DreamLayer: Simultaneous Multi-Layer Generation via Diffusion Model

Junjia Huang[1,2*] Pengxiang Yan[3*] Jinhang Cai[3] Jiyang Liu[3]
Zhao Wang[3] Yitong Wang[3] Xinglong Wu[3] Guanbin Li[1,2,4†]
[1]Sun Yat-sen University, [2]Peng Cheng Laboratory, [3]ByteDance Intelligent Creation
[4]Guangdong Key Laboratory of Big Data Analysis and Processing
huangjj77@mail2.sysu.edu.cn, wantong1017@163.com, liguanbin@mail.sysu.edu.cn
{yanpengxiang.ai, caijinhang, liujiyang.liu, zhaoxu.bit, wuxinglong}@bytedance.com
https://ll3rd.github.io/DreamLayer/

Abstract

Text-driven image generation using diffusion models has recently gained significant attention. To enable more flexible image manipulation and editing, recent research has expanded from single image generation to transparent layer generation and multi-layer compositions. However, existing approaches often fail to provide a thorough exploration of multi-layer structures, leading to inconsistent inter-layer interactions, such as occlusion relationships, spatial layout, and shadowing. In this paper, we introduce DreamLayer, a novel framework that enables coherent text-driven generation of multiple image layers, by explicitly modeling the relationship between transparent foreground and background layers. DreamLayer incorporates three key components, i.e., Context-Aware Cross-Attention (CACA) for global-local information exchange, Layer-Shared Self-Attention (LSSA) for establishing robust inter-layer connections, and Information Retained Harmonization (IRH) for refining fusion details at the latent level. By leveraging a coherent full-image context, DreamLayer builds inter-layer connections through attention mechanisms and applies a harmonization step to achieve seamless layer fusion. To facilitate research in multi-layer generation, we construct a high-quality, diverse multi-layer dataset including 400k samples. Extensive experiments and user studies demonstrate that DreamLayer generates more coherent and well-aligned layers, with broad applicability, including latent-space image editing and image-to-layer decomposition.

1. Introduction

In recent years, text-to-image generation based on diffusion models [1, 3, 4, 22, 24, 27] has demonstrated impressive capabilities to create high-quality, detail-rich images from text prompts. However, most methods focus on generating a single, complete image, significantly limiting their potential in applications like content editing and graphic design, which rely heavily on layered compositions. Layered structures are particularly advantageous for images containing multiple objects, as they allow for more flexible editing and creative modifications. This paper investigates the application of diffusion models to generate coherent, multi-layered images through a simple text-driven process.

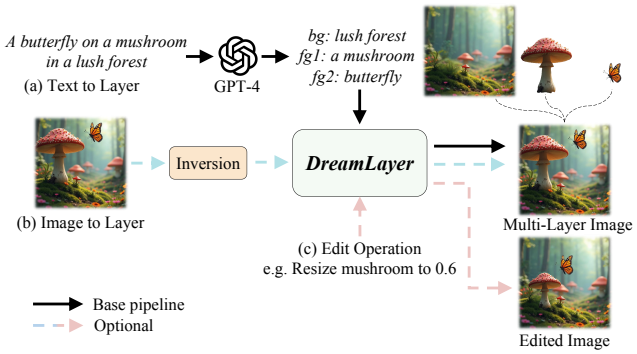


Figure 1. DreamLayer can handle multiple tasks: (a) Text-to-layer: Given a text input, we use GPT-4 to decompose foreground and background elements, feeding them into DreamLayer to generate a multi-layered image. (b) Image-to-layer: By using inversion to initialize starting latent, DreamLayer can decompose an image based on text prompts. (c) Latent-space editing: During denoising, DreamLayer can respond to editing instructions, producing more harmonious and consistent edited images.

Recent methods have started to explore the simultaneous generation of layered image structures to better support AI-driven image editing workflows. Most existing methods [42, 45] are limited to the generation of two-layer structures, *i.e.*, foreground and background layers. While certain approaches [11, 42] attempt to model the multi-layer generation task, they lack consideration for the relationship

*Equal Contribution.
†Corresponding Author.

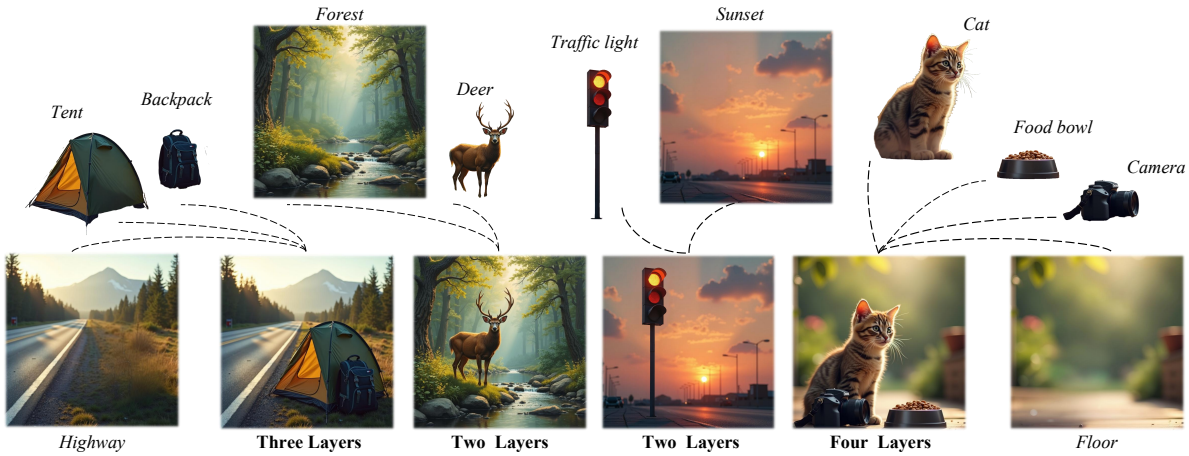


Figure 2. Multi-layer Dataset: Each image consists of a multi-layer structure, including a background and multiple foreground objects, with each foreground object represented as a transparent layer.

between different foreground layers and the background. For instance, LayerDiffusion [42] disregards the spatial relationships between layers when adding new ones, leading to unintended overlaps between layers. LayerDiff [11] attempts to generate multi-layer composite images simultaneously, but it can only generate isolated, non-overlapping layers. These methods typically rely on simple stacking for layer composition, neglecting essential effects like shadows and occlusion, which are important for cohesive multi-layer generation and editing. Furthermore, a significant challenge in multi-layer generation is the lack of large-scale, high-quality open-source datasets. Current approaches and their datasets often rely on randomly stacked segmentation data [37], suffer from limited data volume [32], or lack strictly defined multi-layered images [11].

To address these challenges, we propose a multi-layer data generation pipeline that decomposes images generated by advanced text-to-image models, createing a dataset of $400k$ multi-layer samples, as shown in Fig. 2. In existing text-to-image generation models, when given a text prompt containing a background and multiple foregrounds, the models often demonstrate the ability to automatically arrange objects in a reasonable layout and generate harmonious compositions. Building upon this, we introduce DreamLayer, a framework that utilizes global layer information to guide inter-layer attention and integrates a harmosnization mechanism. To address layout issues in foreground layers,s we begin by generating a cohesive global image from the full-text prompt. Then, we employ Context-Aware Cross-Attention to extract contextual information from the global image, guiding the generation of foreground layers. To establish connections between layers, we adapt Layer-Shared Self-Attention, which further facilitates the sharing of global information across independent layers. Finally, we apply Information Retained Harmonization to fuse the composite image in latent space, ensuring a harmonious final result and improving consistency for subsequent editing tasks. DreamLayer enables the generation of multi-layer images with cohesive layouts and seamless integration across layers. It also supports a variety of tasks. As shown in Fig. 1, (a) DreamLayer can perform text-to-layer to generate multi-layer images by adaptively decomposing user text prompts; (b) DreamLayer supports image-to-layer decomposition by initializing the denoising latent via inversion and directing it based on text prompts in a training-free manner; (c) DreamLayer supports user-driven edits during the denoising process, ensuring harmonious edited images. In summary, our key contributions are threefold:

- We introduce DreamLayer, a simultaneous multi-layer generation framework that enhances harmony and consistency across layers via inter-layer interaction.
- We propose a layer-level harmonization approach to achieve smoother inter-layer blending, making it more adaptable for subsequent editing tasks.
- The release of a large-scale, high-quality multi-layer dataset, containing $400k$ meticulously curated multi-layer images, covering multiple objects and scenes.

2. Related Work

Diffusion based Image Generation. Diffusion models [10, 31] have shown leading performance in generative tasks, including image generation [24, 40, 41, 48], editing [2, 12], inpainting [17], and video generation [7, 47]. These models have evolved from early pixel-space denoising [28] to latent-space denoising [24], with architectures progressing from U-Net [25] to advanced designs like DiT [4, 21]. For multi-layer image generation, Text2Layer [45] uses a latent diffusion model to jointly reconstruct the RGB and alpha channels for two-layer images. LayerDiffusion [42] encodes the alpha channel in the latent manifold and shares

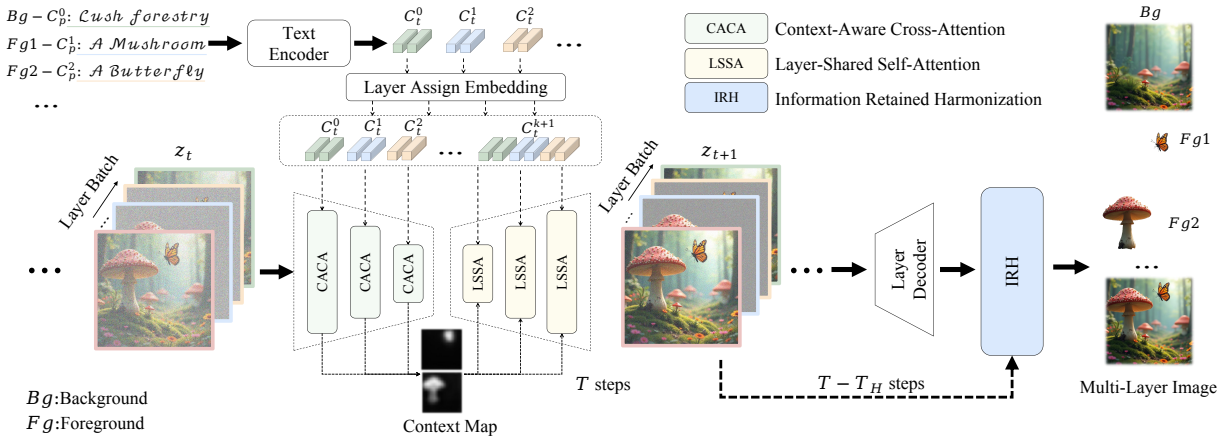
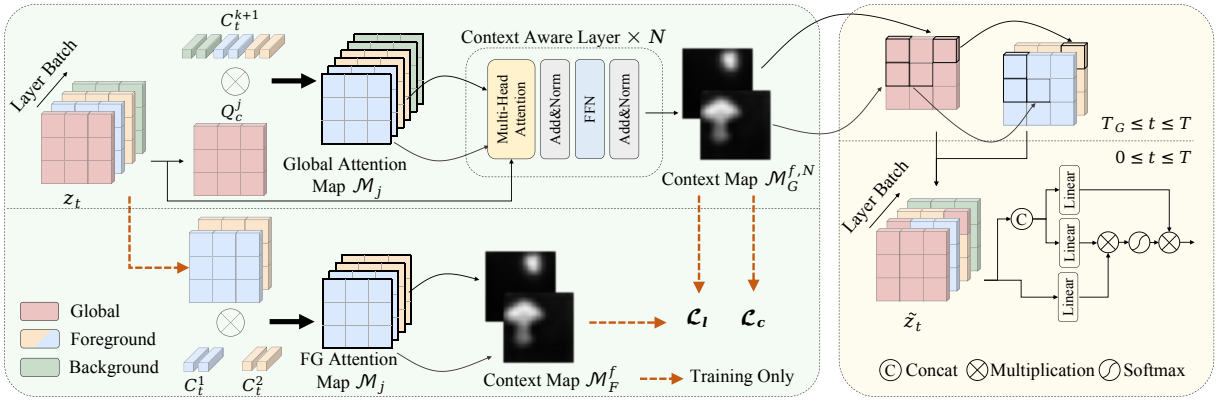


Figure 3. The DreamLayer Framework for Multi-Layer Image Generation: During the generation process, background and foreground prompts are combined via layer assign embeddings to form a global prompt C_t^{k+1}. In the attention phase, CACA extracts a context map from the global layer. Subsequently, the contextual information is fused across layers through LSSA, based on the global context map. Finally, IRH fuses the images using the latent image during the denoising process, achieving a harmonious result.



(a) Context-Aware Cross-Attention (b) LayerShared Self-Attention

Figure 4. Overview of the Attention Mechanism in DreamLayer: (a) Context-Aware Cross-Attention for extracting the global context map and guiding the foreground layer layout; (b) Layer-Shared Self-Attention for establishing inter-layer connections and ensuring consistency.

attention between foreground and background layers to generate multi-layered images. LayerDiff [11] proposes to generate multi-layer composites with layer-collaborative attention. However, these approaches often neglect integrated effects such as shadows and other inter-layer interactions between multiple foreground and background layers.

Controllable Diffusion Model and Image Editing. To enhance controllability in image generation, a range of methods have been developed. Textual Inversion [5] and DreamBooth [26] enable personalized content generation from a small set of example images. ControlNet [43] and T2I-Adapter [20, 39] introduce conditional signals, using reference images as visual prompts for direct guidance. Other methods [35, 46] utilize bounding boxes to control image layout, while P2P [8] and PnP [33] condition attention layers to manage content and style. To enable more customized image content, some methods [6, 12, 19] leverage inversion techniques, converting the input image into a noise latent representation, which is then edited and generated in a controlled manner based on text prompts. DesignEdit [13] further segment the latent representations into multiple layers, allowing more flexible spatial editing. However, existing methods typically restrict layers to non-overlapping structures. In this work, we utilize a harmonious global layer to guide the generation of layers and employ independent layers to achieve seamless fusion in the latent space.

3. Methodology

Definitions. Intuitively, a k-layer image consists of a background layer I^1, $k-1$ foreground layers $\{I^i\}_{i=2}^k$ and a global layer I^{k+1}. Each layer comprises a three-channel color image $I_c \in \mathbb{R}^{H\times W\times 3}$ and an alpha channel $I_\alpha \in \mathbb{R}^{H\times W\times 1}$, where the alpha channel indicates the visibility

of the pixels within the color image. Formally, the global layer image I^{k+1} can be expressed as

$$I^{k+1} = \sum_{i=1}^{k} (I_\alpha^i \cdot I_c^i \cdot \prod_{f=i+1}^{k} (1 - I_\alpha^f)). \tag{1}$$

Each layer is associated with a corresponding textual description as a text prompt $\{C_p^i\}_{i=1}^{k+1}$.

Generation of Alpha Channel. For each layer image, we fill the image with a solid gray background based on its alpha channel to obtain an RGB layer. This RGB layer is then encoded into a latent image $z \in \mathbb{R}^{hw \times D}$ and perturbed with noise for t timesteps to produce a noisy latent image z_t. With the timestep t and a text prompt C as conditions, the diffusion model trains a network ϵ_θ to predict the noise added to the noisy latent image z_t with

$$\mathcal{L}_{noise} = \mathbb{E}_{z_t, t, C, \epsilon \sim \mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2 \right], \tag{2}$$

where $\mathcal{L}_{noise}$ represents the learning objective of the diffusion model. After T denoising steps, the latent image z_0 is decoded by a layer decoder to generate the final transparent layer image with alpha channel. The layer decoder can be diverse, some methods [11, 45] train a 4-channel VAE decoder, while others [42] utilize a VAE decoder combined with a gray-background segmentation model. In this work, we adopt the same layer decoder as used in LayerDiffusion [42]. Notably, we primarily focuses on the layout coherence and overall harmony in multi-layer generation, rather than the accuracy of alpha channel generation.

Overview. As shown in Fig. 3, for multi-layer generation, we simultaneously encode the prompt of background layer and each foreground layer with the text encoder to obtain text embedding $\{C_t^i \in \mathbb{R}^{S \times D}\}_{i=1}^{k}$, where S denotes the sequence length after tokenization. A learnable layer assign embedding is then added to each text embedding. We extract the portion between the [SOS] and [EOS] tokens from each text embedding and concatenate them to form a global embedding $C_t^{k+1} \in \mathbb{R}^{S \times D}$, which captures the essential information of all layers and guides the generation of the global layer. All layers are processed in a batch-wise manner during the attention computation of the diffusion model. To fully utilize the information from the global layer, we design three key components: Context-Aware Cross-Attention (CACA), Layer-Shared Self-Attention (LSSA), and Information Retained Harmonization (IRH). These components leverage guidance from the global layer to ensure consistency across background layer and foreground layers, facilitating the generation of harmonious multi-layer images.

3.1. Context-Aware Cross-Attention

The key to multi-layer generation is maintaining consistency in layout and proportions across all layers. During the generation process, we align the layout positions of other layers with the global layer. Utilizing the text embeddings of each layer, we extract relevant information from the cross-attention of the global layer. Formally, as shown in Fig. 4 (a), the global noisy latent image $z_t^{j,k+1}$ in the j^{th} cross-attention mechanism is projected to a query matrix $Q_c^j = \ell_Q(z_t^{j,k+1})$ and the attention map $\mathcal{M}_j \in \mathbb{R}^{hw \times S}$ is then calculated with global embedding as

$$\mathcal{M}_j = Softmax(\frac{Q_c^j \ell_K (C_t^{k+1})^T}{\sqrt{d}}), \tag{3}$$

where ℓ_Q, ℓ_K are linear projections and d is the latent dimension. The attention map preserves the spatial layout and geometry of the different foreground objects [8, 44]. Therefore, we extract the cross-attention maps corresponding to each foreground object from J layers in the diffusion model. These maps are combined to create f initial spatial-aware global attention maps $\mathcal{M}_G^f$:

$$\mathcal{M}_G^f = Norm(\sum_{s=1}^{S_f} \sum_{j=1}^{J} (\mathcal{M}_j^s)), f = 2, \cdots, k \tag{4}$$

where $Norm(\cdot)$ denotes the Min-Max Normalization and S_f denotes the token length of each foreground's text embedding within the global embedding. To enhance foreground layer context in the extracted attention map, we feed the initial map and the global noisy latent image of J cross-attention mechanism into N context-aware layers $CAL(\cdot, \cdot)$ to generate global context map $\mathcal{M}_G^{f,n+1}$ as:

$$\mathcal{M}_G^{f,n+1} = CAL(\mathcal{M}_G^{f,n}, \sum_{j=1}^{J} z_t^{j,k+1}). \tag{5}$$

Each context-aware layer consists of a multi-head attention [34] followed by a feed-forward network (FFN). The context map is supervised by the alpha channel of the foreground image with

$$\mathcal{L}_c = \sum^{f} ||\mathcal{R}(I_\alpha^f) - \mathcal{M}_G^{f,N}||_2, \tag{6}$$

where $\mathcal{R}(\cdot)$ denotes the resize operation with interpolation.

After extracting the harmonious layout and geometric information of the foreground layer from the global layer I^{k+1}, we apply the same way to extract the corresponding spatial-aware attention maps $\mathcal{M}_F^f$ from the foreground-specific layers $(I^i)_{i=2}^k$. Next, we implement a layout align loss $\mathcal{L}_{layout}$ to enable the global layer to supervise and guide the local foreground layers, facilitating alignment and coherence between them:

$$\mathcal{L}_{layout} = \sum^{f} ||\mathcal{M}_G^{f,N} - \mathcal{M}_F^f||_2. \tag{7}$$

The final objective can be jointly written as

$$\mathcal{L} = \lambda_{noise}\mathcal{L}_{noise} + \lambda_c \mathcal{L}_c + \lambda_{layout}\mathcal{L}_{layout}, \tag{8}$$

where $\lambda_{noise}, \lambda_c$ and λ_{layout} are weight terms.

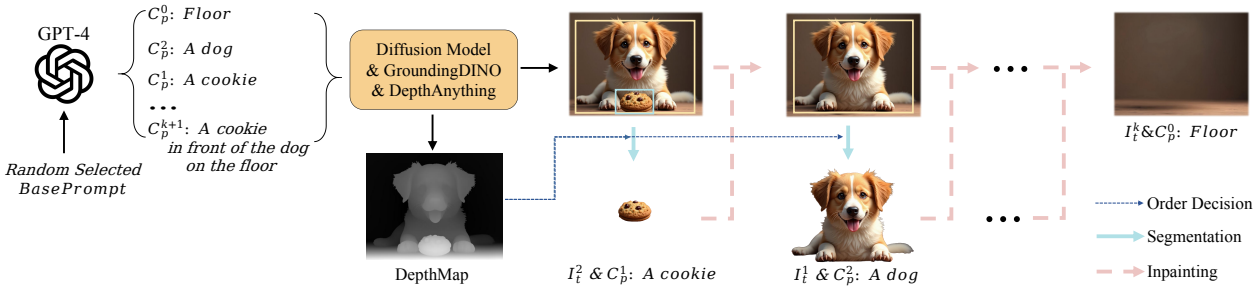


Figure 5. The pipeline of multi-layer data preparation. We utilize GPT-4 to process a randomly selected base prompt, structuring it into a background prompt and multiple foreground prompts. After generating the image using a diffusion model, we apply an open-set detection model GroundingDINO to identify the positions of the foreground objects and use the DepthAnything model to obtain a depth map. Based on the depth order, we sequentially extract the foreground layers and fill in the missing areas with an inpainting model.

3.2. Layer-Shared Self-Attention

To further strengthen the connections between layers, we propose a layer-shared self-attention. This approach first integrates global layer information into the foreground layers through the attention map, then processes information from all layers simultaneously within the self-attention mechanism, reinforcing inter-layer relationships and ensuring consistency throughout the multi-layer generation.

Specifically, as shown in Fig. 4 (b), given a layer batch of noisy latent images $\{z_t^i\}_{i=1}^{k+1}$ and the global context map $\{\mathcal{M}_G^{i,t}\}_{i=2}^{k}$ at time step t, we integrate global information into the foreground layers based on the global context map, which is expressed as:

$$\tilde{z}_t^i = z_t^{k+1} \cdot \mathcal{M}_G^{i,t} + z_t^i \cdot (1 - \mathcal{M}_G^{i,t}). \tag{9}$$

For a diffusion model with T denoising steps, we execute the process during the first T_G steps. Furthermore, to establish interaction between layers, we concatenate all noisy latent images along the sequence dimension to form a joint noise image at each denoising step:

$$\tilde{z}_t^c = \text{concat}(\tilde{z}_t^1, \cdots, \tilde{z}_t^{k+1}). \tag{10}$$

We then perform linear projections to generate the joint key K_s^c and value V_s^c, and apply attention with original layer query Q_s^i, formally as:

$$O_s^i = Softmax(\frac{Q_s^i (K_s^c)^T}{\sqrt{d}}) V_s^c, \tag{11}$$

where d denotes the latent dimension. The weights of the linear projection are initialized from the original weights.

3.3. Information Retained Harmonization

In multi-layer fusion, simply blending layers based on the alpha channel often affects the overall visual quality, as adding foreground objects to the background typically introduces shadow variations in real-world scenarios. To achieve a more harmonious fusion of the composite image, we propose Information Retained Harmonization, which blends latent during the denoising process and incorporates additional denoising steps, resulting in a more coherent and visually consistent final composite image.

Specifically, during the standard T denoising steps, we retain the noisy latent images between step T_H and T'_H, denoted as $\{z_t\}_{t=T_H}^{T'_H}$. After completing the T denoising steps, we obtain the alpha channel $\{I_\alpha^i\}_{i=2}^k$ of the foreground-specific layer through the layer decoder. We then perform re-denoising for $T - T_H$ steps as a harmonization process, and between steps T_H and T'_H, we conduct latent-level layer fusion. The formulation is as follows:

$$\hat{z}_t^m = \hat{z}_t^1 \cdot \prod_{i=2}^{k} (1 - I_\alpha^i) + \sum_{i=2}^{k} (z_t^i \cdot I_\alpha^i \cdot \prod_{f=i+1}^{k} (1 - I_\alpha^f)), \tag{12}$$

where $\hat{z}$ represents the noisy latent image obtained during the harmonization steps. During the IRH process, the fused latent is influenced by the foreground objects throughout the denoising steps, allowing for the generation of corresponding shadow details and enhancing the overall coherence of the image. Simultaneously, information from the foreground layers is gradually preserved during denoising, ensuring the consistency of the generated foreground layers.

Additionally, we can edit the layers within the latent space, ensuring a smoother, more harmonious fusion of the layers. This is expressed as follows:

$$\begin{aligned} \hat{z}_t^m = & \hat{z}_t^1 \cdot \prod_{i=2}^{k} (1 - op(I_\alpha^i)) + \\ & \sum_{i=2}^{k} (op(z_t^i) \cdot op(I_\alpha^i) \cdot \prod_{f=i+1}^{k} (1 - op(I_\alpha^f))), \end{aligned} \tag{13}$$

where $op(\cdot)$ represents the operations such as resizing, flipping, and moving.

Methods	Two Layers			Three Layers			Four Layers		
	AES↑	Clip↑	FID↓	AES↑	Clip↑	FID↓	AES↑	Clip↑	FID↓
SD v1.5 [24]	6.930	34.678	53.950	6.363	34.222	55.198	6.367	35.000	59.149
LayerDiffusion [42]	6.522	32.466	63.481	6.058	30.350	67.118	5.975	29.158	79.997
DreamLayer w/o IRH	6.967	34.587	51.957	6.351	34.696	58.812	6.340	34.993	57.073
DreamLayer	**7.013**	**34.835**	**50.761**	**6.441**	**35.267**	**54.508**	**6.422**	**35.723**	**53.598**

Table 1. Quantitative comparison of multi-layer composite image generation. For SD v1.5, we generate the complete image from the global prompt as a composite image.

Dataset	Images	Resolutions	Classes	Instances
MuLAn [32]	44,860	600∼800	759	101,269
DreamLayer	408,187	896∼1152	1453	525,388
-TwoLayer	305,801	896∼1152	1379	305,801
-ThreeLayer	87,571	896∼1152	1322	175,142
-FourLayer	14,815	896∼1152	1045	44,445

Table 2. Dataset comparison between MuLAn and DreamLayer.

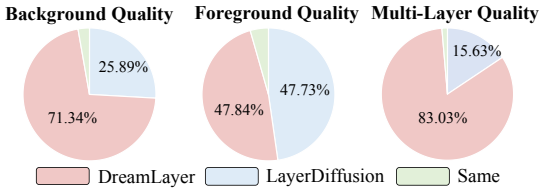


Figure 6. The vote preference percentage in user study. We evaluate our method and LayerDiffusion on three aspects: multi-layer, foreground, and background quality.

3.4. Dataset Preparation

Fig. 5 illustrates the construction process of our multi-layer dataset. To manage complex layer relationships, we begin with the global layer and employ open-set object detection, depth maps, and inpainting to decompose it into multiple layers. First, we randomly sample a prompt from a large-scale prompt dataset [36] as the base prompt. This base prompt it then processed by the GPT-4 model, which breaks it down into a background prompt C_p^0, several foreground prompts, and a complete global prompt C_p^{k+1}. If the base prompt lacks sufficient foreground objects, GPT-4 selects a suitable category from the Object365 [30] dataset. Next, the global prompt is passed through a powerful image generation diffusion model, such as Flux [15], SD3 [16], or SDXL [22], to generate a complete image. We then use the foreground prompts and the open-set detection model, GroundingDINO [18], to match the text with objects in the image. Grounding DINO predicts bounding boxes based on input text. We match each box with a segmentation mask by computing their IoU, and link the corresponding text to the matched mask. Simultaneously, we generate a depth map of the complete image using the DepthAnything [38] model. Based on the depth map, we extract the object at the forefront using a matting model and fill in the missing areas with an inpainting model. Repeating this process, we determine the sequence of layers using the depth map and extract the corresponding foreground layers. We match the objects and text using segmentation masks and detection boxes, ultimately obtaining transparent images for multi-layers. Current generative models still struggle in generating a larger number of objects, resulting in low data retention. Therefore, we set the final output to 4 layers. Details of the pipeline are in the supplementary materials.

Following this pipeline, we generate a dataset containing millions of multi-layer images. To further improve the quality, we apply both VLM-based (e.g., GPT-4o) and human quality inspection to filter out samples with (1) background artifacts, (2) incomplete foreground masks, and (3) text-layer mismatches from Grounding DINO errors. The final dataset comprises $300k$ two-layer images, $85k$ three-layer images, and $15k$ four-layer images. As shown in Tab. 2, our dataset contains more samples and encompasses a broader range of classes compared to existing datasets.

4. Experiments

4.1. Implementation Details

Training. We initialize training with the pre-trained weights of Stable Diffusion v1.5 [24] and employ the Custom Diffusion [14] strategy, fine-tuning the K&V linear layers in all attention layers. For foreground layers, additional K&V layers are trained separately. The Context-Aware Cross-Attention is applied in the downsampling layers at a resolution of 16, while Layer-Shared Self-Attention is used in all upsampling layers. Each layer batch is initialized with same timestep noise, and the Layer Embedding is zero-initialized to minimize interference with the original weights. The training is performed over 4 days on 2 A100 GPUs with a batch size of 4 and a learning rate of 2e-6. More details are available in the supplementary materials.
Evaluation. We evaluate DreamLayer on a test set of 3k multi-layer images from our proposed dataset. Aesthetic quality is assessed using the AES Score [29], text-image alignment with the CLIP-Score [23], and distribution similarity with the FID [9].

4.2. Comparisons of Multi-Layer Image Generation

Quantitative Comparisons. As shown in Tab. 1, We compare DreamLayer's performance in generating complete layers with results from Stable Diffusion [24] (SD15) and

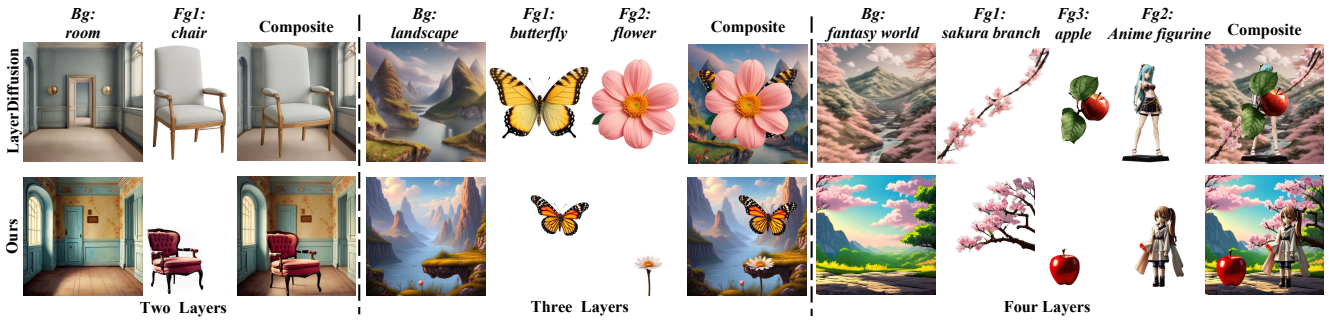


Figure 7. Qualitative comparison of multi-layer image generation. We present the generation results of Layerdiffusion and our method with two-layer, three-layer and four-layer images.

Figure 8. Ablation Study on Context-Aware Cross-Attention: Using $\mathcal{L}_{layout}$ supervision to extract layout information from the global image, guiding the generation of foreground layers and reducing overlapping placements.

DreamLayer		Multi-Layers (Average)		
LSSA	CACA	AES↑	Clip↑	FID↓
		6.438	33.808	57.241
✓		6.471	33.727	56.598
	✓	6.561	34.004	55.788
✓	✓	**6.625**	**35.275**	**52.956**

Table 3. Ablation study on LSSA and CACA.

LayerDiffusion [42]. In this setup, SD15 generates a single complete image based solely on a global prompt. We use LayerDiffusion's background-to-foreground approach for three-layer and four-layer images, sequentially adding foreground elements to simulate multi-layer composition. As shown in the table, our method outperforms LayerDiffusion across all three metrics for multi-layer generation, with a notable improvement of around 0.5 in aesthetic score. For composite multi-layer images, our approach also achieves higher aesthetic quality and better text alignment compared to direct full-image generation by SD15.

Qualitative Comparisons. In Fig. 7, we present the multi-layer image generation results. Compared to Layerdiffusion [42], our method produces more coherent and appropriately sized foreground layers and achieves a more harmonious blending of the foreground and background.

User Study. As shown in Fig. 6, we perform a user study with 20 subjects on 200 samples to evaluate the multi-layer generation quality of our method and LayerDiffusion [42]

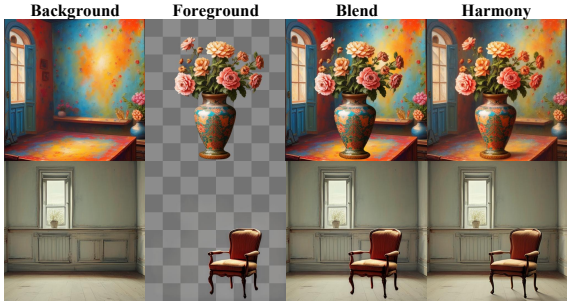


Figure 9. Ablation study on IRH. Our harmonization approach, unlike direct blending, generates appropriate shadows for foreground objects, resulting in a more cohesive overall composition.

T_H	0	200	400	600	800
T'_H	0	0	200	400	600
Avg AES↑	6.553	6.568	6.600	**6.625**	6.615

Table 4. Investigation of T_H and T'_H in IBH.

across three aspects: multi-layer, foreground, and background quality. The results show that our method achieves a preference percentage of 71.34%, 47.84%, 83.03% w.r.t the above three aspects. It indicates our method delivers more cohensive layouts and higher quality, particularly in background and multi-layer images.

4.3. Ablation Study

Context-Aware Cross-Attention. CACA extracts the context map information from the global layer and utilizes $\mathcal{L}_{layout}$ to guide the layout of the foreground layer. As shown in Fig. 8, without the layout alignment loss w/o $\mathcal{L}_l$, foreground objects tend to generate in the same position, leading to overlap and occlusion. We report the qualitative results in Tab. 3. Removing CACA significantly degrades image quality, reducing the overall AES score of the multi-layer generation by 0.154.

Layer-Shared Self-Attention. LSSA is primarily used to maintain consistency across different image layers. As shown in Tab. 3, the absence of LSSA leads to a significant drop in the CLIP score, decreasing by approximately 1.27.

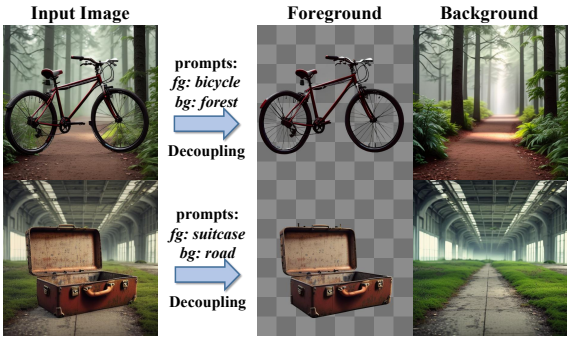


Figure 10. Image to Layer Visualization: By leveraging inversion to transfer the input image as the initial noise latent for all layers, DreamLayer can decompose the input with the text prompt.

Information Retained Harmonization. We further investigate the role of IRH in layer composition. As shown in Fig. 9, simply stacking foreground and background layers (Blend) produces unrealistic composite images, lacking texture details like shadows. For example, the chair in Fig. 9 appears to float without a shadow, disrupting visual harmony. With IRH, however, shadows and other details are generated in the background to reflect the presence of foreground objects, resulting in a more natural and cohesive layer composition. For quantitative results, as shown in Tab. 1, "DreamLayer w/o IRH" shows a noticeable decline in aesthetic scores, dropping by 0.1 without IRH.

T_H and T'_H in IBH. We investigate the values of T_H and T'_H in the IBH. We experiment with T_H from 800 to 0 steps. As Tab. 4 shows, when $T'_H < 600$, IBH is applied near the end of the denoising process, resulting in poor harmonization and low AES score. Conversely, when T_H is large (e.g., $T_H = 800$), IBH over-modifies the background, reducing the AES score. Based on these observations, we selected $T_H = 600$ and $T'_H = 400$.

4.4. Further Application

Image to Layer Within the DreamLayer framework, we can extend it to Image-to-Layer task in a training-free manner. Specifically, we encode the input image into a latent representation as the global latent, then progressively add noise up to the T step latent using an inversion technique [19], which serves as the initial latent for all layers in DreamLayer. To obtain a more accurate initial latent during this inversion process, we isolate global image information using a mask, minimizing the influence of other layers. As shown in Fig. 10, this approach enables us to decompose the input image into separate layers based on text prompts. Detailed steps are provided in the supplementary materials.

Layer Editing In practical applications, DreamLayer can generate multi-layer images and allow users to make harmonious edits to the layers. As described in Eq. (13), we perform these edits within IRH at the latent level, ensuring more cohesive adjustments. For instance, in Fig. 11, when the chair is flipped and moved, the floor shadow is updated to align with its new position, enhancing overall realism. Moreover, when parts of a foreground object extend beyond the image boundary, DreamLayer has the ability of foreground object amodal completion, which can complete the missing sections as needed when repositioned. As shown in Fig. 11, compared to existing methods like DesignEdit [13], DreamLayer successfully restores the out-of-frame areas of objects such as the robot and blue vase after they are moved.

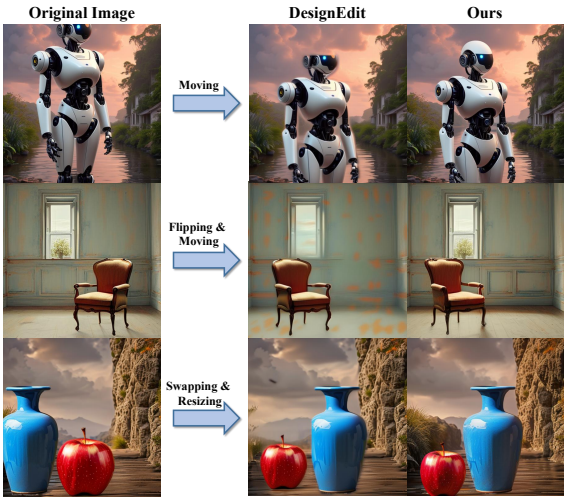


Figure 11. Layer Editing Visualization: Compared to DesignEdit, DreamLayer can complement objects at the image edges and create more cohesive results when they are flipped or moved.

5. Conclusion

In this paper, we introduce a large-scale, high-quality multi-layer dataset featuring diverse foreground objects and backgrounds. Building on this, we propose DreamLayer, a framework for simultaneously generating multi-layer images. To address layout consistency among foreground layers, we introduce Context-Aware Cross-Attention, which guides foreground generation using the harmonious layout of a global image. To enhance inter-layer connections, we present Layer-Shared Self-Attention, enabling effective information exchange between layers. Finally, to generate a cohesive composite image, we propose Information Retained Harmonization, which merges layers at the latent level to achieve seamless fusion. DreamLayer support not only multi-layer generation but also layer decomposition for image-to-layer task with inversion, enabling flexible editing within the latent space for harmonious adjustments. Experimental results demonstrate the effectiveness of DreamLayer in multi-layer generation.

6. Acknowledgment

This work is supported in part by the National Key R&D Program of China (2024YFB3908503), and in part by the National Natural Science Foundation of China (62322608).

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2

[3] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-delta: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 1

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ICLR*, 2024. 1, 2

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[6] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Improving tuning-free real image editing with proximal guidance. *WACV*, 2023. 3

[7] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *ECCV*, pages 123–139. Springer, 2024. 2

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 4

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 6

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[11] 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. *arXiv preprint arXiv:2403.11929*, 2024. 1, 2, 3, 4

[12] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, pages 12469–12478, 2024. 2, 3

[13] Yueru Jia, Yuhui Yuan, Aosong Cheng, Chuke Wang, Ji Li, Huizhu Jia, and Shanghang Zhang. Designedit: Multi-layered latent decomposition and fusion for unified & accurate image editing. *arXiv preprint arXiv:2403.14487*, 2024. 3, 8

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 6

[15] Black Forest Labs. Flux: Inference repository, 2024. Accessed: 2024-10-25. 6

[16] Stability Ai Labs. Stable diffusion 3, 2024. Accessed: 2024-10-25. 6

[17] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *CVPR*, pages 8038–8047, 2024. 2

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 3, 8

[20] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 3

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2

[22] 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. 1, 6

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 6

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 6

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[26] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 1

[28] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021. 2

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 6

[30] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 6

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[32] 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 *CVPR*, pages 22413–22422, 2024. 2, 6

[33] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 3

[34] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 4

[35] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, pages 6232–6242, 2024. 3

[36] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 6

[37] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *CVPR*, pages 9099–9109, 2024. 2

[38] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024. 6

[39] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[40] Jiacheng Zhang, Jie Wu, Huafeng Kuang, Haiming Zhang, Yuxi Ren, Weifeng Chen, Manlin Zhang, Xuefeng Xiao, and Guanbin Li. Treereward: Improve diffusion model via tree-structured feedback learning. In *ACM MM*, pages 4533–4542, 2024. 2

[41] Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Min Zheng, et al. Unifl: Improve stable diffusion via unified feedback learning. *NeurIPS*, 2025. 2

[42] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 1, 2, 4, 6, 7

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3

[44] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023. 4

[45] Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023. 1, 2, 4

[46] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *arXiv preprint arXiv:2407.02329*, 2024. 3

[47] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 2

[48] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models. In *CVPR*, pages 14235–14245, 2023. 2