ART: Anonymous Region Transformer for Variable Multi-Layer Transparent Image Generation

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 Baining Guo
[†]equal technical contribution [‡]project lead

Microsoft Research Asia Tsinghua University Peking University USTC

https://art-msra.github.io

Abstract

*Multi-layer image generation is a fundamental task that enables users to isolate, select, and edit specific image layers, thereby revolutionizing interactions with generative models. In this paper, we introduce the Anonymous Region Transformer (ART), which facilitates the direct generation of variable multi-layer transparent images based on a global text prompt and an anonymous region layout. Inspired by Schema theory[1], this anonymous region layout allows the generative model to autonomously determine which set of visual tokens should align with which text tokens, which is in contrast to the previously dominant semantic layout for the image generation task. In addition, the layer-wise region crop mechanism, which only selects the visual tokens belonging to each anonymous region, significantly reduces attention computation costs and enables the efficient generation of images with numerous distinct layers (*e.g.*, 50+). When compared to the full attention approach, our method is over 12 times faster and exhibits fewer layer conflicts. Furthermore, we propose a high-quality multi-layer transparent image autoencoder that supports the direct encoding and decoding of the transparency of variable multi-layer images in a joint manner. By enabling precise control and scalable layer generation, ART establishes a new paradigm for interactive content creation.*

1. Introduction

Diffusion-based generative models have shown tremendous success in producing high-quality images from given text prompts [4, 16, 21, 52, 58]. These models are typically limited to producing entire images in a single, unified layer, which restricts the ability to edit or manipulate specific elements independently. This limitation presents significant challenges in fields like graphic design and digital art, where creators frequently rely on layer-by-layer control to construct and refine complex compositions.

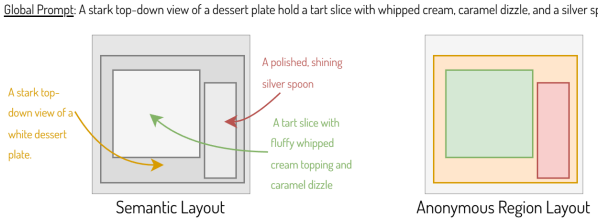


Figure 1. **Semantic Layout *vs*. Anonymous Region Layout**. The conventional semantic layout requires specifying what objects to generate in each given region, whereas our anonymous region layout only identifies where the important regions are. People can leverage the prior knowledge, activated by the global prompt, to intuitively infer the semantic label of each anonymous region. The generative model also learns to harness this capability and autonomously determine what to generate in each region.

This paper presents Anonymous Region Transformer for multi-layer transparent image generation. The key ingredient of the anonymous region transformer is the anonymous region layout, which solely consists of a set of anonymous rectangular regions without any region-wise prompt annotations, as shown in Figure 1. This is unlike the conventional semantic layout for text-to-image generation [45, 78, 81], which requires clearly specify both the global prompt for the entire image and the location and region-wise prompts for each region[2]. The drawback of the conventional layout is that it heavily relies on human labor for creating the layout and this process can be very labor intensive, especially when handling tens or even hundreds of regions on a canvas, a common scenario in graphic design generation. The anonymous region transformer significantly reduces the human labor by allowing the generative model to perform the visual planning task of determining which objects to generate in each anonymous region based on the global prompt. The core insight behind the anonymous region layout is to *enhance generative model control, while preserving user flexibility over manipulating multi-layered outputs.*

[1]Schema theory [3, 57] suggests that knowledge is organized in frameworks (schemas) that enable people to interpret and learn from new information by linking it to prior knowledge.

[2]We use 'region' and 'layer' interchangeably in this paper.

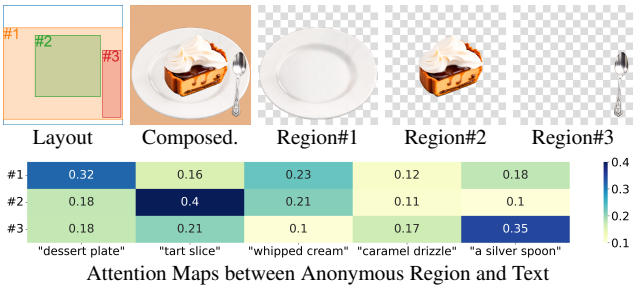


Figure 2. **Visual planning capability of our Anonymous Region Transformer**. We visualize the averaged attention maps of all visual tokens within the same anonymous region (as Query) attending to the entities within the global prompt text tokens (as Key and Value). These attention maps reveal that each anonymous region assigns the majority of attention weights to one of the major objects identified in the given text prompt.

A natural question arises regarding how the anonymous region layout can function effectively without region-wise prompts, especially given that these prompts are central to conventional semantic layout approaches. This effectiveness can be explained by Schema Theory [1, 3, 40, 57], a well-established cognitive framework that helps bridge the gap between abstract concepts (such as *plate* or *spoon*) and specific sensory experiences (such as *layout*). It suggests that people can infer each region's semantic label based on their prior knowledge activated by a global prompt. In our case, we find that the effectiveness of the anonymous-region layout for multi-layer image generation tasks stems from the Transformer model's ability to autonomously identify semantic labels for each layer through interactions between text tokens and visual tokens. The generative model learns to capture the prior knowledge similar to Schema Theory, enabling it to determine which set of visual tokens (from an anonymous region) attends to which text tokens (representing different entities), as shown in Figure 2. Our experiments further demonstrate that adding additional region-wise prompts for each layer does not necessarily improve the results and can even diminish coherence across layers.

The anonymous region transformer offers several key advantages over the conventional approach for multi-layer transparent image generation. First, it ensures better coherence across different layers. We observe that, in the semantic layout, regional visual tokens struggle to balance attention weights between region-wise text tokens (to ensure *prompt following*) and the corresponding global visual tokens located at the same position (ensure *coherence*). This difficulty arises from a semantic gap between the global visual tokens and region-wise visual tokens as they are forced to attend different text tokens. In contrast, our anonymous region layout enables all regional visual tokens and global visual tokens to attend to the same set of global text tokens, thereby closing this gap. Second, annotating the anonymous-region layout is more scalable, especially for native multi-layer graphic design images. We can easily generate a large number of high-quality anonymous-region

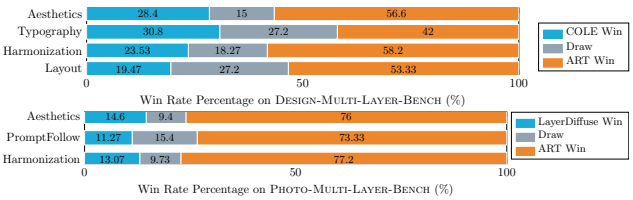


Figure 3. **ART *vs*. previous SOTA** in multi-layer transparent image generation: user study results across different domains. ART significantly outperforms LayerDiffuse [83] in the photorealistic domain and COLE [37] in the graphic-design domain across multiple aspects.

layouts, whereas recaptioning each region is non-trivial and often suffers from significant noise due the semantic gap between captioning a crop conditioned on an entire image and captioning only a small crop. Third, by focusing on the anonymous regions within each layer, we can significantly reduce computation costs and enables the efficient generation of images with numerous distinct layers (e.g., 50+).

Our methodology consists of three key components: the Multi-layer Transparent Image Autoencoder, the Anonymous Region Transformer, and the Anonymous Region Layout Planner. The Multi-layer Transparent Autoencoder encodes and decodes a variable number of transparent layers at different resolutions using a sequence of latent visual tokens. The Anonymous Region Transformer concurrently generates a global reference image, a background image, and multiple cropped transparent foreground layers from Gaussian noise conditioned on the anonymous region layout. The Anonymous Region Layout Planner predicts a set of anonymous bounding boxes based on the user-provided text prompt. Compared existing methods in multi-layer image generation—such as Text2Layer [84], LayerDiff [32], and LayerDiffuse [83]-the key difference is that these methods can produce only a limited number of transparent layers at fixed resolutions. Additionally, unlike the COLE [37] and OpenCOLE [36], which apply a cascade of diffusion models to generate layers sequentially, our method generates all transparent layers and the reference image simultaneously in an *end-to-end* manner, ensuring a better global harmonization across different layers. The experimental results demonstrate the advantages of our approach over previous methods, and we report the user study results in Figure 3.

In summary, this paper not only proposes a novel approach to multi-layer transparent image generation, but also opens up numerous possibilities for future research and applications. Our main contributions are as follows:

1. We are the first to propose a novel pipeline for multi-layer transparent image generation that supports generating a variable number of layers at variable resolution.
2. We introduce the anonymous region layout, which offers several key advantages over conventional semantic layout for multi-layer transparent image generation.
3. Our method empirically outperforms prior state-of-the-art approaches, producing higher-quality multi-layer transparent images with significantly more layers.

2. Related works

Multi-Layer Transparent Image Generation has primarily been approached through two different paths. The first generates all layers simultaneously, as seen in Text2Layer [84], which adapts Stable Diffusion for two-layer image generation, and LayerDiff [32], which uses a layer-collaborative diffusion model to generate up to four layers guided by layer-wise prompts. The second path. generates layers sequentially. LayerDiffuse [83] introducing a background-conditioned model, which generates image layers iteratively. COLE [37] and OpenCOLE [36] utilizing LLMs and diffusion models to iteratively create image elements. In contrast, our method supports generating tens of transparent layers using an anonymous region transformer, outperforming prior methods in photorealistic and design-oriented multi-layer image generation tasks.

Layout Generation and Layout Control have gained attention for their broad applications in image generation. Existing approaches can be grouped into two: (1) designing better layout generation models and (2) controlling image generation with a given layout prior. The first focuses on generating layouts from visual elements. For instance, Graphist [11], Visual Layout Composer [60], and MarkupDM [41] propose methods based on transparent visual layers. For more developments, see [6, 8–10, 17–19, 33–35, 38, 39, 43, 62, 71, 73, 77, 80]. The second enhances diffusion models' compositional capabilities by specifying object placement. Key works include GLIGEN [45], InstanceDiffusion [67], and MS-Diffusion [68], which inject positional information into diffusion models. Other efforts, such as [2, 42, 59, 63, 78, 85], propose training-free, post-training, or harmonization techniques. Closely related works like LayoutGPT [17] and TextLap [9] predict semantic layouts from a global prompts. We demonstrate the advantages of our anonymous region layout planner for multi-layer transparent image generation.

Dynamic Neural Networks adaptively adjust their structures or parameters conditioned on different inputs [23, 70], leading to notable advantages in terms of performance, adaptiveness [20, 79], computational efficiency [56, 82], and representational power [54], thus revolutionizing the paradigm of traditional static models. Dynamic networks are typically categorized into three types: sample-wise [14, 25, 27, 30, 50, 53, 55, 65, 76], spatial-wise [24, 26, 28, 31, 49, 51, 66, 74, 75, 86], and temporal-wise [29, 69]. Viewing image layers as a temporal dimension, our method can be interpreted as a temporal-wise dynamic network. It is conceptually similar to the AdaFocus [69, 72], which leverages reinforcement learning to identify and focus on the most critical regions in each video frame for video understanding. In comparison, our approach utilizes a layout planner to predict the spatial placement of each layer, enabling efficient and harmonious multi-layer image generation.

3. Approach

The conventional text-to-image model [4, 16, 44, 52, 58] supports only a single, unified image generation from a global prompt. Our approach enables diffusion transformer-based models to jointly generate images with multiple transparent layers conditioned on an anonymous region layout provided by the user or predicted by an LLM. The entire framework consists of three key components: the *Multi-layer Transparent Autoencoder* (Section 3.1), which jointly encodes and decodes multi-layer images and their corresponding latent representations; the *Anonymous Region Transformer* (Section 3.2), which concurrently generates a global reference image, a background image, and multiple RGBA transparent foreground image layers from a sequence of layout-guided noisy tokens; and the *Anonymous Region Layout Planner* (Section 3.3), which predicts a set of anonymous bounding boxes given the user-provided text prompt. The technical details are presented as follows.

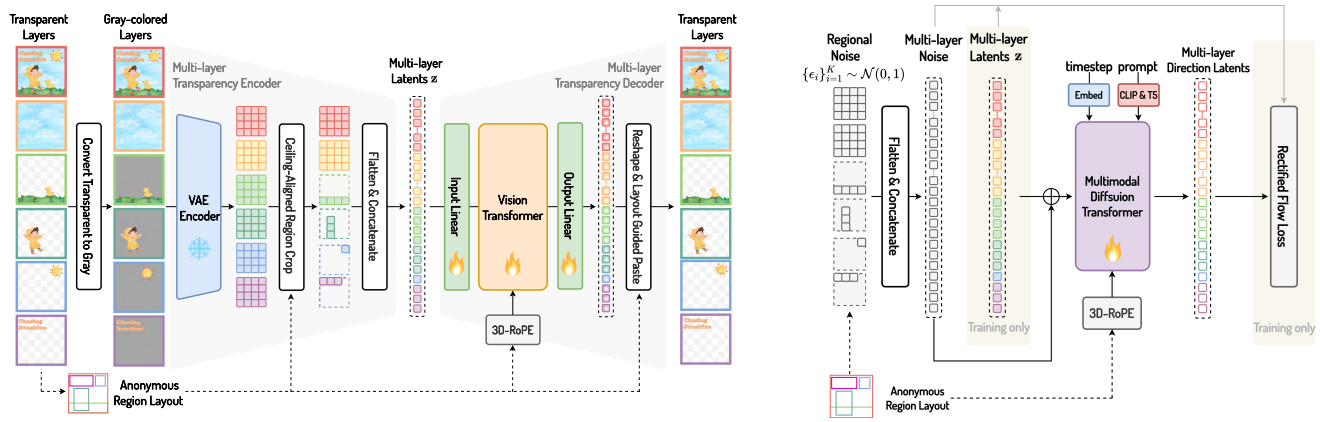
3.1. Multi-Layer Transparent Image Autoencoder

A multi-layer transparent image consists of an RGB background layer $\mathbf{I}_{\text{bg}} \in \mathbb{R}^{H\times W\times 3}$, and a variable number K of RGBA foreground layers, $\{\mathbf{I}^i_{\text{fg}} \in \mathbb{R}^{H_i\times W_i\times 4}\}_{i=1}^K$. The corresponding merged image $\mathbf{I}_{\text{mg}} \in \mathbb{R}^{H\times W\times 3}$ can be obtained by integrating $\mathbf{I}_{\text{bg}}$ as the base layer and overlaying all $\mathbf{I}^i_{\text{fg}}$ layers according to a predefined layout. We use $\mathbf{L} = \{x^i_c, y^i_c, H_i, W_i\}_{i=1}^K$ to represent the anonymous region layout of all K foreground layers. Here, x^i_c, y^i_c and H_i, W_i denote the center coordinates and the height and width of the bounding box that encapsulates the i-th transparent foreground layer, respectively.

Transparency Encoding. Our method integrates the transparency in alpha channel $\mathbf{I}^i_{\text{fg},\alpha}$ directly into the RGB channels $\mathbf{I}^i_{\text{fg,RGB}}$. Specifically, we compute $\hat{\mathbf{I}}^i_{\text{fg}} = (0.5\mathbf{I}^i_{\text{fg},\alpha} + 0.5) \times \mathbf{I}^i_{\text{fg,RGB}}$, converting the transparent-background image $\mathbf{I}^i_{\text{fg}}$ into a gray-background image $\hat{\mathbf{I}}^i_{\text{fg}}$. All channel values are normalized to range between -1 to 1. Empirically, we found that this gray background sufficient to ensure accurate transparency decoding in subsequent stages.

Multi-Layer Transparency Encoder. In the encoder part of the Multi-layer Transparency AutoEncoder (Figure 4a), the merged reference image $\mathbf{I}_{\text{mg}}$, the background layer $\mathbf{I}_{\text{bg}}$, and all the padded gray-background image layers $\{\hat{\mathbf{I}}^i_{\text{fg}}\}_{i=1}^K$ are all concatenated along the batch dimension, and then fed into the VAE encoder $\mathcal{E}_{\text{VAE}}$. This encoder [44] downsamples the spatial dimension with a factor of 8 while obtaining a 16-channel feature dimension. The extracted latent representations of the merged reference image $\mathbf{I}_{\text{mg}}$ and the background image $\mathbf{I}_{\text{bg}}$ are flattened into sequence of tokens:

$$\mathbf{z}_{\text{mg}} = \mathsf{Flatten}(\mathcal{E}_{\text{VAE}}(\mathbf{I}_{\text{mg}})), \mathbf{z}_{\text{bg}} = \mathsf{Flatten}(\mathcal{E}_{\text{VAE}}(\mathbf{I}_{\text{bg}})). \quad (1)$$



(a) Multi-Layer Transparent Image Autoencoder (b) Anonymous Region Transformer

Figure 4. (a) **Multi-layer Transparent Image Autoencoder** directly encodes each layer of the multi-layer image, accompanied by the entire composed image, into latent space and jointly decodes the multi-layer latent tokens into RGBA transparent image layers. (b) **Anonymous Region Transformer (ART)** performs denoising diffusion on the noisy multi-layer latents corresponding to a variable number of transparent layers jointly.

The VAE-downsampled foreground image layers are first subjected to a ceiling-aligned tight crop using $\mathbf{L}_i$ and then flattened into latent tokens with different lengths:

$$\mathbf{z}_{\text{fg}}^i = \text{Flatten}(\text{Crop}(\mathcal{E}_{\text{VAE}}(\hat{\mathbf{I}}_{\text{fg}}^i), \mathbf{L}_i)), \quad i = 1, \cdots, K, \quad (2)$$

where $\mathbf{L}_i$ denotes the foreground area position of layer $\mathbf{I}_{\text{fg}}^i$. The ceiling-aligned tight crop is performed by identifying the tightest bounding box with a height and width divisible by 16 to adapt to the VAE downsample rate of 8 and diffusion transformer patch size 2. Finally, the compressed multi-layer image latent $\mathbf{z}$ is obtained by concatenating the latent of the merged reference image, the background image, and the transparent foreground layers:

$$\mathbf{z} = \text{Concatenate}(\mathbf{z}_{\text{mg}}, \mathbf{z}_{\text{bg}}, \mathbf{z}_{\text{fg}}^1, \mathbf{z}_{\text{fg}}^2, \cdots, \mathbf{z}_{\text{fg}}^K). \quad (3)$$

Multi-Layer Transparency Decoder. The detailed design of our novel multi-layer transparency decoder is illustrated on the right in Figure 4a, which supports the direct decoding of a variable number of transparent layers at varying resolutions from a sequence of concatenated visual tokens in a single forward pass. We implement the multi-layer transparent image decoder based on a standard ViT architecture. The mathematical formulations are shown as follows:

$$\mathbf{v} = \text{ViT}(\text{Linear}_{\text{in}}(\mathbf{z})), \quad (4)$$

$$\mathbf{t} = \text{Reshape}(\text{Linear}_{\text{out}}(\mathbf{v}), \mathbf{L}), \quad (5)$$

where $\text{ViT}(\cdot)$ represents a ViT [12] model, $\text{Linear}_{\text{in}}(\cdot)$ denotes a linear projection that transforms the channel dimension of the latent representation, *i.e.* 16, to the hidden dimension size of ViT, especially 768, $\mathbf{v}$ represents the output representation of the ViT, $\text{Linear}_{\text{out}}(\cdot)$ denotes a linear projection that transforms the output dimension from 768 to 256, where each token can be reshaped to form an RGBA patch of size $8 \times 8 \times 4$. Another key modification in our design is the replacement of the original absolute position embedding with 3D RoPE, which is explained in the following discussion. We simply apply $\mathcal{L}_1$ loss to optimize the parameters of the multi-layer transparency decoder while freezing the parameters of the multi-layer transparency encoder.

The advantages of our multi-layer transparency decoder are twofold, including improved efficiency and enhanced transparency predictions compared to the previous single-layer transparent decoder [83]. We present the qualitative comparison results in the experimental section.

3.2. Anonymous Region Transformer

The Anonymous Region Transformer (ART) generates the visual tokens of a global reference image, a background image and all foreground layers simultaneously. The purpose of generating reference images is twofold: to better leverage the original capabilities of the existing text-to-image generation model and to ensure overall visual harmonization by preventing conflicts and inconsistency across layers. Generating all layers simultaneously also avoids the need for inpainting algorithms to complete missing parts of the occluded layers. We choose the latest multimodal diffusion transformer (MMDiT), *e.g.*, FLUX.1[dev] [44], to build our variable multi-layer image generation model, ART.

MMDiT is an improved variant of DiT framework [16] that uses two different sets of model weights to process text tokens and image tokens separately. The original MMDiT model, which only supports single image generation from a global prompt. We transform it into a multi-layer generation model by modifying the input visual tokens to encode the anonymous region layout information with a novel 3D RoPE design. We present the overall framework of ART in Figure 4b. The input consists of an anonymous region layout $\mathbf{L}$ and a global prompt $\mathbf{T}$. The noisy input is computed by adding Gaussian noise to a sequence of clean multi-layer latents $\mathbf{z}$ that encodes the reference image, background image, and all the transparent layers. We extract multi-layer latents $\mathbf{z}$ with our multi-layer transparency encoder.

Layout Conditional Multi-Layer 3D RoPE. Rotary Position Embedding (RoPE) [61] is a specific type of position embedding that applies a rotation operation to key and query in self-attention layers as channel-wise multiplications. The advantage of RoPE is that it allows the model to handle sequences of varying lengths, making it more flexible and efficient. The key design of our ART is to use a layout conditional multi-layer 3D RoPE to encode the accurate relative position information for all visual tokens, which is also utilized in the multi-layer transparency decoder. We first extract the layer-wise 3D indexing for the given noisy latents according to the anonymous region layout, *i.e.* $\mathbf{p}_n = \{p_n^x, p_n^y, p_n^l\}$ represent the width index, height index, and layer index of the n-th latents, respectively. Then, denoted n-th query and m-th key as $\mathbf{q}_n$ and $\mathbf{k}_m \in \mathbb{R}^{d_{\text{head}}}$, respectively, we split both query and key into 3 parts along channel dimensions, *i.e.* $\mathbf{q}_n = \{\mathbf{q}_n^x, \mathbf{q}_n^y, \mathbf{q}_n^l\}$ and $\mathbf{k}_m = \{\mathbf{k}_m^x, \mathbf{k}_m^y, \mathbf{k}_m^l\}$. Thus, the (n, m) component of the attention matrix is calculated as:

$$\mathbf{A}_{(n,m)} = \sum_{c \in \{x,y,l\}} \text{Re}[\mathbf{q}_n^c (\mathbf{k}_m^c)^* e^{i(p_n^c - p_m^c)\theta}], \quad (6)$$

where $\text{Re}[\cdot]$ is the real part of a complex number and $(\mathbf{k}_m^c)^*$ represents the conjugate complex number of $\mathbf{k}_m^c$. $\theta \in \mathbb{R}$ is a preset non-zero constant. The detailed implementation can be found in the supplementary material.

3.3. Anonymous Region Layout Planner

We propose an anonymous region layout planner, which predicts a set of bounding boxes based on the text input. This planner is implemented by fine-tuning an LLM model on our layout dataset, specifically using the pre-trained LLaMa-3.1-8B [15]. An example of prompts as input and the corresponding predicted layouts is given below. Unlike conventional layout definitions [36, 37, 39, 43] that specify both position and content, our anonymous region layout planner avoids assigning specific semantic labels to regions. In addition, it refrains from asking users to provide explicit layout details by users, offering greater flexibility.

Anonymous Layout Example

Input: The image is a vibrant Ramadan-themed ad featuring a rich blue background with Islamic art-inspired designs and three lit golden lanterns. The white text in the center announces a "special offer Ramadan big sale", with a subtitle that states "Discount up to 30% off". **Output**: [{ "layer": 0, "x": 512, "y": 512, "width": 1024, "height": 1024 }, { "layer": 1, "x": 744, "y": 496, "width": 496, "height": 256 }, { "layer": 2, "x": 856, "y": 704, "width": 240, "height": 96 }, { "layer": 3, "x": 792, "y": 640, "width": 368, "height": 64 }, { "layer": 4, "x": 840, "y": 336, "width": 272, "height": 64 }]

Dataset	# Samples	# Layers	Source Data	Alpha Quality
MAGICK [7]	$\sim$ 150 K	1	generated	good
Multi-layer Dataset [83]	$\sim$ 1 M	2	commercial, generated	good
LAION-L^{2}I [84]	$\sim$ 57 M	2	LAION	normal
MuLAn [64]	$\sim$ 44 K	$2 \sim 6$	COCO, LAION	poor
MLCID [32]	$\sim$ 2 M	[2,3,4]	LAION	poor
Crello [77]	$\sim$ 20 K	$2 \sim 50$	Graphic design website	normal
MLTD (ours)	$\sim$ 1 M	$2 \sim 50$	Graphic design website	good

Table 1. Comparison with existing multi-layer datasets.

3.4. Multi-Layer Transparent Design Dataset

We have collected a private, high-quality, multi-layered transparent design (MLTD) dataset that consists of approximately 1 million instances considering their high-quality alpha channels and coherent spatial arrangements. Each instance comprises multiple transparent layers with variable resolutions. The resolutions of the merged images range from 1024×1024 to 1500×1500. The average number of layers is 11, and 99.9% of designs have fewer than 50 layers. The average number of visual tokens is 11.38K, which is significantly smaller than $20 \times 32 \times 32 = 20.48$K. This indicates that the area of most foregrounds is relatively small.

Comparison with Existing Multi-Layer Data Table 1 provides a comparison between previously existing multi-layer datasets and our proposed Multi-Layer-Design dataset. Our MLTD dataset is the first large-scale dataset that includes a wide range of transparent layers with high-quality alpha channels. We also verified in the experimental section that our method can achieve sufficiently good results with only 8K high-quality data, making our method easy to replicate.

4. Experiment

Implementation details. We conduct all the experiments using the latest FLUX.1[dev] model [44]. For ablation studies, we train the MMDiT with LoRA for 30,000 iterations, with a global batch size of 8 and a learning rate of 1.0 using the Prodigy optimizer [48]. The LoRA rank is set at 64, and the image resolution is at 512×512. To ensure fair comparisons during system-level experiments, we increased the number of iterations to 90,000 and the image resolution to 1024×1024. For the multi-layer transparency decoder, we selected the ViT-Base configuration [12]. This configuration includes 12 layers, a hidden dimension size of 768, an MLP dimension size of 3072, and 12 attention heads, resulting in a total of 86 million parameters.

Training set & validation set. We choose 800K multi-layer graphic design images as the training set and a set of 5K graphic design samples to form the validation set, referred to as DESIGN-MULTI-LAYER-BENCH. Additionally, we also create a set of photorealistic multi-layer image prompts chosen from the COCO dataset [46], forming PHOTO-MULTI-LAYER-BENCH, to evaluate the model's performance on multi-layer real image generation.

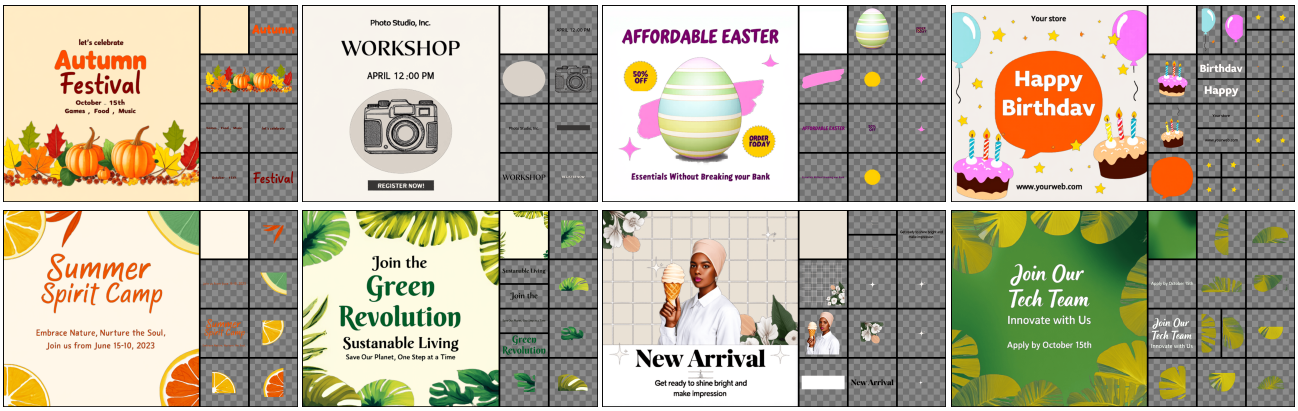


Figure 5. **Variable multi-layer transparent images generated with ART**. Transparent layers for top-left to bottom-right: 7, 8, 11, 30, 8, 10, 12, 13.

↑ COLE [37] *vs.* ART ↓ ↑ LayerDiffuse [83] *vs.* ART ↓

Figure 6. **ART v.s. COLE or LayerDiffuse**: Given the same global prompt, we display the generated multiple transparent layers to the right of their merged entire image separately. The overall aesthetics and layout of our merged image are superior.

Evaluation metric. For the ablation studies, we report standard metrics, including FID [13], PSNR, and SSIM. To assess the quality of the Anonymous Region Transfomer, the FID is computed by comparing the predicted merged images to the ground truth merged images, denoted as $\text{FID}_{\text{merged}}$. The PSNR and SSIM are calculated by comparing the merged image with the predicted reference composed image. To assess the quality of the multi-layer transparency image autoencoder, we report the PSNR for the RGB channels and the alpha channel separately, *i.e.*, $\text{PSNR}^{\text{layer}}_{\text{RGB}}$ and $\text{PSNR}^{\text{layer}}_{\text{alpha}}$, by comparing the reconstructed transparent layers with the ground-truth transparent layers. For the system-level comparisons, we conduct a user study to assess the quality of the composed image and transparent layers from four aspects: visual aesthetics, prompt adherence, typography, and inter-layer harmonization.

For fair comparisons, we use the layout predicted by our anonymous region layout planner model for the system-level comparison experiments, whereas the human-provided anonymous layout is instead used by default for all ablation studies, unless otherwise specified.

4.1. System-level Comparisons

We report the system-level comparisons with state-of-the-art methods in photorealistic image space (evaluated on PHOTO-MULTI-LAYER-BENCH) and graphic design space (evaluated on DESIGN-MULTI-LAYER-BENCH).

Comparison to LayerDiffuse. We first compare our approach with the latest multi-layer generation method, LayerDiffuse [83], in the multi-layer real image generation benchmark, *i.e.*, PHOTO-MULTI-LAYER-BENCH. We conduct a user study involving 30 participants with diverse backgrounds in AI, graphic design, art, and marketing, each evaluating 50 pairs of multi-layer transparent images generated by our ART and LayerDiffuse across three aspects: harmonization, aesthetics, and prompt following. The results of the user study are illustrated in Figure 3, showing our approach outperforms LayerDiffuse in all dimensions.

Comparison to COLE. We further conduct a user study to compare our approach with the multi-layer graphic design image generation method COLE [37]. We also ask the same 30 participants to evaluate the organization of the elements (layout), the visual appeal (aesthetics), the correctness of the text (typography), and the coherence and quality of each layer (harmonization), with each user evaluating 50 image pairs. The results in Figure 3 reveal that our approach achieves significantly better multi-layer image generation results in various aspects, except for typography, as the text in COLE is rendered with typography render.

More results. We present more multi-layer image generation in Figure 5 (up to 30 layers), as well as qualitative comparison results with COLE and LayerDiffuse in Figure 6.

4.2. Ablation Study and Analysis

Anonymous Region Layout is Sufficient. We first address the key question of whether region-specific prompts are necessary for multi-layer image generation tasks by comparing the conventional semantic layout and our anonymous region layout. For the semantic layout, we generate region-specific prompts for each layer using the LLaVA 1.6 model [47] and ensure that the visual tokens of each region mainly attend to their respective regional prompts. To ensure a fair comparison, we utilize the ground-truth layout provided by our DESIGN-MULTI-LAYER-BENCH while maintaining consistency across all other experimental settings, differing only in the use of region-specific prompts.

method	FID_{merged}	PSNR	SSIM	Harmonization Score (GPT-4o)
Semantic Layout	17.51	17.71	0.8443	3.67
Anonymous Region Layout	17.79	22.90	0.9021	3.92

Table 2. Anonymous Region Layout *vs*. Semantic Layout.

composed image pred.	FID_{merged}	PSNR	SSIM	Inference speed (s)
✗	20.44	-	-	19.20
✓	17.79	22.90	0.9021	26.62

Table 3. Composed image prediction improves the image quality.

attention type	FID_{merged}	PSNR	SSIM
Full Att.	41.35	16.87	0.7738
Spatial Att. + Temporal Att.	167.99	16.92	0.7985
Regional Full Att.	17.79	22.90	0.9021

Table 4. Full Att. *vs*. Spatial Att. + Temporal Att. *vs*. *Regional Full Att.*

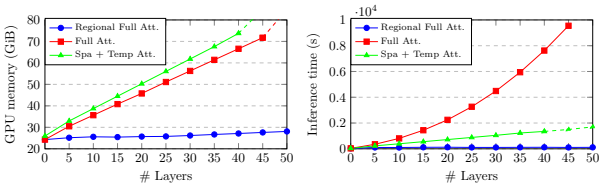


Figure 7. Illustrating the efficiency comparison of three different attention mechanism design: our Regional Full Attention (marked as Regional Full Att.), Full Attention (marked as Full Att.) and Spatial + Temporal Attention (marked as Spa + Temp Att.). The GPU memory consumption and inference time are evaluated and averaged over 100 samples at a resolution of 1024×1024, for each given number of layers. Some data points are represented with dashed lines or are not shown due to the OOM issue.

Table 2 provides a detailed comparison of the results. We find that the FID_{merged} scores for both methods are comparable, while the PSNR score for the anonymous region layout is significantly higher. This suggests superior layer coherence and global harmonization in our approach. Additionally, we employ GPT-4o to evaluate both methods in terms of global harmonization, arriving at the consistent conclusion that our approach yields better layer coherence. One potential reason for the lower coherence in the semantic layout approach is the conflict between local region-specific prompts and global visual tokens. We provide a deeper analysis of these conflicts in the supplementary material.

In addition, we present a statistical analysis comparing the inferred label assignments for the anonymous regions generated by our ART model with the human-annotated region-wise prompts. Our findings reveal that over 80% of the inferred labels align with the human annotations, suggesting that the generative models have acquired prior knowledge akin to Schema Theory. Additional details can be found in the supplementary material.

Benefits of Predicting the Reference Composed Image. We introduced an additional prediction of the reference composed image for two main reasons. First, it improves coherence across multiple image layers by facilitating bidirectional information propagation between the composed image and each transparent layer. Second, it provides a mechanism to evaluate the quality and consistency of the predicted transparent layers by calculating the PSNR and SSIM scores between the reference image and the layer-merged image on the validation set. As illustrated in Table 3, predicting the composed image as a reference significantly enhances image quality, indicated by the improved FID_{merged} score, despite a minor increase in inference time.

PE method	FID_{merged}	PSNR	SSIM
2D-RoPE	124.3	11.99	0.4265
2D-RoPE + LayerPE	20.66	23.23	0.9101
3D-RoPE	17.79	22.90	0.9021

Table 5. Different position embedding scheme in diffusion transformer.

# samples	FID_{merged}	PSNR	SSIM
80	30.38	23.18	0.8893
800	18.89	20.45	0.8609
8k	18.06	22.43	0.8882
80k	18.04	23.13	0.9081
800k	17.79	22.90	0.9021

Table 6. Increasing the dataset scale improves performance.

Method	FID_{merged}	PSNR	SSIM	Inference speed (s)
GPT-4o	20.72	22.80	0.9078	-
LayoutGPT [17]	20.92	23.18	0.9113	-
Semantic Layout Planner	21.45	17.69	0.8382	19.19
Semantic Layout Planner†	20.63	22.90	0.9092	19.19
Anonymous Region Layout Planner	19.90	22.70	0.9038	5.68

Table 7. Anonymous region layout planner *vs*. semantic layout planner and other planner alternatives. † means that we remove the predicted region-specific prompts and only use the predicted bounding boxes.

Regional Full Attention *vs*. Full Attention *vs*. Spatial + Temporal Attention. A key design element of our approach is the ceiling-aligned tight crop for each transparent layer, which removes most transparent pixels and compels the diffusion model to focus on the smallest rectangle encapsulating the non-transparent foreground regions. We refer to this as the Regional Full Attention scheme. This design is crucial for improving efficiency and explicitly constrains layer predictions to align with the positions specified by the anonymous region layout. We also evaluate two additional baselines: the Full Attention scheme, which does not apply regional cropping, and the Spatial Attention + Temporal Attention scheme, which introduces temporal attention to facilitate interactions across different layers, similar to architectural designs in video generation [5, 22]. Detailed comparison results are presented in Table 4, where our method demonstrates superior FID_{merged}, PSNR and SSIM scores. The primary factor behind our improved performance is the use of the anonymous region layout.

Moreover, Figure 7 shows that our method maintains nearly constant computational costs when processing between 10 and 50 layers, whereas the Full Attention scheme, lacking regional cropping, exhibits quadratic growth in both GPU memory usage and inference time consumption.

Layer-aware Position Encoding is Critical. Encoding positional information is essential for the diffusion transformer (especially, Anonymous Region Transformer) to distinguish visual tokens from different transparent layers. Our empirical analysis shows that incorporating layer position information is crucial, with the proposed 3D-RoPE

PE method	$PSNR^{layer}_{rgb}$	$PSNR^{layer}_{alpha}$	PSNR	FID_{merged}
2D-AbsPE	26.91	18.42	26.06	17.04
2D-AbsPE + LayerPE	26.98	18.76	26.11	16.24
2D-RoPE	34.05	23.08	30.09	3.16
2D-RoPE + LayerPE	34.46	23.31	30.13	3.10
3D-RoPE	34.89	23.85	30.48	2.84

Table 8. Position embedding scheme in multi-layer decoder.

composed image	bg image	$PSNR^{layer}_{rgb}$	$PSNR^{layer}_{alpha}$	PSNR	FID_{merged}
✗	✗	33.25	22.82	29.35	3.76
✓	✗	33.25	21.95	29.39	3.53
✗	✓	34.37	23.39	30.20	3.06
✓	✓	34.89	23.85	30.48	2.84

Table 9. Condition choice for the multi-layer decoder.

Method	Multi layer	$PSNR^{layer}_{rgb}$	$PSNR^{layer}_{alpha}$	PSNR	FID_{merged}
LayerDiffuse [83]	✗	20.94	18.48	26.51	4.27
Flux-RGBA decoder	✗	30.25	20.11	27.74	5.23
Ours	✓	34.89	23.85	30.48	2.84

Table 10. Comparison with existing transparency decoder.

scheme outperforming the absolute layer position encoding method. Full comparison results are presented in Table 5.

Multi-layer Data Scaling Enhances Performance. Table 6 shows results from training with varying dataset scales. Our findings clearly demonstrate that performance improves with larger dataset sizes. Notably, ART achieves impressive results with just 8K training samples, demonstrating the data efficiency of our approach.

Anonymous Region Layout Planner v.s. Semantic Layout Planner. We fine-tune both an anonymous layout planner and a semantic layout planner using data sampled from the 800K training dataset and evaluate their performance by integrating them with our ART model. Additionally, we include two strong baselines, GPT-4o and LayoutGPT [17], which support transforming the global prompt into a usable layout. Detailed results are presented in Table 7. Our Anonymous Region Layout Planner not only achieves a better FID_{merged} score but also operates more than 3× faster than the Semantic Layout Planner. Interestingly, removing the region-specific prompts of the semantic layout planner can enhance overall performance by avoiding conflicts among region-wise prompts, especially regarding layer coherence, as reflected by the higher PSNR scores.

RoPE is Critical for Multi-layer Decoder Quality. Table 8 summarizes the results of the comparison experiments involving different position embedding schemes for the multi-layer transparency decoder. The original ViT pretrained on the ImageNet classification task employs absolute position encoding, which is inadequate for capturing positional information across a variable number of transparent layers. We find that simply adding an additional set of layer-wise absolute position embeddings provides minimal improvement; however, replacing the absolute position encoding with the RoPE scheme significantly enhances decoding quality. We observe that the 3D-RoPE scheme achieves the best FID_{merged} score, which aligns with our findings regarding the choice of position encoding scheme for the latent features sent into MMDiT. Consequently, we adopt the 3D-RoPE scheme as the default setting.

(a) Ground-truth (b) LayerDiffuse (c) Flux-RGBA (d) Ours

Figure 8. Comparison with existing transparency decoder.

Composed Image as Condition. Although we only need to decode the transparency for all the foreground transparent layers, we empirically find that sending both the merged entire image and the background image as additional conditions, along with applying supervision on them, leads to even better performance, as shown in Table 9. We hypothesize that the information from the merged and background images is beneficial for the transparency layers to interact more effectively, ensuring a more coherent final composed image with these transparent foreground layers.

Comparison with Previous Transparency Decoder. We compare our multi-layer transparency decoder with the previous transparency decoder and two strong baselines designed for single-layer transparency decoding, as shown in Table 10. We utilize the officially released weights of the transparency decoder proposed by LayerDiffuse [83]. For the Flux-RGBA decoder, we modify the output projection to support an additional alpha layer prediction and fine-tune the decoder using our dataset. Our design achieves the best FID_{merged} score as shown in Table 10. The qualitative comparison results are also presented in Figure 8.

5. Conclusion

In this paper, we introduce the Anonymous Region Transformer, a novel approach for generating multi-layer transparent images from an anonymous region layout. Our results and analysis reveal that our anonymous layout is sufficient for the multi-layer transparent image generation task. Our method offers several key advantages over traditional semantic layout methods, including better coherence across layers and more scalable annotation. Furthermore, our method enables the efficient generation of images with numerous distinct transparent layers, reducing computational costs and generalizing to various distinct anonymous region layouts. However, our approach does have certain limitations, including repeated layer generation and combined layer generation. The generalizability of this capability across all potential layouts requires further exploration. Future work should focus on enhancing the model's ability to autonomously identify semantic labels and improving the quality and flexibility of the generated images. Despite these challenges, our approach shows promising potential for graphic design creation and digital art.

Acknowledgements

The work is supported in part by the National Natural Science Foundation of China under Grants 42327901, 62321005 and 62372015, as well as Shenzhen Key Laboratory of next generation interactive media innovative technology under Grants No. ZDSYS20210623092001004.

References

[1] Robert Axelrod. Schema theory: An information processing model of perception and cognition. *American political science review*, 67(4):1248–1266, 1973. 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 3

[3] Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995. 1, 2

[4] 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*, 2023. 1, 3

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 7

[6] Cameron Braunstein, Hevra Petekkaya, Jan Eric Lenssen, Mariya Toneva, and Eddy Ilg. Slayr: Scene layout generation with rectified flow. *arXiv preprint arXiv:2412.05003*, 2024. 3

[7] Ryan D Burgert, Brian L Price, Jason Kuen, Yijun Li, and Michael S Ryoo. MAGICK: A large-scale captioned dataset from matting generated images using chroma keying. In *CVPR*, 2024. 5

[8] Shang Chai, Liansheng Zhuang, and Fengying Yan. LayoutDM: Transformer-based diffusion model for layout generation. In *CVPR*, 2023. 3

[9] Jian Chen, Ruiyi Zhang, Yufan Zhou, Jennifer Healey, Jiuxiang Gu, Zhiqiang Xu, and Changyou Chen. TextLap: Customizing language models for text-to-layout planning. In *EMNLP Findings*, 2024. 3

[10] Chin-Yi Cheng, Forrest Huang, Gang Li, and Yang Li. Play: Parametrically conditioned layout generation using latent diffusion. In *ICML*, 2023. 3

[11] Yutao Cheng, Zhao Zhang, Maoke Yang, Hui Nie, Chunyuan Li, Xinglong Wu, and Jie Shao. Graphic design with large multimodal model. *arXiv:2404.14368*, 2024. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 5

[13] DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 1982. 6

[14] Chaoqun Du, Yulin Wang, Jiayi Guo, Yizeng Han, Jie Zhou, and Gao Huang. Unitta: Unified benchmark and versatile framework towards realistic test-time adaptation. *arXiv:2407.20080*, 2024. 3

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 5

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 3, 4

[17] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional visual planning and generation with large language models. In *NeurIPS*, 2024. 3, 7, 8

[18] Alessandro Fontanella, Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, and Sarah Parisot. Generating compositional scenes via text-to-image rgba instance generation. *arXiv preprint arXiv:2411.10913*, 2024.

[19] Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. LayoutFlow: Flow matching for layout generation. In *ECCV*, 2024. 3

[20] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023. 3

[21] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *CVPR*, 2024. 1

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 7

[23] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE TPAMI*, 2021. 3

[24] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Yitian Zhang, and Haojun Jiang. Spatially adaptive feature refinement for efficient inference. *IEEE TIP*, 2021. 3

[25] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfen Cao, Wenhui Huang, Chao Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *ECCV*, 2022. 3

[26] Yizeng Han, Zhihang Yuan, Yifan Pu, Chenhao Xue, Shiji Song, Guangyu Sun, and Gao Huang. Latency-aware spatial-wise dynamic networks. In *NeurIPS*, 2022. 3

[27] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *ICCV*, 2023. 3

[28] Yizeng Han, Zeyu Liu, Zhihang Yuan, Yifan Pu, Chaofei Wang, Shiji Song, and Gao Huang. Latency-aware unified

dynamic networks for efficient image recognition. *IEEE TPAMI*, 2024. 3

[29] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. Neural speed reading with structural-jump-lstm. In *ICLR*, 2019. 3

[30] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens Van Der Maaten, and Kilian Q Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018. 3

[31] Gao Huang, Yulin Wang, Kangchen Lv, Haojun Jiang, Wenhui Huang, Pengfei Qi, and Shiji Song. Glance and focus networks for dynamic visual recognition. *IEEE TPAMI*, 2022. 3

[32] 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 *ECCV*, 2024. 2, 3, 5

[33] Mude Hui, Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, Yuwang Wang, and Yan Lu. Unifying layout generation with a decoupled diffusion model. In *CVPR*, 2023. 3

[34] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. LayoutDM: Discrete diffusion model for controllable layout generation. In *CVPR*, 2023.

[35] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Towards flexible multi-modal document models. In *CVPR*, 2023. 3

[36] Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. OpenCOLE: Towards reproducible automatic graphic design generation. In *CVPR Workshops*, 2024. 2, 3, 5

[37] Peidong Jia, Chenxuan Li, Zeyu Liu, Yichao Shen, Xingru Chen, Yuhui Yuan, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, et al. COLE: A hierarchical generation framework for graphic design. *arXiv:2311.16974*, 2023. 2, 3, 5, 6

[38] Zhaoyun Jiang, Shizhao Sun, Jihua Zhu, Jian-Guang Lou, and Dongmei Zhang. Coarse-to-fine generative modeling for graphic layouts. In *AAAI*, 2022. 3

[39] Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. LayoutFormer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *CVPR*, 2023. 3, 5

[40] Immanuel Kant, John Miller Dow Meiklejohn, Thomas Kingsmill Abbott, and James Creed Meredith. *Critique of pure reason*. JM Dent London, 1934. 2

[41] Kotaro Kikuchi, Naoto Inoue, Mayu Otani, Edgar Simo-Serra, and Kota Yamaguchi. Multimodal markup document models for graphic design completion. *arXiv:2409.19051*, 2024. 3

[42] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 3

[43] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. BLT: Bidirectional layout transformer for controllable layout generation. In *ECCV*, 2022. 3, 5

[44] Black Forest Labs. Flux.1 model family, 2024. 3, 4, 5

[45] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*, 2023. 1, 3

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 5

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 6

[48] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv:2306.06101*, 2023. 5

[49] Zanlin Ni, Yulin Wang, Renping Zhou, Yizeng Han, Jiayi Guo, Zhiyuan Liu, Yuan Yao, and Gao Huang. Enat: Rethinking spatial-temporal interactions in token-based image synthesis. In *NeurIPS*, 2024. 3

[50] Zanlin Ni, Yulin Wang, Renping Zhou, Rui Lu, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Yuan Yao, and Gao Huang. Adanat: Exploring adaptive policy for token-based image generation. In *ECCV*, 2024. 3

[51] Xuran Pan, Tianzhu Ye, Zhuofan Xia, Shiji Song, and Gao Huang. Slide-transformer: Hierarchical vision transformer with local self-attention. In *CVPR*, 2023. 3

[52] 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. In *ICLR*, 2024. 1, 3

[53] Yifan Pu, Yizeng Han, Yulin Wang, Junlan Feng, Chao Deng, and Gao Huang. Fine-grained recognition with learnable semantic data augmentation. *IEEE TIP*, 2023. 3

[54] Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wang, Shiji Song, and Gao Huang. Adaptive rotated convolution for rotated object detection. In *ICCV*, 2023. 3

[55] Yifan Pu, Weicong Liang, Yiduo Hao, Yuhui Yuan, Yukang Yang, Chao Zhang, Han Hu, and Gao Huang. Rank-detr for high quality object detection. In *NeurIPS*, 2024. 3

[56] Yifan Pu, Zhuofan Xia, Jiayi Guo, Dongchen Han, Qixiu Li, Duo Li, Yuhui Yuan, Ji Li, Yizeng Han, Shiji Song, et al. Efficient diffusion transformer with step-wise dynamic attention mediators. In *ECCV*, 2024. 3

[57] David E Rumelhart. Schemata: The building blocks of cognition. In *Theoretical issues in reading comprehension*, pages 33–58. Routledge, 2017. 1, 2

[58] 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. In *NeurIPS*, 2022. 1, 3

[59] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. In *WACV*, 2024. 3

[60] Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, and Yasutaka Furukawa. Vi-

sual Layout Composer: Image-vector dual diffusion model for design layout generation. In *CVPR*, 2024. 3

[61] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 5

[62] Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. LayoutNUWA: Revealing the hidden layout expertise of large language models. In *ICLR*, 2023. 3

[63] Omost Team. Omost github page, 2024. 3

[64] 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*, 2024. 5

[65] Jiangshan Wang, Yifan Pu, Yizeng Han, Jiayi Guo, Yiru Wang, Xiu Li, and Gao Huang. Gra: Detecting oriented objects through group-wise rotating and attention. In *ECCV*, 2024. 3

[66] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *CVPR*, 2021. 3

[67] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. InstanceDiffusion: Instance-level control for image generation. In *CVPR*, 2024. 3

[68] X. Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv:2406.07209*, 2024. 3

[69] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *ICCV*, 2021. 3

[70] Yulin Wang, Yizeng Han, Chaofei Wang, Shiji Song, Qi Tian, and Gao Huang. Computation-efficient deep learning for computer vision: A survey. *Cybernetics and Intelligence*, 2023. 3

[71] Yilin Wang, Zeyuan Chen, Liangjun Zhong, Zheng Ding, Zhizhou Sha, and Zhuowen Tu. Dolfin: Diffusion layout transformers without autoencoder. In *ECCV*, 2024. 3

[72] Yulin Wang, Haoji Zhang, Yang Yue, Shiji Song, Chao Deng, Junlan Feng, and Gao Huang. Uni-adafocus: Spatial-temporal dynamic computation for video recognition. *IEEE TPAMI*, 2024. 3

[73] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and CL Chen. Desigen: A pipeline for controllable design template generation. In *CVPR*, 2024. 3

[74] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 3

[75] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023. 3

[76] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024. 3

[77] Kota Yamaguchi. CanvasVAE: Learning to generate vector graphic documents. In *ICCV*, 2021. 3, 5

[78] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *ICML*, 2024. 1, 3

[79] Qisen Yang, Shenzhi Wang, Qihang Zhang, Gao Huang, and Shiji Song. Hundreds guide millions: Adaptive offline reinforcement learning with expert guidance. *IEEE TNNLS*, 2023. 3

[80] Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. PosterLLaVa: Constructing a unified multi-modal layout generator with LLM. *arXiv:2406.02884*, 2024. 3

[81] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. ReCo: Region-controlled text-to-image generation. In *CVPR*, 2023. 1

[82] Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. In *NeurIPS*, 2024. 3

[83] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics*, 2024. 2, 3, 4, 5, 6, 8

[84] Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2Layer: Layered image generation using latent diffusion model. *arXiv:2307.09781*, 2023. 2, 3, 5

[85] Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. IterComp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. *arXiv:2410.07171*, 2024. 3

[86] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. In *ICLR*, 2025. 3