August 4, 2025

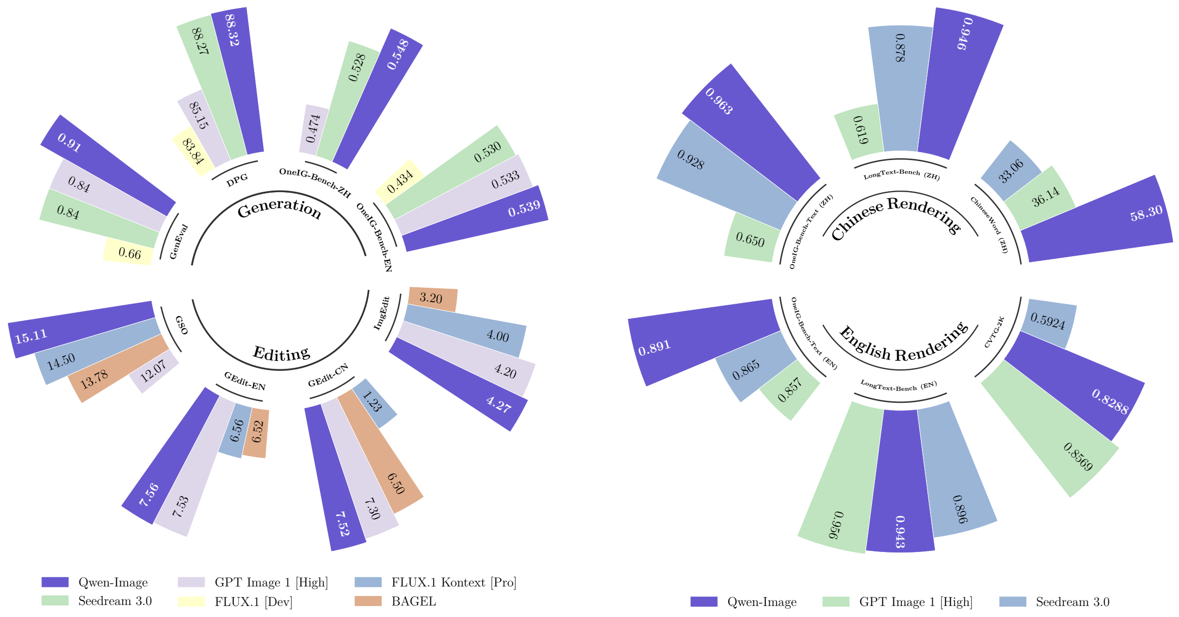
Qwen-Image Technical Report

Qwen Team

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

Abstract

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. 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.



(a) Image generation and editing benchmarks

(b) Text rendering benchmarks

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

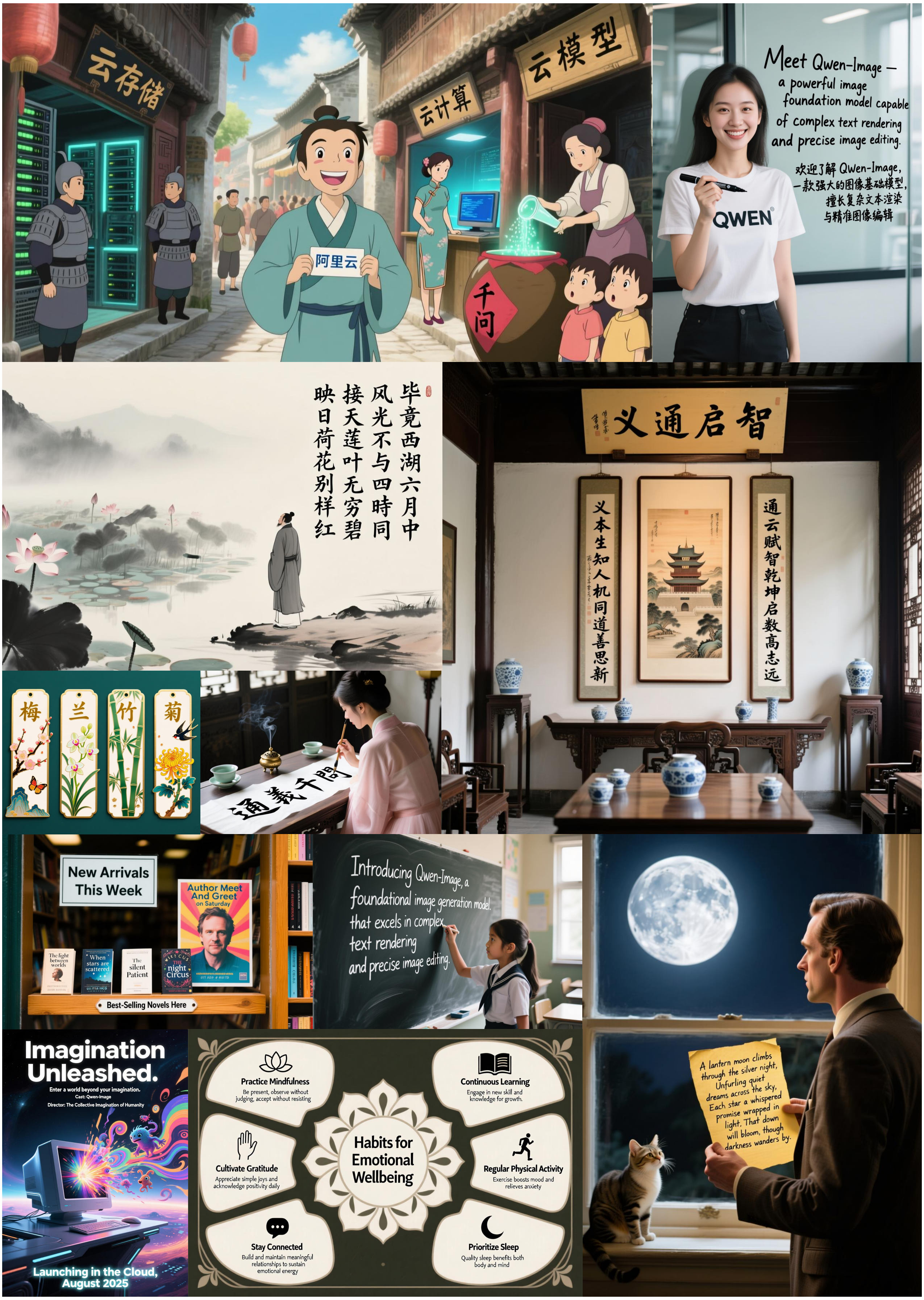


Figure 2: Showcase of Qwen-Image in complex text rendering, including multi-line layouts, paragraph-level semantics, and fine-grained details. Qwen-Image supports both alphabetic languages (e.g., English) and logographic languages (e.g., Chinese) with high fidelity.

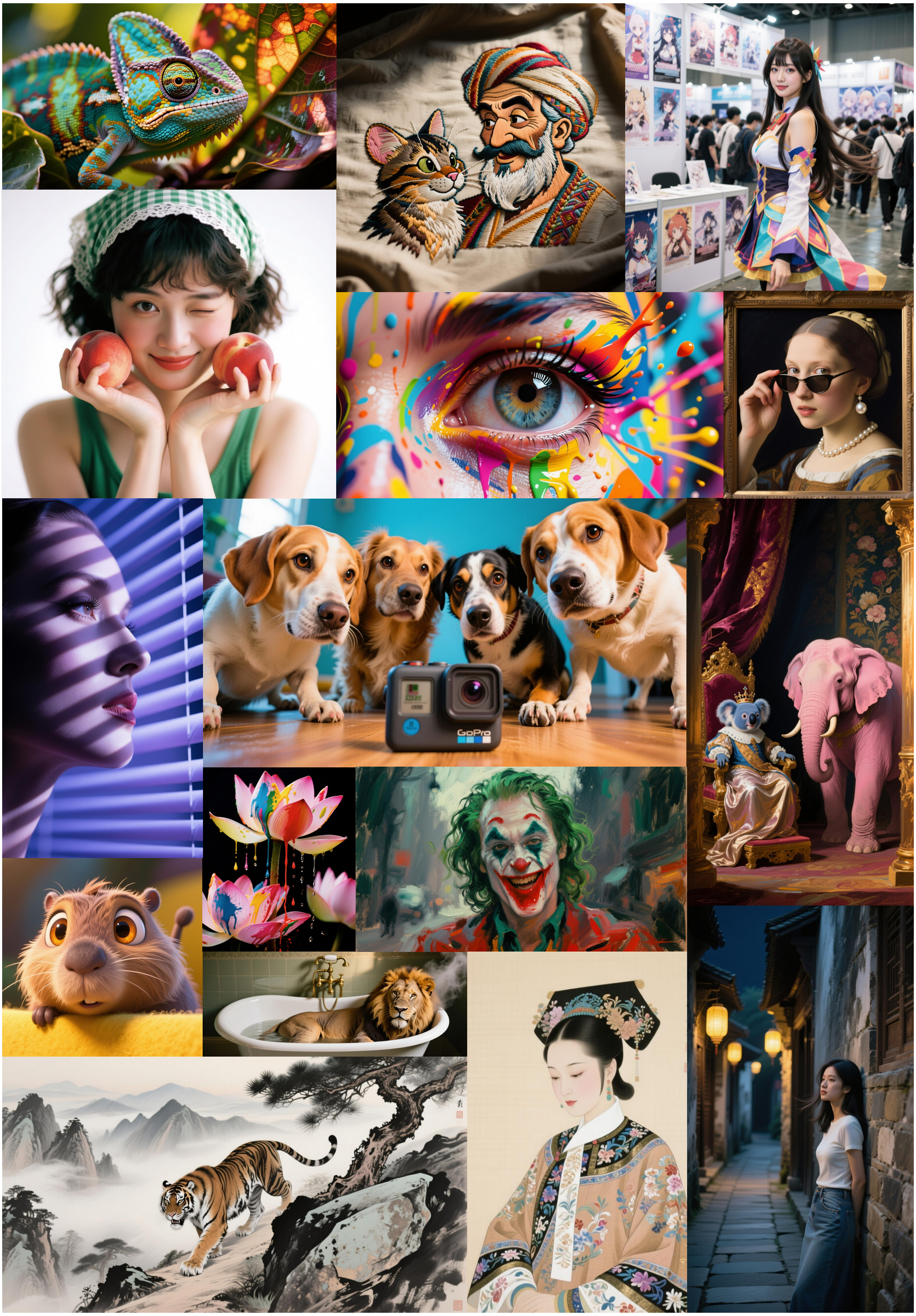


Figure 3: Showcase of Qwen-Image in general image generation, supporting diverse artistic styles.

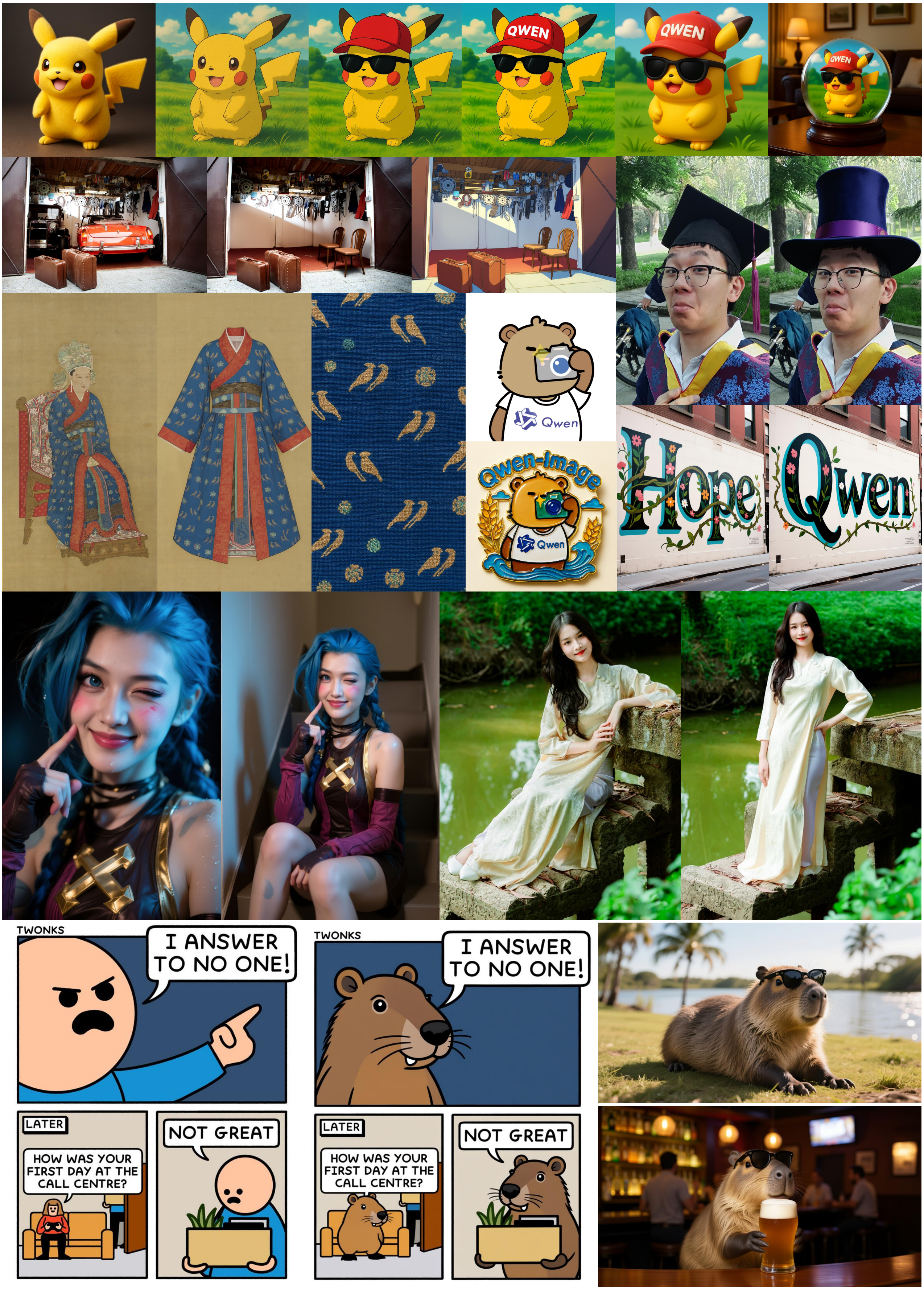


Figure 4: Showcase of Qwen-Image in general image editing, including style transfer, text editing, background change, object addition, removal, and replacement, pose manipulation, and more.

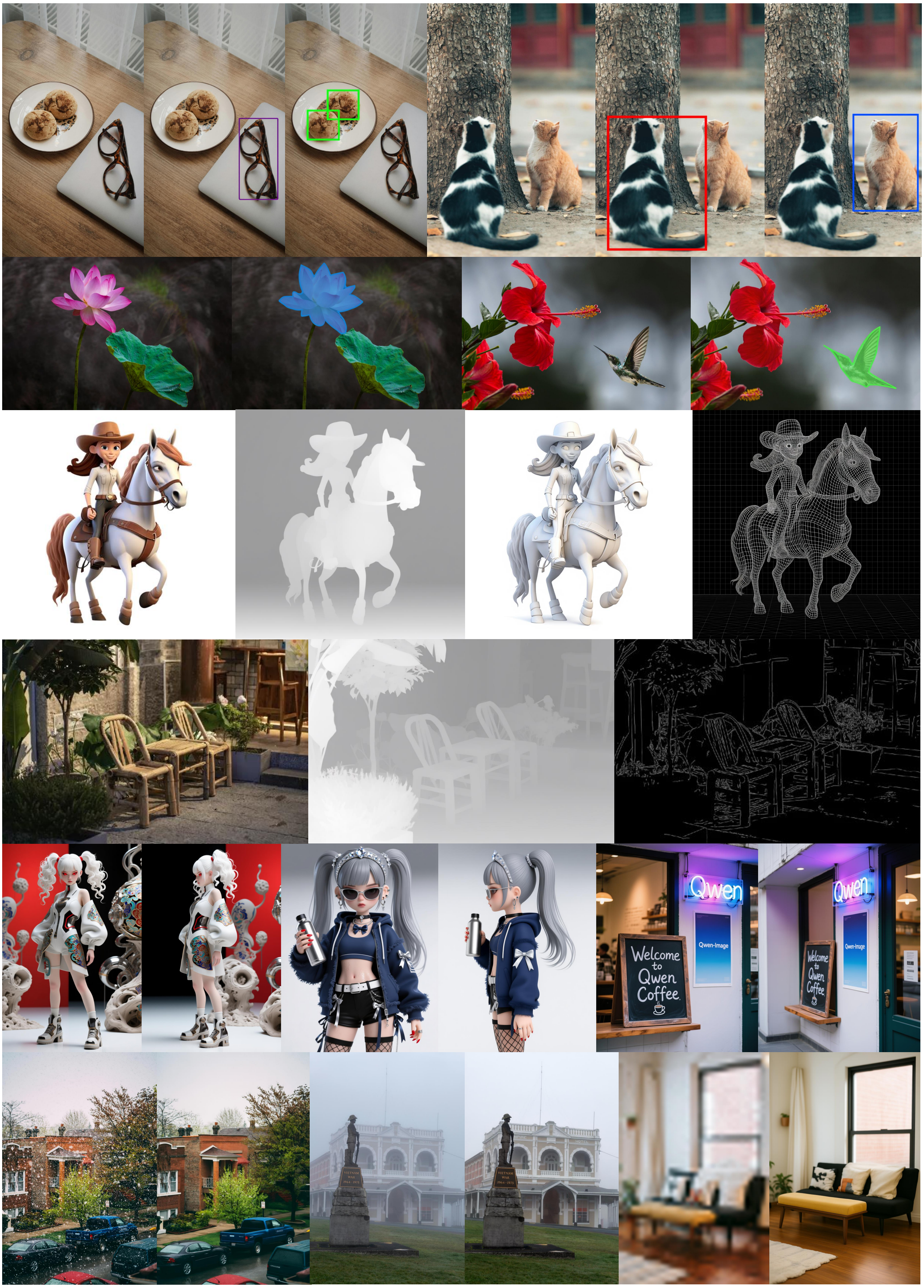


Figure 5: Showcase of Qwen-Image in general image understanding tasks, including detection, segmentation, depth/canny estimation, novel view synthesis, and super resolution—tasks that can all be viewed as specialized forms of image editing.

1 Introduction

Image generation models—encompassing both text-to-image generation (T2I) (Rombach et al., 2021; Wu et al., 2022; Liang et al., 2022; OpenAI, 2023; Podell et al., 2023; Chen et al., 2024c; Li et al., 2024b; Esser et al., 2024; BlackForest, 2024; Gao et al., 2025; Gong et al., 2025; Cai et al., 2025) and image editing (TI2I) (Brooks et al., 2023; Zhang et al., 2023; Wang et al., 2025; Deng et al., 2025; Labs et al., 2025; Wu et al., 2025b; Liu et al., 2025b; Cai et al., 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 et al., 2020; Liu et al., 2022) that enable high-resolution image generation while capturing fine-grained semantic details.

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 et al., 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).

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.

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.

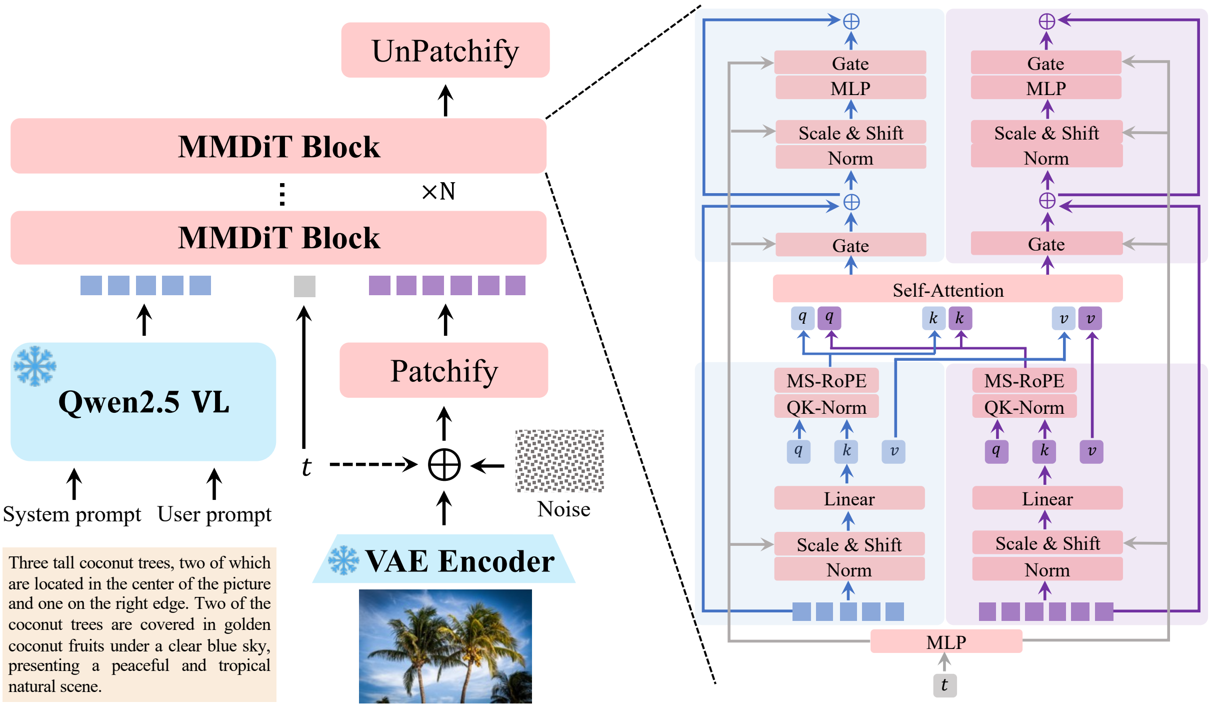
To address the challenge of image alignment, we propose an enhanced multi-task learning framework that seamlessly integrates T2I,I2I and TI2I objectives within a shared latent space. Specifically, the input image is encoded into two distinct yet complementary feature representations: semantic features are extracted via Qwen-VL (Bai et al., 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 et al., 2024) as conditioning signals. This dual-conditioning design enables the model to simultaneously maintain semantic coherence and visual consistency.

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 et al., 2019) framework. We also implement extensive monitoring tools to ensure reliable convergence and debugging capabilities throughout large-scale training.

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.

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.



Three tall coconut trees, two of which are located in the center of the picture and one on the right edge. Two of the coconut trees are covered in golden coconut fruits under a clear blue sky, presenting a peaceful and tropical natural scene.

Figure 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.

2 Model

In this section, we present the architectural design of the Qwen-Image model, along with a comprehensive overview of the training data and training details.

2.1 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.

2.2 Multimodal Large Language Model

Qwen-Image employs the Qwen2.5-VL model (Bai et al., 2025) as the feature extraction module for textual inputs, owing to three key reasons: (1) The language and visual spaces of Qwen2.5-VL have already been aligned, which makes it more suitable for text-to-image tasks compared to language-based models like Qwen3 (Yang et al., 2025); (2) Qwen2.5-VL retains strong language modeling capabilities, without significant degradation compared to language models; (3) Qwen2.5-VL supports multimodal inputs, thereby enabling Qwen-Image to unlock a broader range of functionalities, e.g., image editing (Labs et al., 2025). Let x and y denote the image and textual inputs, respectively. Given the user inputs, such as prompts and images, we adopt the Qwen2.5-VL model to extract features. To better guide the model in generating the refined representation latent, while accounting for the varying input modalities across different tasks, we design distinct system prompts tailored for **pure text input** and **text-and-image input**, respectively. We illustrate the system template in Figure 7 and Figure 15. Finally, we utilize the latent of the last layer's hidden state from Qwen2.5-VL language model backbone as the representation of the user input.

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
```

Figure 7: System prompt for Text-to-Image generation task, where **<|user_text|>** is the user input prompt.

2.3 Variational AutoEncoder

A strong VAE representation is crucial for building a powerful image foundation model. Current image foundation models typically train an image VAE with 2D convolutions on massive image datasets to obtain a high-quality image representation. In contrast, our work aims to develop a more general visual representation compatible with both images and videos. However, existing joint image-video VAEs, such as Wan-2.1-VAE (Wan et al., 2025), typically suffer a performance trade-off that results in degraded image reconstruction capabilities. To this end, we leverage a single-encoder, dual-decoder architecture. This design utilizes a shared encoder compatible with both images and videos, alongside separate, specialized decoders for each modality, which enables our image foundation model to serve as a backbone for future video models. Specifically, we adopt the architecture of Wan-2.1-VAE, freeze its encoder, and exclusively fine-tune the image decoder.

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.

2.4 Multimodal Diffusion Transformer

Qwen-Image adopts Multimodal Diffusion Transformer (MMDiT) (Esser et al., 2024) to jointly model text and images. This approach has proven effective in a range of works, such as the FLUX (BlackForest, 2024; Labs et al., 2025) series and the Seedream (Gong et al., 2025; Gao et al., 2025) series.

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 et al., 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 et al., 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.

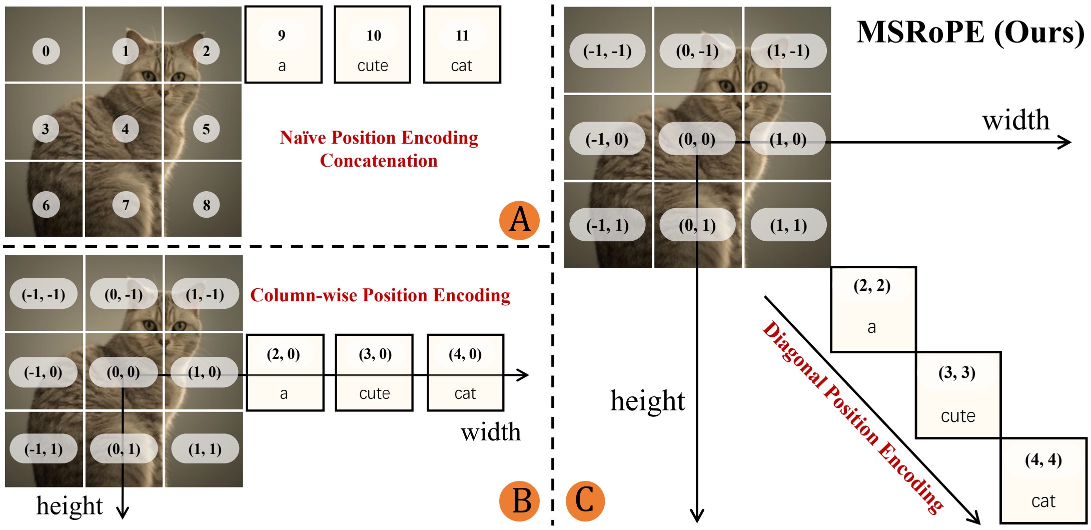


Figure 8: Comparison of different image-text joint positional encoding strategies. We design the Multimodal Scalable RoPE (MSRoPE) strategy, which starts encoding from the image center and positions textual encodings along the diagonal of the grid, enabling better image resolution scaling and improved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

Table 1: Configuration of Qwen-Image architecture.

Configuration	VLM		VAE		MMDiT
	ViT	LLM	Enc	Dec	
# Layers	32	28	11	15	60
# Num Heads (Q / KV)	16 / 16	28 / 4	-	-	24 / 24
Head Size	80	128	-	-	128
Intermediate Size	3,456	18,944	-	-	12,288
Patch / Scale Factor	14	-	8x8	8x8	2
Channel Size	-	-	16	16	-
# Parameters	7B		54M	73M	20B

3 Data

3.1 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.

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.

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 layout design.

Next, the People category makes up about 13% of the dataset, encompassing subcategories such as Portrait,

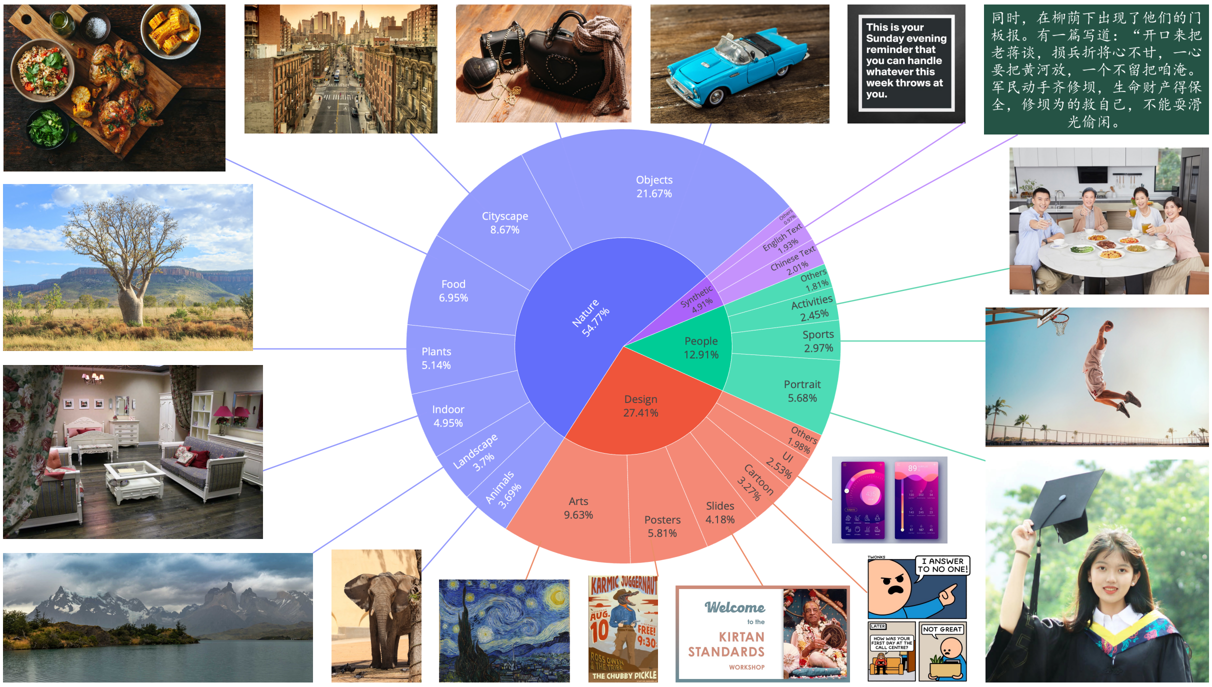


Figure 9: Overview of Data Collection. The dataset includes four main categories: Nature (general-purpose generation), People (human-centric generation), Design (artistic styles, text rendering, and complex layouts), and Synthetic Data (text rendering enhancement). Our collection strategy balances diversity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

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.

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.

3.2 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.

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.

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

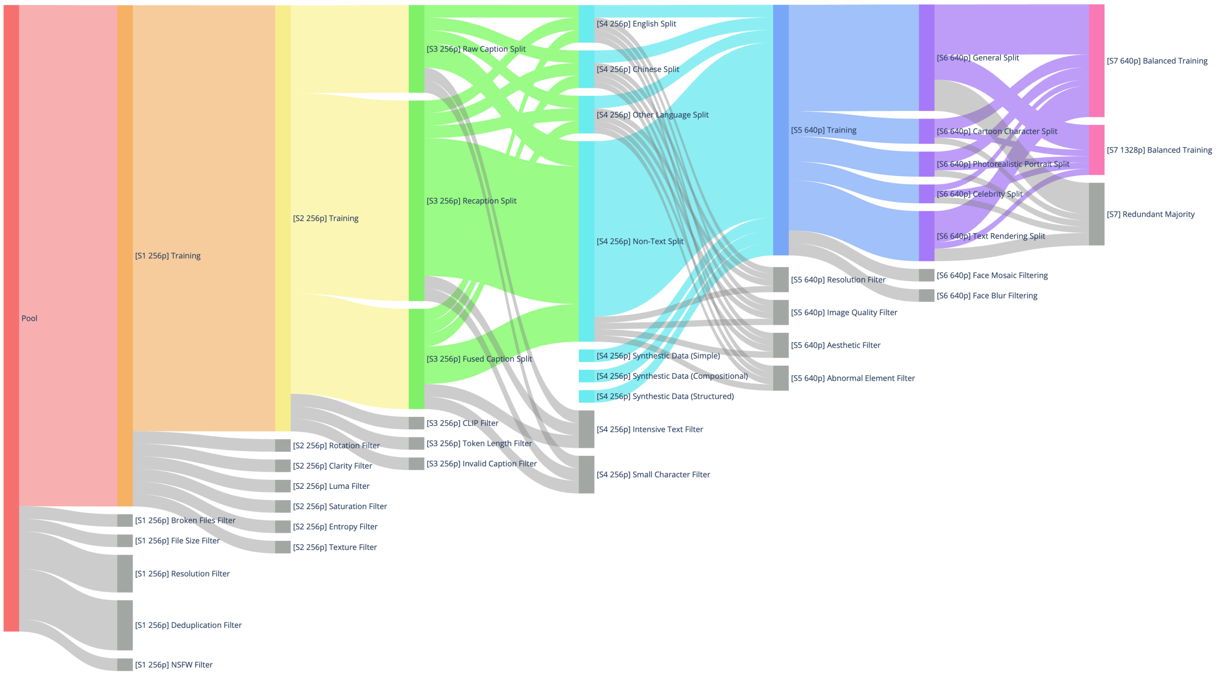

Pool
[S1 256p] Training
[S1 256p] Broken Files Filter
[S1 256p] File Size Filter
[S1 256p] Resolution Filter
[S1 256p] Deduplication Filter
[S1 256p] NSFW Filter
[S2 256p] Training
[S2 256p] Rotation Filter
[S2 256p] Clarity Filter
[S2 256p] Luma Filter
[S2 256p] Saturation Filter
[S2 256p] Entropy Filter
[S2 256p] Texture Filter
[S3 256p] Raw Caption Split
[S3 256p] Recaption Split
[S3 256p] Fused Caption Split
[S3 256p] CLIP Filter
[S3 256p] Token Length Filter
[S3 256p] Invalid Caption Filter
[S4 256p] English Split
[S4 256p] Chinese Split
[S4 256p] Other Language Split
[S4 256p] Non-Text Split
[S4 256p] Synthetic Data (Simple)
[S4 256p] Synthetic Data (Compositional)
[S4 256p] Synthetic Data (Structured)
[S4 256p] Intensive Text Filter
[S4 256p] Small Character Filter
[S5 640p] Training
[S5 640p] Resolution Filter
[S5 640p] Image Quality Filter
[S5 640p] Aesthetic Filter
[S5 640p] Abnormal Element Filter
[S6 640p] General Split
[S6 640p] Cartoon Character Split
[S6 640p] Photorealistic Portrait Split
[S6 640p] Celebrity Split
[S6 640p] Text Rendering Split
[S6 640p] Face Mosaic Filtering
[S6 640p] Face Blur Filtering
[S7 640p] Balanced Training
[S7 1328p] Balanced Training
[S7] Redundant Majority

Figure 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.

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.

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 et al., 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 et al., 2022) Filter and the SigLIP 2 (Tschannen et al., 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.".

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 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.

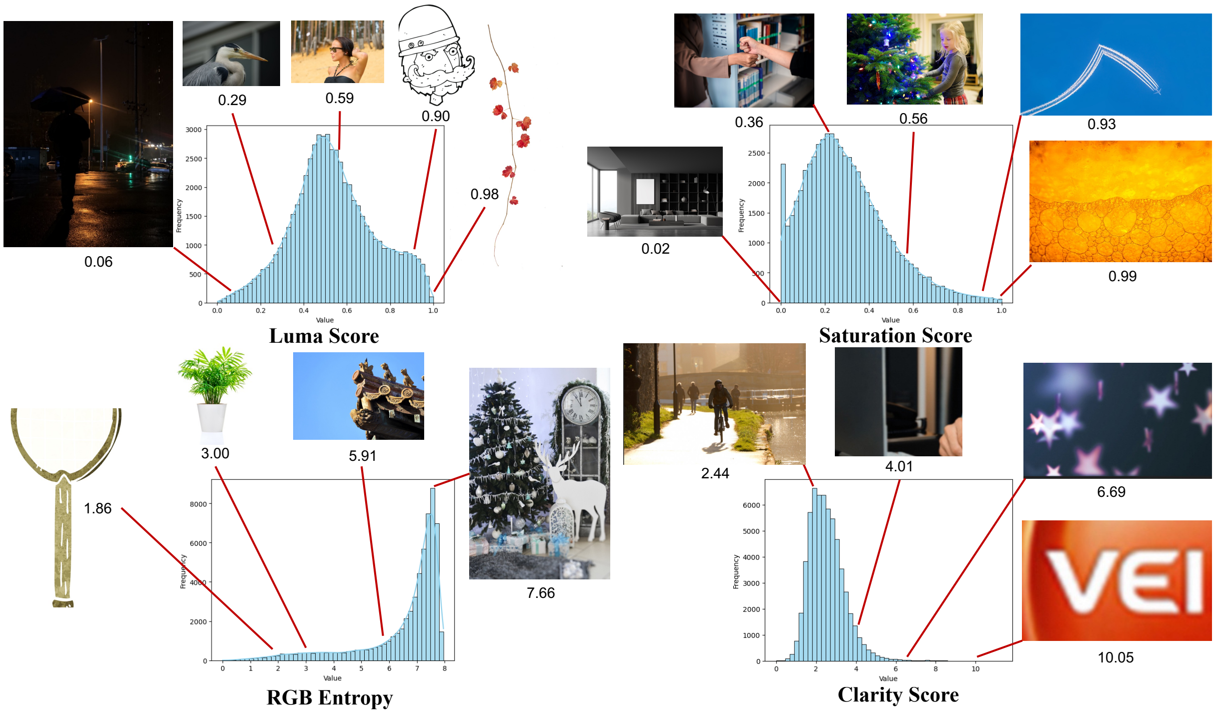


Figure 11: Examples of some filtering operators used in our pipeline. Extreme values in these operators often indicate atypical or low-quality images. For instance, images with excessively high Luma score typically contain large areas of white or overexposed regions.

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.

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.

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.

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",
}
```
~~~

Figure 12: Example of the annotation prompt used in Qwen-Image.

3.3 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.

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 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.

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.

3.4 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 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.

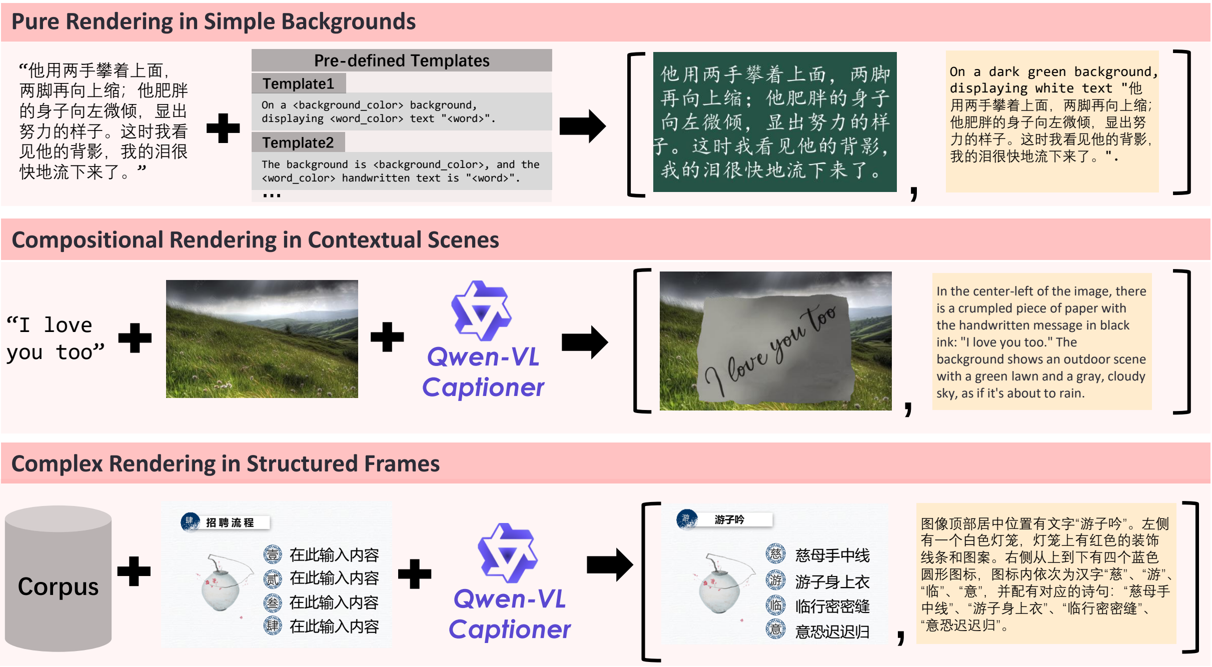

Pure Rendering in Simple Backgrounds
“他用两手攀着上面，两脚再向上缩；他肥胖的身子向左微倾，显出努力的样子。这时我看见他的背影，我的泪很快地流下来了。”
Pre-defined Templates
Template1
On a <background_color> background, displaying <word_color> text "<word>".
Template2
The background is <background_color>, and the <word_color> handwritten text is "<word>".
...
他用两手攀着上面，两脚再向上缩；他肥胖的身子向左微倾，显出努力的样子。这时我看见他的背影，我的泪很快地流下来了。
On a dark green background, displaying white text "他用两手攀着上面，两脚再向上缩；他肥胖的身子向左微倾，显出努力的样子。这时我看见他的背影，我的泪很快地流下来了。".
Compositional Rendering in Contextual Scenes
“I love you too”
Qwen-VL Captioner
I love you too
In the center-left of the image, there is a crumpled piece of paper with the handwritten message in black ink: "I love you too." The background shows an outdoor scene with a green lawn and a gray, cloudy sky, as if it's about to rain.
Complex Rendering in Structured Frames
Corpus
招聘流程
在此输入内容
在此输入内容
在此输入内容
在此输入内容
Qwen-VL Captioner
游子吟
慈母手中线
游子身上衣
临行密密缝
意恐迟迟归
图像顶部居中位置有文字“游子吟”。左侧有一个白色灯笼，灯笼上有红色的装饰线条和图案。右侧从上到下有四个蓝色圆形图标，图标内依次为汉字“慈”、“游”、“临”、“意”，并配有对应的诗句：“慈母手中线”、“游子身上衣”、“临行密密缝”、“意恐迟迟归”。

Figure 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.

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.

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.

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.

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 synthesize 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.

4 Training

4.1 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 $\boldsymbol{S}$, which may comprise text or prompt combined with image, the guidance latent h is obtained from an MLLM ϕ, i.e., $h = \phi(\boldsymbol{S})$, where $\phi : \mathcal{S} \mapsto h$. In addition, a diffusion timestep t is sampled from a logit-normal distribution with $t \in [0, 1]$. According to Rectified Flow (Liu et al., 2022; Esser et al., 2024), the intermediate latent variable at timestep t and its corresponding velocity v_t can be calculated as:

$$\begin{cases} x_t = t x_0 + (1-t) x_1 \\ v_t = \dfrac{dx_t}{dt} = x_0 - x_1 \end{cases} . \tag{1}$$

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:

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

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

4.1.1 Producer-Consumer Framework

To ensure both high throughput and training stability when scaling to large-scale GPU clusters, we adopt a Ray (Moritz et al., 2018)-inspired Producer–Consumer framework that decouples data preprocessing from model training. This design enables both stages to operate asynchronously and at optimal efficiency, while also supporting on-the-fly updates to the data pipeline without interrupting the ongoing training process. On the **Producer** side, raw image-caption pairs are first filtered according to our pre-defined criteria, such as image resolution and detection operators. The selected data is then encoded into latent representations using MLLM models (e.g., Qwen2.5 VL) and VAE. Processed images are subsequently grouped by resolution in fast-access cache buckets and stored in a shared, location-aware store, allowing consumers to fetch them immediately without waiting in line. The connection between the Producer and Consumer is achieved by employing a specific HTTP transport layer, which natively supports the RPC semantics required for asynchronous, zero-copy scheduling between the two endpoints. The **Consumer** is deployed on GPU-dense clusters and is dedicated exclusively to model training. By offloading all data processing to the Producer, the Consumer nodes can devote their entire compute budget to training the MMDiT model. The MMDiT parameters are distributed across these nodes under a 4-way tensor-parallel layout, and every data-parallel group asynchronously pulls pre-processed batches directly from the Producer. Additional consumer-side optimizations are detailed in the following section.

4.1.2 Distributed Training Optimization

Given the large parameter size of the Qwen-Image model, using FSDP (Zhao et al., 2023) alone is insufficient to fit the model on each GPU. Therefore, we leverage Megatron-LM (Shoeybi et al., 2019; Korthikanti et al., 2023) for training and apply the following optimizations:

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[1], 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.

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

[1] https://github.com/NVIDIA/TransformerEngine

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× (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.

4.1.3 Training Strategy

We adopt a multi-stage pre-training strategy aimed at progressively enhancing data quality, image resolution, and model performance. Throughout different training stages, we integrate various training strategies to optimize the learning process. These training strategies are listed below:

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 256×256 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 640×640 pixels, and finally reaching 1328×1328 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.

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.

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 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.

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.

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.

4.2 Post-training

In this section, we present the post-training framework for Qwen-Image, which consists of two stages: supervised fine-tuning (SFT) and reinforcement learning (RL) (Kaelbling et al., 1996).

4.2.1 Supervised Fine-Tuning (SFT)

During the SFT stage, we construct a hierarchically organized dataset of semantic categories and employ meticulous human annotation to address specific shortcomings of the model. We require selected images to be clear, rich in detail, bright, and photorealistic. This approach is designed to guide the model towards producing content with greater realism and finer details.

4.2.2 Reinforcement Learning (RL)

We employ two distinct RL strategies: Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Group Relative Policy Optimization (GRPO) (Shao et al., 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.

(A) Direct Preference Optimization (DPO)

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.

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 1. Then, inspired by previous work (Wallace et al., 2024), we construct the DPO objective based on the flow matching training criterion, which can be formulated as follows:

$$\begin{cases} \text{Diff}_{\text{policy}} = \left(\left\| v_\theta(x_t^{win}, h, t) - v_t^{win} \right\|_2^2 - \left\| v_\theta(x_t^{lose}, h, t) - v_t^{lose} \right\|_2^2 \right) \\ \text{Diff}_{\text{ref}} = \left(\left\| v_{\text{ref}}(x_t^{win}, h, t) - v_t^{win} \right\|_2^2 - \left\| v_{\text{ref}}(x_t^{lose}, h, t) - v_t^{lose} \right\|_2^2 \right) \\ \mathcal{L}_{DPO} = -\mathbb{E}_{h, (x_0^{win}, x_0^{lose}) \sim \mathcal{D}, t \sim \mathcal{U}(0,1)} \left[\log \sigma \Big(-\beta \left(\text{Diff}_{\text{policy}} - \text{Diff}_{\text{ref}} \right) \Big) \right], \end{cases} \tag{3}$$

where $\text{Diff}_{\text{policy}}$ and Diff_{ref} denote the preference differences computed by the policy model and the reference model, respectively, β is a scaling parameter, and $\sigma(\cdot)$ denotes the sigmoid function.

(B) Group Relative Policy Optimization (GRPO)

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

$$A_i = \frac{R(x_0^i, h) - mean(\{R(x_0^i, h)\}_{i=1}^G)}{std(\{R(x_0^i, h)\}_{i=1}^G)}, \tag{4}$$

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

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = & \mathbb{E}_{h \sim \mathcal{D}, \{x_T^i, ..., x_0^i\}_{i=1}^G \sim \pi_\theta} \\ & \frac{1}{G} \sum_{i=1}^{G} \frac{1}{T} \sum_{t=0}^{T-1} \Big(\min(r_t^i(\theta) A_i, \text{clip}(r_t^i(\theta), 1 - \epsilon, 1 + \epsilon) A_i) - \beta D_{KL}(\pi_\theta || \pi_{\text{ref}}) \Big), \end{aligned} \tag{5}$$

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

When sampling the trajectories $\{x_T^i, ..., x_0^i\}_{i=1}^G \sim \pi_\theta$, we have $\mathrm{d}x_t = v_t \mathrm{d}t$ for flow-matching sampling (following Eq. 1), 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:

$$\mathrm{d}x_t = \left(v_t + \frac{\sigma_t^2}{2t} (x_t + (1 - t) v_t) \right) \mathrm{d}t + \sigma_t \mathrm{d}w, \tag{6}$$

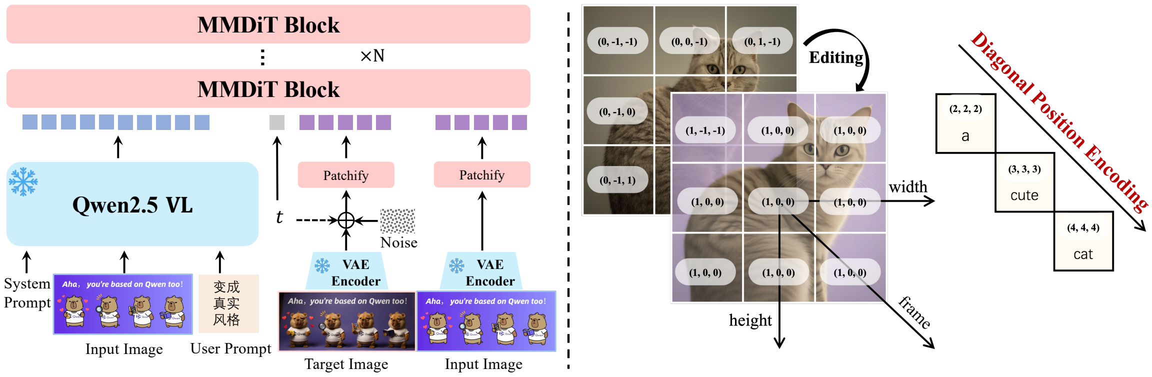


Figure 14: Overview of the Image Editing (TI2I) task. Left: Illustration of the TI2I task training input format. The user prompt is "Turn into realistic style" in English. Right: The corresponding modification to MSRoPE for TI2I, where a new dimension (frame) is introduced to distinguish between the images before and after editing.

System Prompt for TI2I task

<|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

Figure 15: System prompt for Image Editing (TI2I) task, where **<|user_image|>** is the user input image and **<|user_text|>** is the user input prompt.

where σ_t denotes the magnitude of randomness. Using Euler-Maruyama discretization, we have

$$x_{t+\Delta t} = x_t + \left[v_\theta(x_t, t, h) + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta(x_t, t, h)) \right] \Delta t + \sigma_t \sqrt{\Delta t}\epsilon. \tag{7}$$

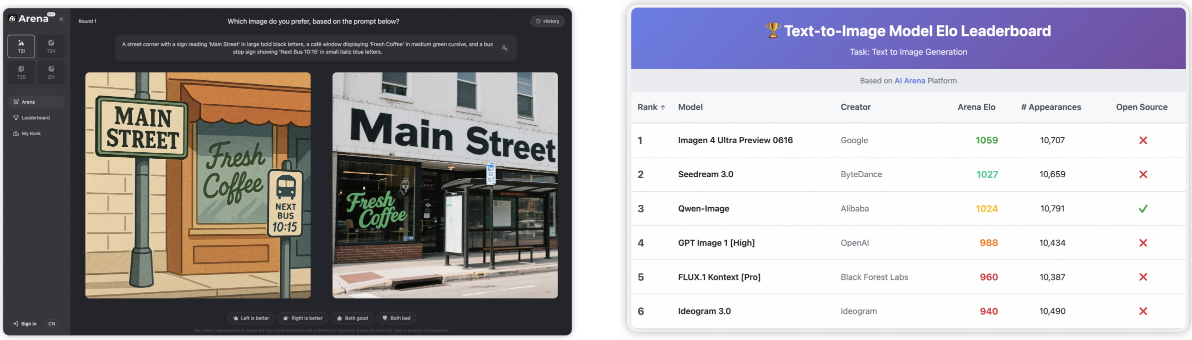
We use the above equation for sampling the trajectories. The KL-divergence in Eq. (5) can be solved in a closed form

$$D_{KL}(\pi_\theta || \pi_{\text{ref}}) = \frac{\Delta t}{2} \left(\frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t} \right)^2 ||v_\theta(x_t, t, h) - v_{\text{ref}}(x_t, t, h)||^2. \tag{8}$$

4.3 Multi-task training

In addition to text-to-image (T2I) generation, we extend our base model to explore multi-modal image generation tasks that incorporate both text and image inputs, including instruction-based image editing (Wang et al., 2025), novel view synthesis (Wang et al., 2024b), and computer vision tasks such as depth estimation (Bochkovskii et al., 2024). We can broadly regard these as general image editing tasks. Building on the capabilities of Qwen2.5-VL, our model natively supports image inputs: visual patches extracted from the user-provided image are encoded by a Vision Transformer (ViT) and concatenated with text tokens to form the input sequence. We design the system prompt shown in Figure 15 and extract both the input image and textual instructions as inputs to the text stream of Qwen-Image MMDiT.

Inspired by prior work (Labs et al., 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



Which image do you prefer, based on the prompt below?

A street corner with a sign reading 'Main Street' in large bold black letters, a café window displaying 'Fresh Coffee' in medium green cursive, and a bus stop sign showing 'Next Bus 10:15' in small italic blue letters.

The content is generated by AI models and may include information that is offensive or inaccurate. It does not reflect the views or positions of the platform.

🏆 Text-to-Image Model Elo Leaderboard

Task: Text to Image Generation

Based on AI Arena Platform

Rank ↑	Model	Creator	Arena Elo	# Appearances	Open Source
1	Imagen 4 Ultra Preview 0616	Google	1059	10,707	✗
2	Seedream 3.0	ByteDance	1027	10,659	✗
3	Qwen-Image	Alibaba	1024	10,791	✓
4	GPT Image 1 [High]	OpenAI	988	10,434	✗
5	FLUX.1 Kontext [Pro]	Black Forest Labs	960	10,387	✗
6	Ideogram 3.0	Ideogram	940	10,490	✗

(a) Front end of AI Arena platform

(b) Text-to-Image Elo Leaderboard

Figure 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.

pixel-level VAE embeddings further enhances the model's ability to preserve visual fidelity and maintain structural consistency with the user-provided image.

5 Experiments

5.1 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[2], an open benchmarking platform built upon the Elo rating system (Elo &Sloan, 2008), as shown in Figure 16.

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

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 et al., 2025), GPT Image 1 [High] (OpenAI, 2025), FLUX.1 Kontext [Pro] (Labs et al., 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.

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 $\tilde{3}0$ 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.

5.2 Quantitative Results

To comprehensively evaluate the visual generation capabilities of our model, we first report its performance on VAE reconstruction in §§ 5.2.1, which serves to demonstrate the upper bound of the model's

[2] https://aiarena.alibaba-inc.com

Table 2: Quantitative Evaluation results of VAE.

Model	# Params		# Image Params		Imagenet_256x256		Text_256x256	
	Enc	Dec	Enc	Dec	PSNR(↑)	SSIM(↑)	PSNR(↑)	SSIM(↑)
Wan2.1-VAE (Wan et al., 2025)	54M	73M	19M	25M	31.29	0.8870	26.77	0.9386
Hunyuan-VAE (Kong et al., 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 et al., 2025)	31M	46M	31M	46M	32.23	0.9010	30.62	0.9664
SD-3.5-VAE (Esser et al., 2024)	34M	50M	34M	50M	31.22	0.8839	29.93	0.9658
Qwen-Image-VAE	54M	73M	19M	25M	**33.42**	**0.9159**	**36.63**	**0.9839**

generation quality. We further conduct evaluations on two fundamental visual generation tasks, text-to-image (T2I) in §§ 5.2.2 and image editing (TI2I) in §§ 5.2.3, to provide a thorough assessment of the model's foundational generative abilities.

5.2.1 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 et al., 2025), and SD-3.5-VAE (Esser et al., 2024) are image tokenizers, while Wan2.1-VAE (Wan et al., 2025), Hunyuan-VAE (Kong et al., 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 et al., 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 encoder and 25M in the decoder, achieving an optimal balance between reconstruction quality and computational efficiency.

5.2.2 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 et al., 2024b), GenEval (Ghosh et al., 2023), OneIG-Bench (Chang et al., 2025), and TIIF (Wei et al., 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 et al., 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 et al., 2025), a benchmark designed to evaluate the performance on rendering longer texts in both English and Chinese.

DPG Table 3 shows the performance comparison on **DPG** (Hu et al., 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.

GenEval Table 4 presents a comparison of model performance on the **GenEval** (Ghosh et al., 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-

Table 3: Quantitative evaluation results on DPG (Hu et al., 2024b).

Model	**Global**	**Entity**	**Attribute**	**Relation**	**Other**	**Overall↑**
SD v1.5 (Rombach et al., 2021)	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-α (Chen et al., 2024c)	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next (Zhuo et al., 2024)	82.82	88.65	86.44	80.53	81.82	74.63
SDXL (Podell et al., 2023)	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 (Li et al., 2024a)	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT (Li et al., 2024b)	84.59	80.59	88.01	74.36	86.41	78.87
Janus (Wu et al., 2025a)	82.33	87.38	87.70	85.46	86.41	79.68
PixArt-Σ (Chen et al., 2024b)	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen (Wang et al., 2024a)	85.21	86.68	86.84	90.22	83.15	80.60
Janus-Pro-1B (Chen et al., 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 et al., 2024)	87.90	91.01	88.83	80.70	88.68	84.08
Janus-Pro-7B (Chen et al., 2025b)	86.90	88.90	89.40	89.32	89.48	84.19
HiDream-I1-Full (Cai et al., 2025)	76.44	90.22	89.48	93.74	91.83	85.89
Lumina-Image 2.0 (Qin et al., 2025)	-	91.97	90.20	**94.85**	-	87.20
Seedream 3.0 (Gao et al., 2025)	**94.31**	**92.65**	91.36	92.78	88.24	88.27
GPT Image 1 [High] (OpenAI, 2025)	88.89	88.94	89.84	92.63	90.96	85.15
Qwen-Image	91.32	91.56	**92.02**	94.31	**92.73**	**88.32**

Table 4: Quantitative Evaluation results on GenEval (Ghosh et al., 2023).

Model	**Single Object**	**Two Object**	**Counting**	**Colors**	**Position**	**Attribute Binding**	**Overall↑**
Show-o (Xie et al., 2024)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3-Gen (Wang et al., 2024a)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
PixArt-α (Chen et al., 2024c)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD3 Medium (Esser et al., 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 et al., 2024)	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow (Ma et al., 2025)	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Lumina-Image 2.0 (Qin et al., 2025)	-	0.87	0.67	-	-	0.62	0.73
Janus-Pro-7B (Chen et al., 2025b)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream-I1-Full (Cai et al., 2025)	**1.00**	0.98	0.79	0.91	0.60	0.72	0.83
GPT Image 1 [High] (OpenAI, 2025)	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 3.0 (Gao et al., 2025)	0.99	**0.96**	0.91	**0.93**	0.47	0.80	0.84
Qwen-Image	0.99	0.92	0.89	0.88	0.76	0.77	0.87
Qwen-Image-RL	**1.00**	0.95	**0.93**	0.92	**0.87**	**0.83**	**0.91**

enhanced model against other leading foundation models. Notably, our base model already surpasses the state of the art, outperforming Seedream 3.0 (Gao et al., 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.

OneIG-Bench Table. 5 and Table. 6 report the quantitative results on OneIG-Bench (Chang et al., 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, 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.

TIIF Table 7 shows the performance comparison on **TIIF Bench mini** (Wei et al., 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.

Table 5: Quantitative evaluation results on OneIG-EN (Chang et al., 2025). The overall score is the average of the five dimensions.

Model	Alignment	Text	Reasoning	Style	Diversity	Overall↑
Janus-Pro (Chen et al., 2025b)	0.553	0.001	0.139	0.276	0.365	0.267
BLIP3-o (Chen et al., 2025a)	0.711	0.013	0.223	0.361	0.229	0.307
BAGEL (Deng et al., 2025)	0.769	0.244	0.173	0.367	0.251	0.361
BAGEL+CoT (Deng et al., 2025)	0.793	0.020	0.206	0.390	0.209	0.324
Show-o2-1.5B (Xie et al., 2025b)	0.798	0.002	0.219	0.317	0.186	0.304
Show-o2-7B (Xie et al., 2025b)	0.817	0.002	0.226	0.317	0.177	0.308
OmniGen2 (Wu et al., 2025b)	0.804	0.680	0.271	0.377	0.242	0.475
SD 1.5 (Rombach et al., 2021)	0.565	0.010	0.207	0.383	**0.429**	0.319
SDXL (Podell et al., 2023)	0.688	0.029	0.237	0.332	0.296	0.316
SD3.5 Large (Esser et al., 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 et al., 2025a)	0.762	0.054	0.209	0.387	0.222	0.327
SANA-1.5 4.8B (PAG) (Xie et al., 2025a)	0.765	0.069	0.217	0.401	0.216	0.334
Lumina-Image 2.0 (Qin et al., 2025)	0.819	0.106	0.270	0.354	0.216	0.353
HiDream-I1-Full (Cai et al., 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 et al., 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	**0.345**	**0.462**	0.151	0.533
Qwen-Image	**0.882**	**0.891**	0.306	0.418	0.197	**0.539**

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

Model	Alignment	Text	Reasoning	Style	Diversity	Overall↑
Janus-Pro (Chen et al., 2025b)	0.324	0.148	0.104	0.264	**0.358**	0.240
BLIP3-o (Chen et al., 2025a)	0.608	0.092	0.213	0.369	0.233	0.303
BAGEL (Deng et al., 2025)	0.672	0.365	0.186	0.357	0.268	0.370
BAGEL+CoT (Deng et al., 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 et al., 2025)	0.731	0.136	0.221	0.343	0.240	0.334
HiDream-I1-Full (Cai et al., 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 et al., 2025)	0.793	0.928	0.281	0.397	0.243	0.528
GPT Image 1 [High] (OpenAI, 2025)	0.812	0.650	**0.300**	**0.449**	0.159	0.474
Qwen-Image	**0.825**	**0.963**	0.267	0.405	0.279	**0.548**

CVTG-2K Table 8 reports the quantitative results of English rendering on **CVTG-2K** (Du et al., 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.

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.

LongText-Bench Table 10 reports the quantitative results on **LongText-Bench** (Geng et al., 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.

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

Model	Overall		Basic Following								Advanced Following												Designer	
			Avg		Attribute		Relation		Reasoning		Avg		Attribute +Relation		Attribute +Reasoning		Relation +Reasoning		Style		Text		Real World	
	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 et al., 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 et al., 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 (Zhuo et al., 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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 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	81.22	71.59	60.72	64.59	83.33	80.00	24.83	20.83	68.83	63.61
Seedream 3.0 (Gao et al., 2025)	86.02	84.31	87.07	84.93	90.50	90.00	**89.85**	85.94	80.86	78.86	79.16	80.60	79.76	81.82	77.23	78.85	75.64	78.64	**100.00**	93.33	**97.17**	87.78	83.21	83.58
GPT Image 1 [High] (OpenAI, 2025)	**89.15**	**88.29**	**90.75**	**89.66**	**91.33**	87.08	84.57	84.57	**96.32**	**97.32**	**88.55**	**88.35**	**87.07**	**89.44**	**87.22**	**83.96**	**85.59**	**83.21**	90.00	93.33	89.83	86.83	89.73	**93.46**
Qwen-Image	86.14	86.83	86.18	87.22	90.50	**91.50**	88.22	**90.78**	79.81	79.38	79.30	80.88	79.21	78.94	78.85	81.69	75.57	78.59	**100.00**	**100.00**	92.76	**89.14**	**90.30**	91.42

Table 8: Quantitative evaluation results of English text rendering on CVTG-2K (Du et al., 2025).

Model	Word Accuracy↑					NED↑	CLIPScore↑
	2 regions	3 regions	4 regions	5 regions	average		
SD3.5 Large (Esser et al., 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 et al., 2024)	0.0513	0.1739	0.1948	0.2249	0.1804	0.4675	0.7432
TextDiffuser-2 (Chen et al., 2024a)	0.5322	0.3255	0.1787	0.0809	0.2326	0.4353	0.6765
RAG-Diffusion (Chen et al., 2024d)	0.4388	0.3316	0.2116	0.1910	0.2648	0.4498	0.7797
3DIS (Zhou et al., 2024)	0.4495	0.3959	0.3880	0.3303	0.3813	0.6505	0.7767
TextCrafter (Du et al., 2025)	0.7628	0.7628	0.7406	0.6977	0.7370	0.8679	0.7868
Seedream 3.0 (Gao et al., 2025)	0.6282	0.5962	0.6043	0.5610	0.5924	0.8537	0.7821
GPT Image 1 [High] (OpenAI, 2025)	**0.8779**	**0.8659**	**0.8731**	**0.8218**	**0.8569**	**0.9478**	0.7982
Qwen-Image	0.8370	0.8364	0.8313	0.8158	0.8288	0.9116	**0.8017**

Table 9: Quantitative comparison results of Chinese text rendering.

Model	Level-1 Acc	Level-2 Acc	Level-3 Acc	Overall↑
Seedream 3.0 (Gao et al., 2025)	53.48	26.23	1.25	33.05
GPT Image 1 [High] (OpenAI, 2025)	68.37	15.97	3.55	36.14
Qwen-Image	**97.29**	**40.53**	**6.48**	**58.30**

5.2.3 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 et al., 2025b) and ImgEdit (Ye et al., 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 et al., 2022) and depth estimation (Bochkovskii et al., 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.

GEdit Table 11 reports the results on the **GEdit-Bench** (Liu et al., 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-

Table 10: Quantitative evaluation results on LongText-Bench (Geng et al., 2025). The best result is in bold and the second best result is underlined.

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

Table 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.

Model	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 et al., 2023)	3.58	5.49	3.68	-	-	-
AnyEdit (Yu et al., 2025)	3.18	5.82	3.21	-	-	-
MagicBrush (Zhang et al., 2023)	4.68	5.66	4.52	-	-	-
UniWorld-v1 (Lin et al., 2025)	4.93	7.43	4.85	-	-	-
OmniGen (Xiao et al., 2025)	5.96	5.89	5.06	-	-	-
OmniGen2 (Wu et al., 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 et al., 2025)	7.36	6.83	6.52	7.34	6.85	6.50
FLUX.1 Kontext [Pro] (Labs et al., 2025)	7.02	7.60	6.56	1.11	7.36	1.23
Step1X-Edit (Liu et al., 2025b)	7.66	7.35	6.97	7.20	6.87	6.86
GPT Image 1 [High] (OpenAI, 2025)	7.85	7.62	7.53	7.67	7.56	7.30
Qwen-Image	**8.00**	**7.86**	**7.56**	**7.82**	**7.79**	**7.52**

Table 12: Quantitative comparison results on ImgEdit (Ye et al., 2025). All metrics are evaluated by GPT-4.1. "Overall" is calculated by averaging all scores across tasks.

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

Image ranks at the top of both the English and Chinese leaderboards, demonstrating strong editing capability and generalization to multilingual user instructions.

ImgEdit Table 12 presents the results on the **ImgEdit** benchmark (Ye et al., 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.

Novel view synthesis Table 13 shows the results of novel view synthesis on GSO (Downs et al., 2022) dataset. We compare the similarity of the generated novel view image of a 3D object given its front view

Table 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 et al., 2022) dataset.

Model	**PSNR↑**	**SSIM↑**	**LPIPS↓**
Zero123 (Liu et al., 2023)	13.48	0.854	0.166
ImageDream (Wang &Shi, 2023)	15.22	0.883	0.164
CRM (Wang et al., 2024b)	**15.93**	**0.891**	**0.152**
GPT Image 1 [High] (OpenAI, 2025)	12.07	0.804	0.361
BAGEL (Deng et al., 2025)	13.78	0.825	0.237
FLUX.1 Kontext [Pro] (Labs et al., 2025)	14.50	0.859	0.201
Qwen-Image	**15.11**	**0.884**	**0.153**

Table 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.

Model	**KITTI**		**NYUv2**		**ScanNet**		**DIODE**		**ETH3D**	
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
MiDaS (Ranftl et al., 2020)	0.236	0.630	0.111	0.885	0.121	0.846	0.332	0.715	0.184	0.752
DPT-large (Ranftl et al., 2021)	0.100	0.901	0.098	0.903	0.082	0.934	0.182	0.758	0.078	0.946
DepthAnything (Yang et al., 2024a)	0.080	0.946	0.043	0.980	0.043	0.981	0.261	0.759	0.058	**0.984**
DepthAnything v2 (Yang et al., 2024b)	0.080	0.943	0.043	0.979	0.042	0.979	0.321	0.758	0.066	0.983
Depth Pro (Bochkovskii et al., 2024)	0.055	0.974	0.042	0.977	0.041	0.978	0.217	0.764	0.043	0.974
Metric3D v2 (Hu et al., 2024a)	**0.052**	**0.979**	**0.039**	**0.979**	**0.023**	**0.989**	**0.147**	**0.892**	**0.040**	0.983
GeoWizard (Fu et al., 2024)	0.097	0.921	**0.052**	0.966	0.061	0.953	0.297	0.792	**0.064**	0.961
DepthFM (Gui et al., 2024)	0.083	0.934	0.065	0.956	-	-	0.225	0.800	-	-
Marigold (Ke et al., 2024)	0.099	0.916	0.055	0.964	0.064	0.951	0.308	0.773	0.065	0.960
DMP (Lee et al., 2024)	0.240	0.622	0.109	0.891	0.146	0.814	0.361	0.706	0.128	0.857
Qwen-Image	**0.078**	**0.951**	0.055	**0.967**	**0.047**	**0.974**	**0.197**	**0.832**	0.066	**0.962**

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.

Depth Estimation Table 14 summarizes performance on five widely used datasets: NYUv2 (Nathan Silberman &Fergus, 2012), KITTI (Geiger et al., 2013), ScanNet (Dai et al., 2017), DIODE (Vasiljevic et al., 2019), and ETH3D (Schops et al., 2017). During training, we adopt DepthPro (Bochkovskii et al., 2024) as the teacher model to provide supervisory depth signals, following the protocol used in previous work such as DICEPTION (Zhao et al., 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.

5.3 Qualitative Results

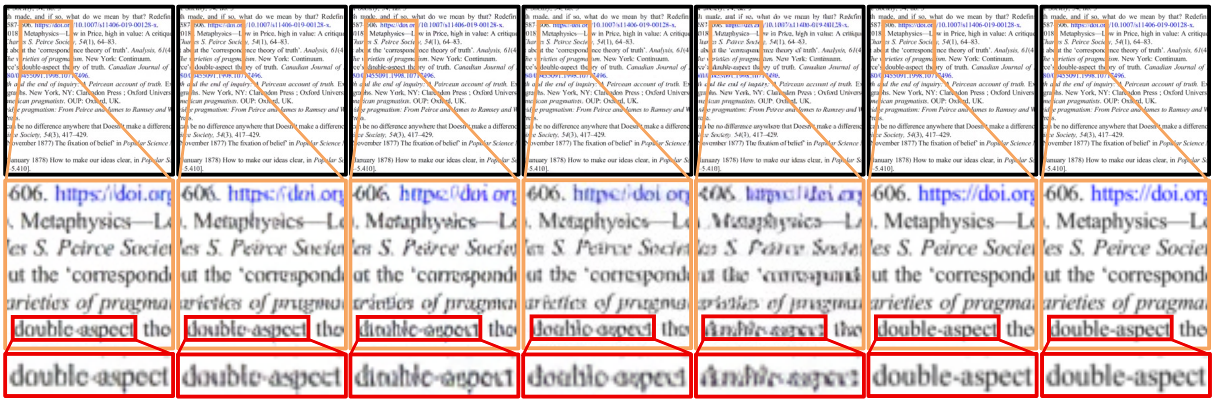
To comprehensively compare the visual generation capabilities of Qwen-Image and state-of-the-art models, we first qualitatively compared the reconstruction performance of VAE on text-rich images in §§ 5.3.1. We further conducted qualitative comparison on two basic visual generation tasks, text-to-image (T2I) generation in §§ 5.3.2 and image editing (TI2I) in §§ 5.3.3, to comprehensively evaluate the basic generation capabilities of the models.

5.3.1 Qualitative Results on VAE Reconstruction

Figure 17 presents qualitative results of reconstructing text-rich images with the state-of-the-art image VAEs. The first row illustrates the reconstruction of a PDF image containing English text. In our result, the phrase "double-aspect" remains clearly legible, whereas it is unrecognizable in the reconstructions produced by other models. Overall, Qwen-Image-VAE delivers more precise reconstructions for images with small texts.

5.3.2 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



FLUX Hunyuan SD-3.5 Cosmos-CI Wan2.1 Qwen-Image Original

Figure 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.

against leading Text-to-Image foundation models, including both close-source models (GPT Image 1 [High] (OpenAI, 2025), Seedream 3.0 (Gao et al., 2025), Recraft V3 (Recraft, 2024) and open-source models (Hidream-I1-Full (Cai et al., 2025) and Lumina-Image 2.0 (Qin et al., 2025)).

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.

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.

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.

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.

5.3.3 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 et al., 2025)[3], FLUX.1 Kontext [Pro] (Labs et al., 2025), and GPT Image 1 [High] (OpenAI, 2025).

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.

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.

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.

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.

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.

[3] SeedEdit 3.0 does not provide an official API, all watermarked images were obtained via their web interface.

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.

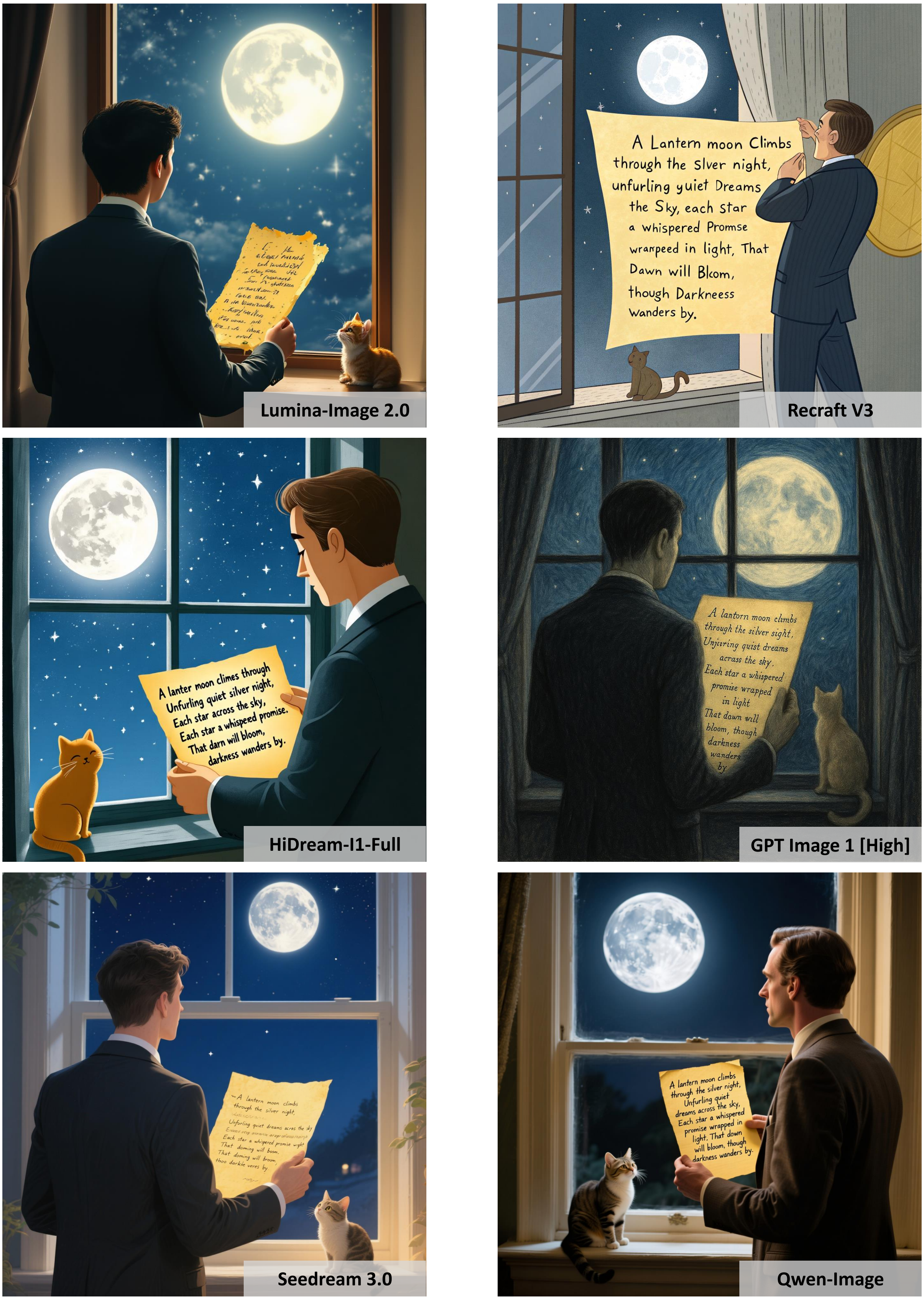
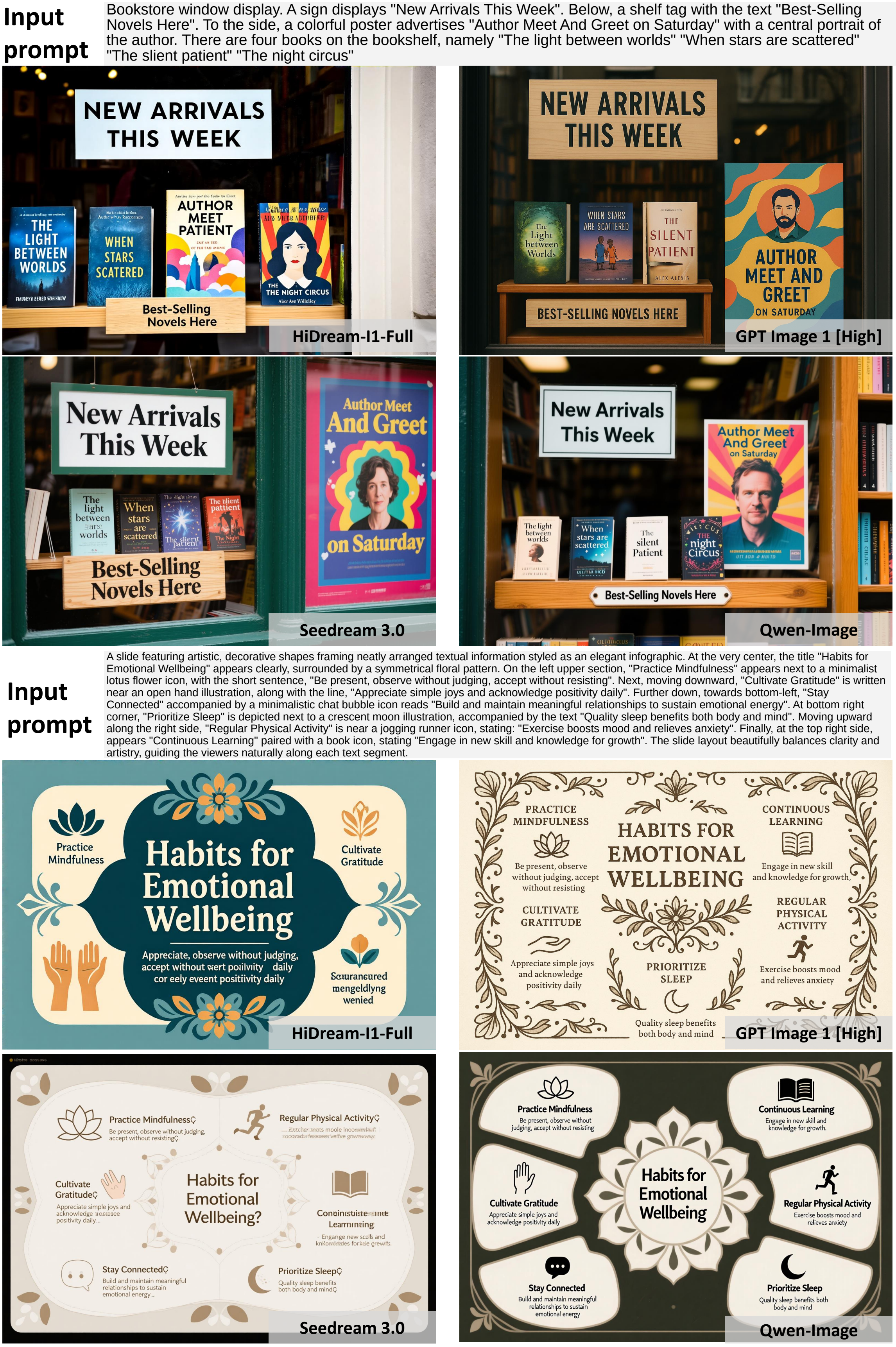

Lumina-Image 2.0
A Lantern moon Climbs through the Slver night, unfurling yuiet Dreams the Sky, each star a whispered Promse wrarpeed in light, That Dawn will Bkom, though Darkneess wanders by.
Recraft V3
A lanter moon climes through Unfurling quiet silver night, Each star across the sky, Each star a whispeed promise. That darn will bloom, darkness wanders by.
HiDream-I1-Full
A lantorn moon climbs through the silver sight, Unjiuring quist dreams acrass the sky. Each star a whispered promise wrapped in light That down will bloom, though darkness wanders by
GPT Image 1 [High]
A lantern moon climbs through the silver night, Unfurling quiet dreams acros the sky Each star a whispered promise wright That doming will bosm, That doming will broom, thoo darkle veres by
Seedream 3.0
A lantern moon climbs through the silver night, Unfurling quiet dreams across the sky, Each star a whispered promise wrapped in light. That down will bloom, though darkness wanders by.
Qwen-Image

Figure 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.



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 slient patient" "The night circus"

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 minimalistic 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.

Practice Mindfulness
Habits for Emotional Wellbeing
Appreciate, observe without judging, accept without wert poilvity daily cor eely eveent positivity daily
Cultivate Gratitude
Scauranoured mengeldlyng wenied
HiDream-I1-Full
PRACTICE MINDFULNESS
Be present, observe without judging, accept without resisting
HABITS FOR EMOTIONAL WELLBEING
CONTINUOUS LEARNING
Engage in new skill and knowledge for growth.
CULTIVATE GRATITUDE
Appreciate simple joys and acknowledge positivity daily
REGULAR PHYSICAL ACTIVITY
Exercise boosts mood and relieves anxiety
PRIORITIZE SLEEP
Quality sleep benefits both body and mind
GPT Image 1 [High]
Practice Mindfulness
Be present, observe without judging, accept without resisting
Regular Physical Activity
Cultivate Gratitude
Appreciate simple joys and acknowledge positivity daily
Habits for Emotional Wellbeing?
Continuous Learning
Stay Connected
Build and maintain meaningful relationships to sustain emotional energy
Prioritize Sleep
Quality sleep benefits both body and mind
Seedream 3.0
Practice Mindfulness
Be present, observe without judging, accept without resisting
Continuous Learning
Engage in new skill and knowledge for growth.
Cultivate Gratitude
Appreciate simple joys and acknowledge positivity daily
Habits for Emotional Wellbeing
Regular Physical Activity
Exercise boosts mood and relieves anxiety
Stay Connected
Build and maintain meaningful relationships to sustain emotional energy
Prioritize Sleep
Quality sleep benefits both body and mind
Qwen-Image

Figure 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.

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。

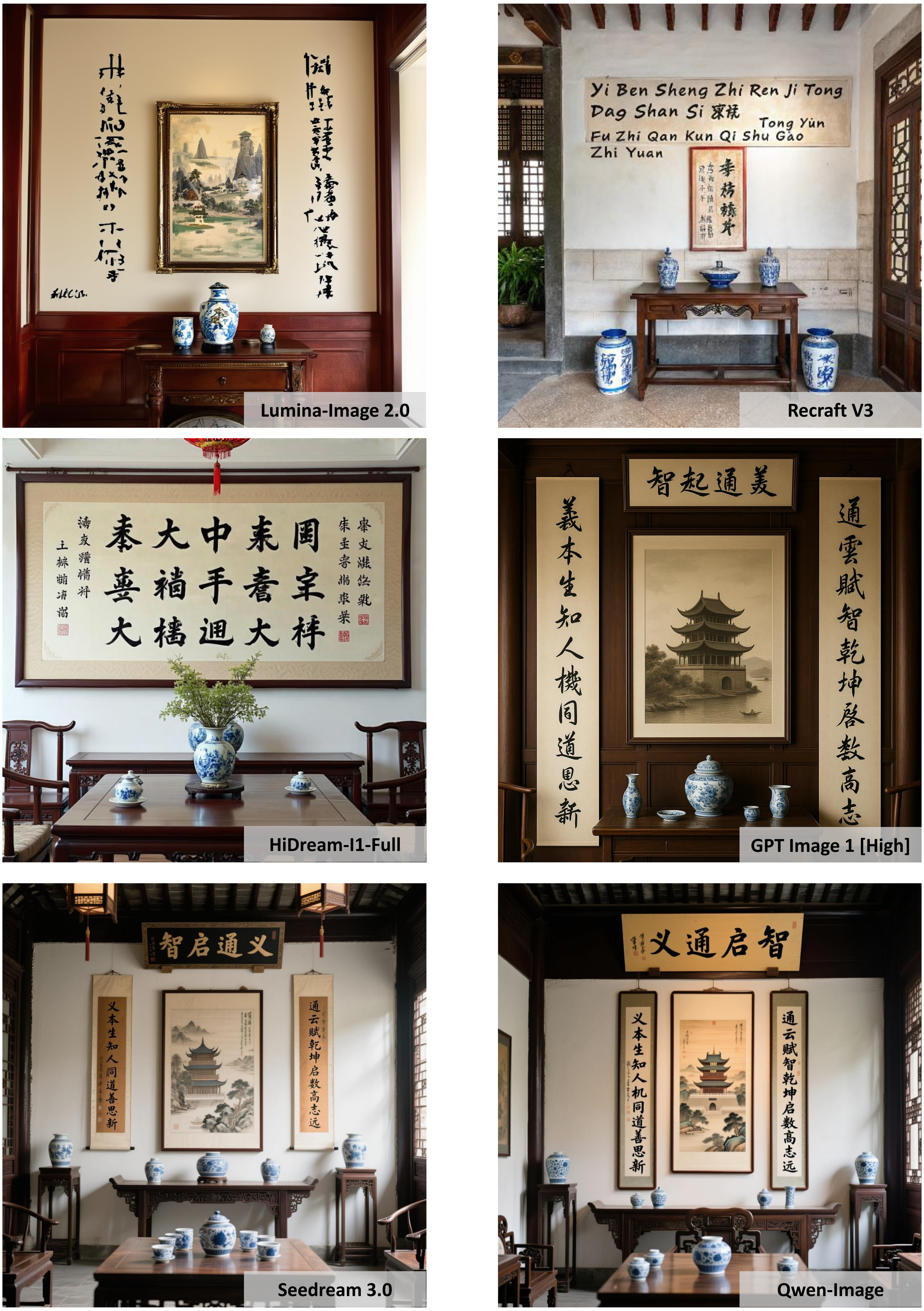
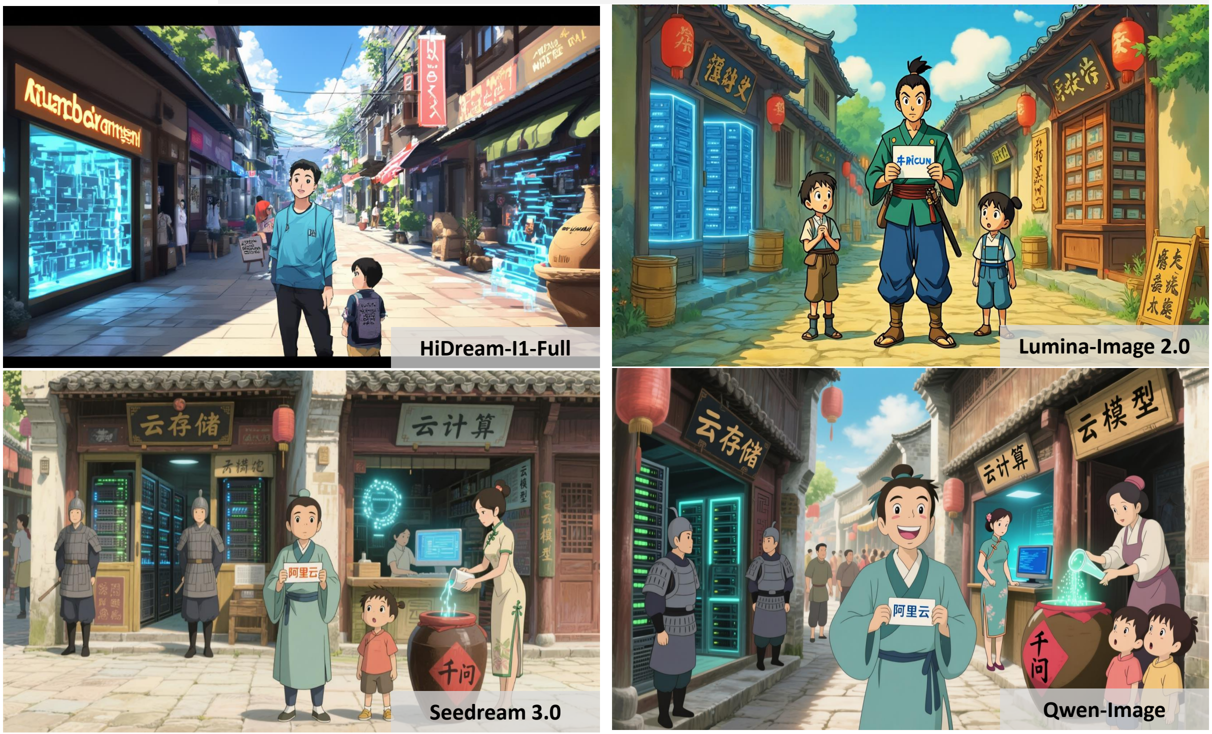


Figure 20: Comparison of Chinese text rendering capability in image generation. Qwen-Image accurately generates the expected Chinese couplet. In contrast, GPT Image 1 [high] and Seedream 3.0 miss or generate distorted characters. While other model in comparison cannot generate correct Chinese couplets.

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面相镜头微笑。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。"

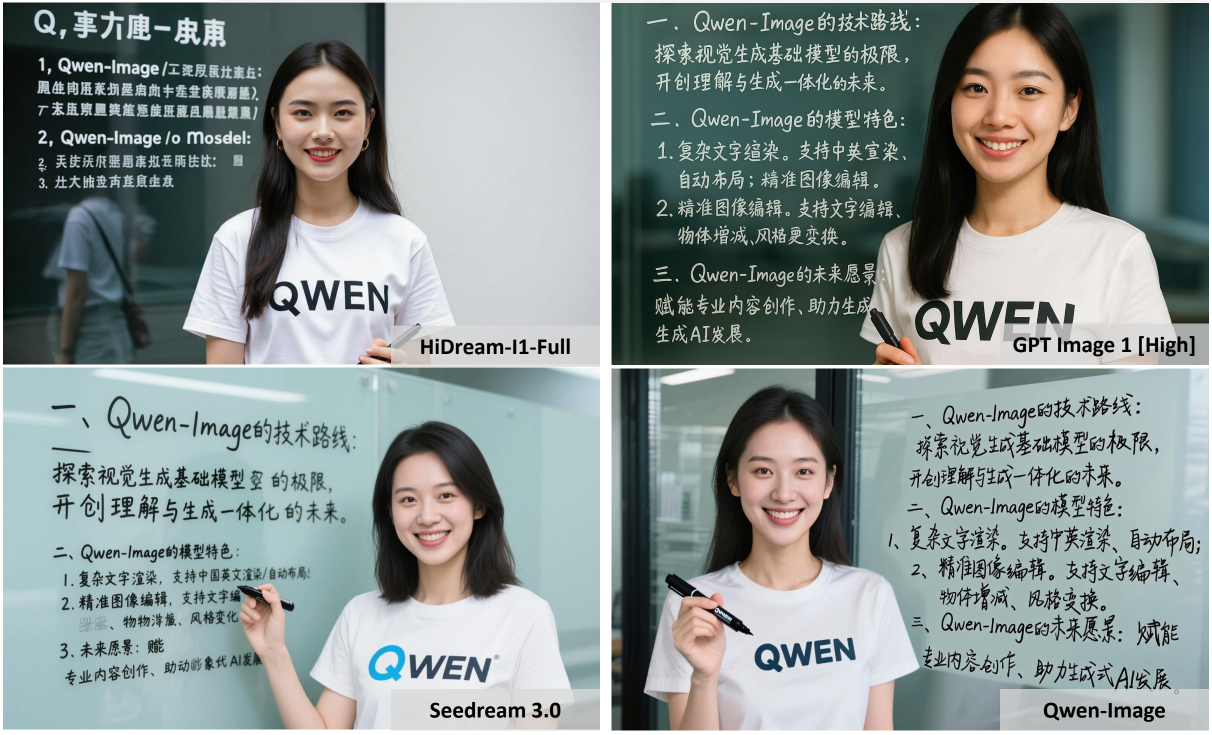
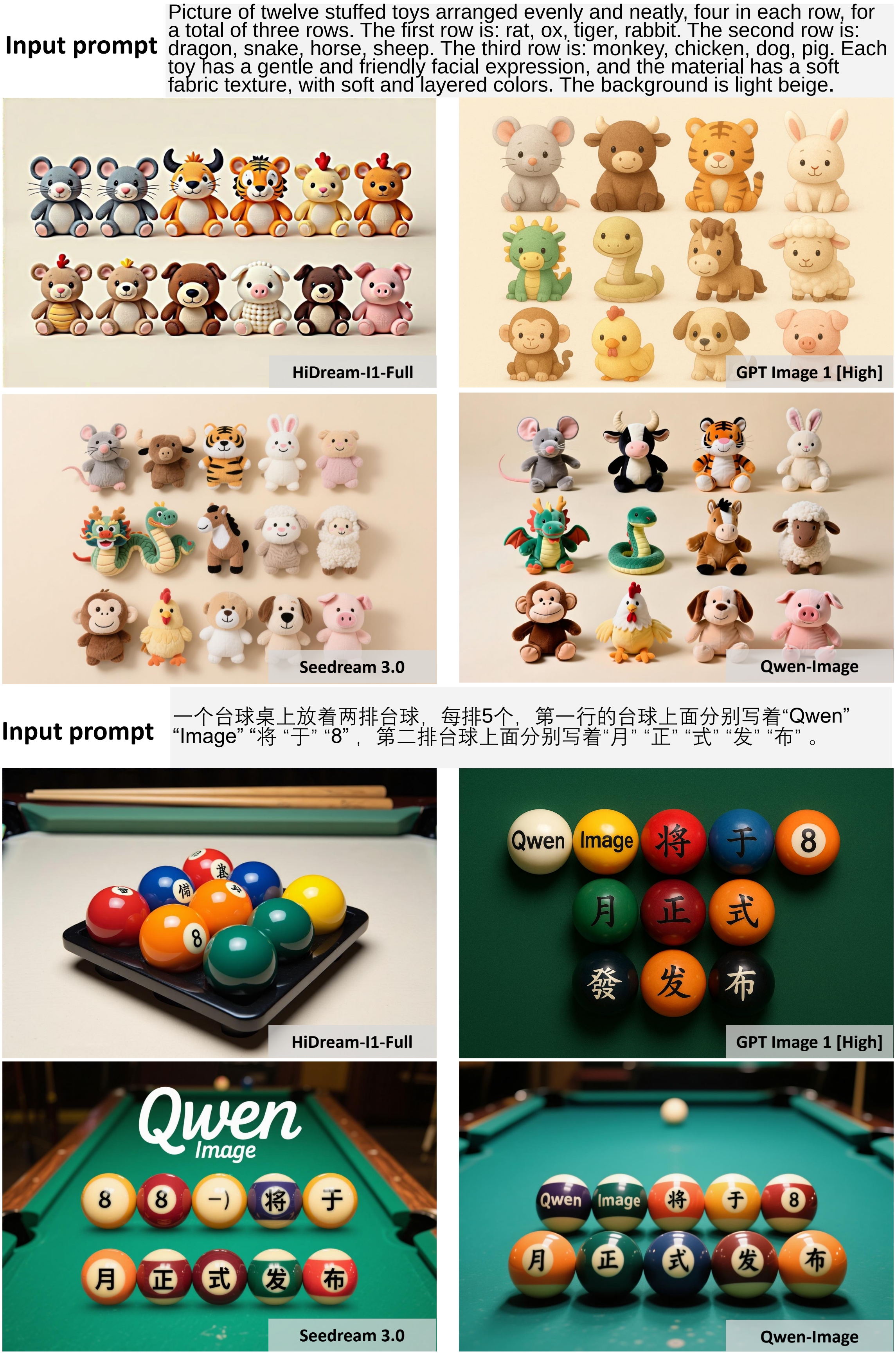

QWEN
HiDream-I1-Full
一、Qwen-Image的技术路线：
探索视觉生成基础模型的极限，
开创理解与生成一体化的未来。
二、Qwen-Image的模型特色：
1.复杂文字渲染。支持中英文染、
自动布局；精准图像编辑。
2.精准图像编辑。支持文字编辑、
物体增减、风格更变换。
三、Qwen-Image的未来愿景：
赋能专业内容创作、助力生成
生成AI发展。
QWEN
GPT Image 1 [High]
一、Qwen-Image的技术路线：
探索视觉生成基础模型 的极限，
开创理解与生成一体化的未来。
QWEN
Seedream 3.0
一、Qwen-Image的技术路线：
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
QWEN
Qwen-Image

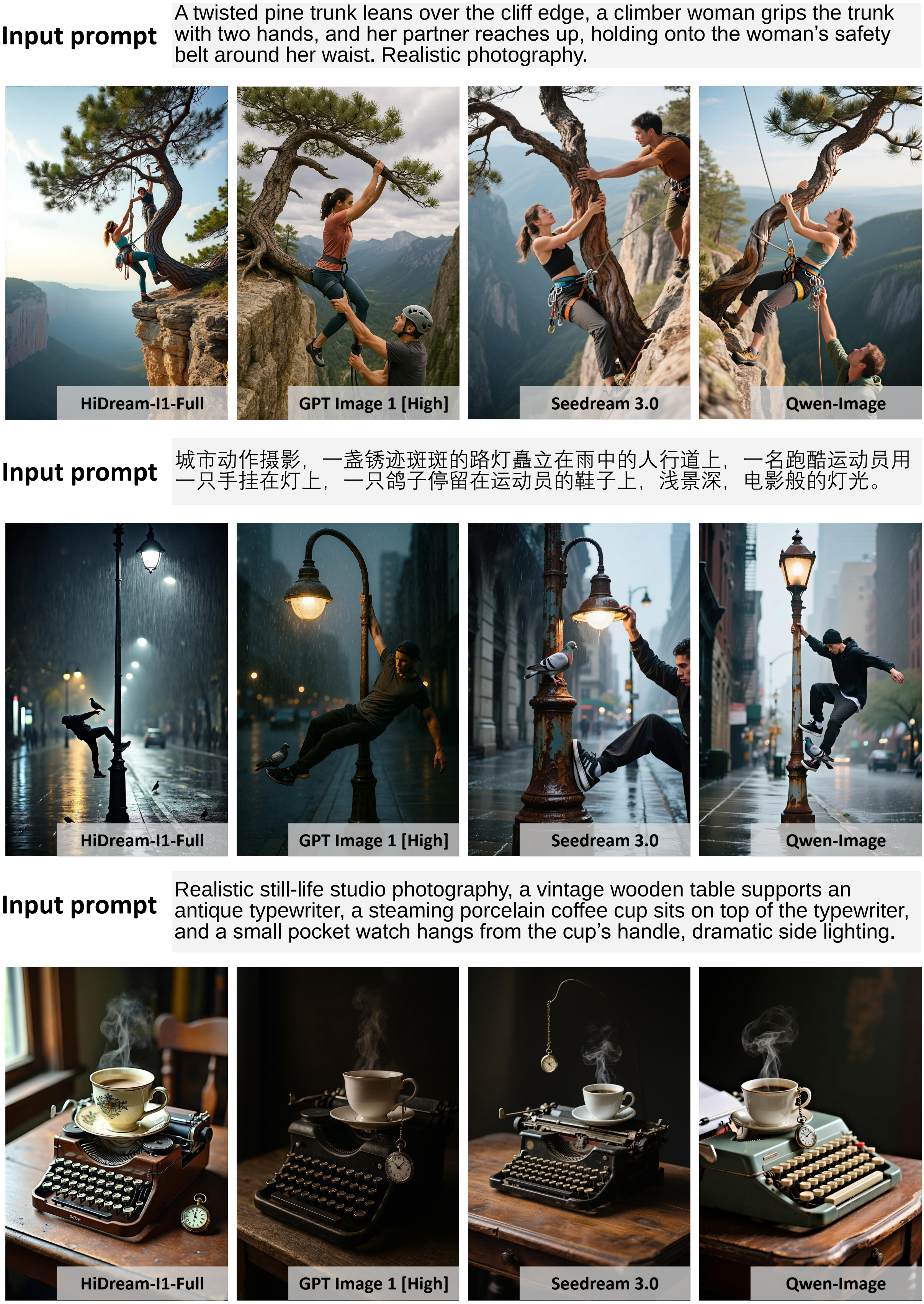
Figure 21: Comparison of complex Chinese text rendering in image generation: the first case shows that Qwen-Image can render text on multiple objects while maintaining consistency with the real scene, such as aligning text with the depth and tilt of each plaque; the second case demonstrates its ability to render structured paragraph text in a glass panel. Qwen-Image is the only model capable of accurately rendering long text.



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” 。

Figure 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.



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.

Figure 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.

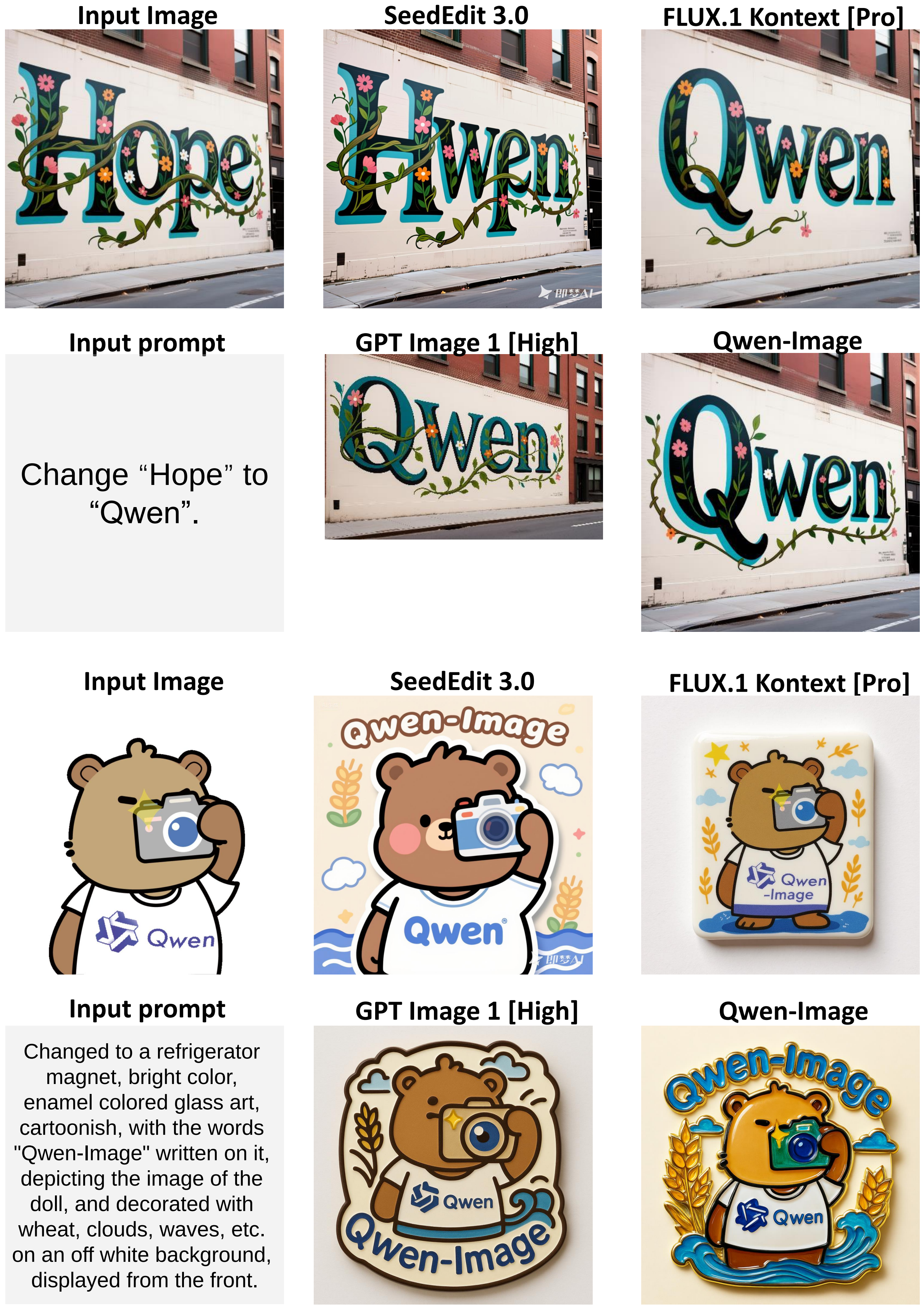

Input Image
SeedEdit 3.0
FLUX.1 Kontext [Pro]
Input prompt
Change “Hope” to “Qwen”.
GPT Image 1 [High]
Qwen-Image
Input Image
SeedEdit 3.0
FLUX.1 Kontext [Pro]
Input prompt
Changed to a refrigerator magnet, bright color, enamel colored glass art, cartoonish, with the words "Qwen-Image" written on it, depicting the image of the doll, and decorated with wheat, clouds, waves, etc. on an off white background, displayed from the front.
GPT Image 1 [High]
Qwen-Image

Figure 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.

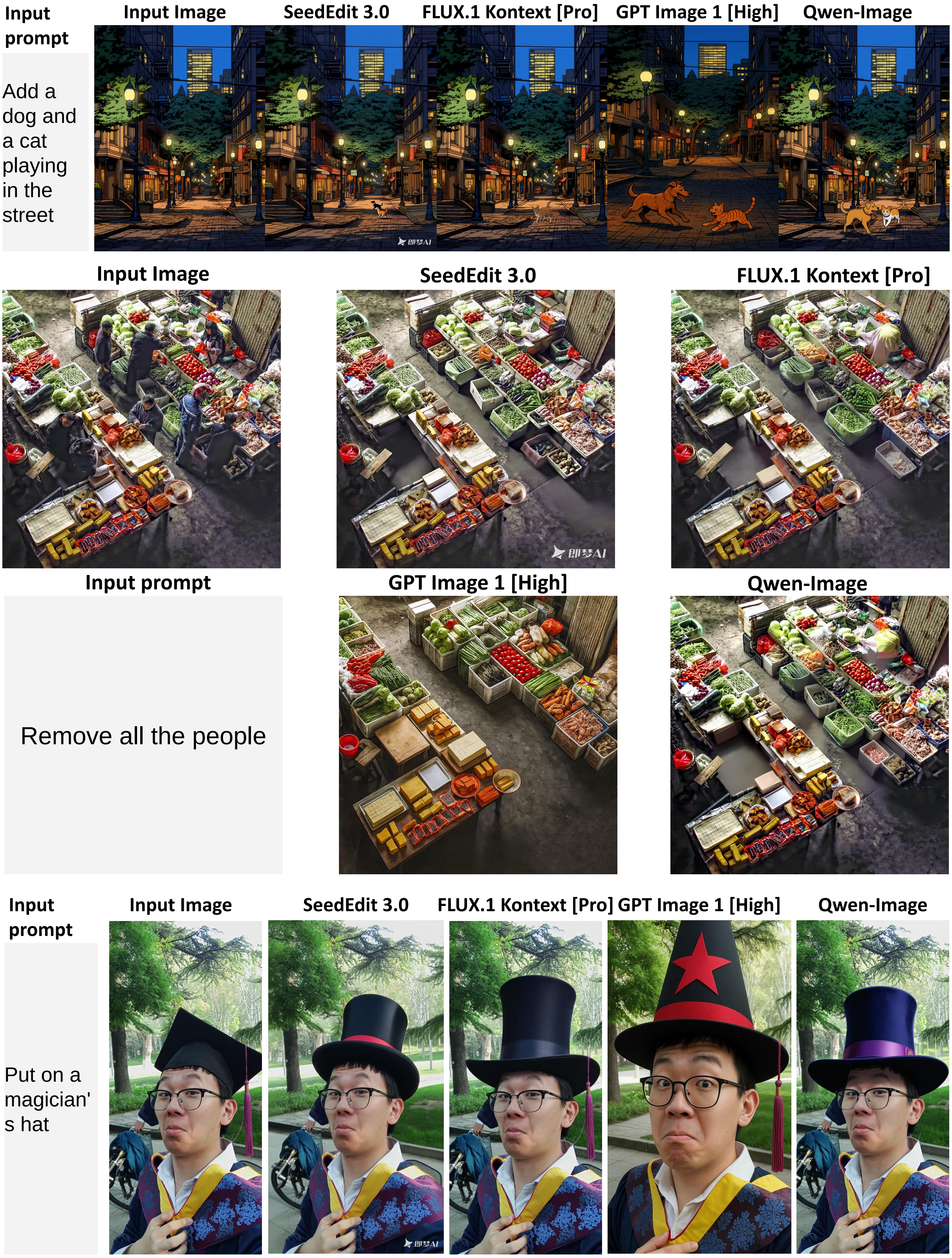


Figure 25: Qualitative comparison on object editing (addition, removal, and replacement): Object editing is a relatively stable capability for all models. Qwen-Image demonstrates superior consistency in unmodified regions and achieves better style alignment for the newly generated objects.

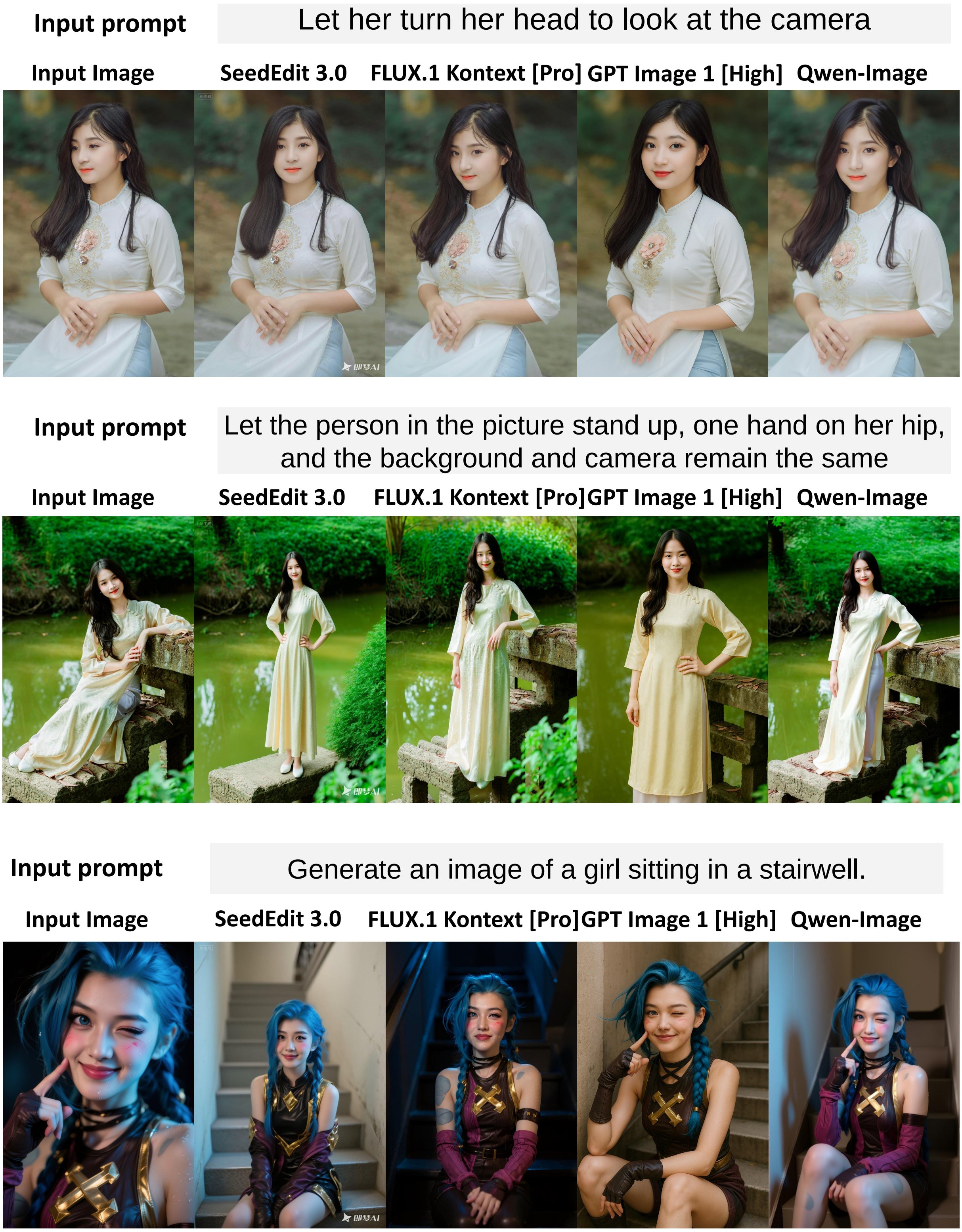
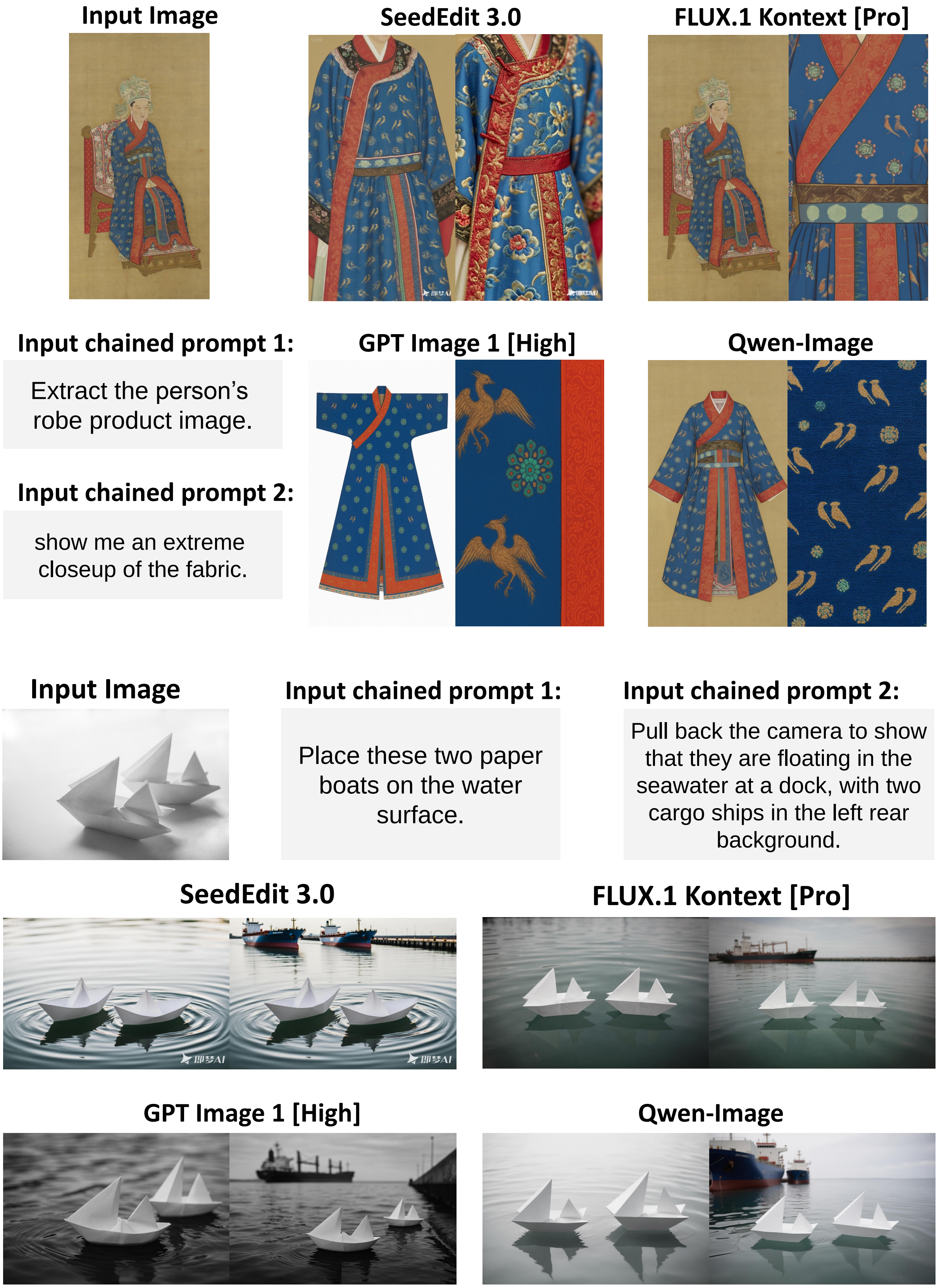


Figure 26: Qualitative comparison on pose manipulation. Qwen-Image is able to accurately follow pose manipulation instructions while preserving fine details of the person (such as hair strands) and maintaining consistency in the background (e.g., the stone steps behind the subject).



Extract the person's robe product image.

show me an extreme closeup of the fabric.

Place these two paper boats on the water surface.

Pull back the camera to show that they are floating in the seawater at a dock, with two cargo ships in the left rear background.

Figure 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.

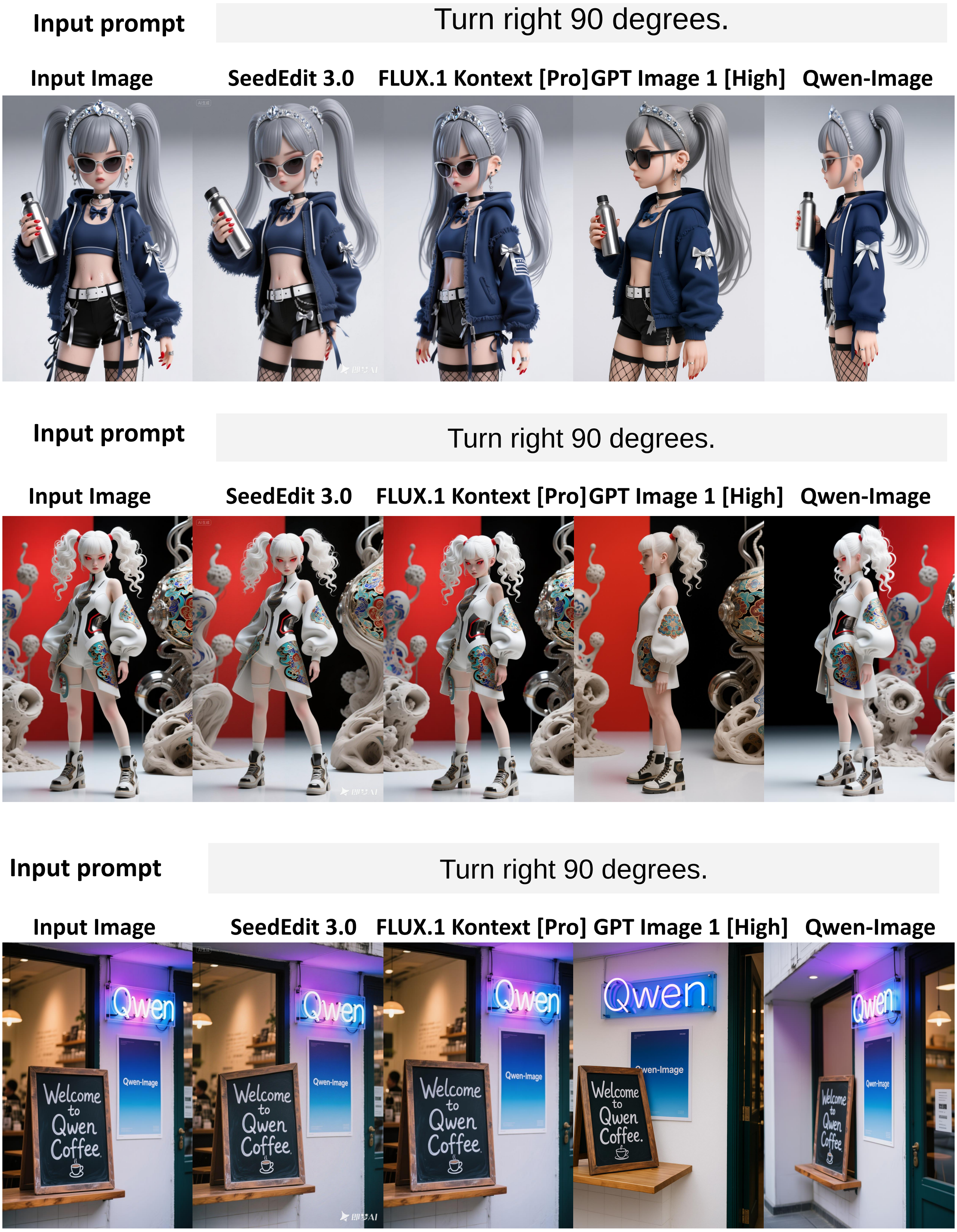


Figure 28: Qualitative comparison on novel view synthesis. We evaluate three progressively challenging rotation tasks under the same text prompt "Turn right 90 degrees": (1) single-person rotation, (2) co-rotation of person and background, and (3) real-world scenario. In all cases, Qwen-Image achieves 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.

6 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.

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

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.

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.

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.

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.

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.

7 Authors

Core Contributors[4]**:** 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

Contributors[5]**:** 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

[4] Alphabetical order.

[5] Alphabetical order.

References

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

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-i1: 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-σ: 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-α: 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, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 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 rgbd 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 Tschannen, 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, Feiwu 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. Seededit 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/THUDM/CogView4.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

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

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.