

Qwen-Image-Layered 1.5: Unified Image Generation and Editing with Native Transparency Awareness

Shengming Yin¹, Zekai Zhang², Yixian Xu², Kaiyuan Gao², Zhendong Wang²
Lihan Jiang², Ningyuan Tang², Tianhe Wu², Xiaoyue Chen², Yanran Zhang², Zihao Liu²
Zikai Zhou², Jiahao Li², Xiao Xu², Kun Yan², Yuxiang Chen², Yilei Chen², Yan Shu²
Jie Zhang², Heung-Yeung Shum³, Lionel M. Ni¹, Jingren Zhou², Chenfei Wu^{2†}
¹HKUST(GZ) ²Alibaba ³HKUST

syin460@connect.hkust-gz.edu.cn fulai.hr@alibaba-inc.com

Abstract

Recent advances in image generation and editing have been driven by powerful pretrained unified models, yet most existing approaches remain fundamentally raster-centric: they represent images as opaque pixels without natively modeling transparency. This limitation restricts their applicability to a broad range of real-world applications. In this work, we present **Qwen-Image-Layered 1.5**, a pretrained unified image generation and editing model with native transparency awareness. By jointly modeling visual appearance and transparency, our framework supports a wide range of both raster-centric and transparency-aware tasks within a single model. Notably, we observe that the model can perform transparent image editing in a zero-shot manner. To systematically evaluate this new task, we introduce **TransEdit-Bench**, the first benchmark specifically designed for transparent image editing. Extensive experiments across various tasks demonstrate strong performance and highlight the great potential of transparency-native image modeling as a unified paradigm for future image synthesis and editing.

1 Introduction

Recent image generative models have achieved remarkable progress toward unifying image generation and editing within a single architecture (Labs et al., 2025; Wu et al., 2025a; Cai et al., 2025a; Kou et al., 2026; Feng et al., 2026). However, existing unified visual generation models remain predominantly raster-centric, focusing exclusively on opaque images and assuming that all visual content can be represented as three-channel pixels.

This assumption overlooks transparency, a fundamental component of visual content in many real-world applications such as graphic design, digital asset creation, and compositional image editing, where appearance and transparency

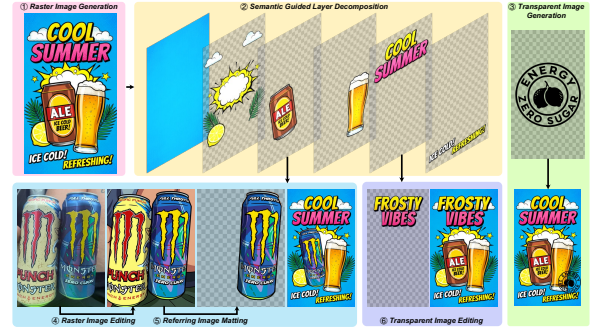


Figure 1: Qwen-Image-Layered 1.5 supports a wide range of both raster-centric and transparency-aware tasks. **All the tasks are performed in one model within a share latent space.**

jointly determine not only how an object looks, but also how it can be edited or recomposed. In contrast, RGB images entangle foreground appearance, opacity, and background into a single raster observation. As a result, current unified image models cannot natively support transparency-aware tasks that require explicit reasoning over alpha and compositional structure.

Meanwhile, existing efforts on transparency modeling are relatively fragmented. Tasks involving alpha channels often rely on task-specific architectures (Huang et al., 2025; Pu et al., 2025; Huang et al., 2026), additional fine-tuning (Zhang and Agrawala, 2024; Dai et al., 2025), or post-processing pipelines (Quattrini et al., 2024). Moreover, different tasks are often studied in isolation under highly specialized model designs, such as image matting (Yao et al., 2024a; Kim et al., 2025; Hu et al., 2024; Xia et al., 2024; Yao et al., 2024b; Xiang et al., 2025) and image decomposition (Yang et al., 2025; Suzuki et al., 2025; Wang et al., 2025c; Liu et al., 2025c; Yin et al., 2025).

In this work, we argue that transparency should be treated as a first-class property of visual generative models, rather than an auxiliary attribute of raster images. Based on this insight, we present a unified framework for image generation and editing with native transparency awareness. Our

[†]Corresponding author.

core idea is to jointly model appearance and transparency within a shared latent space, enabling a single generative model to seamlessly support both raster-centric and transparency-aware tasks.

To this end, we propose RGBA Visual foundation model Aligned Variational AutoEncoder (RGBA-VAAE), which learns a unified latent representation for both RGB and RGBA images. A key challenge is that transparent images are both less abundant and statistically different from standard RGB images, making a naive four-channel extension (Wang et al., 2025b) insufficient for learning transferable semantics. To address this issue, we introduce a visual foundation model alignment loss (Yao et al., 2025) that aligns the learned latent space with strong pretrained visual priors (Siméoni et al., 2025), thereby improving both optimization efficiency and cross-format generalization. Built on this unified representation, our transparency-aware framework supports a broad range of tasks without requiring task-specific architectural modifications, as illustrated in Fig. 1. Notably, the model demonstrates zero-shot transparent image editing capability, suggesting that transparency-native modeling can unlock capabilities beyond those of existing unified models. To systematically evaluate this new task, we further introduce TransEdit-Bench, the first benchmark for text-guided transparent image editing.

We summarize our contributions as follows:

- We propose **Qwen-Image-layered 1.5**, a pretrained unified image generation and editing framework with native transparency awareness, seamlessly supporting both raster-centric and transparency-aware tasks within a single architecture.
- We introduce RGBA-VAAE, a 4-channel VAE that learns a shared latent space for RGB and RGBA images. Furthermore, we incorporate a visual foundation model alignment loss to improve training efficiency and cross-format generalization.
- Our transparency-aware framework enables zero-shot transparent image editing. For rigorous evaluation, we present **TransEdit-Bench**, the first benchmark dedicated to text-guided transparent image editing.
- Extensive experiments demonstrate state-of-the-art performance across a broad range of tasks, establishing a new paradigm for unified visual generative modeling.

2 Related Work

2.1 Raster-Centric Image Synthesis

The quality of image generation has improved remarkably through the combination of diffusion-based generative frameworks (Ho et al., 2020; Song et al., 2020; Lipman et al., 2022) and the scalability of Transformer backbones (Vaswani et al., 2017). Early approaches relied on latent diffusion models (Rombach et al., 2022; Podell et al., 2024), which were subsequently evolved to diffusion Transformer backbones (Peebles and Xie, 2023; Chen et al., 2024; Li et al., 2024b; Esser et al., 2024). More recent paradigms further incorporate vision-language models as condition encoders (Wu et al., 2025a; Team, 2025), enabling finer-grained semantic control.

Parallel to advances in generation, image editing has also made significant progress, particularly in enhancing instruction-following capability and maintaining visual consistency (Brooks et al., 2023; Zhang et al., 2023a; Wang et al., 2025a; Liu et al., 2025b). For instance, Qwen-Image-Edit (Wu et al., 2025a) leverages both high-level semantic features from vision-language models (Bai et al., 2025b) and low-level reconstructive features from VAEs (Kingma and Welling, 2013) to improve edit fidelity and semantic alignment.

More recently, unifying image generation and editing within a single framework has emerged as a dominant paradigm (Xia et al., 2025; Labs et al., 2025; Wu et al., 2025a; Feng et al., 2026), fostering cross-task enhancement and improving deployment efficiency. Despite these advances, existing methods remain fundamentally raster-centric, operating exclusively on opaque pixels. This architectural constraint precludes the explicit modeling of alpha channels and layered structure, making transparency-aware tasks inherently infeasible.

2.2 Transparent-Aware Image Synthesis

Transparent image synthesis and processing, essential for numerous real-world applications, have garnered growing attention. Early studies adapted RGB generators to produce RGBA outputs via additional fine-tuning (Zhang and Agrawala, 2024) or post-processing (Quattrini et al., 2024). A closely related task is image matting (Li et al., 2021b; Yao et al., 2024a; Kim et al., 2025; Xia et al., 2024; Yao et al., 2024b), which focuses on estimating the alpha channel from a given image. Along this line, Matting Anything (Li et al., 2024a)

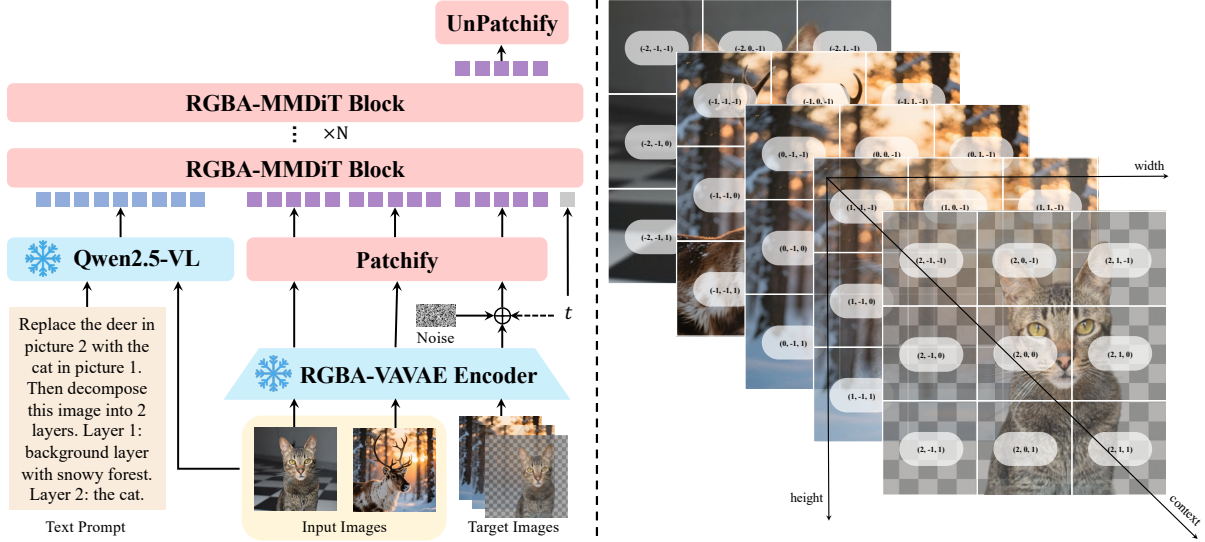


Figure 2: Overview of Qwen-Image-Layered 1.5 architecture. Left: Illustration of RGBA-MMDiT, where all images are represented in four channels and encoded by the RGBA-VAAE Encoder, while input images are simultaneously encoded by Qwen2.5-VL to extract semantic embeddings. Right: Illustration of Bi3D RoPE, which introduces a new context dimension to distinguish between input and output images.

leverages the Segment Anything Model (Kirillov et al., 2023) as its backbone, enabling flexible and interactive user guidance. More recently, diffusion models (Xu et al., 2023; Hu et al., 2024; Xiang et al., 2025) have been successfully applied to image matting. For instance, DiffMatte (Hu et al., 2024) introduces a lightweight diffusion decoder to improve efficiency and employs a modified self-aligned strategy to tackle sampling drift.

Meanwhile, multilayer image generation has seen rapid advancements (Zhang et al., 2023b; Tudosi et al., 2024; Fontanella et al., 2024; Huang et al., 2024; Chen et al., 2025a; Huang et al., 2025; Pu et al., 2025; Liu et al., 2025a; Huang et al., 2026). Among these, LayerDiff (Huang et al., 2024) explicitly models cross-layer dependencies through carefully designed inter- and intra-layer attention mechanisms. Concurrently, several studies have also explored decomposing images into layers. For instance, LayerD (Suzuki et al., 2025) iteratively extracts the topmost unoccluded layer and completes the background, while LayeringDiff (Kang et al., 2025) leverages mask conditioning to separate foreground and background layers. More recently, Qwen-Image-Layered (Yin et al., 2025) introduces an end-to-end diffusion model capable of directly decomposing RGB images into a variable number of layers.

Despite substantial progress, existing methods typically address each transparency-related task in isolation, lacking a unified framework. In particular, transparent image editing remains underexplored, representing a critical gap in the field.

3 Method

3.1 RGBA-VAAE

Latent diffusion models (LDMs) (Rombach et al., 2022) leverage variational autoencoders (VAEs) (Kingma and Welling, 2013) to reduce data dimensionality, thereby improving both training and sampling efficiency. However, most existing VAEs primarily focus on three-channel RGB image reconstruction (Yao et al., 2025; Zheng et al., 2025). To accommodate transparency, LayerDiff (Zhang and Agrawala, 2024) introduces a shift in the latent space to encode the alpha channel; ART (Pu et al., 2025) maps transparency into a grayscale to facilitate encoding; AlphaVAE (Wang et al., 2025b) extends the VAE to four channels, though it focuses exclusively on transparent image reconstruction. These discrepancies hinder the development of a unified generation and editing model capable of seamlessly handling both opaque and transparent imagery.

To address this limitation, we propose jointly modeling opaque and transparent images within a shared latent space. To this end, we introduce RGBA Visual foundation model Aligned Variational AutoEncoder (RGBA-VAAE). In order to learn a unified representation, we train the model on both types of images. However, because transparent images are both less abundant and statistically different from standard RGB images, a naive four-channel extension is insufficient to learn transferable semantics. To bridge this gap, we propose an alpha-aware visual foundation

model alignment loss (α -VF loss). Since vision foundation models accept only three-channel inputs, we composite RGBA images over both white and black backgrounds to extract semantic features. Formally, the alignment loss is defined as:

$$\mathcal{L}_{\alpha\text{-VF}} = \frac{1}{2} \left[\text{VF}(\mathcal{E}(x), \text{DINOv3}(\mathcal{A}(x, \mathbf{0}))) + \text{VF}(\mathcal{E}(x), \text{DINOv3}(\mathcal{A}(x, \mathbf{1}))) \right]$$

Here, $\mathcal{A}$ denotes the alpha compositing operation, x is the input image, and $\mathcal{E}$ is the VAE encoder. For the visual foundation model, we employ DINOv3 (Siméoni et al., 2025). Following VAAE (Yao et al., 2025), $\text{VF}(\cdot)$ incorporates both a marginal cosine similarity loss and a marginal distance matrix similarity loss. Notably, for opaque images ($\alpha = 1$), alpha compositing leaves the visual content unchanged, and the formulation thereby degenerates to the standard visual foundation model alignment loss. The overall training objective is formulated as a weighted sum of reconstruction, perceptual, KL regularization, and the proposed alignment loss.

3.2 RGBA-MMDiT

Based on this unified representation, we propose RGBA-MMDiT, a framework designed to accommodate both variable-length inputs (e.g., multi-image editing) and outputs (e.g., multi-layer image decomposition), thereby enabling a wide range of tasks across different image formats.

Formally, the model accepts M input images, $I \in \mathbb{R}^{M \times H \times W \times 4}$, where all images are represented in four channels and the alpha channel of RGB images is set to 1. The model then generates N output images $O \in \mathbb{R}^{N \times H \times W \times 4}$. Following Qwen-Image (Wu et al., 2025a), we adopt a flow-matching (Lipman et al., 2022) training objective. Let $x_0 \in \mathbb{R}^{N \times h \times w \times c}$ denote the latent representation of the target images. We sample initial noise x_1 from a standard Gaussian distribution and a timestep $t \in [0, 1]$ from a logit-normal distribution. The intermediate state x_t and velocity field v_t at timestep t are defined as:

$$x_t = tx_1 + (1 - t)x_0$$

$$v_t = \frac{dx_t}{dt} = x_1 - x_0$$

For the input images I , we employ Qwen2.5-VL (Bai et al., 2025b) to extract semantic embeddings. Specifically, images are first encoded

by a Vision Transformer (ViT), then concatenated with text tokens and fed into a large language model; the output hidden states serve as the semantic condition h . For transparent images, we pre-composite them over a white background before feeding them into ViT. Furthermore, we incorporate reconstructive features to preserve visual consistency: the input images are encoded by the RGBA-VAAE Encoder, and the latent representations z_I are concatenated with the noisy image latent x_t along the sequence dimension. The model is optimized to predict the target velocity by minimizing the mean squared error between the predicted value $v_\theta(x_t, t, z_I, h)$ and the ground-truth v_t , where $\mathcal{D}$ denotes the training dataset.

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{(x_t, t, z_I, h) \sim \mathcal{D}} \|v_\theta(x_t, t, z_I, h) - v_t\|^2$$

To model the relationship between input and output images, we adopt a multi-modal attention mechanism, as shown on the left of Fig. 2. Specifically, we apply $2 \times$ patchification to both z_I and x_t along spatial dimensions. Within each RGBA-MMDiT block, two separate sets of parameters are utilized to process the semantic condition h and the visual inputs x_t, z_I , respectively. During attention computation, these representations are concatenated to explicitly model the relationship between the conditions and the targets.

As illustrated on the right of Fig. 2, we propose a Bi3D RoPE for each RGBA-MMDiT block to distinguish between conditions and targets. Building on Qwen-Image, we adopt the MSRoPE design, where the positional encodings are shifted toward the center. To accommodate variable numbers of input and output images, we introduce an additional context dimension. Specifically, for the intermediate state x_t , context indices start from 0 and increase sequentially. Conversely, for the conditional images z_I , we assign negative context indices, where the first input image is indexed as -1 , followed by subsequent images in descending order. This bidirectional indexing scheme ensures a clear distinction between conditions and targets.

4 Experiments

4.1 Data Collection

To train a unified image generation and editing model that simultaneously supports raster and transparent imagery, we curated the following datasets. It is worth noting that no dedicated datasets exist for transparent image editing.

Table 1: Quantitative evaluation of transparent image editing on TransEdit-Bench-EN and TransEdit-Bench-CN. SC: Semantic Consistency. PQ: Perceptual Quality. TC: Transparency Consistency. O: Overall.

Method	TransEdit-Bench-EN $\uparrow$				TransEdit-Bench-CN $\uparrow$			
	SC	PQ	TC	O	SC	PQ	TC	O
AlphaVAE (Wang et al., 2025b)+SDEdit (Meng et al., 2022)	1.64	7.49	5.43	1.94	1.46	7.38	5.21	1.92
Editing (Wu et al., 2025a)+Matting (Zheng et al., 2024)	8.94	7.68	8.93	8.30	8.92	7.69	8.85	8.25
Qwen-Image-Layered 1.5	9.23	7.72	9.45	8.65	9.08	7.72	9.39	8.54

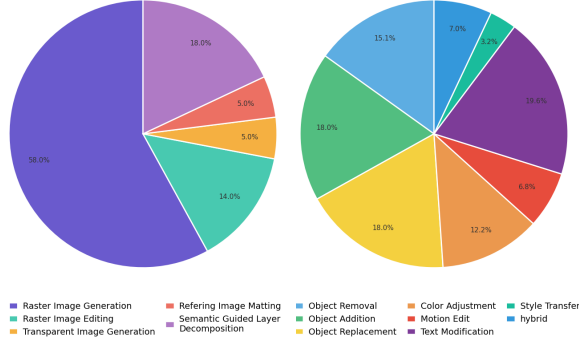


Figure 3: Left: Distribution of training tasks. Right: Category distribution in the proposed TransEdit-Bench.

Transparent Image Generation We curated a large corpus of RGBA images from the web. To ensure data quality, we first excluded images with insufficient resolution or an excessive proportion of opaque areas. We then utilized ArtiMuse (Cao et al., 2025) to filter the dataset based on aesthetic quality, yielding a final collection of 200K high-quality RGBA images. For annotation, we leveraged Qwen3-VL (Bai et al., 2025a) to generate detailed descriptions for each image, placing specific emphasis on transparency attributes to distinguish between opaque and transparent images.

Referring Image Matting We collected 150K image-alpha pairs from open-source matting datasets (Zeng et al., 2019; Li et al., 2021b,a; Xie et al., 2022; Qin et al., 2022; Li et al., 2022; Sun et al., 2022; Deng et al., 2023). Subsequently, we employed Qwen3-VL to generate referring matting prompts for each pair. Finally, the source images were combined with their corresponding alpha mattes to construct 4-channel RGBA outputs.

Semantic Guided Layer Decomposition We leveraged the multilayer dataset proposed in Qwen-Image-Layered (Yin et al., 2025) and employed a Vision-Language Model (VLM) to generate descriptions for each layer. During training, we concatenated the descriptions of all layers to form a decomposition instruction.

Raster Image Generation and Editing To maintain robust raster image generation and editing capabilities, we utilize the dataset from Qwen-Image, spanning diverse domains and scenarios.



Figure 4: Qualitative evaluation of transparent image editing. The leftmost column is the input image; prompts are listed above each row.

4.2 Implementation Details

We initialize Qwen-Image-Layered 1.5 based on Qwen-Image-Edit (Wu et al., 2025a). Specifically, the RGBA-VAE is trained for 400K steps using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 1×10^{-4} . Subsequently, the RGBA-MMDiT is optimized using Adam optimizer (Kingma and Ba, 2014) for 600K steps at a learning rate of 5×10^{-5} . The distribution of different tasks during training is shown in the left of Fig. 3.

4.3 TransEdit-Bench

Transparent image editing remains a largely underexplored task. To bridge this gap, we introduce TransEdit-Bench, the first benchmark dedicated to this task. We first collected diverse transparent images, spanning typographic designs, cartoons, and portraits. For each image, we manually crafted an editing instruction, resulting in a total of 556 samples across 8 categories, as illustrated in the right of Fig. 3. Regarding evaluation protocol, we adopt semantic consistency and per-

Table 2: Quantitative evaluation of transparent image generation on GenEval (Ghosh et al., 2023).

Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall↑
LayerDiffuse (Zhang and Agrawala, 2024)	<u>0.95</u>	0.71	0.40	0.84	0.16	0.27	0.55
AlphaVAE (Wang et al., 2025b)	<u>0.95</u>	0.66	0.70	0.75	0.18	0.38	0.60
OmniAlpha (Yu et al., 2025)	0.88	0.92	0.79	0.81	<u>0.52</u>	0.70	<u>0.77</u>
Qwen-Image-Layered 1.5	0.98	<u>0.87</u>	<u>0.77</u>	<u>0.83</u>	0.74	<u>0.56</u>	0.79

Table 3: Quantitative evaluation of semantic guided layer decomposition.

Model	DreamLayer (Huang et al., 2025)		PrismLayer (Chen et al., 2025c)	
	RGB L1↓	Alpha Soft IoU↑	RGB L1↓	Alpha Soft IoU↑
LayerD (Suzuki et al., 2025)	0.2125	0.5722	0.0490	0.6346
Qwen-Image-Layered (Yin et al., 2025)	0.0376	0.8000	0.0157	0.8845
Qwen-Image-Layered 1.5	0.0201	0.9103	0.0106	0.9219



Figure 5: Qualitative evaluation of transparent image generation. The text prompts are listed above each row. ceptual quality, following GEdit-Bench (Liu et al., 2025b). Furthermore, to assess the preservation of transparency, we introduce a new metric: transparency consistency. To ensure reproducibility, we employ Qwen3-VL-32B for automatic evaluation. Notably, to enable the VLM to perceive transparency, we composite all transparent images over a gray-white checkerboard background.

4.4 Quantitative and Qualitative Results

4.4.1 Transparent Image Editing

We compare our approach against two baselines on transparent image editing: (1) the first combines AlphaVAE with SDEdit, which employs noise injection followed by denoising; (2) the sec-

ond blends the transparent image onto a solid-color background for editing and then performs image matting. As shown in Tab. 1, Qwen-Image-Layered 1.5 achieves state-of-the-art performance across all metrics. Notably, it demonstrates superior transparency consistency, underscoring the advantages of unified image generation and editing with native transparency. We also provide a qualitative comparison in Fig. 4: AlphaVAE+SDEdit struggles to follow the textual prompts, while Editing+Matting suffers from transparency corruption and noticeable matte contamination. In contrast, Qwen-Image-Layered 1.5 not only accurately follows instructions but also maintains both visual and transparency consistency.

4.4.2 Transparent Image Generation

To quantitatively evaluate transparent image generation, we employ GenEval, an object-centric text-to-image benchmark. As shown in Tab. 2, our method outperforms existing models in overall performance, demonstrating strong instruction-following capabilities. Furthermore, Fig. 5 presents a qualitative comparison: as shown in the first row, while LayerDiffuse and AlphaVAE struggle to correctly render the text, our model accurately generates the required text. The subsequent rows further highlight our model’s superior controllability in style, position, and object counting.

4.4.3 Semantic Guided Layer Decomposition

To quantitatively evaluate semantic guided layer decomposition, we curated a test set containing 1,000 instances (500 each from DreamLayer and PrismLayer) with detailed layer-wise annotations. We report two metrics in Tab. 3: RGB L1 (the L1 distance after blending layers onto a white back-



Figure 6: Qualitative evaluation of semantic guided layer decomposition. The leftmost column is the input image; the subsequent columns are decomposed layers.

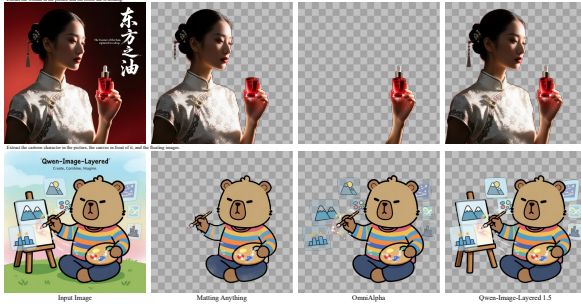


Figure 7: Qualitative evaluation of referring image matting. The matting prompts are shown above each row.

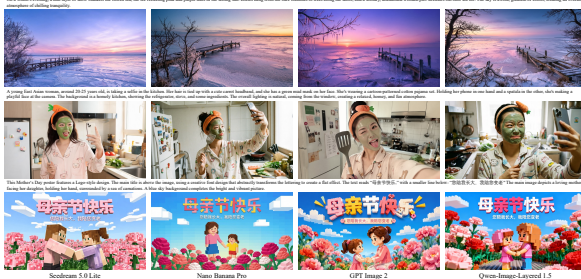


Figure 8: Qualitative evaluation of raster image generation. The prompts are shown above each row.

ground) and Alpha Soft IoU (the soft IoU between predicted and ground-truth alpha channels). As shown in Tab. 3, our method achieves higher decomposition accuracy. Furthermore, Fig. 6 provides a qualitative comparison. While LayerD and Qwen-Image-Layered struggle to follow decomposition instructions, Qwen-Image-Layered 1.5 can accurately control the content of each layer, significantly improving its practical utility.

4.4.4 Referring Image Matting

Tab. 4 presents a quantitative comparison of referring image matting on RefMatte-RW100, where Qwen-Image-Layered 1.5 achieves state-of-the-art matting accuracy in MSE and MAD. Furthermore,

Table 4: Quantitative evaluation of referring image matting on RefMatte-RW100 (Li et al., 2023).

Model	SAD↓	MSE↓	MAD↓
MDETR (Kamath et al., 2021)	131.580	0.068	0.075
CLIPSeg (Lüddecke and Ecker, 2022)	211.860	0.118	0.122
SAM (Kirillov et al., 2023)	122.760	0.068	0.069
MAM (Li et al., 2024a)	120.100	0.066	0.068
CLIPMat (Li et al., 2023)	85.830	0.047	0.050
OmniAlpha (Yu et al., 2025)	14.768	0.022	0.026
Qwen-Image-layered 1.5	37.942	0.019	0.022

Table 5: Quantitative evaluation of raster image generation on OneIG-EN (Chang et al., 2025).

Model	Alignment	Text	Reasoning	Style	Diversity	Overall↑
Janus-Pro (Chen et al., 2025d)	0.553	0.001	0.139	0.276	0.365	0.267
BLIP3-o (Chen et al., 2025b)	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
Show-o2-7B (Xie et al., 2025)	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., 2022)	0.565	0.010	0.207	0.383	0.429	0.319
SDXL (Podell et al., 2024)	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] (Labs, 2024)	0.786	0.523	0.253	0.368	0.238	0.434
CogView4 (Zai, 2025)	0.786	0.641	0.246	0.353	0.205	0.446
Lumina-Image 2.0 (Qin et al., 2025)	0.819	0.106	0.270	0.354	0.216	0.353
HiDream-11-Full (Cai et al., 2025b)	0.829	0.707	0.317	0.347	0.186	0.477
Imagen3 (Baldridge et al., 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 (OpenAI, 2025)	0.851	0.857	0.345	0.462	0.151	0.533
Qwen-Image (Wu et al., 2025a)	0.882	0.891	0.306	0.418	0.197	0.539
Nano-Banana (DeepMind, 2025a)	0.878	0.894	0.346	0.450	0.182	0.550
Qwen-Image-Layered 1.5	0.872	0.960	0.287	0.396	0.276	0.558

Table 6: Quantitative evaluation of raster image editing on GEdit-Bench-EN (Liu et al., 2025b). G_SC: Semantic Consistency. G_PQ: Perceptual Quality. G_O: Overall.

Model	G_SC↑	G_PQ↑	G_O↑
OmniGen (Xiao et al., 2025)	5.96	5.89	5.06
OmniGen2 (Wu et al., 2025b)	7.16	6.77	6.41
BAGEL (Deng et al., 2025)	7.36	6.83	6.52
FLUX.1 Kontext [Dev] (Labs et al., 2025)	6.52	7.38	6.00
FLUX.1 Kontext [Pro] (Labs et al., 2025)	7.02	7.60	6.56
FLUX.2 [Dev] (Labs, 2025)	7.84	8.06	7.41
Step1X-Edit (Liu et al., 2025b)	7.66	7.35	6.97
GPT Image 1 (OpenAI, 2025)	7.85	7.62	7.53
Qwen-Image (Wu et al., 2025a)	8.00	7.86	7.56
Nano-Banana (DeepMind, 2025a)	7.40	8.45	7.29
Nano-Banana-Pro (DeepMind, 2025b)	8.10	8.34	7.74
Seedream 4.0 (Team et al., 2025)	8.14	8.12	7.70
Emu3.5 (Cui et al., 2025)	8.11	7.70	7.59
Qwen-Image-Layered 1.5	8.39	8.02	7.76

Fig. 7 provides a qualitative comparison, highlighting our model’s superior semantic controllability. In contrast, other methods produce inaccurate or incomplete results (e.g., the woman in the first row and the canvas in the second row).

4.4.5 Raster Image Generation

Tab. 5 presents a quantitative evaluation of raster image generation on the OneIG-EN benchmark, where Qwen-Image-Layered 1.5 achieves state-of-the-art overall performance. Notably, the model demonstrates exceptional text rendering capabilities. Furthermore, Fig. 8 provides a qualitative comparison, highlighting that while the model natively supports transparency-aware tasks, it maintains competitive performance in standard raster image generation, on par with leading models.

Table 7: Quantitative comparison of image reconstruction on RGBA and RGB images.

(a) RGBA Image Reconstruction on the AIM-500.

Model	Base Model	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
LayerDiffuse (Zhang and Agrawala, 2024)	SDXL	32.09	0.9436	0.0418
AlphaVAE (Wang et al., 2025b)	SDXL	35.74	0.9576	0.0495
	FLUX	36.94	0.9737	0.0283
Qwen-Image-Layered (Yin et al., 2025)	Qwen-Image	38.83	0.9802	0.0123
RGBA-VAAE	Qwen-Image	40.81	0.9844	0.0091

(b) RGB Image Reconstruction on the ImageNet.

Model	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SD-3.5 (Esser et al., 2024)	31.22	0.8839	0.0459
FLUX.1-dev (Labs, 2024)	32.84	0.9155	0.0445
HunyuanImage-3.0 (Team, 2025)	31.08	0.8655	0.0812
Cosmos-CI8x8 (Agarwal et al., 2025)	32.23	0.9010	0.1086
RGBA-VAAE	33.10	0.9170	0.0302

Table 8: Ablation study on transparency-aware tasks. TIE: Transparent Image Editing. TIG: Transparent Image Generation. SGLD: Semantic Guided Layer Decomposition. RIM: Referring Image Matting. Unless otherwise noted, higher values indicate better performance.

Task	TIE(TransEdit-Bench-EN)				TIG(GenEval)							SGLD(DreamLayer)		RIM(RefMatte-RW100)		
Metric	SC	PQ	TC	O	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall	RGB L1↓	Alpha Soft IoU	SAD↓	MSE↓	MAD↓
Qwen-Image-Layered 1.5 w/o α -VF loss	8.38	7.58	9.22	7.83	0.76	0.51	0.51	0.51	0.57	0.19	0.51	0.0294	0.8591	60.699	0.033	0.035
Qwen-Image-Layered 1.5 w/o Bi3D	9.06	7.67	9.35	8.49	0.87	0.69	0.63	0.66	0.64	0.45	0.66	0.3280	0.5522	97.659	0.059	0.061
Qwen-Image-Layered 1.5	9.23	7.72	9.45	8.65	0.98	0.87	0.77	0.83	0.74	0.56	0.79	0.0201	0.9103	37.942	0.019	0.022



Figure 9: Qualitative evaluation of raster image editing.

4.4.6 Raster Image Editing

We conducted a quantitative evaluation of raster image editing on the GEdit-Bench-EN benchmark. As shown in Tab. 6, Qwen-Image-Layered 1.5 achieves outstanding overall performance, notably excelling in semantic consistency. Furthermore, Fig. 9 presents a qualitative comparison, demonstrating that our model attains performance comparable to top-tier models in both visual consistency and instruction adherence.

4.4.7 Image Reconstruction

We quantitatively evaluated the image reconstruction performance of RGBA-VAAE on both RGBA and RGB images. For RGBA images, we adopted the evaluation protocol proposed by AlphaVAE and report results on the AIM-500 dataset (Li et al., 2021b) in Tab. 7a. As demonstrated in the table, our method achieves state-of-the-art performance across all metrics. Furthermore, we evaluated RGB image reconstruction on

ImageNet (Deng et al., 2009) in Tab. 7b. Notably, despite being a 4-channel VAE, our model delivers superior reconstruction quality on RGB images, thereby establishing a solid foundation for unifying raster-centric and transparency-aware tasks.

4.5 Ablation Study

To validate the effectiveness of our model design, we conducted an ablation study on four transparency-aware tasks, with results presented in Tab. 8. First, we removed the α -VF loss during VAE training. As evidenced by the comparison between the first and third rows, this removal leads to a slight performance decline, demonstrating that our proposed RGBA-VAAE effectively enhances cross-format generalization. Furthermore, we replaced Bi3D RoPE with standard 2D MSRoPE positional encoding. The comparison between the second and third rows shows noticeable performance degradation across all tasks, especially in semantic guided layer decomposition, underscoring the necessity of Bi3D RoPE in distinguishing between different input and output images.

5 Conclusion

We present Qwen-Image-Layered 1.5, a unified image generation and editing framework that natively supports both opaque and transparent images by unifying them into a shared latent space. Notably, our approach unlocks zero-shot transparent image editing capabilities. We further propose TransEdit-Bench to systematically evaluate this new task. Extensive experiments demonstrate that our method achieves state-of-the-art performance across a broad range of raster-centric and transparency-aware tasks, establishing a new paradigm for unified visual generative modeling.

Limitations

While Qwen-Image-Layered 1.5 unifies the generation and editing of both opaque and transparent images, the generation quality for transparent images still lags behind that of opaque ones. This limitation is primarily attributed to the scarcity and inferior quality of available transparency-related datasets compared to standard raster imagery.

References

- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, and 58 others. 2025. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluís Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, and 241 others. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.
- Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, and 4 others. 2025a. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*.
- Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, and 3 others. 2025b. Hidream-1l: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*.
- Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajun Yao, and Yihao Liu. 2025. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*.
- Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. 2025. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*.
- Jingye Chen, Zhaowen Wang, Nanxuan Zhao, Li Zhang, Difan Liu, Jimei Yang, and Qifeng Chen. 2025a. Rethinking layered graphic design generation with a top-down approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16861–16870.
- Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. 2025b. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pages 57611–57640.
- Junwen Chen, Heyang Jiang, Yanbin Wang, Keming Wu, Ji Li, Chao Zhang, Keiji Yanai, Dong Chen, and Yuhui Yuan. 2025c. Prismlayers: Open data for high-quality multi-layer transparent image generative models. *arXiv preprint arXiv:2505.22523*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025d. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, and 4 others. 2025. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*.

- Yuekun Dai, Haitian Li, Shangchen Zhou, and Chen Change Loy. 2025. Trans-adapter: A plug-and-play framework for transparent image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15015–15024.
- Google DeepMind. 2025a. [Gemini 2.5 flash image](#).
- Google DeepMind. 2025b. [Gemini 3 pro image](#).
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Xinhao Deng, Pingping Zhang, Wei Liu, and Huchuan Lu. 2023. Recurrent multi-scale transformer for high-resolution salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7413–7423.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Kailai Feng, Yuxiang Wei, Bo Chen, Yang Pan, Hu Ye, Songwei Liu, Chenqian Yan, and Yuan Gao. 2026. Dreamlite: A lightweight on-device unified model for image generation and editing. *arXiv preprint arXiv:2603.28713*.
- Alessandro Fontanella, Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, and Sarah Parisot. 2024. Generating compositional scenes via text-to-image rgba instance generation. *Advances in Neural Information Processing Systems*, 37:43864–43893.
- Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, and 12 others. 2025. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*.
- Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.
- Google. 2025. [Imagen 4](#).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi. 2024. Diffusion for natural image matting. In *European Conference on Computer Vision*, pages 181–199. Springer.
- Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. 2026. Psdiffusion: Harmonized multi-layer image generation via layout and appearance alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3233–3242.
- Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. 2025. Dreamlayer: Simultaneous multi-layer generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3357–3366.
- Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. 2024. Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *European Conference on Computer Vision*, pages 144–160. Springer.
- Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790.
- Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. 2025. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*.
- Beomyoung Kim, Chanyong Shin, Joonhyun Jeong, Hyungsik Jung, Se-Yun Lee, Sewhan Chun, Dong-Hyun Hwang, and Joonsang Yu. 2025. Zim: Zero-shot image matting for anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23828–23838.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

- Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yungang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, and Zhijie Deng. 2026. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. *arXiv preprint arXiv:2601.10332*.
- Black Forest Labs. 2024. Flux. <https://github.com/black-forest-labs/flux>.
- Black Forest Labs. 2025. Flux.2: Frontier visual intelligence.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*.
- Jiachen Li, Jitesh Jain, and Humphrey Shi. 2024a. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785.
- Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. 2021a. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509.
- Jizhizi Li, Jing Zhang, Stephen J. Maybank, and Dacheng Tao. 2022. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266.
- Jizhizi Li, Jing Zhang, and Dacheng Tao. 2021b. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*.
- Jizhizi Li, Jing Zhang, and Dacheng Tao. 2023. Referring image matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22448–22457.
- Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, and 26 others. 2024b. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.
- Cheng Liu, Yiren Song, Haofan Wang, and Mike Zheng Shou. 2025a. Omnipsd: Layered psd generation with diffusion transformer. *arXiv preprint arXiv:2512.09247*.
- Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, and 5 others. 2025b. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.
- Zihao Liu, Zunnan Xu, Shi Shu, Jun Zhou, Ruicheng Zhang, Zhenchao Tang, and Xiu Li. 2025c. Controllable layer decomposition for reversible multi-layer image generation. *arXiv preprint arXiv:2511.16249*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.
- OpenAI. 2025. [Gpt-image-1](#).
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874.
- Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, and Bainig Guo. 2025. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7952–7962.
- Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhui Wang, Conghui He, Bo Zhang, Xiaohong Liu, and 4 others. 2025. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*.
- Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. 2022. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer.

- Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. 2024. Alfie: Democratising rgba image generation with no \$\$\$\$. In *European Conference on Computer Vision*, pages 38–55. Springer.
- Recraft. 2024. Recraft v3. <https://www.recraft.ai/>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2022. Human instance matting via mutual guidance and multi-instance refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2656.
- Tomoyuki Suzuki, Kang-Jun Liu, Naoto Inoue, and Kota Yamaguchi. 2025. Layerd: Decomposing raster graphic designs into layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17783–17792.
- Kuaishou Kolors team. 2025. Kolors2.0. <https://app.klingai.com/cn/>.
- Seedream Team, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, and 31 others. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.
- Tencent Hunyuan Foundation Model Team. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*.
- Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. 2024. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22413–22422.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. 2025a. Seedit3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*.
- Zile Wang, Hao Yu, Jiabo Zhan, and Chun Yuan. 2025b. Alphavae: Unified end-to-end rgba image reconstruction and generation with alpha-aware representation learning. *arXiv preprint arXiv:2507.09308*.
- Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. 2025c. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.
- Chenyuan Wu, Pengfei Zheng, Ruirao Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyao Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, and 3 others. 2025b. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*.
- Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. 2025. Dreamomni: Unified image generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28533–28543.
- Ruihao Xia, Yu Liang, Peng-Tao Jiang, Hao Zhang, Qianru Sun, Yang Tang, Bo Li, and Pan Zhou. 2024. Towards natural image matting in the wild via real-scenario prior. *arXiv preprint arXiv:2410.06593*.
- Tianyi Xiang, Weiying Zheng, Yutao Jiang, Tingrui Shen, Hwei Yu, Yangyang Xu, and Shengfeng He. 2025. Teaching diffusion models to ground alpha matte. *Transactions on Machine Learning Research*.
- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruirao Yan, Chaofan Li, Shutong Wang, Tiejun Huang, and Zheng Liu. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304.
- Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. 2022. Pyramid grafting

- network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11717–11726.
- Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o2: Improved native unified multi-modal models. *arXiv preprint arXiv:2506.15564*.
- Yangyang Xu, Shengfeng He, Wenqi Shao, Kwan-Yee K Wong, Yu Qiao, and Ping Luo. 2023. Diffusionmat: Alpha matting as sequential refinement learning. *arXiv preprint arXiv:2311.13535*.
- Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. 2025. Generative image layer decomposition with visual effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7643–7653.
- Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. 2024a. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103:102091.
- Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. 2024b. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147:105067.
- Jingfeng Yao, Bin Yang, and Xinggang Wang. 2025. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712.
- Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, Lionel M. Ni, Jingren Zhou, Junyang Lin, and Chenfei Wu. 2025. Qwen-image-layered: Towards inherent editability via layer decomposition. *arXiv preprint arXiv:2512.15603*.
- Hao Yu, Jiabo Zhan, Zile Wang, Jinglin Wang, Huaisong Zhang, Hongyu Li, Xinrui Chen, Yongxian Wei, and Chun Yuan. 2025. Omnialpha: A sequence-to-sequence framework for unified multi-task rgba generation. *arXiv preprint arXiv:2511.20211*.
- THUKEG Z.ai. 2025. [Cogview4](#).
- Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. 2019. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7234–7243.
- Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.
- Lvmin Zhang and Maneesh Agrawala. 2024. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*.
- Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. 2023b. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*.
- Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. 2025. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*.
- Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. 2024. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*.