1、2:3、3:2、3:4、4:3、9:16、16:9、1:3 和 3:1）开始，然后增加到  $640\times 640$  像素，最后达到  $1328\times 1328$  像素。通过增强图像分辨率，模型能捕捉更多细节特征，产生更好的效果。更丰富的特征空间促进了对未见数据的泛化。例如，从低分辨率花卉图像过渡到高分辨率图像时，模型能识别花瓣纹理和色彩梯度等细节。

##### 融合文字渲染：从无文字到有文字 (Integrating Textual Rendering: From Non-text to Text)

To address the limited availability of textual content in conventional vision datasets and the consequent suboptimal glyph generation performance, particularly for Chinese characters, we progressively introduce images containing rendered text superimposed on natural backgrounds. This strategy enables the model to initially learn general visual representations and subsequently acquire text rendering capability.

为解决常规视觉数据集中文字内容有限以及由此导致的字形生成性能不佳（特别是对中文字符）的问题，我们逐步引入包含叠加在自然背景上的渲染文本的图像。此策略使模型先学习一般视觉表示，随后获得文本渲染能力。

##### 数据质量精化：从大规模数据到精选数据 (Refining Data Quality: From Massive Data to Refined Data)

During the early stages of pre-training, we utilize large-scale datasets to enable the model to acquire fundamental

---

<sup>2</sup><https://github.com/NVIDIA/TransformerEngine>

---

visual generation capabilities. As training progresses, we gradually employ increasingly stringent data filtering mechanisms to select higher-quality data. This progressive data refinement ensures that only the most relevant and high-quality samples are leveraged to ensure the training efficiency and model performance.

预训练初期，我们利用大规模数据集使模型获得基础视觉生成能力。随着训练进展，我们逐渐采用更严格的数据过滤机制以选择高质量数据。此渐进式数据精化确保仅利用最相关和高质量的样本，进而保证训练效率和模型性能。

**数据分布平衡：从不平衡到平衡 (Balancing Data Distribution: From Unbalanced to Balanced)** Throughout the training process, we progressively balance the dataset with respect to domain and image resolution distributions. This adjustment mitigates the risk of the model overfitting to particular domains or resolutions, which could otherwise compromise the fidelity and fine-grained details of generated images in underrepresented settings. By maintaining a more uniform data distribution, we promote robust generalization across diverse domains and resolutions.

在整个训练过程中，我们逐步平衡数据集在领域和图像分辨率分布上的比例。此调整降低了模型对特定领域或分辨率过拟合的风险，否则可能损害生成图像在代表不足设置中的保真度和细节。通过维持更均匀的数据分布，我们促进了在多个领域和分辨率上的稳健泛化。

**合成数据增强：从真实数据到合成数据 (Augmenting with Synthetic Data: From Real-World Data to Synthetic Data)** Certain data distributions, such as surrealistic styles or high-resolution images containing extensive textual content, are underrepresented or even absent in real-world datasets. Additionally, the availability of some high-quality data samples is inherently limited. To address these gaps, we employ data synthesis techniques to generate supplementary samples, thereby enriching the dataset and ensuring more comprehensive coverage of diverse visual domains. This augmentation strategy enhances the model's ability to generalize and perform robustly across a wider range of scenarios.

某些数据分布（如超现实主义风格或包含大量文字的高分辨率图像）在真实数据集中代表不足甚至完全缺失。此外，某些高质量数据样本的可用性本身就受限。为解决这些空缺，我们采用数据合成技术生成补充样本，从而丰富数据集并确保对多种视觉领域的全面覆盖。此增强策略提升了模型在更广泛场景中的泛化和鲁棒性能。

## 4.2 后训练

Post-training

本节介绍 Qwen-Image 的后训练框架，包含两个阶段：监督微调 (SFT) 和强化学习 (RL) (Kaelbling 等, 1996)。

### 4.2.1 监督微调 (SFT)

Supervised Fine-Tuning (SFT)

在 SFT 阶段，我们构建了语义分类层次化组织的数据集，并采用细致的人工标注来针对模型的具体不足进行改进。我们要求选中的图像清晰、细节丰富、亮度高且具有照相逼真感。此方法旨在引导模型生成具有更高逼真度和更精细细节的内容。

### 4.2.2 强化学习 (RL)

Reinforcement Learning (RL)

We employ two distinct RL strategies: Direct Preference Optimization (DPO) (Rafailov 等, 2023) and Group Relative Policy Optimization (GRPO) (Shao 等, 2024). DPO excels at flow-matching (one step) online preference modeling and is computationally efficient, whereas GRPO performs on-policy sampling during training and evaluates each trajectory with a reward model. To leverage the scalability advantages of offline preference



learning, we conduct relative large-scale RL with DPO and reserve GRPO for small fine-grained RL refinement. Details of both algorithms are provided below.

我们采用两种不同的强化学习策略：直接偏好优化（DPO）（Rafailov 等, 2023）和组相对策略优化（GRPO）（Shao 等, 2024）。DPO 擅长流匹配（单步）在线偏好建模且计算高效，而 GRPO 在训练时执行同策略采样并用奖励模型评估每条轨迹。为充分利用离线偏好学习的可扩展性优势，我们采用 DPO 进行相对大规模的强化学习，将 GRPO 保留用于小规模细粒度强化学习精化。两种算法的详细信息见下文。

#### (A) 直接偏好优化（DPO） (A) Direct Preference Optimization (DPO)

**数据准备 (Data Preparation)** For DPO training data, given the same prompt, multiple images are generated with different random initialization seeds. Human annotators are then tasked with selecting the best and the worst images from these candidates. The data is divided into two categories: prompts associated with reference (gold) images, and prompts without reference images. For data with reference images, annotators first compare the generated outputs to the reference. If there is a significant discrepancy, the annotators are instructed to designate the worst generation as the rejected sample. For prompts without reference images, annotators are asked to select the best and worst samples among the generated images, or to indicate if all generated results are of unsatisfactory quality.

对于 DPO 训练数据，给定相同的提示，用不同的随机初始化种子生成多张图像。人工标注员随后被要求从这些候选中选择最佳和最差的图像。数据分为两类：与参考（金标准）图像关联的提示，以及无参考图像的提示。对于有参考图像的数据，标注员首先将生成的输出与参考进行比较。如果存在显著差异，标注员被指示将最差生成标记为被拒绝样本。对于无参考图像的提示，标注员被要求在生成的图像中选择最佳和最差样本，或指出所有生成结果质量均不满意。

**算法 (Algorithm)** Given the text hidden state  $h$ , chosen generated image (or golden image)  $x_0^{win}$  and rejected generated image  $x_0^{lose}$ , we sample timestep  $t \sim (0, 1)$  to construct the input latent variable  $x_t^{win}$  and  $x_t^{lose}$  as well as their corresponding velocity  $v_t^{win}$  and  $v_t^{lose}$  following Equation 2. Then, inspired by previous work (Wallace 等, 2024), we construct the DPO objective based on the flow matching training criterion, which can be formulated as follows:

给定文本隐状态  $h$ 、被选中的生成图像（或金标准图像） $x_0^{win}$  和被拒绝的生成图像  $x_0^{lose}$ ，我们采样时间步  $t \sim (0, 1)$  以构造输入潜在变量  $x_t^{win}$  和  $x_t^{lose}$  及其对应速度  $v_t^{win}$  和  $v_t^{lose}$ ，按照方程 2。然后，受之前工作的启发 (Wallace 等, 2024)，我们基于流匹配训练准则构造 DPO 目标，其可表述如下：

$$\begin{cases} \text{Diff}_{\text{policy}} = \left( \|v_\theta(x_t^{win}, h, t) - v_t^{win}\|_2^2 - \|v_\theta(x_t^{lose}, h, t) - v_t^{lose}\|_2^2 \right) \\ \text{Diff}_{\text{ref}} = \left( \|v_{\text{ref}}(x_t^{win}, h, t) - v_t^{win}\|_2^2 - \|v_{\text{ref}}(x_t^{lose}, h, t) - v_t^{lose}\|_2^2 \right) \\ \mathcal{L}_{DPO} = -\mathbb{E}_{h, (x_0^{win}, x_0^{lose}) \sim \mathcal{D}, t \sim \mathcal{U}(0,1)} \left[ \log \sigma \left( -\beta (\text{Diff}_{\text{policy}} - \text{Diff}_{\text{ref}}) \right) \right], \end{cases} \quad (5)$$

其中  $\text{Diff}_{\text{policy}}$  和  $\text{Diff}_{\text{ref}}$  分别表示由策略模型和参考模型计算的偏好差异， $\beta$  是缩放参数， $\sigma(\cdot)$  表示 sigmoid 函数。

#### (B) 组相对策略优化（GRPO） (B) Group Relative Policy Optimization (GRPO)

**算法 (Algorithm)** After training with DPO, we perform further fine-grained training using GRPO following the Flow-GRPO (Liu 等, 2025a) 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  $\{x_T^i, x_{T-1}^i, \dots, x_0^i\}_{i=1}^G$ . Within each group, the advantage function can be formulated as:

$$A_i = \frac{R(x_0^i, h) - \text{mean}(\{R(x_0^i, h)\}_{i=1}^G)}{\text{std}(\{R(x_0^i, h)\}_{i=1}^G)}, \quad (6)$$

在使用 DPO 训练后，我们按照 Flow-GRPO (Liu 等, 2025a) 框架进行进一步的细粒度训练。给定文本隐状态  $h$ ，流模型预测一组  $G$  张图像  $\{x_0^i\}_{i=1}^G$  及其对应轨迹  $\{x_T^i, x_{T-1}^i, \dots, x_0^i\}_{i=1}^G$ 。在每组内，优势函数可表述为：

$$A_i = \frac{R(x_0^i, h) - \text{mean}(\{R(x_0^i, h)\}_{i=1}^G)}{\text{std}(\{R(x_0^i, h)\}_{i=1}^G)}, \quad (7)$$

where  $R$  is the reward model. Then, the training objective of GRPO is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{h \sim \mathcal{D}, \{x_T^i, \dots, x_0^i\}_{i=1}^G \sim \pi_\theta} \left( \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^{T-1} \left( \min(r_t^i(\theta) A_i, \text{clip}(r_t^i(\theta), 1 - \epsilon, 1 + \epsilon) A_i) - \beta D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right) \right), \quad (8)$$

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

其中  $R$  是奖励模型。GRPO 的训练目标为：

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{h \sim \mathcal{D}, \{x_T^i, \dots, x_0^i\}_{i=1}^G \sim \pi_\theta} \left( \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^{T-1} \left( \min(r_t^i(\theta) A_i, \text{clip}(r_t^i(\theta), 1 - \epsilon, 1 + \epsilon) A_i) - \beta D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right) \right), \quad (9)$$

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

When sampling the trajectories  $\{x_T^i, \dots, x_0^i\}_{i=1}^G \sim \pi_\theta$ , we have  $dx_t = v_t dt$  for flow-matching sampling (following Eq. 2), where  $v_t = v_\theta(x_t, t, h)$  is the predicted velocity. However, this sampling strategy exhibits no randomness, which is not suitable for exploration. Thus, we reformulate the sampling process as an SDE process for more randomness. The SDE sampling process can be written as:

$$dx_t = \left( v_t + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_t) \right) dt + \sigma_t dw, \quad (10)$$

where  $\sigma_t$  denotes the magnitude of randomness. Using Euler-Maruyama discretization, we have

采样轨迹  $\{x_T^i, \dots, x_0^i\}_{i=1}^G \sim \pi_\theta$  时，流匹配采样有  $dx_t = v_t dt$  (按照方程 2)，其中  $v_t = v_\theta(x_t, t, h)$  是预测速度。但此采样策略不含随机性，不适于探索。因此，我们将采样过程重新表述为随机微分方程 (SDE) 过程以引入更多随机性。SDE 采样过程可写为：

$$dx_t = \left( v_t + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_t) \right) dt + \sigma_t dw, \quad (11)$$

其中  $\sigma_t$  表示随机性的幅度。采用 Euler-Maruyama 离散化，我们得到

$$x_{t+\Delta t} = x_t + \left[ v_\theta(x_t, t, h) + \frac{\sigma_t^2}{2t} (x_t + (1-t)v_\theta(x_t, t, h)) \right] \Delta t + \sigma_t \sqrt{\Delta t} \epsilon. \quad (12)$$

We use the above equation for sampling the trajectories. The KL-divergence in Eq. (9) can be solved in a closed form

$$D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) = \frac{\Delta t}{2} \left( \frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t} \right)^2 \|v_\theta(x_t, t, h) - v_{\text{ref}}(x_t, t, h)\|^2. \quad (13)$$

我们用上述方程进行轨迹采样。方程 (9) 中的 KL 散度可以求得闭式解

$$D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) = \frac{\Delta t}{2} \left( \frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t} \right)^2 \|v_\theta(x_t, t, h) - v_{\text{ref}}(x_t, t, h)\|^2. \quad (14)$$

### 4.3 多任务训练

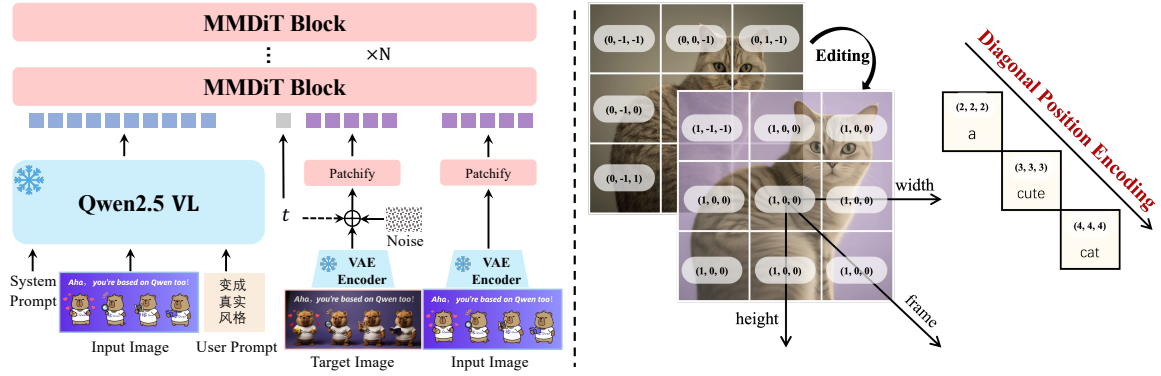


图 14: Overview of the Image Editing (TI2I) task. Left: Illustration of the TI2I task training input format. The user prompt is “Turn into realistic style” in English. Right: The corresponding modification to MSRoPE for TI2I, where a new dimension (frame) is introduced to distinguish between the images before and after editing. 图像编辑 (TI2I) 任务概览。左图: TI2I 任务训练输入格式的示意图。用户提示为英文“Turn into realistic style”。右图: TI2I 对应的 MSRoPE 修改, 引入新的维度 (帧) 以区分编辑前后的图像。

#### 图像编辑任务系统提示

```
<|im_start|>system
Describe the key features of the input image (color, shape, size, texture, objects, background), then
explain how the user's text instruction should alter or modify the image. Generate a new image that
meets the user's requirements while maintaining consistency with the original input where appropriate.
<|im_end|>
<|im_start|>user
<|vision_start|><|user_image|><|vision_end|><|user_text|><|im_end|>
<|im_start|>assistant
```

图 15: System prompt for Image Editing (TI2I) task, where  $\langle |user\_image| \rangle$  is the user input image and  $\langle |user\_text| \rangle$  is the user input prompt. 图像编辑 (TI2I) 任务系统提示, 其中  $\langle |user\_image| \rangle$  是用户输入图像,  $\langle |user\_text| \rangle$  是用户输入提示。

#### Multi-task training

除文本到图像 (T2I) 生成外, 我们扩展基础模型以探索多模态图像生成任务, 其包含文本和图像输入, 包括基于指令的图像编辑 (Wang 等, 2025)、新视角合成 (Wang 等, 2024b) 以及深度估计 (Bochkovskii 等, 2024) 等计算机视觉任务。我们广义上将这些任务视为图像编辑任务。基于 Qwen2.5-VL 的能力, 我们的模型原生支持图像输入: 从用户提供的图像中提取的视觉图块通过视觉 Transformer (ViT) 编码后与文本标记拼接, 形成输入序列。我们设计了图 15 所示的系统提示, 并将输入图像和文本指令作为输入提供给 Qwen-Image MMDiT 的文本流。

Inspired by prior work (Labs 等, 2025), which demonstrates that incorporating VAE embeddings helps maintain character and scene consistency, we additionally feed the VAE-encoded latent representation of the input image into the image stream, concatenating it with the noised image latent along the sequence dimension. To enable the model to distinguish between multiple images, we extend MSRoPE by introducing an additional frame dimension, in addition to the height and width used to locate image patches within a single image (see the right part of Figure 14). Empirically, we find that providing the visual semantic embeddings from the MLLM enables better instruction following, while introducing pixel-level VAE embeddings further enhances the model's ability to preserve visual fidelity and maintain structural consistency with the user-provided image.

受先前工作的启发 (Labs 等, 2025), 该工作证明了融入 VAE 嵌入有助于保持字符和场景一致性, 我们还将输入图像的 VAE 编码潜在表示喂入图像流, 沿序列维度将其与噪声图像潜在表示拼接。为使模型能够区分多张图像, 我们扩展 MSRoPE 并引入额外的帧维度, 加上用于定位单个图像内图块位置的高度和宽度 (见图 14 的右部分)。实验表明, 提供来自 MLLM 的视觉语义嵌入能够提升指令遵循能力, 而引



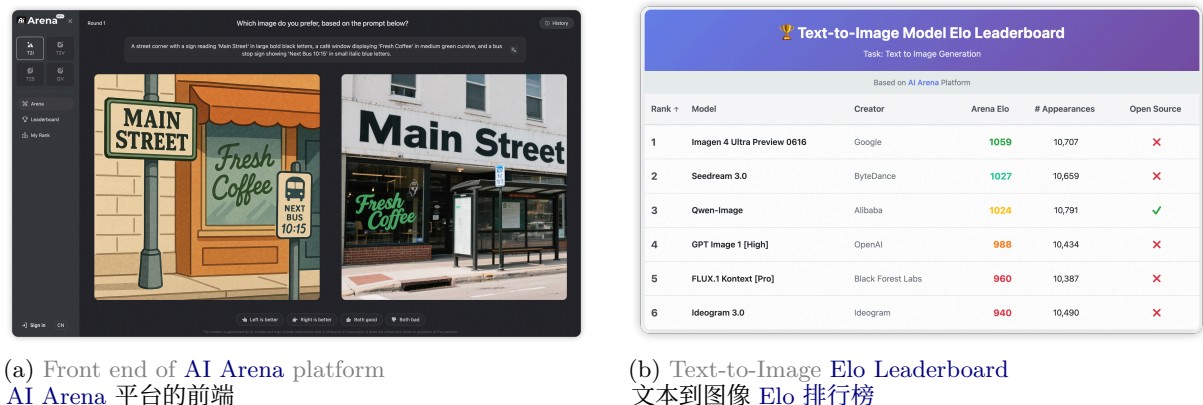


图 16: Comparison of Qwen-Image and leading closed-source APIs on the AI Arena platform. Users can compare images generated by two anonymous models based on the prompt and choose which one is better, both are good, or both are bad. Elo Leaderboard is powered by Qwen3-Coder (Qwen, 2025) from chat.qwen.ai. Qwen-Image 与主流闭源 API 在 AI Arena 平台上的对比。用户可以基于提示词比较两个匿名模型生成的图像，选择哪一个更优、两个都好或两个都不好。Elo 排行榜由来自 chat.qwen.ai 的 Qwen3-Coder (Qwen, 2025) 驱动。

入像素级 VAE 嵌入进一步增强了模型保持视觉保真度和与用户提供图像的结构一致性的能力。

## 5 实验

### Experiments

#### 5.1 人工评测

##### Human Evaluation

To comprehensively evaluate the general image generation capabilities of Qwen-Image and objectively compare it with state-of-the-art closed-source APIs, we have developed AI Arena<sup>3</sup>, an open benchmarking platform built upon the Elo rating system (Elo & Sloan, 2008), as shown in Figure 16.

为全面评估 Qwen-Image 的通用图像生成能力，并与最先进的闭源 API 进行客观对比，我们开发了 AI Arena<sup>4</sup>，一个基于 Elo 评分系统 (Elo & Sloan, 2008) 的开放基准平台，如图 16 所示。

AI Arena serves as a fair and dynamic open competition platform. In each round, two images generated by randomly selected models using the same prompt are anonymously presented to users for pairwise comparison. Users vote for the superior image, and the results are used to update personal and global leaderboards through the Elo algorithm, allowing developers, researchers, and users to holistically assess the performance of models. Specifically, we curated about 5,000 diverse prompts that span a wide range of evaluation dimensions, including subject, style, photographic perspective, and more. More than 200 evaluators from various professional backgrounds were invited to participate in the assessment process.

AI Arena 是一个公平、动态的开放竞争平台。每轮评测中，系统随机选择两个模型用相同的提示词生成图像，以匿名方式提交给用户进行成对比较。用户投票选择更优的图像，结果经由 Elo 算法更新个人和全球排行榜，让开发者、研究者和用户能够全面评估模型性能。我们精心设计了约 5000 个多样提示词，覆盖对象、风格、摄影视角等广泛的评测维度。邀请了来自不同行业的 200 多名评测员参与。

AI Arena is now open to the public. Anyone can participate in the comparison of different models. In the future, the platform will further expand from text-to-image generation to various multimodal generation tasks such as image editing, text-to-audio, text-to-video, and image-to-video generation. The platform strictly adheres

<sup>3</sup><https://aiarena.alibaba-inc.com>

<sup>4</sup><https://aiarena.alibaba-inc.com>

---

to the standards of objectivity and independence, and will detect and eliminate cheating or invalid data through various techniques.

AI Arena 现已向公众开放，任何人都可以参与不同模型的对比。未来，平台将从文本到图像生成进一步扩展到各种多模态生成任务，如图像编辑、文本到音频、文本到视频和图像到视频生成。平台严格遵守客观性和独立性标准，并将通过各种技术检测和消除作弊或无效数据。

We selected five state-of-the-art closed-source APIs as competitors for Qwen-Image in the arena: Imagen 4 Ultra Preview 0606 (Google, 2025), Seedream 3.0 (Gao 等, 2025), GPT Image 1 [High] (OpenAI, 2025), FLUX.1 Kontext [Pro] (Labs 等, 2025), and Ideogram 3.0 (Ideogram, 2025). To date, each model has participated in at least 10,000 pairwise comparisons, ensuring robustness and fairness in the evaluation. Considering that most closed-source APIs do not reliably support Chinese text generation, we excluded prompts involving Chinese text to maintain objectivity in the comparative results.

我们选择了五个最先进的闭源 API 作为竞技场中 Qwen-Image 的竞争对手: Imagen 4 Ultra Preview 0606 (Google, 2025)、Seedream 3.0 (Gao 等, 2025)、GPT Image 1 [High] (OpenAI, 2025)、FLUX.1 Kontext [Pro] (Labs 等, 2025) 和 Ideogram 3.0 (Ideogram, 2025)。迄今为止，每个模型已参与至少 10000 次成对比较，确保评测的鲁棒性和公平性。考虑到大多数闭源 API 不能可靠地支持中文文本生成，我们排除了涉及中文文本的提示词，以保持对比结果的客观性。

As shown in Figure 16, Qwen-Image, as the only open-source image generation model, ranks third in the AI Arena. Although Qwen-Image trails the leading Imagen 4 Ultra Preview 0606 by 30 Elo points, it demonstrates a significant advantage of over 30 Elo points compared to models such as GPT Image 1 [High] and FLUX.1 Kontext [Pro]. These results establish Qwen-Image as a powerful open-source image generation model, providing strong performance and broad utility for developers, researchers, and users.

如 Figure 16 所示，Qwen-Image 作为唯一的开源图像生成模型，在 AI Arena 中排名第三。虽然 Qwen-Image 落后于领先的 Imagen 4 Ultra Preview 0606 约 30 个 Elo 分，但相比 GPT Image 1 [High] 和 FLUX.1 Kontext [Pro] 等模型，它展现出超过 30 个 Elo 分的显著优势。这些结果证实了 Qwen-Image 是一个强大的开源图像生成模型，为开发者、研究者和用户提供了强劲的性能和广泛的实用性。

## 5.2 定量结果

### Quantitative Results

为了全面评估我们模型的视觉生成能力，我们首先在 §5.2.1 中报告其在 VAE 重建中的性能，以展示模型生成质量的上界。我们进一步在两个基础视觉生成任务上进行评估，包括 §5.2.2 中的文本到图像 (T2I) 和 §5.2.3 中的图像编辑 (TI2I)，以提供对模型基础生成能力的全面评估。

### 5.2.1 VAE 重建性能

#### Performance of VAE Reconstruction

We quantitatively evaluate several state-of-the-art image tokenizers, reporting Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) to assess reconstruction quality. All compared VAEs operate at an 8x8 compression rate using a latent channel dimension of 16. Notably, FLUX-VAE (BlackForest, 2024), Cosmos-CI-VAE (Agarwal 等, 2025), and SD-3.5-VAE (Esser 等, 2024) are image tokenizers, while Wan2.1-VAE (Wan 等, 2025), Hunyuan-VAE (Kong 等, 2024), and Qwen-Image-VAE function as joint image and video tokenizers. For a fair comparison, we report effective image parameters (see "Image Params" in Table 2). This accounts for converting the 3D convolutions in joint models to equivalent 2D convolutions for image processing. Following prior art, evaluation is conducted on the ImageNet-1k (Deng 等, 2009) validation set at 256x256 resolution for general domain performance. To further evaluate reconstruction capabilities on small texts, we also include reconstruction results on an in-house text-rich corpus covering diverse text sources (PDFs, PPT slides, posters, and synthetic texts) and languages. For improved numerical precision, we evaluate tokenizers using float32. As shown in Table 2, Qwen-Image-VAE achieves state-of-the-art reconstruction performance across all evaluated metrics. Significantly, when processing images, Qwen-Image-VAE activates only 19M parameters in the

表 2: Quantitative Evaluation results of VAE.  
VAE 的定量评估结果。

模型	参数量		图像参数量		Imagenet_256x256		Text_256x256	
	Enc	Dec	Enc	Dec	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
Wan2.1-VAE (Wan 等, 2025)	54M	73M	19M	25M	31.29	0.8870	26.77	0.9386
Hunyuan-VAE (Kong 等, 2024)	100M	146M	34M	50M	33.21	0.9143	32.83	0.9773
FLUX-VAE (BlackForest, 2024)	34M	50M	34M	50M	32.84	0.9155	32.65	0.9792
Cosmos-CI-VAE (Agarwal 等, 2025)	31M	46M	31M	46M	32.23	0.9010	30.62	0.9664
SD-3.5-VAE (Esser 等, 2024)	34M	50M	34M	50M	31.22	0.8839	29.93	0.9658
<b>Qwen-Image-VAE</b>	54M	73M	19M	25M	<b>33.42</b>	<b>0.9159</b>	<b>36.63</b>	<b>0.9839</b>

encoder and 25M in the decoder, achieving an optimal balance between reconstruction quality and computational efficiency.

我们定量评估了几个最先进的图像 tokenizer，报告峰值信噪比 (PSNR) 和结构相似性指数 (SSIM) 来评估重建质量。所有比较的 VAE 都以 8x8 压缩率运行，潜在通道维度为 16。值得注意的是，FLUX-VAE (BlackForest, 2024)、Cosmos-CI-VAE (Agarwal 等, 2025) 和 SD-3.5-VAE (Esser 等, 2024) 是图像 tokenizer，而 Wan2.1-VAE (Wan 等, 2025)、Hunyuan-VAE (Kong 等, 2024) 和 Qwen-Image-VAE 是联合图像和视频 tokenizer。为公平起见，我们报告有效的图像参数（见表 2 中的”Image Params”），将联合模型的 3D 卷积换算为等效的 2D 卷积。按照先前的工作，我们在 ImageNet-1k (Deng 等, 2009) 验证集上以 256x256 分辨率进行评估，以评测通用域性能。为进一步评估小文本的重建能力，我们还在公司内部文本丰富语料库上进行评估，涵盖多种文本来源（PDF、PPT 幻灯片、海报和合成文本）和语言。为提高数值精度，我们使用 float32 对所有 tokenizer 进行评估。如表 2 所示，Qwen-Image-VAE 在所有评估指标上都达到了最先进的重建性能。值得注意的是，处理图像时，Qwen-Image-VAE 在编码器中仅使用 19M 参数，在解码器中使用 25M 参数，在重建质量和计算效率之间取得最优平衡。

### 5.2.2 文本到图像生成性能

#### Performance of Text-to-Image Generation

We evaluate Qwen-Image’s performance on the text-to-image (T2I) task from two perspectives: general generation capability and text rendering capability. To assess the model’s general generation performance, we conduct evaluations on four publicly available benchmarks —DPG (Hu 等, 2024b), GenEval (Ghosh 等, 2023), OneIG-Bench (Chang 等, 2025), and TIIF (Wei 等, 2025). These benchmarks provide a comprehensive measurement of the model’s ability to generate high-quality and semantically consistent images from text prompts. To further evaluate the model’s text rendering capability, we separately evaluate its performance on English and Chinese text generation. For English text rendering, we use the CVTG-2K (Du 等, 2025) benchmark, which is specifically designed to assess the readability of the rendered English text. To address the lack of standardized evaluation for Chinese text rendering, we introduce a new benchmark named ChineseWord, which evaluates the model’s ability to render Chinese characters, allowing us to systematically assess the model’s text rendering performance. Furthermore, to fully evaluate Qwen-Image’s ability to accurately render long texts, we conducted an evaluation on LongText-Bench (Geng 等, 2025), a benchmark designed to evaluate the performance on rendering longer texts in both English and Chinese.

我们从两个角度评估 Qwen-Image 在文本到图像 (T2I) 任务上的性能：通用生成能力和文本渲染能力。为了评估模型的通用生成性能，我们在四个公开可用的基准上进行评估——DPG (Hu 等, 2024b)、GenEval (Ghosh 等, 2023)、OneIG-Bench (Chang 等, 2025) 和 TIIF (Wei 等, 2025)。这些基准全面衡量了模型从文本提示词生成高质量、语义一致图像的能力。为进一步评估模型的文本渲染能力，我们分别评估其在英文和中文文本生成中的性能。对于英文文本渲染，我们使用 CVTG-2K (Du 等, 2025) 基准，该基准专门设计用于评估渲染英文文本的可读性。为了解决中文文本渲染缺乏标准化评估的问题，我们引入了一个名为 ChineseWord 的新基准，用于评估模型渲染中文字符的能力，使我们能够系统地评测模型的文本渲染性能。此外，为充分评估 Qwen-Image 准确渲染长文本的能力，我们在 LongText-Bench (Geng



等, 2025) 上进行了评估, 该基准设计用于评估英文和中文长文本的渲染性能。

表 3: Quantitative evaluation results on DPG (Hu 等, 2024b).  
DPG 上的定量评估结果。

模型	全局	实体	属性	关系	其他	总分 ↑
SD v1.5 (Rombach 等, 2021)	74.63	74.23	75.39	73.49	67.81	63.18
PixArt- $\alpha$ (Chen 等, 2024c)	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next (Zhuo 等, 2024)	82.82	88.65	86.44	80.53	81.82	74.63
SDXL (Podell 等, 2023)	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 (Li 等, 2024a)	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT (Li 等, 2024b)	84.59	80.59	88.01	74.36	86.41	78.87
Janus (Wu 等, 2025a)	82.33	87.38	87.70	85.46	86.41	79.68
PixArt- $\Sigma$ (Chen 等, 2024b)	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen (Wang 等, 2024a)	85.21	86.68	86.84	90.22	83.15	80.60
Janus-Pro-1B (Chen 等, 2025b)	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] (BlackForest, 2024)	74.35	90.00	88.96	90.87	88.33	83.84
SD3 Medium (Esser 等, 2024)	87.90	91.01	88.83	80.70	88.68	84.08
Janus-Pro-7B (Chen 等, 2025b)	86.90	88.90	89.40	89.32	89.48	84.19
HiDream-I1-Full (Cai 等, 2025)	76.44	90.22	89.48	93.74	91.83	85.89
Lumina-Image 2.0 (Qin 等, 2025)	-	91.97	90.20	<b>94.85</b>	-	87.20
Seedream 3.0 (Gao 等, 2025)	<b>94.31</b>	<b>92.65</b>	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
<b>Qwen-Image</b>	91.32	91.56	<b>92.02</b>	94.31	<b>92.73</b>	<b>88.32</b>

**DPG** Table 3 shows the performance comparison on DPG (Hu 等, 2024b). This benchmark consists of 1K dense prompts, enabling fine-grained assessment of different aspects of prompt adherence. In general, Qwen-Image achieves the highest overall score, indicating its superior prompt-following capability. In particular, Qwen-Image excels at interpreting prompts involving attributes and other facets, outperforming all other models in the comparison.

表 3 展示了在 DPG (Hu 等, 2024b) 上的性能对比。该基准由 1K 个密集提示词组成, 能够对提示词遵循的不同方面进行细粒度评估。总体上, Qwen-Image 取得最高总分, 表明其优越的提示词遵循能力。特别是, Qwen-Image 在解释涉及属性和其他方面的提示词方面表现出色, 优于所有其他对比模型。

**GenEval** Table 4 presents a comparison of model performance on the GenEval (Ghosh 等, 2023) benchmark, which focuses on object-centric text-to-image generation using compositional prompts with diverse object attributes. We separately assess the performance of both the SFT model and the RL-enhanced model against other leading foundation models. Notably, our base model already surpasses the state of the art, outperforming Seedream 3.0 (Gao 等, 2025) and GPT Image 1 [High] (OpenAI, 2025). Following reinforcement learning (RL) fine-tuning, our model achieves an even higher score of 0.91, making it the only foundation model on the leaderboard to exceed the 0.9 threshold. These results demonstrate the superior controllable generation capabilities of Qwen-Image.

表 4 展示了在 GenEval (Ghosh 等, 2023) 基准上的模型性能对比, 该基准侧重于使用具有多种对象属性的组合式提示词的以物体为中心的文本到图像生成。我们分别评估了 SFT 模型和 RL 增强模型与其他领先基础模型的性能。值得注意的是, 我们的基础模型已经超越了最先进的水平, 优于 Seedream 3.0 (Gao 等, 2025) 和 GPT Image 1 [High] (OpenAI, 2025)。经过强化学习 (RL) 微调后, 我们的模型取得了更高的 0.91 分, 成为排行榜上唯一超过 0.9 阈值的基础模型。这些结果展示了 Qwen-Image 优越的可控生成能力。

表 4: Quantitative Evaluation results on GenEval (Ghosh 等, 2023).  
GenEval 上的定量评估结果。

模型	单 物体	双 物体	计数	颜色	位置	属性 绑定	总分 $\uparrow$
Show-o (Xie 等, 2024)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3-Gen (Wang 等, 2024a)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
PixArt- $\alpha$ (Chen 等, 2024c)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD3 Medium (Esser 等, 2024)	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1 [Dev] (BlackForest, 2024)	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD3.5 Large (Esser 等, 2024)	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow (Ma 等, 2025)	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Lumina-Image 2.0 (Qin 等, 2025)	-	0.87	0.67	-	-	0.62	0.73
Janus-Pro-7B (Chen 等, 2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream-I1-Full (Cai 等, 2025)	<b>1.00</b>	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 等, 2025)	0.99	<b>0.96</b>	0.91	<b>0.93</b>	0.47	0.80	0.84
<b>Qwen-Image</b>	0.99	0.92	0.89	0.88	0.76	0.77	<u>0.87</u>
<b>Qwen-Image-RL</b>	<b>1.00</b>	0.95	<b>0.93</b>	0.92	<b>0.87</b>	<b>0.83</b>	<b>0.91</b>

表 5: Quantitative evaluation results on OneIG-EN (Chang 等, 2025). The overall score is the average of the five dimensions.  
OneIG-EN 上的定量评估结果。总分为五个维度的平均值。

模型	对齐	文本	推理	风格	多样性	总分 $\uparrow$
Janus-Pro (Chen 等, 2025b)	0.553	0.001	0.139	0.276	0.365	0.267
BLIP3-o (Chen 等, 2025a)	0.711	0.013	0.223	0.361	0.229	0.307
BAGEL (Deng 等, 2025)	0.769	0.244	0.173	0.367	0.251	0.361
BAGEL+CoT (Deng 等, 2025)	0.793	0.020	0.206	0.390	0.209	0.324
Show-o2-1.5B (Xie 等, 2025b)	0.798	0.002	0.219	0.317	0.186	0.304
Show-o2-7B (Xie 等, 2025b)	0.817	0.002	0.226	0.317	0.177	0.308
OmniGen2 (Wu 等, 2025b)	0.804	0.680	0.271	0.377	0.242	0.475
SD 1.5 (Rombach 等, 2021)	0.565	0.010	0.207	0.383	<b>0.429</b>	0.319
SDXL (Podell 等, 2023)	0.688	0.029	0.237	0.332	0.296	0.316
SD3.5 Large (Esser 等, 2024)	0.809	0.629	0.294	0.353	0.225	0.462
FLUX.1 [Dev] (BlackForest, 2024)	0.786	0.523	0.253	0.368	0.238	0.434
CogView4 (Z.ai, 2025)	0.786	0.641	0.246	0.353	0.205	0.446
SANA-1.5 1.6B (PAG) (Xie 等, 2025a)	0.762	0.054	0.209	0.387	0.222	0.327
SANA-1.5 4.8B (PAG) (Xie 等, 2025a)	0.765	0.069	0.217	0.401	0.216	0.334
Lumina-Image 2.0 (Qin 等, 2025)	0.819	0.106	0.270	0.354	0.216	0.353
HiDream-I1-Full (Cai 等, 2025)	0.829	0.707	0.317	0.347	0.186	0.477
Imagen3 (Google, 2024)	0.843	0.343	0.313	0.359	0.188	0.409
Recraft V3 (Recraft, 2024)	0.810	0.795	0.323	0.378	0.205	0.502
Kolors 2.0 (team, 2025)	0.820	0.427	0.262	0.360	0.300	0.434
Seedream 3.0 (Gao 等, 2025)	0.818	0.865	0.275	0.413	0.277	0.530
Imagen4 (Google, 2025)	0.857	0.805	0.338	0.377	0.199	0.515
GPT Image 1 [High] (OpenAI, 2025)	0.851	0.857	<b>0.345</b>	<b>0.462</b>	0.151	0.533
<b>Qwen-Image</b>	<b>0.882</b>	<b>0.891</b>	0.306	0.418	0.197	<b>0.539</b>

**OneIG-Bench** Table. 5 and Table. 6 report the quantitative results on OneIG-Bench (Chang 等, 2025), a comprehensive benchmark designed for fine-grained evaluation of T2I models across multiple dimensions. For a fair overall comparison, we average the scores across all dimensions to obtain the final overall score. In general,

表 6: Quantitative evaluation results on OneIG-ZH (Chang 等, 2025). The overall score is the average of the five dimensions.

OneIG-ZH 上的定量评估结果。总分为五个维度的平均值。

模型	对齐	文本	推理	风格	多样性	总分 ↑
Janus-Pro (Chen 等, 2025b)	0.324	0.148	0.104	0.264	<b>0.358</b>	0.240
BLIP3-o (Chen 等, 2025a)	0.608	0.092	0.213	0.369	0.233	0.303
BAGEL (Deng 等, 2025)	0.672	0.365	0.186	0.357	0.268	0.370
BAGEL+CoT (Deng 等, 2025)	0.719	0.127	0.219	0.385	0.197	0.329
Cogview4 (Z.ai, 2025)	0.700	0.193	0.236	0.348	0.214	0.338
Lumina-Image 2.0 (Qin 等, 2025)	0.731	0.136	0.221	0.343	0.240	0.334
HiDream-I1-Full (Cai 等, 2025)	0.620	0.205	0.256	0.304	0.300	0.337
Kolors 2.0 (team, 2025)	0.738	0.502	0.226	0.331	0.333	0.426
Seedream 3.0 (Gao 等, 2025)	0.793	0.928	0.281	0.397	0.243	0.528
GPT Image 1 [High] (OpenAI, 2025)	0.812	0.650	<b>0.300</b>	<b>0.449</b>	0.159	0.474
<b>Qwen-Image</b>	<b>0.825</b>	<b>0.963</b>	0.267	0.405	0.279	<b>0.548</b>

Qwen-Image achieves the highest overall score on both the Chinese and English tracks, demonstrating its strong general-purpose generation capability. Notably, it ranks first in the Alignment and Text categories, evidencing its superior prompt following and text rendering capabilities.

表 5 和表 6 报告了在 OneIG-Bench (Chang 等, 2025) 上的定量结果，这是一个为 T2I 模型跨多维度细粒度评估而设计的综合基准。为了进行公平的整体对比，我们对所有维度的分数进行平均以获得最终总分。总体上，Qwen-Image 在中文和英文赛道上都取得最高总分，展示了其强大的通用生成能力。值得注意的是，它在对齐和文本类别中排名第一，证明了其优越的提示词遵循和文本渲染能力。

表 7: Quantitative evaluation results on TIIF Bench testmini (Wei 等, 2025). The best result is in bold and the second best result is underlined.

TIIF Bench testmini 上的定量评估结果。最优结果加粗，次优结果加下划线。

模型	总分		基础跟随								高级跟随										设计			
			平均		属性		关系		推理		平均		属性 + 关系		属性 + 推理		关系 + 推理		风格		文本		真实世界	
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
FLUX.1 [dev] (BlackForest, 2024)	71.09	71.78	83.12	78.65	87.05	83.17	87.25	80.39	75.01	72.39	65.79	68.54	67.07	73.69	73.84	73.34	69.09	71.59	66.67	66.67	43.83	52.83	70.72	71.47
FLUX.1 [Pro] (BlackForest, 2024)	67.32	69.89	79.08	78.91	78.83	81.33	82.82	83.82	75.57	71.57	61.10	65.37	62.32	65.57	69.84	71.47	65.96	67.72	63.00	63.00	35.83	55.83	71.80	68.80
DALL-E 3 (OpenAI, 2023)	74.96	70.81	78.72	78.50	79.50	79.83	80.82	78.82	75.82	76.82	73.39	67.27	73.45	67.20	72.01	71.34	63.59	60.72	89.66	86.67	66.83	54.83	72.93	60.99
SD 3 (Esser 等, 2024)	67.46	66.09	78.32	77.75	83.33	79.83	82.07	78.82	71.07	74.07	61.46	59.56	61.07	64.07	68.84	70.34	50.96	57.84	66.67	76.67	59.83	20.83	63.23	67.34
PixArt-Σ (Chen 等, 2024b)	62.00	58.12	70.66	75.25	69.33	78.83	75.07	77.32	67.57	69.57	57.65	49.50	65.20	56.57	66.96	61.72	66.59	54.59	83.33	70.00	1.83	1.83	62.11	52.41
Lumina-Next (Zhao 等, 2024)	50.93	52.46	64.58	66.08	56.83	59.33	67.57	71.82	69.32	67.07	44.75	45.63	51.44	43.20	51.09	59.72	44.72	54.46	70.00	66.67	0.00	0.83	47.56	49.05
Hunyuan-DiT (Li 等, 2024b)	51.38	53.28	69.33	69.00	65.83	69.83	78.07	73.82	64.07	63.32	42.62	45.45	50.20	41.57	59.22	61.84	47.84	51.09	56.67	73.33	0.00	0.83	40.10	44.20
Show-o (Xie 等, 2024)	59.72	58.86	73.08	75.83	74.83	79.83	78.82	78.32	65.57	69.32	53.67	50.38	60.95	56.82	68.59	68.96	66.46	56.22	63.33	66.67	3.83	2.83	55.02	50.92
LightGen (Wu 等, 2025c)	53.22	43.41	66.58	47.91	55.83	47.33	74.82	45.82	69.07	50.57	46.74	41.53	62.44	40.82	61.71	50.47	50.34	45.34	53.33	53.33	0.00	6.83	50.92	50.55
SANA 1.5 (Xie 等, 2025a)	67.15	65.73	79.66	77.08	79.83	77.83	85.57	83.57	73.57	69.82	61.50	60.67	65.32	56.57	69.96	73.09	62.96	65.84	80.00	80.00	17.83	15.83	71.07	68.83
Infinity (Han 等, 2025)	62.07	62.32	73.08	75.41	74.33	76.83	72.82	77.57	72.07	71.82	56.64	54.98	60.44	55.57	74.22	64.71	60.22	59.71	80.00	73.33	10.83	23.83	54.28	56.89
Janus-Pro-7B (Chen 等, 2025b)	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	80.32	79.07	59.71	58.82	66.07	56.20	70.46	70.84	67.22	59.97	60.00	70.00	28.83	33.83	65.84	60.25
MidJourney v7 (Midjourney, 2025)	68.74	65.69	77.41	76.00	77.58	81.83	82.07	76.82	72.57	69.32	64.66	60.53	67.20	62.70	<u>81.22</u>	71.59	60.72	64.59	83.33	80.00	24.83	20.83	68.83	63.61
Seedream 3.0 (Gao 等, 2025)	86.02	84.31	<u>87.07</u>	84.93	<u>90.50</u>	<u>90.00</u>	<b>89.85</b>	<u>85.94</u>	<u>80.86</u>	78.86	79.16	80.60	<u>79.76</u>	<u>81.82</u>	77.23	78.85	<u>75.64</u>	<u>78.64</u>	<b>100.00</b>	<u>93.33</u>	<b>97.17</b>	<u>87.78</u>	83.21	83.58
GPT Image 1 [High] (OpenAI, 2025)	<b>89.15</b>	<b>88.29</b>	<b>90.75</b>	<b>89.66</b>	<b>91.33</b>	87.08	84.57	84.57	<b>96.32</b>	<b>97.32</b>	<b>88.55</b>	<b>88.35</b>	<b>87.07</b>	<b>89.44</b>	<b>87.22</b>	<b>83.96</b>	<b>85.59</b>	<b>83.21</b>	<u>90.00</u>	<u>93.33</u>	89.83	86.83	<u>89.73</u>	<b>93.46</b>
Qwen-Image	<u>86.14</u>	<u>86.83</u>	86.18	<u>87.22</u>	<u>90.50</u>	<b>91.50</b>	<u>88.22</u>	<b>90.78</b>	79.81	<u>79.38</u>	<u>79.30</u>	<u>80.88</u>	79.21	78.94	78.85	<u>81.69</u>	75.57	78.59	<b>100.00</b>	<b>100.00</b>	<u>92.76</u>	<u>89.14</u>	<b>90.30</b>	<u>91.42</u>

**TIIF** Table 7 shows the performance comparison on TIIF Bench mini (Wei 等, 2025), a benchmark designed to systematically evaluate T2I model’s ability to interpret and follow intricate textual instructions. Overall, Qwen-Image ranks second, surpassed only by GPT Image 1 (OpenAI, 2025), underscoring its strong instruction-following capabilities.

表 7 展示了在 TIIF Bench mini (Wei 等, 2025) 上的性能对比，这是一个设计用于系统评估 T2I 模型解释和遵循复杂文本指令能力的基准。总体上，Qwen-Image 排名第二，仅次于 GPT Image 1 (OpenAI,



2025), 充分展示了其强大的指令遵循能力。

表 8: Quantitative evaluation results of English text rendering on CVTG-2K (Du 等, 2025).  
CVTG-2K 上英文文本渲染的定量评估结果。

模型	词准确率 ↑					NED↑	CLIPScore↑
	2 regions	3 regions	4 regions	5 regions	平均		
SD3.5 Large (Esser 等, 2024)	0.7293	0.6825	0.6574	0.5940	0.6548	0.8470	0.7797
FLUX.1 [dev] (BlackForest, 2024)	0.6089	0.5531	0.4661	0.4316	0.4965	0.6879	0.7401
AnyText (Tuo 等, 2024)	0.0513	0.1739	0.1948	0.2249	0.1804	0.4675	0.7432
TextDiffuser-2 (Chen 等, 2024a)	0.5322	0.3255	0.1787	0.0809	0.2326	0.4353	0.6765
RAG-Diffusion (Chen 等, 2024d)	0.4388	0.3316	0.2116	0.1910	0.2648	0.4498	0.7797
3DIS (Zhou 等, 2024)	0.4495	0.3959	0.3880	0.3303	0.3813	0.6505	0.7767
TextCrafter (Du 等, 2025)	0.7628	0.7628	0.7406	0.6977	0.7370	0.8679	0.7868
Seedream 3.0 (Gao 等, 2025)	0.6282	0.5962	0.6043	0.5610	0.5924	0.8537	0.7821
GPT Image 1 [High] (OpenAI, 2025)	<b>0.8779</b>	<b>0.8659</b>	<b>0.8731</b>	<b>0.8218</b>	<b>0.8569</b>	<b>0.9478</b>	<u>0.7982</u>
<b>Qwen-Image</b>	<u>0.8370</u>	<u>0.8364</u>	<u>0.8313</u>	<u>0.8158</u>	<u>0.8288</u>	<u>0.9116</u>	<b>0.8017</b>

**CVTG-2K** Table 8 reports the quantitative results of English rendering on CVTG-2K (Du 等, 2025). This benchmark contains 2K prompts, each requiring 2–5 regions of English to be rendered on the generated image. Three metrics: Word Accuracy, NED, CLIPScore are introduced to measure the accuracy of text rendering. As shown in the table, Qwen-Image achieves performance comparable to that of state-of-the-art image generation models, underscoring its powerful English text rendering capability.

表 8 报告了在 CVTG-2K (Du 等, 2025) 上的英文渲染定量结果。该基准包含 2K 个提示词，每个都需要在生成的图像上渲染 2-5 个英文区域。引入了三个指标：词准确率、NED 和 CLIPScore，以衡量文本渲染的准确度。如表所示，Qwen-Image 达到了与最先进的图像生成模型相当的性能，充分展示了其强大的英文文本渲染能力。

表 9: Quantitative comparison results of Chinese text rendering.  
中文文本渲染的定量对比结果。

模型	一级准确率	二级准确率	三级准确率	总分 ↑
Seedream 3.0 (Gao 等, 2025)	53.48	<u>26.23</u>	1.25	33.05
GPT Image 1 [High] (OpenAI, 2025)	<u>68.37</u>	15.97	<u>3.55</u>	<u>36.14</u>
<b>Qwen-Image</b>	<b>97.29</b>	<b>40.53</b>	<b>6.48</b>	<b>58.30</b>

**ChineseWord** Table 9 reports the quantitative results on ChineseWord, our new benchmark for character-level Chinese text rendering. In accordance with the List of Commonly Used Standard Chinese Characters, we group the characters into three difficulty tiers: Level-1 (3500 characters), Level-2 (3000 characters) and Level-3 (1605 characters). We craft several prompt templates that instruct text-to-image models to generate an image containing a single Chinese character. Across all three tiers, Qwen-Image attains the highest rendering accuracy, underscoring its superior ability to render Chinese text.

表 9 报告了在 ChineseWord（我们新推出的字符级中文文本渲染基准）上的定量结果。按照《常用汉字表》，我们将字符分为三个难度等级：1 级（3500 个字符）、2 级（3000 个字符）和 3 级（1605 个字符）。我们精心设计了多个提示词模板，指导文本到图像模型生成包含单个中文字符的图像。在所有三个等级上，Qwen-Image 都达到了最高的渲染准确率，充分体现了其优越的中文文本渲染能力。

**LongText-Bench** Table 10 reports the quantitative results on LongText-Bench (Geng 等, 2025), a benchmark specifically designed to assess a model’s ability to precisely render lengthy texts. The dataset contains 160 prompts spanning eight distinct scenarios. As shown in the table, Qwen-Image attains the highest accuracy on long Chinese text and the second-highest accuracy on long English text, illustrating Qwen-Image’s superior long text rendering capability.

表 10: Quantitative evaluation results on LongText-Bench (Geng 等, 2025). The best result is in bold and the second best result is underlined.  
LongText-Bench 上的定量评估结果。最优结果加粗，次优结果加下划线。

模型	LongText-Bench-EN	LongText-Bench-ZH
Janus-Pro (Chen 等, 2025b)	0.019	0.006
BLIP3-o (Chen 等, 2025a)	0.021	0.018
Kolors 2.0 (team, 2025)	0.258	0.329
BAGEL (Deng 等, 2025)	0.373	0.310
OmniGen2 (Wu 等, 2025b)	0.561	0.059
X-Omni (Geng 等, 2025)	0.900	0.814
HiDream-I1-Full (Cai 等, 2025)	0.543	0.024
FLUX.1 [Dev] (BlackForest, 2024)	0.607	0.005
Seedream 3.0 (Gao 等, 2025)	0.896	<u>0.878</u>
GPT Image 1 [High] (OpenAI, 2025)	<b>0.956</b>	0.619
<b>Qwen-Image</b>	<u>0.943</u>	<b>0.946</b>

表 10 报告了在 LongText-Bench (Geng 等, 2025) 上的定量结果，这个基准专门设计用于评估模型准确渲染长文本的能力。数据集包含跨越八个不同场景的 160 个提示词。如表所示，Qwen-Image 在长中文文本上达到最高准确率，在长英文文本上达到第二高准确率，充分展示了 Qwen-Image 优越的长文本渲染能力。

### 5.2.3 图像编辑性能

#### Performance of Image Editing

We further train a multi-task version of Qwen-Image for image editing (TI2I) tasks, seamlessly integrating both text and image as conditioning inputs. We evaluate our model across two categories of TI2I tasks: First, for general-purpose image editing, we assess the instruction-based editing capability of our model on the GEdit (Liu 等, 2025b) and ImgEdit (Ye 等, 2025) benchmarks. These benchmarks test the model’s ability to perform open-ended edits based on textual and visual instructions. Second, to evaluate the model’s spatial understanding and generative capability in 3D vision tasks, we benchmark its performance on novel view synthesis (Downs 等, 2022) and depth estimation (Bochkovskii 等, 2024). These tasks require the model to infer and generate coherent spatial information conditioned on input images and corresponding textual descriptions. All these tasks can be unified within the TI2I scope, showcasing the general applicability of our approach to diverse multimodal tasks.

我们进一步训练了 Qwen-Image 的多任务版本用于图像编辑 (TI2I) 任务，无缝集成文本和图像作为条件输入。我们在两类 TI2I 任务上评估了我们的模型：首先，对于通用目的的图像编辑，我们在 GEdit (Liu 等, 2025b) 和 ImgEdit (Ye 等, 2025) 基准上评估了模型的基于指令的编辑能力。这些基准测试了模型基于文本和视觉指令执行开放式编辑的能力。其次，为了评估模型在 3D 视觉任务中的空间理解和生成能力，我们在新视角合成 (Downs 等, 2022) 和深度估计 (Bochkovskii 等, 2024) 上对其性能进行了基准测试。这些任务要求模型在输入图像和文本描述的约束下，推断并生成连贯的空间信息。所有这些任务都可以统一在 TI2I 范围内，展示了我们方法对多种多模态任务的通用适用性。

**GEdit** Table 11 reports the results on the GEdit-Bench (Liu 等, 2025b), which evaluates image editing models on real-world user instructions across 11 diverse categories. We adopt three metrics—Semantic Consistency (SQ), Perceptual Quality (PQ), and Overall Score (O)—each ranging from 0 to 10. Qwen-Image ranks at the top of both the English and Chinese leaderboards, demonstrating strong editing capability and generalization to multilingual user instructions.

表 11 报告了 GEdit-Bench (Liu 等, 2025b) 的结果，该基准在 11 个不同类别的真实用户指令上评估图像编辑模型。我们采用三个指标——语义一致性 (SQ)、感知质量 (PQ) 和总分 (O)——每个都在 0 到 10 之间。Qwen-Image 在英文和中文排行榜上都排名第一，展示了强大的编辑能力和对多语言用户指令的泛化能力。

表 11: Comparison of Semantic Consistency (G\_SC), Perceptual Quality (G\_PQ), and Overall Score (G\_O) on GEdit-Bench. All metrics are evaluated by GPT-4.1. G\_O is computed as the geometric mean of G\_SC and G\_PQ, averaged over all samples. Note: FLUX.1 Kontext [Pro] underperforms on GEdit-Bench-CN due to its limited Chinese language capability.

GEdit-Bench 上语义一致性 (G\_SC)、感知质量 (G\_PQ) 和总分 (G\_O) 的对比。所有指标由 GPT-4.1 评估。G\_O 为 G\_SC 和 G\_PQ 的几何平均值，取所有样本的平均。注：FLUX.1 Kontext [Pro] 在 GEdit-Bench-CN 上表现较差，因为其中文能力有限。

模型	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Instruct-Pix2Pix (Brooks 等, 2023)	3.58	5.49	3.68	-	-	-
AnyEdit (Yu 等, 2025)	3.18	5.82	3.21	-	-	-
MagicBrush (Zhang 等, 2023)	4.68	5.66	4.52	-	-	-
UniWorld-v1 (Lin 等, 2025)	4.93	7.43	4.85	-	-	-
OmniGen (Xiao 等, 2025)	5.96	5.89	5.06	-	-	-
OmniGen2 (Wu 等, 2025b)	7.16	6.77	6.41	-	-	-
Gemini 2.0 (DeepMind, 2025)	6.73	6.61	6.32	5.43	6.78	5.36
BAGEL (Deng 等, 2025)	7.36	6.83	6.52	7.34	6.85	6.50
FLUX.1 Kontext [Pro] (Labs 等, 2025)	7.02	7.60	6.56	1.11	7.36	1.23
Step1X-Edit (Liu 等, 2025b)	7.66	7.35	6.97	7.20	6.87	6.86
GPT Image 1 [High] (OpenAI, 2025)	<u>7.85</u>	<u>7.62</u>	<u>7.53</u>	<u>7.67</u>	<u>7.56</u>	<u>7.30</u>
<b>Qwen-Image</b>	<b>8.00</b>	<b>7.86</b>	<b>7.56</b>	<b>7.82</b>	<b>7.79</b>	<b>7.52</b>

表 12: Quantitative comparison results on ImgEdit (Ye 等, 2025). All metrics are evaluated by GPT-4.1. “Overall” is calculated by averaging all scores across tasks.

ImgEdit 上的定量对比结果。所有指标由 GPT-4.1 评估。” 总分” 通过对所有任务的分数求平均得到。

模型	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	总分 ↑
MagicBrush (Zhang 等, 2023)	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix (Brooks 等, 2023)	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit (Yu 等, 2025)	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit (Zhao 等, 2024)	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen (Xiao 等, 2025)	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit (Zhang 等, 2025)	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit (Liu 等, 2025b)	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL (Deng 等, 2025)	3.56	3.31	1.70	3.3	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1 (Lin 等, 2025)	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 (Wu 等, 2025b)	3.57	3.06	1.77	3.74	3.20	3.57	<u>4.81</u>	2.52	4.68	3.44
FLUX.1 Kontext [Pro] (Labs 等, 2025)	4.25	4.15	2.35	<u>4.56</u>	3.57	4.26	4.57	3.68	4.63	4.00
GPT Image 1 [High] (OpenAI, 2025)	<u>4.61</u>	<u>4.33</u>	<u>2.90</u>	4.35	<u>3.66</u>	<b>4.57</b>	<b>4.93</b>	<b>3.96</b>	<b>4.89</b>	<u>4.20</u>
<b>Qwen-Image</b>	<u>4.38</u>	<u>4.16</u>	<b>3.43</b>	<b>4.66</b>	<b>4.14</b>	<u>4.38</u>	<u>4.81</u>	<u>3.82</u>	<u>4.69</u>	<b>4.27</b>

**ImgEdit** Table 12 presents the results on the **ImgEdit** benchmark (Ye 等, 2025), which covers nine common editing tasks across diverse semantic categories with a total of 734 real-world test cases. Evaluation metrics include instruction adherence, image-editing quality, and detail preservation, all scored from 1 to 5. Qwen-Image ranks highest overall, closely followed by GPT Image 1 [High] and demonstrating competitive instruction-based editing performance.

表 12 展示了 **ImgEdit** 基准 (Ye 等, 2025) 的结果，该基准涵盖跨多个语义类别的九项常见编辑任务，共 734 个真实测试用例。评估指标包括指令遵循度、图像编辑质量和细节保留，所有指标的评分范围都是 1 到 5。Qwen-Image 排名整体最高，紧随其后的是 GPT Image 1 [High]，展示了极具竞争力的基于指令的编辑性能。

**新视角合成 (Novel view synthesis)** Table 13 shows the results of novel view synthesis on GSO (Downs 等, 2022) dataset. We compare the similarity of the generated novel view image of a 3D object given its front view with the ground truth image. We prompt Qwen-Image with prompts like “turn left 90 degrees, a dog” to instruct the model to perform novel view synthesis. Qwen-Image demonstrates highly competitive results among the baselines, achieving state-of-the-art performance of novel view synthesis.

表 13 展示了在 GSO (Downs 等, 2022) 数据集上的新视角合成结果。我们对比了给定 3D 对象正视图生成的新视角图像与真实标注图像的相似度。我们使用” 向左转 90 度，一只狗” 这样的提示词来指导



表 13: Quantitative comparison of novel view synthesis with both specialized models and general image generation models. We report PSNR, SSIM, LPIPS on the GSO (Downs 等, 2022) dataset.  
专用模型与通用图像生成模型在新视角合成上的定量对比。在 GSO 数据集上报告 PSNR、SSIM、LPIPS 指标。

模型	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Zero123 (Liu 等, 2023)	13.48	0.854	0.166
ImageDream (Wang & Shi, 2023)	15.22	0.883	0.164
CRM (Wang 等, 2024b)	<b>15.93</b>	<b>0.891</b>	<b>0.152</b>
GPT Image 1 [High] (OpenAI, 2025)	12.07	0.804	0.361
BAGEL (Deng 等, 2025)	13.78	0.825	0.237
FLUX.1 Kontext [Pro] (Labs 等, 2025)	14.50	0.859	0.201
<b>Qwen-Image</b>	<b>15.11</b>	<b>0.884</b>	<b>0.153</b>

表 14: Quantitative comparison of depth estimation with both specialized models and multi-task models on zero-shot datasets. Qwen-Image can perform *on par* with state-of-the-art models.  
深度估计中专用模型和多任务模型的定量对比（零样本数据集）。Qwen-Image 能与最先进的模型相媲美。

模型	KITTI		NYUv2		ScanNet		DIODE		ETH3D	
	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$	AbsRel $\downarrow$	$\delta_1\uparrow$
MiDaS (Ranftl 等, 2020)	0.236	0.630	0.111	0.885	0.121	0.846	0.332	0.715	0.184	0.752
DPT-large (Ranftl 等, 2021)	0.100	0.901	0.098	0.903	0.082	0.934	0.182	0.758	0.078	0.946
DepthAnything (Yang 等, 2024a)	0.080	0.946	0.043	0.980	0.043	0.981	0.261	0.759	0.058	<b>0.984</b>
DepthAnything v2 (Yang 等, 2024b)	0.080	0.943	0.043	0.979	0.042	0.979	0.321	0.758	0.066	0.983
Depth Pro (Bochkovskii 等, 2024)	0.055	0.974	0.042	0.977	0.041	0.978	0.217	0.764	0.043	0.974
Metric3D v2 (Hu 等, 2024a)	<b>0.052</b>	<b>0.979</b>	<b>0.039</b>	<b>0.979</b>	<b>0.023</b>	<b>0.989</b>	<b>0.147</b>	<b>0.892</b>	<b>0.040</b>	0.983
GeoWizard (Fu 等, 2024)	0.097	0.921	<b>0.052</b>	<u>0.966</u>	<u>0.061</u>	<u>0.953</u>	0.297	0.792	<b>0.064</b>	<u>0.961</u>
DepthFM (Gui 等, 2024)	<u>0.083</u>	<u>0.934</u>	0.065	0.956	-	-	0.225	0.800	-	-
Marigold (Ke 等, 2024)	0.099	0.916	0.055	0.964	0.064	0.951	0.308	0.773	<u>0.065</u>	0.960
DMP (Lee 等, 2024)	0.240	0.622	0.109	0.891	0.146	0.814	0.361	0.706	0.128	0.857
<b>Qwen-Image</b>	<b>0.078</b>	<b>0.951</b>	<u>0.055</u>	<b>0.967</b>	<b>0.047</b>	<b>0.974</b>	<b>0.197</b>	<b>0.832</b>	0.066	<b>0.962</b>

模型执行新视角合成。Qwen-Image 在基线模型中展示了极具竞争力的结果，在新视角合成上达到了最先进的性能。

**深度估计(Depth Estimation)** Table 14 summarizes performance on five widely used datasets: NYUv2 (Nathan Silberman & Fergus, 2012), KITTI (Geiger 等, 2013), ScanNet (Dai 等, 2017), DIODE (Vasiljevic 等, 2019), and ETH3D (Schops 等, 2017). During training, we adopt DepthPro (Bochkovskii 等, 2024) as the teacher model to provide supervisory depth signals, following the protocol used in previous work such as DICEPTION (Zhao 等, 2025). Notably, these results are achieved with standalone supervised fine-tuning (SFT), in order to probe the model’s intrinsic task understanding capability. Qwen-Image demonstrates highly competitive results among diffusion-based models, achieving state-of-the-art performance on several key metrics across these benchmarks.

表 14 总结了在五个广泛使用的数据集上的性能：NYUv2 (Nathan Silberman & Fergus, 2012)、KITTI (Geiger 等, 2013)、ScanNet (Dai 等, 2017)、DIODE (Vasiljevic 等, 2019) 和 ETH3D (Schops 等, 2017)。在训练过程中，我们采用 DepthPro (Bochkovskii 等, 2024) 作为教师模型来提供监督信号，遵循了 DICEPTION (Zhao 等, 2025) 等先前工作所使用的协议。值得注意的是，这些结果是通过独立的有监督微调 (SFT) 实现的，以探测模型的内在任务理解能力。Qwen-Image 在基于扩散的模型中展示了极具竞争力的结果，在这些基准上的多个关键指标上达到了最先进的性能。

### 5.3 定性结果

#### Qualitative Results

为了全面比较 Qwen-Image 和最先进模型的视觉生成能力，我们首先在 §§ 5.3.1 中定性比较了 VAE 在文本丰富图像上的重建性能。我们进一步在两个基本视觉生成任务上进行了定性比较，包括 §§ 5.3.2 中的文本到图像 (T2I) 生成和 §§ 5.3.3 中的图像编辑 (TI2I)，以全面评估模型的基本生成能力。

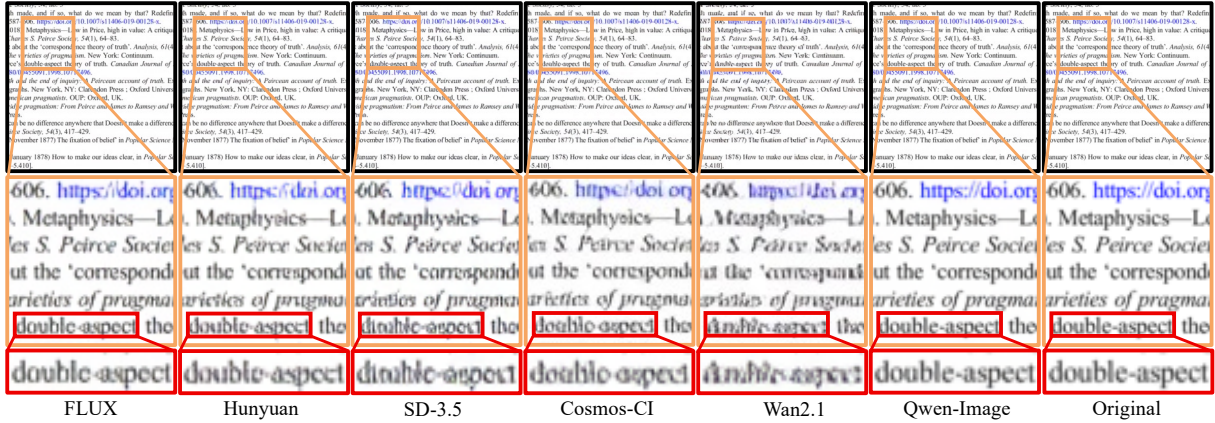


图 17: Visualization of VAE reconstruction. We progressively zoom into the details across three rows (black, orange, red) to compare how different VAEs reconstruct small text in dense document images. VAE 重建的可视化。我们逐步放大三行（黑色、橙色、红色）的细节，以比较不同的 VAE 如何在密集文档图像中重建小文本。

### 5.3.1 VAE 重建的定性结果

#### Qualitative Results on VAE Reconstruction

图 17 展示了用最先进的图像 VAE 重建文本丰富图像的定性结果。第一行展示了包含英文文本的 PDF 图像的重建。在我们的结果中，短语“double-aspect”仍然清晰可读，而在其他模型的重建中则无法识别。总体上，Qwen-Image-VAE 为包含小文本的图像提供了更精确的重建。

### 5.3.2 图像生成的定性结果

#### Qualitative Results on Image Generation

In order to comprehensively evaluate Qwen-Image’s text-to-image generation capability, we conduct qualitative evaluation from four aspects: English Text Rendering, Chinese Text Rendering, Multi-Object Generation, and Spatial Relationship Generation. For comparative analysis, we benchmark our model against leading Text-to-Image foundation models, including both close-source models (GPT Image 1 [High] (OpenAI, 2025), Seedream 3.0 (Gao 等, 2025), Recraft V3 (Recraft, 2024) and open-source models (Hidream-I1-Full (Cai 等, 2025) and Lumina-Image 2.0 (Qin 等, 2025)).

为全面评估 Qwen-Image 的文本到图像生成能力，我们从四个方面进行定性评估：英文文本渲染、中文文本渲染、多物体生成和空间关系生成。为了进行对比分析，我们将我们的模型与主流文本到图像基础模型进行基准对比，包括闭源模型 (GPT Image 1 [High] (OpenAI, 2025)、Seedream 3.0 (Gao 等, 2025)、Recraft V3 (Recraft, 2024)) 和开源模型 (Hidream-I1-Full (Cai 等, 2025) 和 Lumina-Image 2.0 (Qin 等, 2025))。

**英文文本渲染 (English Text Rendering)** Figure 18 and Figure 19 show the qualitative comparison of English text rendering. As shown in Figure 18, Qwen-Image achieves a more realistic visual style and better rendering quality for a long English paragraph. With respect to text rendering, our model demonstrates greater fidelity to the given prompts, effectively avoiding issues such as missing, wrong, or duplicate characters (e.g. wrong “lantern” and “Unfurling” in GPT Image 1 [High], wrong “silver” and “quiet” in Recraft V3, redundant and distorted text in Seedream 3.0). In the upper half of Figure 19, Qwen-Image correctly renders text in seven different locations, demonstrating its complex text rendering capabilities. In contrast, GPT Image 1 misses “The night circus”, and the text rendered by Seedream 3.0 and Hidream-I1-Full is distorted. For the bottom half of Figure 19, Qwen-Image not only successfully renders each text segment, but also presents a slide with reasonable layout and visually aesthetic. In comparison, GPT Image 1 misses “Stay Connected”, Hidream-I1-Full and seedream 3.0 fail to render correct characters.

图 18 和图 19 展示了英文文本渲染的定性对比。如图 18 所示，Qwen-Image 为长英文段落实现了更真实的视觉风格和更好的渲染质量。在文本渲染方面，我们的模型表现出对给定提示词的更高保真度，有效避免了遗漏、错误或重复字符的问题（如 GPT Image 1 [High] 中错误的”lantern”和”Unfurling”、Recraft V3 中错误的”silver”和”quiet”、Seedream 3.0 中冗余和扭曲的文本）。在图 19 的上半部分，Qwen-Image 正确渲染了七个不同位置的文本，展示了其复杂文本渲染能力。相比之下，GPT Image 1 缺少了”The night circus”，Seedream 3.0 和 Hidream-I1-Full 渲染的文本是扭曲的。对于图 19 的下半部分，Qwen-Image 不仅成功渲染了每个文本段落，还呈现了具有合理布局和视觉美感的幻灯片。相比之下，GPT Image 1 缺少了”Stay Connected”，Hidream-I1-Full 和 Seedream 3.0 无法渲染正确的字符。

**中文文本渲染 (Chinese Text Rendering)** Figure 20 and Figure 21 show a qualitative comparison of Chinese text rendering. In Figure 20, Qwen-Image accurately generates the expected Chinese couplet, faithfully reproducing the content and style of the text, and accurately depicts the required room layout and placement. In contrast, GPT Image 1 and Seedream 3.0 miss or generate distorted characters (missed ”远” and ”善” in GPT Image 1, missed ”智” and ”机” in Seedream 3.0), while other models cannot generate correct Chinese couplets. In Figure 21, the upper part shows an anime style scene, where Qwen-Image can correctly generate multiple characters and store plaques, perfectly following the requirements of input prompt for spatial layout and text rendering. Seedream 3.0 encounters difficulties in complex spatial layouts, missing some scenes and characters, while other models cannot correctly understand complex text and spatial instructions. The bottom part shows that Qwen-Image can generate realistic and beautifully typeset handwritten text in complex scenes, which perfectly following the input prompt, while other models struggle to generate structured paragraph text.

图 20 和图 21 展示了中文文本渲染的定性对比。在图 20 中，Qwen-Image 准确生成了中文对联，忠实再现了文本的内容和风格，准确描绘了所需的房间布局。相比之下，GPT Image 1 和 Seedream 3.0 遗漏或生成了扭曲的字符（GPT Image 1 中遗漏了”远”和”善”，Seedream 3.0 中遗漏了”智”和”机”），其他模型更是无法生成正确的中文对联。在图 21 的上半部分，Qwen-Image 在动漫风格场景中正确生成了多个字符和商店牌匾，严格遵循了输入提示词对空间布局和文本渲染的要求。而 Seedream 3.0 在复杂空间布局中表现困难，其他模型也无法正确理解复杂的文本和空间指令。下半部分则展示了 Qwen-Image 能在复杂场景中生成排版精美的手写文本，忠实遵循输入提示词，而其他模型难以生成结构化段落文本。

**多物体生成 (Multi-Object Generation)** As shown in the upper half of Figure 22, Qwen-Image accurately generates all required animals, faithfully preserves their specified positions, and consistently applies the correct plush style. In contrast, GPT Image 1 fails to generate images in the plush style, while Recraft V3 and Seedream 3.0 produce incorrect animals that do not match the prompt. For the bottom part of Figure 22, Qwen-Image not only correctly renders mixed-language texts on the billiards, but also strictly follows the instruction to arrange the billiards in two rows. GPT Image 1 cannot perfectly follow the layout requirements of instruction and also incorrectly generate a Chinese character ”發”, while other models cannot correctly generate most Chinese characters.

如图 22 的上半部分所示，Qwen-Image 准确生成了所有需要的动物，忠实保留了它们指定的位置，并始终采用了正确的毛绒风格。相比之下，GPT Image 1 无法生成毛绒风格图像，而 Recraft V3 和 Seedream 3.0 生成的动物与提示词不符。对于下半部分，Qwen-Image 不仅正确渲染了台球上的混合语言文本，还严格按指令将台球排列成两行。GPT Image 1 无法完美遵循布局要求，还生成了错误的中文字符”發”，其他模型更是难以正确生成大多数中文字符。

**空间关系生成 (Spatial Relationship Generation)** In the first part of Figure 23, Qwen-Image generates an image that accurately reflects the prompt, capturing both the correct climbing scene and the specified interaction between the two people. In contrast, GPT Image 1, Seedream 3.0, and Recraft V3 fail to fully follow the prompt: these models produce incorrect interactions between the climbers. This comparison demonstrates the strong ability of Qwen-Image to understand and precisely follow complex prompts. For the remaining two parts, only Qwen-Image and GPT Image 1 can accurately depict the spatial relationship between the character and the pigeon, as well as the pocket watch and the cup handle.



在图 23 的第一部分中，Qwen-Image 生成的图像准确反映了提示词，捕捉了正确的攀登场景和两个人之间指定的交互。相比之下，GPT Image 1、Seedream 3.0 和 Recraft V3 未能完全遵循提示词：这些模型在登山者之间产生了不正确的交互。这个对比展示了 Qwen-Image 理解和精确遵循复杂提示词的强大能力。对于其余两部分，只有 Qwen-Image 和 GPT Image 1 可以准确描绘人物与鸽子之间以及怀表与杯柄之间的空间关系。

### 5.3.3 图像编辑的定性结果

#### Qualitative Results on Image Editing

To comprehensively assess the image editing (TI2I) capability of Qwen-Image, we conduct a qualitative evaluation focusing on five key aspects: text and material editing, object addition/removal/replacement, pose manipulation, chained editing, and novel view synthesis. For comparative analysis, we benchmark our model against several leading instruction-based image editing models, including SeedEdit 3.0 (Wang 等, 2025)<sup>5</sup>, FLUX.1 Kontext [Pro] (Labs 等, 2025), and GPT Image 1 [High] (OpenAI, 2025).

为全面评估 Qwen-Image 的图像编辑 (TI2I) 能力，我们进行了定性评估，重点关注五个关键方面：文本和材质编辑、物体添加/移除/替换、姿态操控、链式编辑和新视角合成。为了进行对比分析，我们将我们的模型与多个主流的基于指令的图像编辑模型进行基准对比，包括 SeedEdit 3.0 (Wang 等, 2025)<sup>6</sup>、FLUX.1 Kontext [Pro] (Labs 等, 2025) 和 GPT Image 1 [High] (OpenAI, 2025)。

**文本和材质编辑(Text and Material Editing)** Figure 24 presents a qualitative comparison of text and material editing capabilities. In the top example, which involves editing text with a complex style, Seedream 3.0 fails to change the letter "H" to "Q", and GPT Image 1 [High] does not preserve the original style. Both Qwen-Image and FLUX.1 Kontext [Pro] successfully modify the text while maintaining style consistency. In the lower example, all models except FLUX.1 Kontext [Pro] accurately add the required text and related elements. Notably, Qwen-Image is the only model that correctly generates the requested enamel colored glass art, demonstrating superior material rendering and instruction-following ability.

图 24 展示了文本和材质编辑能力的定性比较。在涉及编辑复杂风格文本的上面示例中，Seedream 3.0 未能将字母“H”改为“Q”，GPT Image 1 [High] 没有保留原始风格。Qwen-Image 和 FLUX.1 Kontext [Pro] 都成功修改了文本，同时保持了风格一致性。在下面的示例中，除了 FLUX.1 Kontext [Pro] 外，所有模型都准确添加了所需的文本和相关元素。值得注意的是，Qwen-Image 是唯一正确生成所要求的珐琅彩色玻璃艺术品的模型，展示了优越的材质渲染和指令遵循能力。

**物体添加/移除/替换 (Object Addition/Removal/Replacement)** Object addition, removal, and replacement are among the most common tasks in instruction-based image editing. In Figure 25, we compare the performance of various models on challenging real-world scenarios. With the exception of GPT Image 1 [High], which often fails to maintain overall image consistency, all other models generally perform well in preserving unedited regions. In the top case, where the task is to add both a cat and a dog in a cartoon style, the model must ensure that the new objects match the overall artistic style. We observe that FLUX.1 Kontext [Pro] struggles with consistency when editing non-photorealistic images, while both SeedEdit 3.0 and Qwen-Image produce coherent results that align well with the desired cartoon style. For the middle example, which involves removing all people from a crowded scene—a relatively complex instruction—all models complete the task accurately, differing only in minor details. Additionally, we notice occasional zoom-in and zoom-out effects across different models during editing.

物体添加、移除和替换是基于指令的图像编辑中最常见的任务。在图 25 中，我们比较了各种模型在具有挑战性的真实场景上的性能。除了 GPT Image 1 [High]（经常无法保持整体图像一致性）外，所有其他模型通常在保留未编辑区域方面表现良好。在顶部案例中，任务是以卡通风格添加猫和狗，模型必须确保新对象与整体艺术风格匹配。我们观察到 FLUX.1 Kontext [Pro] 在编辑非写实图像时难以保持一致性，而 SeedEdit 3.0 和 Qwen-Image 都产生了与所需卡通风格相符的连贯结果。对于涉及从拥挤场景

<sup>5</sup>SeedEdit 3.0 does not provide an official API, all watermarked images were obtained via their [web interface](#).

<sup>6</sup>SeedEdit 3.0 没有提供官方 API，所有带水印的图像都是通过他们的 [网页界面](#) 获得的。

---

中移除所有人物的中间示例（相对复杂的指令），所有模型都准确完成了任务，仅在细节上有所不同。此外，我们注意到不同模型在编辑过程中偶尔会出现放大和缩小效果。

**姿态操控 (Pose Manipulation)** Figure 26 presents a qualitative comparison of pose manipulation abilities across different models. In the first example, only FLUX.1 Kontext [Pro] and Qwen-Image are able to preserve fine details such as the subject’s hair strands during pose editing. In the second case, which requires maintaining clothing consistency and scene stability during pose changes, Qwen-Image is the only model that keeps both the background and character unchanged. Remarkably, Qwen-Image accurately infers from the input that the person is wearing a side-slit dress over silk trousers, and faithfully reveals the silk trousers in the standing pose. In the third example, Qwen-Image again outperforms other models by better preserving the original pose and maintaining consistency in clothing decorations.

图 26 展示了不同模型姿态操控能力的定性比较。在第一个示例中，只有 FLUX.1 Kontext [Pro] 和 Qwen-Image 能够在姿态编辑过程中保留精细节，如被摄体的头发丝。在第二个案例中，需要在姿态变化过程中保持服装一致性和场景稳定性，Qwen-Image 是唯一保持背景和人物都未改变的模型。值得注意的是，Qwen-Image 准确推断出输入中的人物穿着开叉连衣裙覆盖丝绸裤，并在站立姿态中忠实地呈现了丝绸裤。在第三个示例中，Qwen-Image 再次优于其他模型，更好地保留了原始姿态，并保持了服装装饰的一致性。

**链式编辑 (Chained Editing)** Chained editing refers to scenarios where generated images are iteratively used as context for subsequent editing steps. In the first case of Figure 27, the task requires extracting a clothing item and depicting the close-up of its fabric details. We select a Chinese traditional painting as the source image. SeedEdit 3.0 and FLUX.1 Kontext [Pro] fail from the first prompt, while both GPT Image 1 [High] and Qwen-Image accurately extract the paired birds. Qwen-Image can better preserve fine texture details against GPT Image 1 [High]. In the second case, the input image features a boat with a double-opening stern. Both Qwen-Image and FLUX.1 Kontext [Pro] are able to preserve this structural feature throughout the entire chained editing process. However, FLUX.1 Kontext [Pro] fails to add two cargo ships as instructed, whereas Qwen-Image successfully completes the complete editing chain.

链式编辑是指生成的图像被迭代地用作后续编辑步骤的上下文的场景。在图 27 的第一个案例中，任务需要提取服装项目并描绘其面料细节的特写。我们选择了一幅中国传统绘画作为源图像。SeedEdit 3.0 和 FLUX.1 Kontext [Pro] 从第一个提示词就失败了，而 GPT Image 1 [High] 和 Qwen-Image 都准确提取了成对的鸟。Qwen-Image 可以比 GPT Image 1 [High] 更好地保留精细纹理细节。在第二个案例中，输入图像呈现一艘具有双开船尾的船。Qwen-Image 和 FLUX.1 Kontext [Pro] 都能够在整个链式编辑过程中保留这个结构特征。然而，FLUX.1 Kontext [Pro] 未能按指令添加两艘货船，而 Qwen-Image 成功完成了整个编辑链。

**新视角合成 (Novel View Synthesis)** Figure 28 evaluates the spatial reasoning and novel view synthesis capabilities of different models. SeedEdit 3.0 and FLUX.1 Kontext [Pro] cannot perform view rotation well under the same instruction. While GPT Image 1 [High] can generate new perspectives when a clear subject is present, it fails to generalize to real-world scenes with complex multiple objects. Only Qwen-Image maintains global consistency—including text fidelity and lighting structure—demonstrating superior spatial and semantic coherence in complex editing tasks.

图 28 评估了不同模型的空间推理和新视角合成能力。SeedEdit 3.0 和 FLUX.1 Kontext [Pro] 在相同指令下无法很好地执行视角旋转。虽然 GPT Image 1 [High] 在有清晰对象时可以生成新的视角，但它无法推广到具有复杂多个对象的真实场景。只有 Qwen-Image 保持全局一致性——包括文本保真度和光照结构——在复杂编辑任务中展示了优越的空间和语义一致性。

## Input prompt

A man in a suit is standing in front of the window, looking at the bright moon outside the window. The man is holding a yellowed paper with handwritten words on it: "A lantern moon climbs through the silver night, Unfurling quiet dreams across the sky, Each star a whispered promise wrapped in light, That dawn will bloom, though darkness wanders by." There is a cute cat on the windowsill.

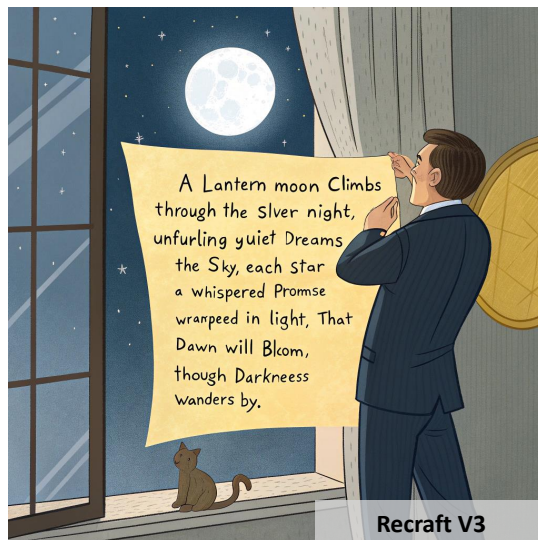
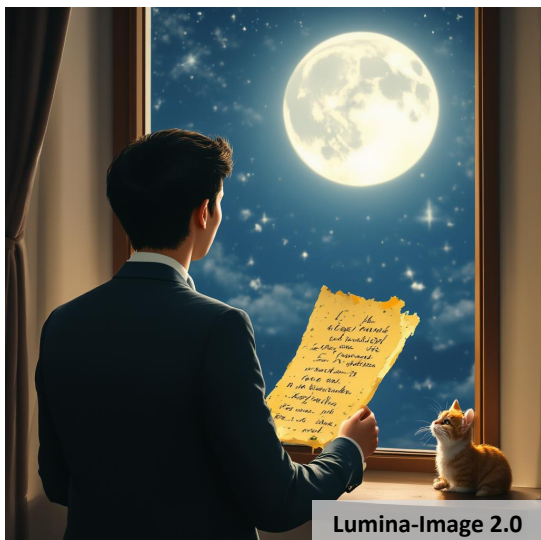


图 18: Comparison of long English rendering capability in image generation. This case requires rendering a long paragraph, and only Qwen-Image and GPT Image 1 [High] manage to render such a long text clearly and almost perfectly. The other models either omit words or produce duplicates.

图像生成中长英文渲染能力的比较。这个案例需要渲染一个长段落，只有 Qwen-Image 和 GPT Image 1 [High] 能够清晰且几近完美地渲染这样长的文本。其他模型要么省略词汇，要么产生重复。



**Input prompt**

Bookstore window display. A sign displays "New Arrivals This Week". Below, a shelf tag with the text "Best-Selling Novels Here". To the side, a colorful poster advertises "Author Meet And Greet on Saturday" with a central portrait of the author. There are four books on the bookshelf, namely "The light between worlds" "When stars are scattered" "The silent patient" "The night circus"



HiDream-I1-Full



GPT Image 1 [High]



Seedream 3.0



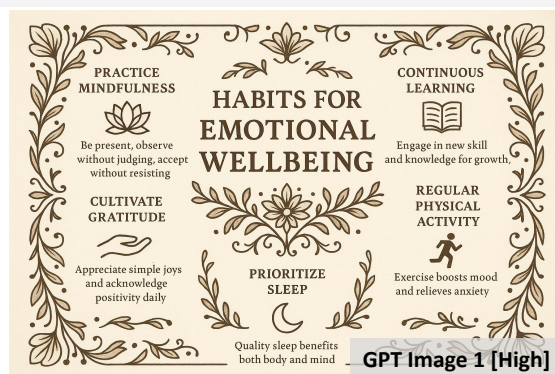
Qwen-Image

**Input prompt**

A slide featuring artistic, decorative shapes framing neatly arranged textual information styled as an elegant infographic. At the very center, the title "Habits for Emotional Wellbeing" appears clearly, surrounded by a symmetrical floral pattern. On the left upper section, "Practice Mindfulness" appears next to a minimalist lotus flower icon, with the short sentence, "Be present, observe without judging, accept without resisting". Next, moving downward, "Cultivate Gratitude" is written near an open hand illustration, along with the line, "Appreciate simple joys and acknowledge positivity daily". Further down, towards bottom-left, "Stay Connected" accompanied by a minimalist chat bubble icon reads "Build and maintain meaningful relationships to sustain emotional energy". At bottom right corner, "Prioritize Sleep" is depicted next to a crescent moon illustration, accompanied by the text "Quality sleep benefits both body and mind". Moving upward along the right side, "Regular Physical Activity" is near a jogging runner icon, stating: "Exercise boosts mood and relieves anxiety". Finally, at the top right side, appears "Continuous Learning" paired with a book icon, stating "Engage in new skill and knowledge for growth". The slide layout beautifully balances clarity and artistry, guiding the viewers naturally along each text segment.



HiDream-I1-Full



GPT Image 1 [High]



Seedream 3.0



Qwen-Image

图 19: Comparison of complex English rendering capability in image generation. We present two cases to illustrate the model's ability to generate multiple English texts in different locations of the real scene and the slide. Only Qwen-image can follow the complex prompts to successfully render the text in reasonable location. 图像生成中复杂英文渲染能力的比较。我们呈现两个案例来展示模型在真实场景和幻灯片的不同位置生成多个英文文本的能力。只有 Qwen-Image 能够遵循复杂提示词在合理位置成功渲染文本。



## Input prompt

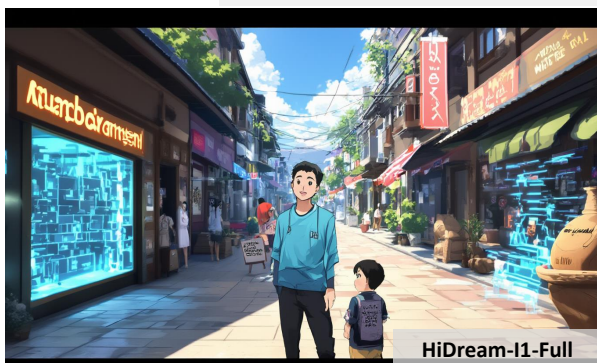
一副典雅庄重的对联悬挂于厅堂之中，房间是个安静古典的中式布置，桌子上放着一些青花瓷，对联上左书“义本生知人机同道善思新”，右书“通云赋智乾坤启数高志远”，横批“智启通义”，字体飘逸，中间挂在一着一副中国风的画作，内容是岳阳楼。



图 20: Comparison of Chinese text rendering capability in image generation. Qwen-Image accurately generates the expected Chinese couplet. In contrast, GPT Image 1 [high] and Seedream 3.0 miss or generate distorted characters. While other model in comparison cannot generate correct Chinese couplets. 图像生成中中文文本渲染能力的比较。Qwen-Image 准确生成了预期的中文对联。相比之下，GPT Image 1 [High] 和 Seedream 3.0 遗漏或生成了扭曲的字符。而其他对比模型无法生成正确的中文对联。



**Input prompt** 宫崎骏的动漫风格。平视角拍摄，阳光下的古街热闹非凡。一个穿着青衫、手里拿着写着“阿里云”卡片的逍遥派弟子站在中间。旁边两个小孩惊讶的看着他。左边有一家店铺挂着“云存储”的牌子，里面摆放着发光的服务器机箱，门口两个侍卫守护者。右边有两家店铺，其中一家挂着“云计算”的牌子，一个穿着旗袍的美丽女子正看着里面闪闪发光的电脑屏幕；另一家店铺挂着“云模型”的牌子，门口放着一个大酒缸，上面写着“千问”，一位老板娘正在往里面倒发光的代码溶液。



**Input prompt** 一个穿着“QWEN”标志的T恤的中国美女正拿着黑色的马克笔相镜头微笑。她身后的玻璃板上手写体写着“一、Qwen-Image的技术路线：探索视觉生成基础模型的极限，开创理解与生成一体化的未来。二、Qwen-Image的模型特色：1、复杂文字渲染。支持中英渲染、自动布局；2、精准图像编辑。支持文字编辑、物体增减、风格变换。三、Qwen-Image的未来愿景：赋能专业内容创作、助力生成式AI发展。”

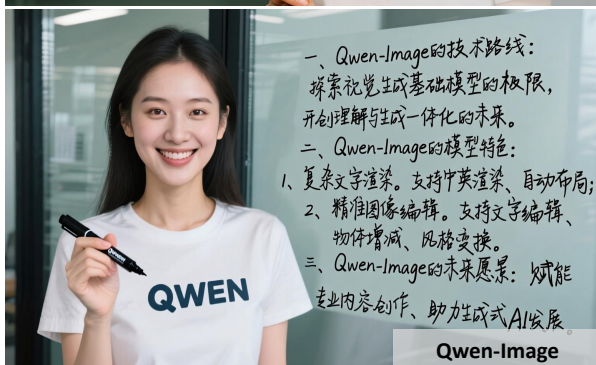


图 21: Comparison of complex Chinese text rendering in image generation: the first case shows that Qwen-Image can render text on multiple objects while maintaining consistency with the real scene, such as aligning text with the depth and tilt of each plaque; the second case demonstrates its ability to render structured paragraph text in a glass panel. Qwen-Image is the only model capable of accurately rendering long text.  
图像生成中复杂中文文本渲染的比较：第一个案例展示了 Qwen-Image 可以在多个对象上渲染文本，同时与真实场景保持一致，例如将文本与每个牌匾的深度和倾斜对齐；第二个案例展示了其在玻璃面板中渲染结构化段落文本的能力。Qwen-Image 是唯一能够准确渲染长文本的模型。



**Input prompt** Picture of twelve stuffed toys arranged evenly and neatly, four in each row, for a total of three rows. The first row is: rat, ox, tiger, rabbit. The second row is: dragon, snake, horse, sheep. The third row is: monkey, chicken, dog, pig. Each toy has a gentle and friendly facial expression, and the material has a soft fabric texture, with soft and layered colors. The background is light beige.



**Input prompt** 一个台球桌上放着两排台球，每排5个，第一行的台球上面分别写着“Qwen”“Image”“将”“于”“8”，第二排台球上面分别写着“月”“正”“式”“发”“布”。



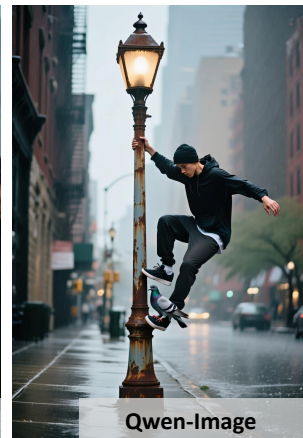
图 22: Comparison of multi-object modeling in image generation: Qwen-Image accurately renders the 12 Chinese zodiac animals and materials in the first case, and handles complex bilingual text across multiple objects in the second.  
图像生成中多物体建模的比较：Qwen-Image 在第一个案例中准确渲染了 12 个中国生肖动物和材质，在第二个案例中处理了跨多个对象的复杂双语文本。



**Input prompt** A twisted pine trunk leans over the cliff edge, a climber woman grips the trunk with two hands, and her partner reaches up, holding onto the woman's safety belt around her waist. Realistic photography.



**Input prompt** 城市动作摄影，一盏锈迹斑斑的路灯矗立在雨中的人行道上，一名跑酷运动员用一只手挂在灯上，一只鸽子停留在运动员的鞋子上，浅景深，电影般的灯光。



**Input prompt** Realistic still-life studio photography, a vintage wooden table supports an antique typewriter, a steaming porcelain coffee cup sits on top of the typewriter, and a small pocket watch hangs from the cup's handle, dramatic side lighting.



图 23: Comparison of spatial relationship modeling capability in image generation. We present three cases to demonstrate interactions involving multiple people and multiple objects. We find that both Qwen-Image and GPT Image 1 [High] exhibit strong ability to understand relationships.  
图像生成中空间关系建模能力的比较。我们呈现三个案例来展示涉及多个人和多个对象的交互。我们发现 Qwen-Image 和 GPT Image 1 [High] 都表现出强大的关系理解能力。



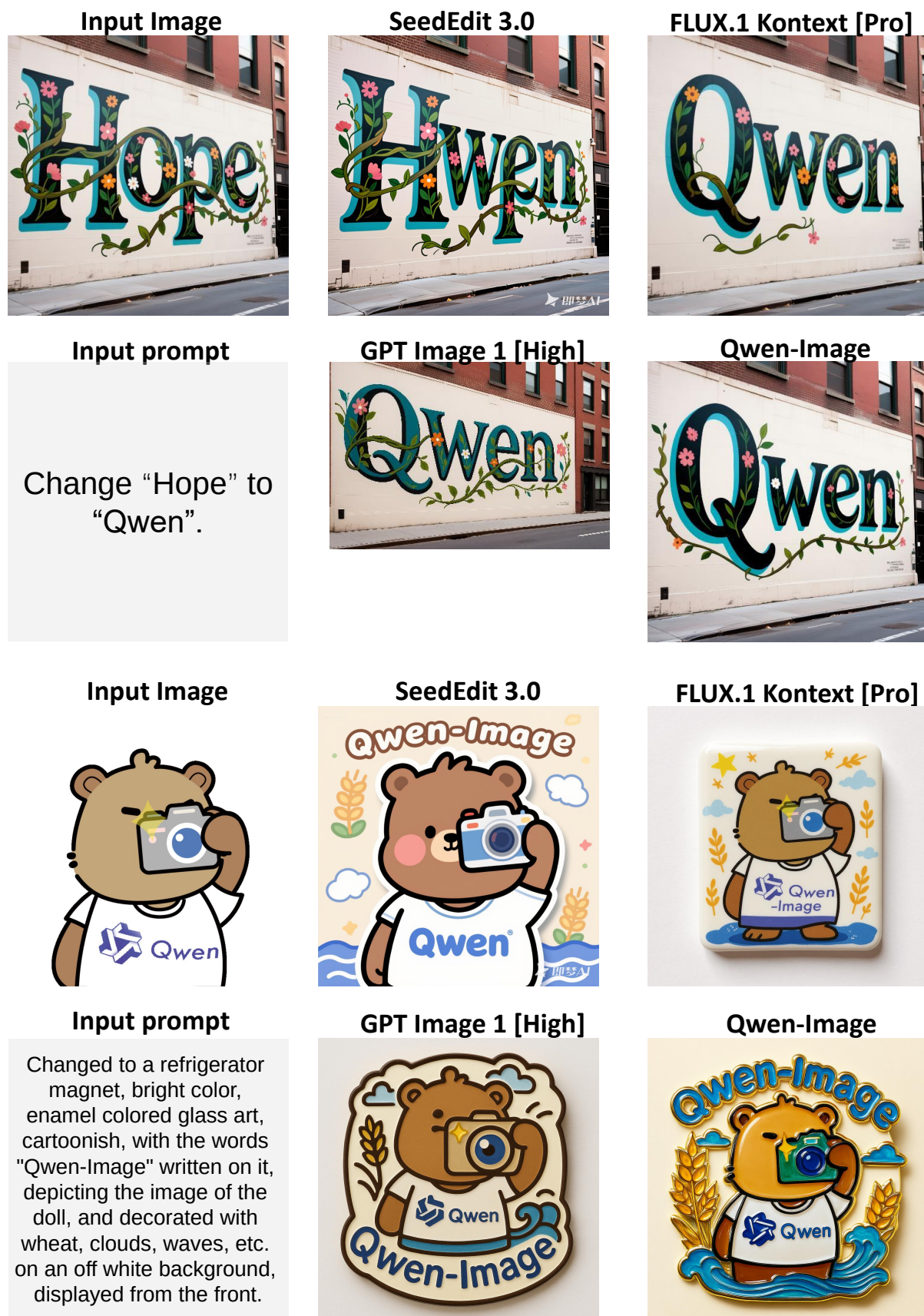


图 24: Qualitative comparison on text and material modification. Both FLUX.1 Kontext [Pro] and Qwen-Image are able to accurately modify text while preserving the original style. In the example below, Qwen-Image is the only model that successfully presents the enamel material.

文本和材质编辑的定性比较。FLUX.1 Kontext [Pro] 和 Qwen-Image 都能够准确修改文本，同时保留原始风格。在下面的示例中，Qwen-Image 是唯一成功呈现珐琅材质的模型。





图 25: Qualitative comparison on object editing (addition, removal, and replacement): Object editing is a relatively stable capability for all models. Qwen-Image demonstrates superior consistency in unmodified regions and achieves better style alignment for the newly generated objects.  
 物体编辑的定性比较（添加、移除和替换）：物体编辑是所有模型都相对稳定的能力。Qwen-Image 在未修改区域展示了优越的一致性，并为新生成的对象实现了更好的风格对齐。



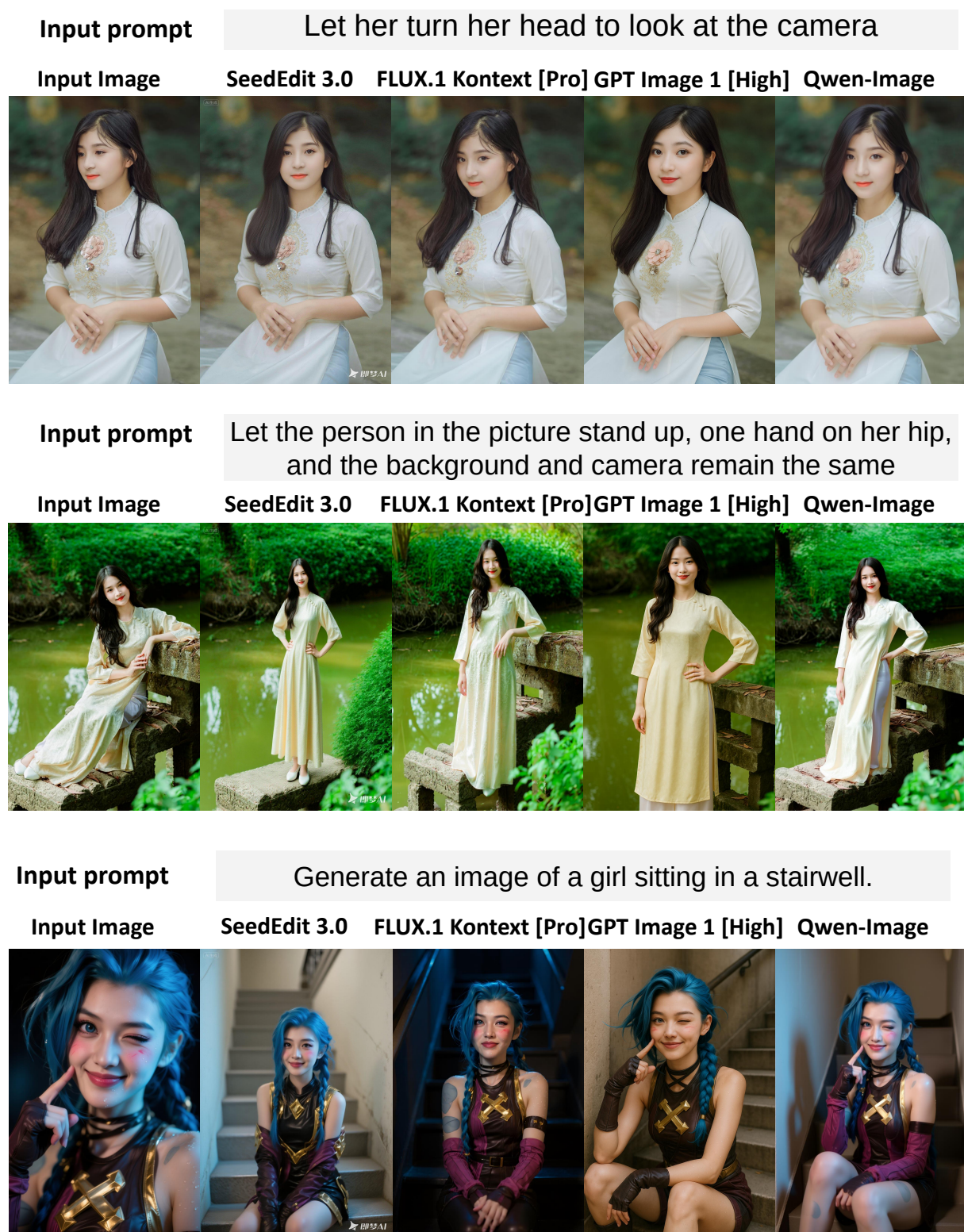


图 26: Qualitative comparison on pose manipulation. Qwen-Image is able to accurately follow pose manipulation instructions while preserving fine details of the person (such as hair strands) and maintaining consistency in the background (e.g., the stone steps behind the subject).

姿态操控的定性比较。Qwen-Image 能够准确遵循姿态操控指令，同时保留人物的精细节（如头发丝），并保持背景的一致性（例如，被摄体后面的石阶）。

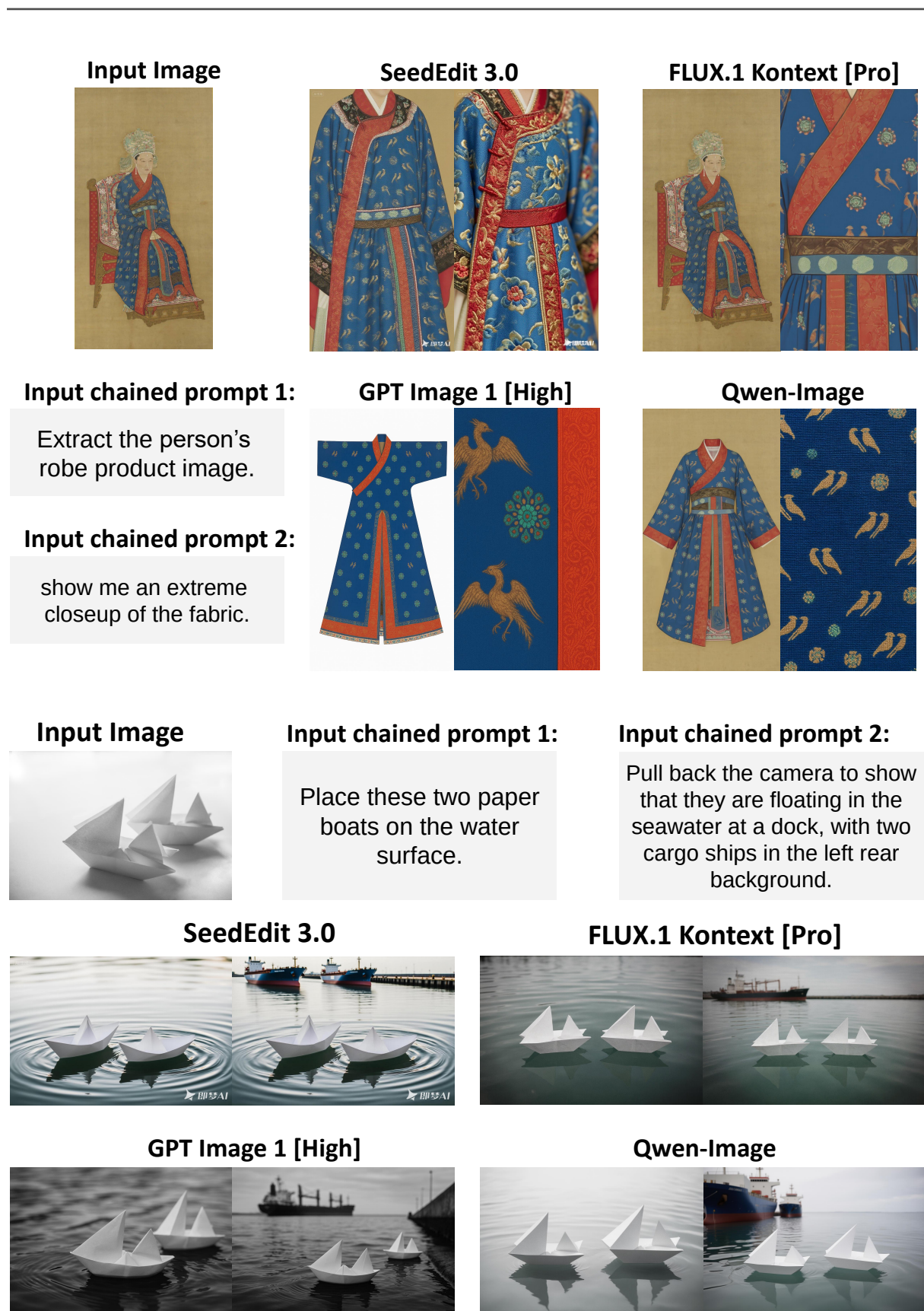


图 27: Examples of two types of chained editing: extract + zoom-in (top) and placement + zoom-out (bottom). GPT Image 1 [High] and Qwen-Image correctly understand the extract operation, but only Qwen-Image accurately captures and magnifies the garment's texture. In the second case, Qwen-Image preserves the open-ended stern of the paper boat throughout the entire chained editing process.

两种链式编辑的示例：提取 + 放大（顶部）和放置 + 缩小（底部）。GPT Image 1 [High] 和 Qwen-Image 正确理解了提取操作，但只有 Qwen-Image 准确捕捉并放大了服装纹理。在第二个案例中，Qwen-Image 在整个链式编辑过程中保留了纸船的开放式船尾。



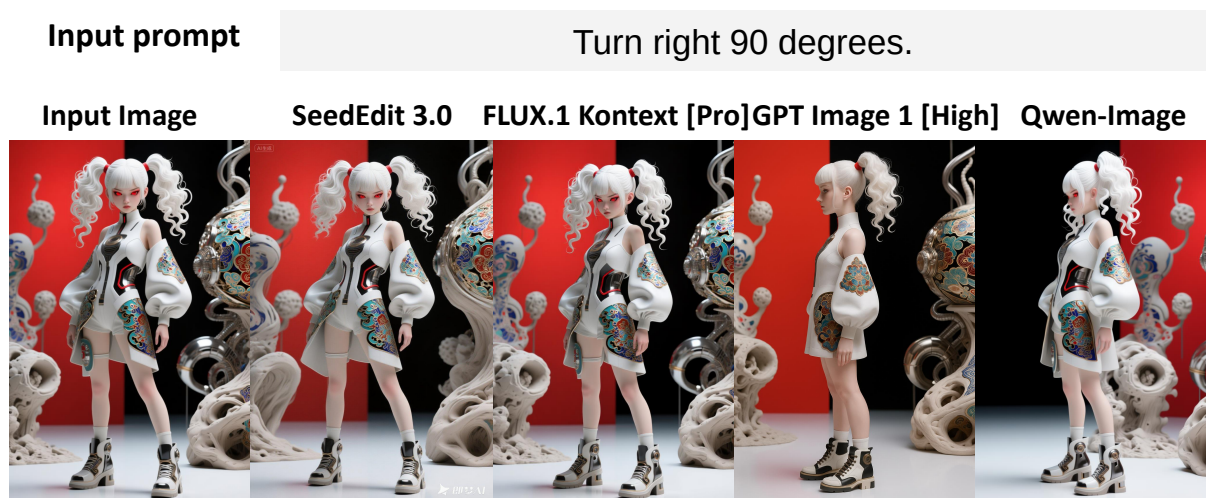
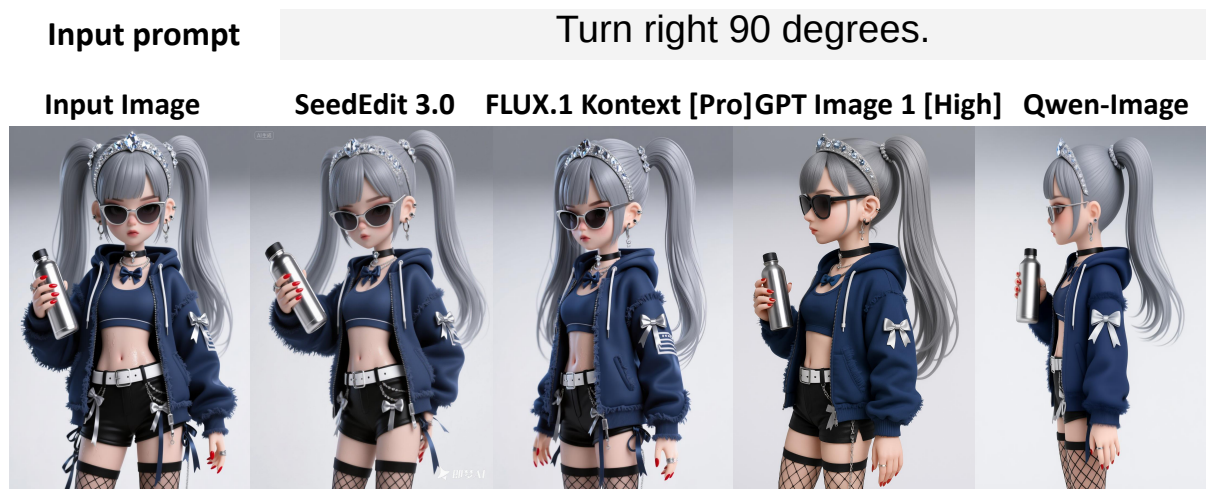


图 28: Qualitative comparison on novel view synthesis. We evaluate three progressively challenging rotation tasks under the same text prompt "Turn right 90 degrees": (1) single-person rotation, (2) co-rotation of person and background, and (3) real-world scenario. In all cases, Qwen-Image achieves the most accurate and consistent results. While other models (e.g., GPT Image 1 [High]) handle basic subject rotation, they fail to rotate the background or preserve scene details.

新视角合成的定性比较。我们在相同的文本提示词“向右转 90 度”下评估三个逐步增加难度的旋转任务：(1) 单人旋转，(2) 人物和背景的共同旋转，(3) 真实场景。在所有情况下，Qwen-Image 都获得了最准确和最一致的结果。虽然其他模型（如 GPT Image 1 [High]）可以处理基本的主体旋转，但它们无法旋转背景或保留场景细节。

---

## 6 结论

### Conclusion

In this paper, we introduce Qwen-Image, an image generation foundation model within the Qwen series, achieving major advancements in both complex text rendering and precise image editing. By constructing a comprehensive data pipeline and adopting a progressive curriculum learning strategy, Qwen-Image substantially improves its capability to render intricate text within generated images. Moreover, our improved multi-task training paradigm and dual-encoding mechanism significantly enhance the consistency and quality of image editing, effectively improving both semantic coherence and visual fidelity. Extensive experiments on public benchmarks consistently demonstrate the state-of-the-art performance of Qwen-Image across a wide range of image generation and editing tasks. These results underscore not only the technical robustness of Qwen-Image but also its broad applicability in real-world multimodal scenarios, marking a significant milestone in the evolution of large foundation models.

本文提出 Qwen-Image，这是 Qwen 系列中的图像生成基础模型，在复杂文本渲染和精确图像编辑两方面都取得了重大进展。借助一套完整的数据流水线和渐进式课程学习策略，Qwen-Image 大幅提升了在生成图像中渲染复杂文本的能力。此外，改进的多任务训练范式和双重编码机制显著增强了图像编辑的一致性与质量，有效兼顾了语义连贯性与视觉保真度。在公开基准上的大量实验一致表明，Qwen-Image 在广泛的图像生成与编辑任务上都达到了最先进的性能。这些结果既印证了 Qwen-Image 的技术鲁棒性，也说明它在真实多模态场景中具有广泛的适用性，标志着大型基础模型演进中的一个重要里程碑。

We now turn to a deeper discussion on the broader implications and significance of Qwen-Image:

下面，我们进一步讨论 Qwen-Image 更广泛的意义与价值：

Qwen-Image as an "image generation" model, in the context of image generation, redefines the priorities in generative modeling. Rather than merely optimizing for photorealism or aesthetic quality ("AI look"), Qwen-Image emphasizes precise alignment between text and image—especially in the challenging task of text rendering. We envision that by strengthening this capability in foundation models, future interfaces can evolve from purely language-based LUIs (Language User Interfaces) to vision-language VLUIs (Vision-Language User Interfaces). When LLMs struggle to convey visual attributes such as color, spatial relationships, or structural layouts, a VLUI empowered by Qwen-Image can generate richly illustrated, text-integrated imagery—enabling structured visual explanations and effective knowledge externalization, where complex ideas are transformed into comprehensible, multimodal representations.

作为一个“图像生成”模型，Qwen-Image 在图像生成的语境下重新定义了生成式建模的优先级。它并非只追求照片级真实感或美学质量（即所谓“AI 味”），而是强调文本与图像之间的精确对齐，尤其在文本渲染这一挑战性任务上。我们设想，通过在基础模型中强化这一能力，未来的交互界面可以从纯语言的 LUI（语言用户界面）演进为视觉-语言的 VLUI（视觉-语言用户界面）。当大语言模型难以表达颜色、空间关系或结构布局等视觉属性时，由 Qwen-Image 赋能的 VLUI 可以生成图文并茂的图像，从而实现结构化的视觉解释与有效的知识外化——把复杂概念转化为易于理解的多模态表示。

Qwen-Image as an image "generation" model, in the context of image understanding, demonstrates that generative frameworks can effectively perform classical understanding tasks. For instance, in depth estimation, although Qwen-Image does not surpass specialized discriminative models, it achieves performance remarkably close to them. Crucially, while traditional expert models rely on discriminative understanding—directly mapping inputs to outputs without modeling underlying distributions—Qwen-Image leverages generative understanding: it first constructs a holistic distribution over visual content, from which depth is naturally inferred. This shift from direct inference to distributional reasoning opens new avenues for unified multimodal understanding.

作为一个“图像生成”模型，Qwen-Image 在图像理解的语境下表明，生成式框架也能有效完成经典的理解任务。以深度估计为例，Qwen-Image 虽未超越专用的判别式模型，但性能已相当接近。关键在于，传统专家模型依赖判别式理解——直接把输入映射到输出，不对底层分布建模；而 Qwen-Image 采用生成式理解：先对视觉内容构建一个整体分布，深度信息由此自然推断得到。这种从直接推断到分布式推

---

理的转变，为统一的多模态理解开辟了新途径。

Qwen-Image as an "image" generation model, in the context of 3D and video generation, shows strong generalization beyond 2D image synthesis. Through the lens of image editing, we apply Qwen-Image to novel view synthesis and find that, as a general-purpose image foundation model, it outperforms dedicated 3D models in several challenging rendering scenarios, exhibiting exceptional consistency across views. In pose editing tasks, Qwen-Image maintains remarkable coherence in both subject identity and background structure despite significant motion changes—an essential requirement for video generation. Moreover, unlike most image generation approaches that rely on image VAEs, we adopt a video VAE for visual representation. While this introduces greater modeling complexity, it aligns with our core objective: to build a foundation model that generalizes across diverse visual modalities, not just static images.

作为一个“图像”生成模型，Qwen-Image 在 3D 与视频生成的语境下，展现出超越 2D 图像合成的强大泛化能力。我们从图像编辑的视角把 Qwen-Image 用于新视角合成，发现作为通用图像基础模型，它在若干挑战性渲染场景中胜过专门的 3D 模型，跨视角一致性尤为出色。在姿态编辑任务中，即便主体发生大幅运动，Qwen-Image 仍能在主体身份和背景结构上保持出色的连贯性——这正是视频生成的关键要求。此外，多数图像生成方法依赖图像 VAE，我们则采用视频 VAE 来做视觉表示。这虽然带来更高的建模复杂度，却契合我们的核心目标：构建一个能泛化到多种视觉模态、而不局限于静态图像的基础模型。

Qwen-Image as a "visual generation" model, in the context of integrated understanding and generation, advances the vision of seamless integration between perception and creation. We argue that achieving true understanding-generation unity rests on three foundational pillars: (1) mastering understanding, (2) mastering generation, and (3) mastering their synergistic integration. As the first work in the Qwen series dedicated to visual generation, Qwen-Image fills a critical gap in the second pillar—generation capability—complementing Qwen2.5-VL, which excels in visual understanding (the first pillar). Together, they form a balanced foundation for the next generation of multimodal AI, paving the way toward Visual-Language Omni systems that are not only capable of perceiving and reasoning but also of generating text-rich, visually coherent imagery—where language and vision are seamlessly fused into illustrative, readable, and semantically faithful visual outputs.

作为一个“视觉生成”模型，Qwen-Image 在理解与生成一体化的语境下，推进了感知与创作无缝融合的愿景。我们认为，要实现真正的理解-生成统一，须依托三大基石：(1) 精通理解，(2) 精通生成，(3) 精通两者的协同融合。作为 Qwen 系列中首个专注于视觉生成的工作，Qwen-Image 填补了第二根支柱——生成能力——上的关键空白，与擅长视觉理解（第一根支柱）的 Qwen2.5-VL 形成互补。二者共同构成下一代多模态 AI 的均衡基础，为视觉-语言全能（Visual-Language Omni）系统铺平道路：这类系统不仅能感知和推理，还能生成富含文本、视觉连贯的图像——语言与视觉在其中无缝融合，输出图文并茂、清晰易读、语义忠实的视觉结果。

In summary, Qwen-Image is more than a state-of-the-art image generation model—it represents a paradigm shift in how we conceptualize and build multimodal foundation models. Its contributions extend beyond technical benchmarks, challenging the community to rethink the roles of generative models in perception, interface design, and cognitive modeling. By emphasizing complex text rendering in image generation and addressing classical understanding tasks such as depth estimation through the lens of image editing, Qwen-Image points toward a future in which: (1) generative models do not merely produce images, but genuinely understand them; and (2) understanding models go beyond passive discrimination, achieving comprehension through intrinsic generative processes. As we continue to scale and refine such systems, the boundary between visual understanding and generation will blur further, paving the way for truly interactive, intuitive, and intelligent multimodal agents.

总之，Qwen-Image 不只是一个最先进的图像生成模型——它代表了我们的构思和构建多模态基础模型方式上的范式转变。它的贡献超越了技术基准，促使学界重新审视生成式模型在感知、界面设计和认知建模中的角色。通过在图像生成中强调复杂文本渲染，并从图像编辑的视角处理深度估计等经典理解任务，Qwen-Image 指向这样一个未来：(1) 生成式模型不再只是产出图像，而是真正理解图像；(2) 理解式模型超越被动判别，通过内在的生成过程实现理解。随着我们不断扩展和优化这类系统，视觉理解与生成之间的界限将进一步模糊，为真正可交互、直观、智能的多模态智能体铺平道路。



---

## 7 作者

Authors

**核心贡献者<sup>7</sup>:** Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang

**贡献者<sup>8</sup>:** An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, Zenan Liu

---

<sup>7</sup>按字母顺序排列。

<sup>8</sup>按字母顺序排列。

---

## 参考文献

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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.
- BlackForest. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.
- Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-1l: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.
- Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pp. 386–402. Springer, 2024a.
- Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.
- Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\sigma$ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024c.
- 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*, 2025b.
- Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024d.

- 
- Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.
- Google DeepMind. Gemini 2.0. <https://gemini.google.com/>, 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.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022. URL <https://arxiv.org/abs/2204.11918>.
- 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.
- Arpad E. Elo and Sam Sloan. *The Rating of Chess Players, Past and Present*. Ishi Press, 2008.
- 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.
- Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.
- 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.
- Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *CoRR*, abs/2507.22058, 2025.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.
- 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.
- Google. Imagen, 2025. URL <https://deepmind.google/models/imagen/>.
- Imagen Team Google. Imagen 3, 2024.
- Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024.



- 
- Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.
- Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024a.
- 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*, 2024b.
- Ideogram. Ideogram 3.0, 2025. URL <https://about.ideogram.ai/3.0>.
- Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5:341–353, 2023.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.
- Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7861–7871, 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*, 2024a.
- Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng 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, 2024b.

- 
- Jian Liang, Chenfei Wu, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *Advances in Neural Information Processing Systems*, 35:15420–15432, 2022.
- Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.
- Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.
- Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.
- 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*, 2025b.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7739–7751, 2025.
- Midjourney. midjourney v7, 2025. URL <https://github.com/midjourney>.
- George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL <https://doi.org/10.1145/219717.219748>.
- Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX symposium on operating systems design and implementation (OSDI 18)*, pp. 561–577, 2018.
- Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012.
- OpenAI. DALL·E 3. <https://openai.com/research/dall-e-3>, September 2023.
- OpenAI. Gpt-image-1, 2025. URL <https://openai.com/index/introducing-4o-image-generation/>.
- 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.
- 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.
- Qwen. Qwen3-coder is the code version of qwen3, the large language model series developed by qwen team, alibaba cloud., 2025. URL <https://github.com/QwenLM/Qwen3-Coder>.

- 
- 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. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.
- René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.
- Recraft. Recraft v3. <https://www.recraft.ai/>, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Kuaishou Kolors team. Kolors2.0. <https://app.klingai.com/cn/>, 2025.
- Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*, 2024.
- Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.



- 
- Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seedit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024a.
- Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European conference on computer vision*, pp. 57–74. Springer, 2024b.
- Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.
- Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pp. 720–736. Springer, 2022.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025a.
- Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.
- Xianfeng Wu, Yajing Bai, Haoze Zheng, Harold Haodong Chen, Yexin Liu, Zihao Wang, Xuran Ma, Wen-Jie Shu, Xianzu Wu, Harry Yang, et al. Lightgen: Efficient image generation through knowledge distillation and direct preference optimization. *arXiv preprint arXiv:2503.08619*, 2025c.
- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.
- Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025a.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025b.
- An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv:2211.01335*, 2022.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

- 
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024b.
- Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.
- Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.
- THUKEG Z.ai. Cogview4, 2025. URL <https://github.com/THUHM/CogView4>.
- Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.
- Kai Zhang, Lingbo Mo, Wenhui 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.
- 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.
- Canyu Zhao, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint arXiv:2502.17157*, 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.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.
- Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024.
- Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.