Qwen-Image-Layered: Towards Inherent Editability via Layer Decomposition

Shengming Yin[1] Zekai Zhang[2] Zecheng Tang[2] Kaiyuan Gao[2]
Xiao Xu[2] Kun Yan[2] Jiahao Li[2] Yilei Chen[2] Yuxiang Chen[2]
Heung-Yeung Shum[3] Lionel M. Ni[1] Junyang Lin[2] Chenfei Wu[2*]
[1]HKUST(GZ) [2]Alibaba [3]HKUST

Figure 1. Qwen-Image-Layered is capable of decomposing an input image into multiple semantically disentangled RGBA layers, thereby enabling inherent editability, where each layer can be independently manipulated without affecting other content.

Abstract

*Recent visual generative models often struggle with consistency during image editing due to the entangled nature of raster images, where all visual content is fused into a single canvas. In contrast, professional design tools employ layered representations, allowing isolated edits while preserving consistency. Motivated by this, we propose **Qwen-Image-Layered**, an end-to-end diffusion model that decomposes a single RGB image into multiple semantically disentangled RGBA layers, enabling **inherent editability**, where each RGBA layer can be independently manipulated without affecting other content. To support variable-length decomposition, we introduce three key components: (1) an RGBA-VAE to unify the latent representations of RGB and RGBA images; (2) a VLD-MMDiT (Variable Layers Decomposition MMDiT) architecture capable of decomposing a variable number of image layers; and (3) a Multi-stage Training strategy to adapt a pretrained image generation model into a multilayer image decomposer. Furthermore, to address the scarcity of high-quality multilayer training images, we build a pipeline to extract and annotate multilayer images from Photoshop documents (PSD). Experiments demonstrate that our method significantly surpasses existing approaches in decomposition quality and establishes a new paradigm for consistent image editing. Our code and models are released on https://github.com/QwenLM/Qwen-Image-Layered*

*Corresponding author.

1. Introduction

Recent advances in visual generative models have enabled impressive image synthesis capabilities [5, 10–12, 24, 31, 34, 41, 42]. However, in the context of image editing, achieving precise modifications while preserving the structure and semantics of unedited regions remains a significant challenge. This issue typically appears as semantic drift (*e.g.* unintended changes to a person's identity) and geometric misalignment (*e.g.* shifts in object position or scale).

Existing editing approaches fail to fundamentally address this problem. Global editing methods [4, 9, 21, 26, 39, 43, 48], which resample the entire image in the latent space of generative models, are inherently limited by the stochastic nature of probabilistic generation and thus cannot ensure consistency in unedited regions. Meanwhile, mask-guided local editing methods [8, 29, 35] restrict modification within user-specified masks. However, in complex scenes, especially those involving occlusion or soft boundaries, the actual editing region is often ambiguous, thus failing to fundamentally solve the consistency problem.

Rather than tackling this issue purely through model design or data engineering, we argue that the core challenge lies in the representation of images themselves. Traditional raster images are flat and entangled: all visual content is fused into a single canvas, with semantics and geometry tightly coupled. Consequently, any edit inevitably propagates through this entangled pixel space, leading to the aforementioned inconsistencies.

To overcome this fundamental limitation, we advocate for a naturally disentangled image representation. Specifically, we propose representing an image as a stack of semantically decomposed RGBA layers, as illustrated in the upper part of Fig. 1. This layered structure enables inherent editability with built-in consistency: edits are applied exclusively to the target layer, physically isolating them from the rest of the content, and thereby eliminating semantic drift and geometric misalignment. Moreover, such a layer-wise representation naturally supports high-fidelity elementary operations—such as resizing, repositioning, and recoloring, as demonstrated in the lower part of Fig. 1.

Based on this insight, we introduce Qwen-Image-Layered, an end-to-end diffusion model that directly decomposes a single RGB image into multiple semantically disentangled RGBA layers. Once decomposed, each layer can be independently manipulated while leaving all other content exactly unchanged—enabling truly consistent image editing. To support variable-length decomposition, our image decomposer is built upon three key designs: (1) an RGBA-VAE that establishes a shared latent space for both RGB and RGBA images; (2) a VLD-MMDiT (Variable Layers Decomposition MMDiT) architecture that enables training with a variable number of layers; and (3) a Multi-stage Training strategy that progressively adapts a pretrained image generation model into an multilayer image decomposer. Furthermore, to address the scarcity of high-quality multilayer image data, we develop a data pipeline to filter and annotate multilayer images from real-world Photoshop documents (PSD).

We summarize our contributions as follows:

- We propose **Qwen-Image-Layered**, an end-to-end diffusion model that decomposes an image into multiple high-quality, semantically disentangled RGBA layers, thereby enabling inherently consistent image editing.
- We design the image decomposer from three aspects: 1) an RGBA-VAE to provide shared latent space for RGB and RGBA images. 2) a VLD-MMDiT architecture to facilitate decomposition with variable number of layers. 3) a Multi-stage Training strategy to adapt a pretrained image generation model to a multilayer image decomposer.
- We develop a data processing pipeline to extract and annotate multilayer images from Photoshop documents, addressing the lack of high-quality multilayer images.
- Extensive experiments demonstrate that Qwen-Image-Layered not only outperforms existing methods in decomposition quality but also unlocks new possibilities for consistent, layer-based image editing and synthesis.

2. Related Work

2.1. Image Editing

Image editing has made significant progress in recent years and can be broadly categorized into two paradigms: global editing and mask-guided local editing. Global editing methods [4, 9, 21, 26, 39, 42, 43, 48] regenerate the entire image to achieve holistic modifications, such as expression editing and style transfer. Among these, Qwen-Image-Edit [42] leverages two distinct yet complementary feature representations—semantic features from Qwen-VL [3] and reconstructive features from VAE [19]—to enhance consistency. However, due to the inherent stochasticity of generative models, these approaches cannot ensure consistency in unedited regions. In contrast, mask-guided local editing methods [8, 29, 35] constrain modifications within a specified mask to preserve global consistency. DiffEdit [8], for instance, first automatically generates a mask to identify regions requiring modification and then edits the target area. Although intuitive, these approaches struggle with occlusions and soft boundaries, making it difficult to precisely identify the actual editing region and thus failing to fundamentally resolve the consistency issue. Unlike these works, we propose decomposing the image into semantically disentangled RGBA layers, where each layer can be independently modified while keeping the others unchanged, thereby fundamentally ensuring consistent across edits.

2.2. Image Decomposition

Numerous studies have attempted to decompose images into layers. Early approaches addressed this problem by performing segmentation in color space [2, 20, 37]. Subsequent work has focused on object-level decomposition in natural scenes [28, 30, 47]. Among these, PCNet [47] learns to recover fractional object masks and contents in a self-supervised manner. More recent research has explored decomposing images into multiple RGBA layers [7, 18, 36, 38, 45]. One class of these methods leverages segmentation [33] or matting [23] to extract foreground objects, followed by image inpainting [46] to reconstruct the background. For instance, LayerD [36] iteratively extracts the topmost unoccluded foreground layer and completes the background. Accordion [7] proposes using Vision-Language Models [25] to guide this decomposition process. Another category of work introduces mask-guided, object-centric image decomposition [18, 45], which decomposes an image into foreground and background layers based on a provided mask. These methods generally require segmentation to provide initial mask. However, segmentation often struggles with complex spatial layouts and the presence of multiple semi-transparent layers, resulting in low-quality layers. Moreover, multilayer decomposition typically requires recursive inference, leading to error propagation. Consequently, existing methods fail to produce complete, high-fidelity RGBA layers suitable for editing. In contrast to the aforementioned approaches, Qwen-Image-Layered employs an end-to-end framework to decompose input images directly into multiple high-quality RGBA layers, thereby enhancing decomposition quality and enabling consistency-preserving image editing.

2.3. Multilayer Image Synthesis

Multilayer image synthesis has also garnered sustained attention [6, 15–18, 32, 49, 50]. As a pioneer in layered image generation, Text2Layer [50] first trains a two-layer image autoencoder [19] and subsequently trains a diffusion model [13] on the latent representations, enabling the creation of two-layer images. LayerDiffusion [49] introduces latent transparency into VAE and employs two different LoRA [14] with shared attention to generate foreground and background. Through carefully designed inter-layer and intra-layer attention mechanisms, LayerDiff [17] is able to synthesize semantically consistent multilayer images. To achieve controllable multilayer image generation, ART [32] proposes an anonymous region layout to explicitly control the layout. LayeringDiff [18] first generates a raster image using existing text-to-image models, and then decomposes it into foreground and background based on a mask. Qwen-Image-Layered is capable of decomposing AI-generated raster images into multiple RGBA layers, thus enabling multilayer image generation.

3. Method

We propose an end-to-end layering approach that directly decomposes an input RGB image $I \in \mathbb{R}^{H\times W\times 3}$ into N RGBA layers $L \in \mathbb{R}^{N\times H\times W\times 4}$, where each layer L_i comprises a color component RGB_i and an alpha matte α_i, i.e. $L_i = [RGB_i; \alpha_i]$. The original image can be reconstructed by sequential alpha blending as follows:

$$
\begin{aligned}
&C_0 = \mathbf{0} \\
&C_i = \alpha_i \cdot RGB_i + (1-\alpha_i)\cdot C_{i-1} \quad \mathrm{i} = 1, \ldots, \mathrm{N}
\end{aligned}
$$

where C_i denotes the composite of the first i layers, and the final composite satisfies $I = C_N$. Building upon Qwen-Image [42], we develop Qwen-Image-Layered from the following three aspects:

- 1) In contrast to previous decomposer [45] that employs separate VAEs, we propose an RGBA-VAE that encodes both RGB and RGBA images. This approach narrows the latent distribution gap between the input RGB image and the output RGBA layers.
- 2) Unlike prior methods that decompose images into foreground and background [18, 45], we propose a VLD-MMDiT (Variable Layers Decomposition MMDiT), which supports decomposition into a variable number of layers and is compatible with multi-task training.
- 3) To progressively adapt pretrained image generation model into a multilayer image decomposer, we design a multi-stage, multi-task training scheme that progressively evolves from simpler tasks to more complex ones.

3.1. RGBA-VAE

Variational Autoencoders (VAEs) [19] are commonly employed in diffusion models [34] to reduce the dimensionality of the latent space, thereby improving both training and sampling efficiency. In previous work, LayeringDiff [18] utilized an RGB VAE to first generate the foreground layer and subsequently applied an additional module to obtain transparency. LayerDecomp [45] adopted separate VAEs for the input RGB image and the output RGBA layers, resulting in a distribution gap between the input and output representations. To address these limitations, we propose RGBA VAE, a four-channel VAE designed to process both RGB and RGBA images.

Inspired by AlphaVAE [40], we extend the first convolution layer of the Qwen-Image VAE encoder $\mathcal{E}$ and the last convolution layer of the decoder $\mathcal{D}$ from three to four channels. To enable reconstruction of both RGB and RGBA images, we train it using both types of images. For RGB images, the alpha channel is set to 1. To maintain RGB reconstruction performance during initialization, we employ the following initialization strategy. Let $W^0_{\mathcal{E}} \in \mathbb{R}^{D_0\times 4\times k\times k\times k}$ and $b^0_{\mathcal{E}} \in \mathbb{R}^{D_0}$ denote the weight and bias of the first convolution layer in the encoder, and $W^l_{\mathcal{D}} \in \mathbb{R}^{4\times D_l\times k\times k\times k}$ and

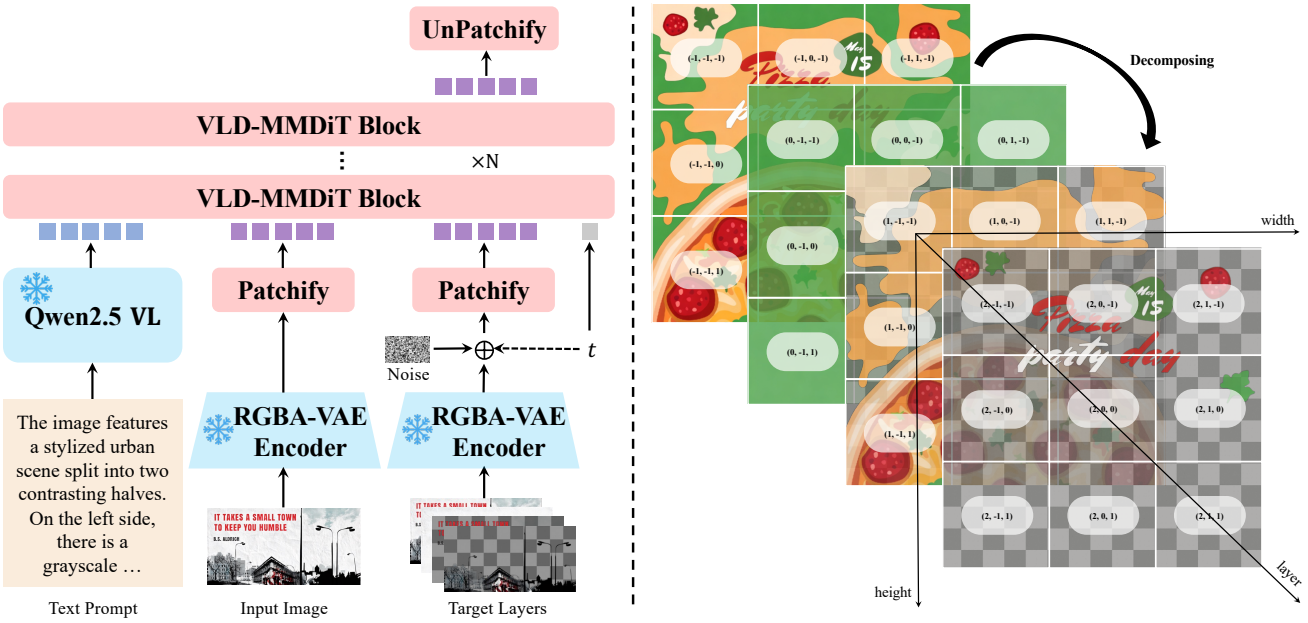


Figure 2. Overview of Qwen-Image-Layered. Left: Illustration of our proposed VLD-MMDiT (Variable Layers Decomposition MMDiT), where the input RGB image and the target RGBA layers are both encoded by our proposed RGBA-VAE. During attention computation, these two sequences are concatenated along the sequence dimension, thereby enhancing inter-layer and intra-layer interactions. Right: Illustration of Layer3D RoPE, where a new layer dimension is introduced to support a variable number of layers.

$b^l_{\mathcal{D}} \in \mathbb{R}^4$ denote those of the last convolution layer in the decoder, where k is the kernel size. We copy the parameters from the pretrained RGB VAE into the first three channels and set the newly initialized parameters as

$$W^0_{\mathcal{E}}[:, 3, :, :, :] = 0 \quad W^l_{\mathcal{D}}[3, :, :, :, :] = 0 \quad b^l_{\mathcal{D}}[3] = 1$$

For the training objective, we use a combination of reconstruction loss, perceptual loss, and regularization loss. After training, both the input RGB image and the output RGBA layers are encoded into a shared latent space, where each RGBA layer is encoded independently. Notably, these layers exhibit no cross-layer redundancy; consequently, no compression is applied along the layer dimension.

3.2. Variable Layers Decomposition MMDiT

Previous studies [7, 18, 36, 45] typically decompose images into background and foreground, requiring recursive inference to perform multilayer decomposition. Instead, Qwen-Image-Layered proposes VLD-MMDiT (Variable Layers Decomposition MMDiT) to facilitate the decomposition of a variable number of layers.

For Qwen-Image-Layered, it tasks an RGB image $I \in \mathbb{R}^{H \times W \times 3}$ as input and decomposes it into multiple RGBA layers $L \in \mathbb{R}^{N \times H \times W \times 4}$. Following Qwen-Image, we adopt the Flow Matching training objective. Formally, let $x_0 \in \mathbb{R}^{N \times h \times w \times c}$ denote the latent representation of the target RGBA layers L, i.e., $x_0 = \mathcal{E}(L)$. Then we sample noise x_1 from standard multivariate normal distribution and a timestep $t \in [0, 1]$ from a logit-normal distribution. According to Rectified Flow [27], the intermediate state x_t and velocity v_t at timestep t is defined as

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

For the input RGB image I, we also use RGBA-VAE to encode it as a latent representation $z_I \in \mathbb{R}^{h \times w \times c}$. Following Qwen-Image, the text prompt is encoded into text condition h with MLLM. In practice, we can use Qwen2.5-VL [3] to automatically generate the caption for the input image. Then, the model is trained to predict the target velocity with loss function defined as the mean squared error between the predicted velocity $v_\theta(x_t, t, z_I, h)$ and the ground truth v_t:

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

where $\mathcal{D}$ denotes the training dataset.

Previous studies [16, 17] have achieved multilayer image generation through sophisticatedly designed inter-layer and intra-layer attention mechanisms. In contrast, we employ a Multi-Modal attention [10] to directly model these relationships, as shown in the left part of Fig. 2. Specifically, we apply $2\times$ patchification to the noise-free input image z_I and the intermediate state x_t along the height and width dimensions. In each VLD-MMDiT block, two separate sets of parameters are used to process textual h and visual information z_I, x_t respectively. During attention computation, we concatenate these three sequences, thereby directly modeling both intra-layer and inter-layer interactions.

As shown in the right part of Fig. 2, we propose a Layer3D RoPE within each VLD-MMDiT block to enable the decomposition of a variable number of layers, while supporting various tasks. Our design is inspired by the MSRoPE from Qwen-Image [42], where the positional encoding in each layer is shifted towards the center. To accommodate a variable number of layers, we introduce an additional layer dimension. For the intermediate state x_t, the layer index starts from 0, and increases accordingly. For conditional image input z_I, we assign a layer index of -1, ensuring a clear distinction from any positive layer indices used in other tasks, *e.g.* text-to-multilayer image generation.

3.3. Multi-stage Training

Directly finetuning a pretrained image generation model to perform image decomposition poses significant challenges, as it not only requires adapting to a new VAE but also involves learning new tasks. To address this issue, we propose a multi-stage, multi-task training scheme that progressively evolves from simpler tasks to more complex ones.

Stage 1: From Text-to-RGB to Text-to-RGBA. We begin by adapting MMDiT to the latent space of RGBA VAE. At this stage, we replace the original VAE and train the model jointly on both text-to-RGB and text-to-RGBA generation tasks. This enables the model to generate not only standard raster images (RGB) but also images with transparency (RGBA).

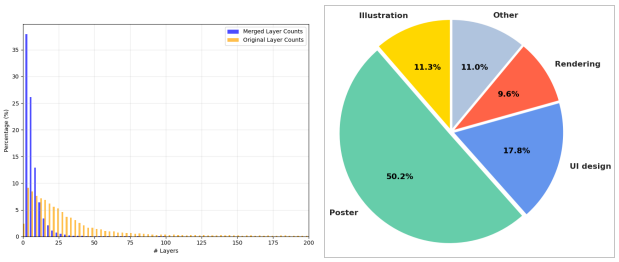
Stage 2: From Text-to-RGBA to Text-to-Multi-RGBA. Initially, the image generator is capable of producing only a single image. To support multilayer generation and adapt to the newly initialized layer dimension, we introduce a text-to-multiple-RGBA generation task. Following ART [32], the model is trained to jointly predict both the final composite image and its corresponding transparent layers, thereby facilitating information propagation between the composite image and its layers. We refer to this model as Qwen-Image-Layered-T2L.

Stage 3: From Text-to-Multi-RGBA to Image-to-Multi-RGBA. Up to this point, all tasks have been conditioned exclusively on textual prompts. In this stage, we introduce an additional image input, as detailed in Sec. 3.2, extending the model's capability to decompose a given RGB image into multiple RGBA layers. We refer to this model as Qwen-Image-Layered-I2L.

4. Experiment

4.1. Data Collection and Annotation

Due to the scarcity of high-quality multilayer images, previous studies [17, 18, 36, 50] have largely relied on either synthetic data [38] or simple graphic design datasets (*e.g.*, Crello [44]), which typically lack complex layouts or semi-transparent layers. To bridge this gap, we developed a data pipeline to filter and annotate multilayer images derived from real world PSD (Photoshop Document) files.



(a) Distribution of Layer Counts (b) Category Distribution

Figure 3. Statistics of the processed multilayer image dataset. (a) Distribution of layer counts before and after merging. (b) Category distribution in the final dataset.

We began by collecting a large corpus of PSD files and extracting all layers using `psd-tools`, an open-source Python library for parsing Adobe Photoshop documents. To ensure data quality, we filtered out layers containing anomalous elements, such as blurred faces. To improve decomposition performance, we removed non-contributing layers that do not influence the final composite image. Furthermore, given that some PSD files contain hundreds of layers—thereby increasing model complexity—we merged spatially non-overlapping layers to reduce the total layer count. As shown in Fig. 3a, this operation substantially reduces the number of layers. Finally, we employed Qwen2.5-VL [3] to generate text descriptions for the composite images, enabling Text-to-Multi-RGBA generation.

4.2. Implementation Details

Building upon Qwen-Image [42], we developed Qwen-Image-Layered. The model was trained using the Adam optimizer [1] with a learning rate of 1×10^{-5}. For Text-to-RGB and Text-to-RGBA generation, training was performed on an internal dataset. For both Text-to-Multi-RGBA and Image-to-Multi-RGBA generation, the model was optimized on our proposed multilayer image dataset, with the maximum number of layers set to 20. The training process was conducted in three stages, comprising 500K, 400K, and 400K optimization steps, respectively.

4.3. Quantitative Results

4.3.1. Image Decomposition

To quantitatively evaluate image decomposition, we adopt the evaluation protocol introduced by LayerD [36]. This protocol aligns layer sequences of varying lengths using order-aware Dynamic Time Warping and allows for the merging of adjacent layers to account for inherent ambiguities in decomposition (i.e., a single image may have multiple plausible decompositions). Quantitative results on Crello dataset [44] are reported in Tab. 1. Following LayerD [36], we report two metrics: RGB L1 (the L1 distance

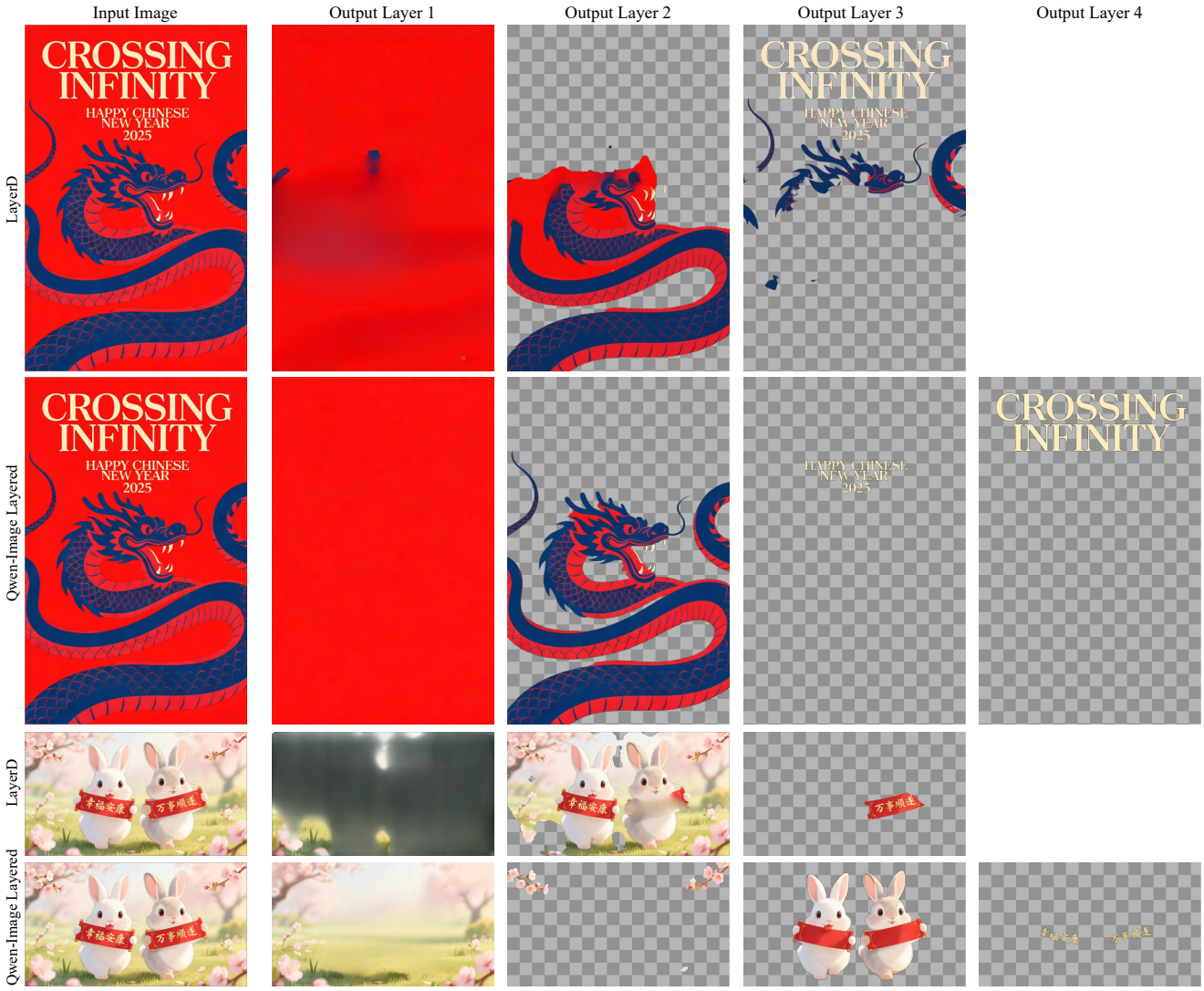


Figure 4. Qualitative comparison of Image-to-Multi-RGBA (I2L). The leftmost column shows the input image; the subsequent columns present the decomposed layers. Notably, LayerD [36] exhibits inpainting artifacts (Output Layer 1) and inaccurate segmentation (Output Layer 2 and 3), while our method produces high-quality, semantically disentangled layers, suitable for inherently consistent image editing.

Table 1. Quantitative comparison of Image-to-Multi-RGBA (I2L) on Crello dataset [44]. RGB L1: L1 distance between RGB channels weighted by the ground-truth alpha. Alpha soft IoU: soft IoU between predicted and ground-truth alpha channel.

Metric	RGB L1↓						Alpha soft IoU↑					
# Max Allowed Layer Merge	0	1	2	3	4	5	0	1	2	3	4	5
VLM Base + Hi-SAM [7]	0.1197	0.1029	0.0892	0.0807	0.0755	0.0726	0.5596	0.6302	0.6860	0.7222	0.7465	0.7589
Yolo Base + Hi-SAM	0.0962	0.0833	0.0710	0.0630	0.0592	0.0579	0.5697	0.6537	0.7169	0.7567	0.7811	0.7897
LayerD [36]	0.0709	0.0541	0.0457	0.0419	0.0403	0.0396	0.7520	0.8111	0.8435	0.8564	0.8622	0.8650
Qwen-Image-Layered-I2L	**0.0594**	**0.0490**	**0.0393**	**0.0377**	**0.0364**	**0.0363**	**0.8705**	**0.8863**	**0.9105**	**0.9121**	**0.9156**	**0.9160**

of the RGB channels weighted by the ground-truth alpha) and Alpha soft IoU (the soft IoU between predicted and ground-truth alpha channels). Due to a significant distribution gap between the Crello dataset and our proposed multilayer dataset—such as differences in the number of layers and the presence of semi-transparent layers—we finetune our model on Crello training set. As shown in Tab. 1, our method achieves the highest decomposition accuracy, notably achieving a significantly higher Alpha soft IoU score, underscoring its superior ability in generating high-fidelity alpha channels.

Figure 5. Qualitative comparison of image editing. The leftmost column is the input image; prompts are listed above each row. Qwen-Image-Edit-2509 [42] struggles with resizing and repositioning, tasks inherently supported by Qwen-Image-Layered. Meanwhile, Qwen-Image-Edit-2509 introduces pixel-level shifts (last row), while Qwen-Image-Layered can ensure consistency by editing specific layers.

Table 2. Ablation study on Crello dataset [44]. L: Layer3D Rope, R: RGBA-VAE, M: Multi-stage Training.

Metric	Component			RGB L1↓						Alpha soft IoU↑					
# Max Allowed Layer Merge	L	R	M	0	1	2	3	4	5	0	1	2	3	4	5
Qwen-Image-Layered-I2L-w/o LRM	×	×	×	0.2809	0.2567	0.2467	0.2449	0.2439	0.2435	0.3725	0.4540	0.5281	0.5746	0.5957	0.6031
Qwen-Image-Layered-I2L-w/o RM	✓	×	×	0.1894	0.1430	0.1255	0.1173	0.1138	0.1126	0.5844	0.6927	0.7576	0.7847	0.7954	0.7984
Qwen-Image-Layered-I2L-w/o M	✓	✓	×	0.1649	0.1178	0.1048	0.0992	0.0966	0.0959	0.6504	0.7583	0.8074	0.8243	0.8310	0.8331
Qwen-Image-Layered-I2L	✓	✓	✓	**0.0594**	**0.0490**	**0.0393**	**0.0377**	**0.0364**	**0.0363**	**0.8705**	**0.8863**	**0.9105**	**0.9121**	**0.9156**	**0.9160**

Table 3. Quantitative comparison of RGBA image reconstruction on the AIM-500 dataset [22].

Model	Base Model	PSNR↑	SSIM↑	rFID↓	LPIPS↓
LayerDiffuse [49]	SDXL	32.0879	0.9436	17.7023	0.0418
AlphaVAE [40]	SDXl	35.7446	0.9576	10.9178	0.0495
	FLUX	36.9439	0.9737	11.7884	0.0283
RGBA-VAE	Qwen-Image	**38.8252**	**0.9802**	**5.3132**	**0.0123**

4.3.2. Ablation Study

We conducted an ablation study on Crello dataset [44] to validate the effectiveness of our proposed method. The results are presented in Tab. 2. For settings without multi-stage training, we initialize the model directly from pre-trained text-to-image weights. For experiments without RGBA-VAE, we employ the original RGB VAE to encode the input RGB image while retaining RGBA-VAE for output RGBA layers. For variants without Layer3D RoPE, we replace it with standard 2D RoPE for positional encoding. All ablation experiments follow the same evaluation protocol as described in Sec. 4.3.1. As shown in the third and fourth rows, multi-stage training effectively improves decomposition quality. Comparing the second and third rows, the superior performance in the third row indicates that RGBA VAE effectively eliminates the distribution gap, thereby improving overall performance. Furthermore, the comparison between the first and second rows illustrates the necessity of Layer3D Rope: without it, the model can not distinguish between different layers, thus failing to decompose images into multiple meaningful layers.

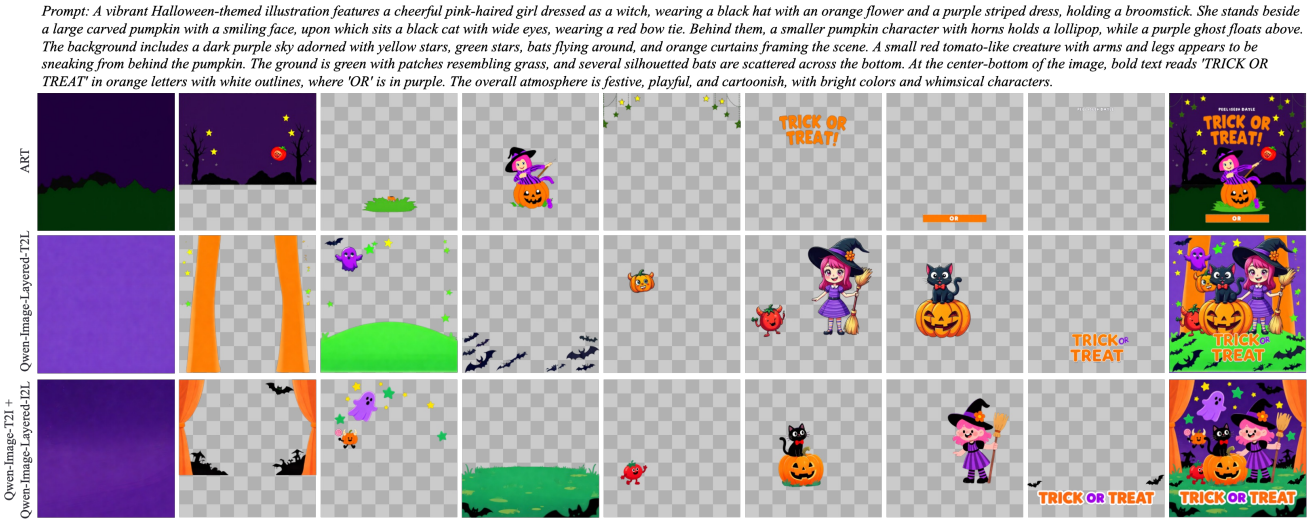


Figure 6. Qualitative comparison of Text-to-Multi-RGBA (T2L). The rightmost column shows the composite image. The second row directly generates layers from text (Qwen-Image-Layered-T2L); the third row first generates a raster image (Qwen-Image-T2I) then decomposes it into layers (Qwen-Image-Layered-I2L). ART [32] fails to follow the prompt, while Qwen-Image-Layered-T2L produces semantically coherent layers, and Qwen-Image-T2I + Qwen-Image-Layered-I2L further improves visual aesthetics.

4.3.3. RGBA Image Reconstruction

Following AlphaVAE [40], we quantitatively evaluate RGBA image reconstruction by blending the reconstructed images over a solid-color background. Quantitative results on AIM-500 dataset [22] are presented in Tab. 3, where we compare our proposed RGBA VAE against LayerDiffuse [49] and AlphaVAE [40] in terms of PSNR, SSIM, rFID, and LPIPS. As shown in Tab. 3, RGBA VAE achieves the highest scores across all four metrics, demonstrating its outstanding reconstruction capability.

4.4. Qualitative Results

4.4.1. Image Decomposition

We present a qualitative comparison of image decomposition with LayerD [36] in Fig. 4. Notably, LayerD produces low-quality decomposition layers due to inaccurate segmentation (layers 2 and 3) and inpainting artifacts (layer 1), rendering its results unsuitable for editing. In contrast, our model performs image decomposition in an end-to-end manner without relying on external modules, yielding more coherent and semantically plausible decompositions, thereby facilitating inherently consistent image editing.

4.4.2. Image Editing

In Fig. 5, we present a qualitative comparison with Qwen-Image-Edit-2509 [42]. For Qwen-Image-Layered, we first decompose the input image into multiple semantically disentangled RGBA layers and then apply simple manual edits. As illustrated, Qwen-Image-Edit-2509 struggles to follow instructions involving layout modifications, resizing, or repositioning. In contrast, Qwen-Image-Layered inherently supports these elementary operations with high fidelity. Moreover, Qwen-Image-Edit-2509 introduces noticeable pixel-level shifts, as shown in the bottom row. By contrast, layered representation enables precise editing of individual layers while leaving others exactly untouched, thereby achieving consistency-preserving editing.

4.4.3. Multilayer Image Synthesis

In Fig. 6, we present a qualitative comparison of Text-to-Multi-RGBA generation. In the second row, we directly employ Qwen-Image-Layered-T2L for text-conditioned multilayer image synthesis. Alternatively, we first generate a raster image from text using Qwen-Image-T2I [42] and then decompose it into multiple layers using Qwen-Image-Layered-I2L. As illustrated, ART [32] struggles to generate semantically coherent multilayer images (*e.g.* missing bats and cat). In contrast, Qwen-Image-Layered-T2L produces semantically coherent multilayer compositions. Moreover, the pipeline combining Qwen-Image-T2I and Qwen-Image-Layered-I2L further leverages the knowledge embedded in the text-to-image generator, enhancing both semantic alignment and visual aesthetics.

5. Conclusion

In this paper, we introduce Qwen-Image-Layered, an end-to-end diffusion model that decomposes a single RGB image into multiple semantically disentangled RGBA layers. By representing images as a stack of layers, our approach enables inherent editability: each layer can be independently manipulated while leaving all other content exactly unchanged, thereby fundamentally ensuring consistency across edits. Extensive experiments demonstrate that our method significantly outperforms existing approaches in decomposition quality and establishes a new paradigm for consistency-preserving image editing.

References

[1] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014. 5

[2] Yağiz Aksoy, Tunç Ozan Aydin, Aljoša Smolić, and Marc Pollefeys. Unmixing-based soft color segmentation for image manipulation. *ACM Transactions on Graphics (TOG)*, 36(2):1–19, 2017. 3

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4, 5

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2

[5] Qi Cai, Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Hidream-i1: An open-source high-efficient image generative foundation model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13636–13639, 2025. 2

[6] Junwen Chen, Heyang Jiang, Yanbin Wang, Keming Wu, Ji Li, Chao Zhang, Keiji Yanai, Dong Chen, and Yuhui Yuan. Prismlayers: Open data for high-quality multi-layer transparent image generative models. *arXiv preprint arXiv:2505.22523*, 2025. 3

[7] Jingye Chen, Zhaowen Wang, Nanxuan Zhao, Li Zhang, Difan Liu, Jimei Yang, and Qifeng Chen. Rethinking layered graphic design generation with a top-down approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16861–16870, 2025. 3, 4, 6

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[9] 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. 2

[10] 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 *Forty-first international conference on machine learning*, 2024. 2, 4

[11] 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.

[12] 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. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[15] Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. Psdiffusion: Harmonized multi-layer image generation via layout and appearance alignment. *arXiv preprint arXiv:2505.11468*, 2025. 3

[16] Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. Dreamlayer: Simultaneous multi-layer generation via diffusion mode. *arXiv preprint arXiv:2503.12838*, 2025. 4

[17] Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024. 3, 4, 5

[18] Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025. 3, 4, 5

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[20] Yuki Koyama and Masataka Goto. Decomposing images into layers with advanced color blending. In *Computer Graphics Forum*, pages 397–407. Wiley Online Library, 2018. 3

[21] 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. 2

[22] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 7, 8

[23] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785, 2024. 3

[24] 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. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[26] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2

[27] 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. 4

[28] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or,

and Chi-Wing Fu. Object-level scene deocclusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[29] Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6842–6850, 2024. 2

[30] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised layered image decomposition into object prototypes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8640–8650, 2021. 3

[31] 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. 2

[32] Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7952–7962, 2025. 3, 5, 8

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[35] Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Lime: localized image editing via attention regularization in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 222–231. IEEE, 2025. 2

[36] Tomoyuki Suzuki, Kang-Jun Liu, Naoto Inoue, and Kota Yamaguchi. Layerd: Decomposing raster graphic designs into layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17783–17792, 2025. 3, 4, 5, 6, 8

[37] Jianchao Tan, Jyh-Ming Lien, and Yotam Gingold. Decomposing digital paintings into layers via rgb-space geometry. *arXiv preprint arXiv:1509.03335*, 2015. 3

[38] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. 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, 2024. 3, 5

[39] 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. 2

[40] Zile Wang, Hao Yu, Jiabo Zhan, and Chun Yuan. Alphavae: Unified end-to-end rgba image reconstruction and generation with alpha-aware representation learning. *arXiv preprint arXiv:2507.09308*, 2025. 3, 7, 8

[41] 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*, pages 720–736. Springer, 2022. 2

[42] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 3, 5, 7, 8

[43] 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*, 2025. 2

[44] Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5481–5489, 2021. 5, 6, 7

[45] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7643–7653, 2025. 3, 4

[46] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 3

[47] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3784–3792, 2020. 3

[48] 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. 2

[49] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 3, 7, 8

[50] Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023. 3, 5