Animated-ART: Multi-Layer Transparent Video Generation

Ziqiang Li[1,4*] Yunnan Wang[1,4*] Dong Chen[2] Yue Dong[2]
Ji Li[2] Yuhui Yuan[3] Xin Jin[4,5] ✉
[1] Shanghai Jiao Tong University [2] Microsoft Research Asia [3] Canva Research
[4] Eastern Institute of Technology, Ningbo [5] Zhongguancun Academy

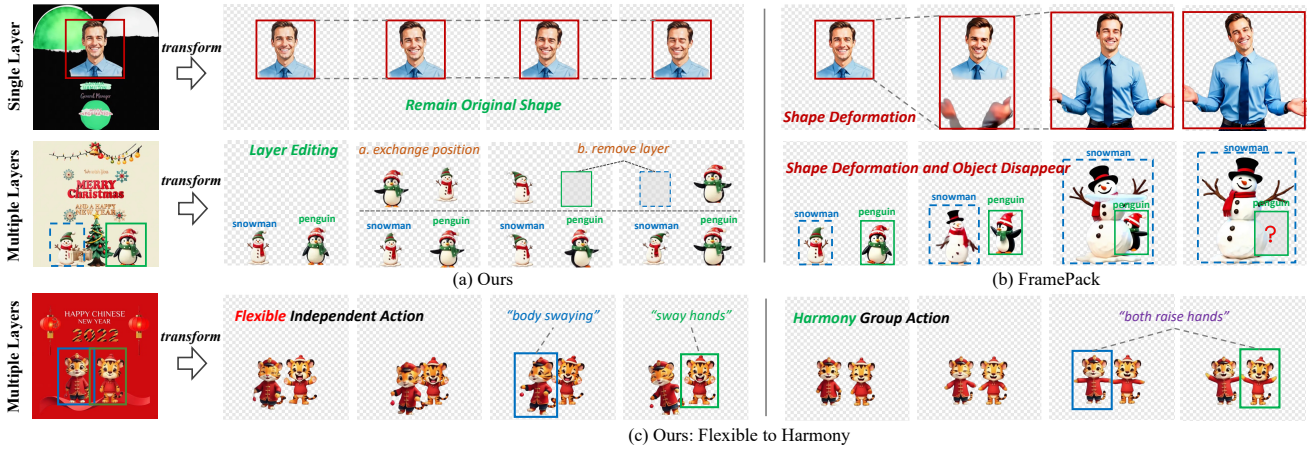


Figure 1. **Animated Generation Comparison with Dynamic Layer**: (a) Given static layers, our DMDL generates a high-fidelity transparent dynamic layer, suitable for flexible editing. (b) Using the same input, general I2V model (e.g., Framepack [39]) frequently causes artifacts like shape deformation and content degradation, and requires matting model for the RGBA format transform. (c) For multi-layer synthesis, DMDL supports two modes: flexible animation for individual layer control and coherent animation for harmony group action.

Abstract

Layered image design is fundamental to professional creative workflows, and the generation of layered images has attracted significant research interest. In this work, we extend layered image generation to the video domain, introducing the novel task of multi-layer transparent video generation—synthesizing multiple transparent dynamic layers that compose into coherent video sequences. We first construct the Transparent Dynamic Layer (TDL) dataset, specifically designed for training and evaluating models on animated RGBA layer sequences. Building on this dataset, we propose the Diffusion Multiple Dynamic Layers (DMDL) model, which generates variable-resolution transparent dynamic layers from static layer inputs, text prompts, and motion region layouts. DMDL supports both single-layer and multi-layer animated generation. Our approach makes two key technical contributions: (i) a latent multiple dynamic-layer diffusion model with static-to-dynamic generation capability, and (ii) a transparent dynamic-layer autoencoder. The diffusion model incorporates layer-aware spatial-temporal 4D-RoPE positional embeddings, enabling cross-layer interaction of visual tokens across layer, spatial, and temporal dimensions. Our autoencoder employs a specialized ViT-based decoder that leverages a two-stage, layout-conditional 3D-RoPE strategy to reconstruct transparent dynamic layers, effectively handling the varying levels of temporal upsampling inherent in the process. Extensive experiments validate our method's effectiveness, establishing a strong baseline for dynamic layer generation.

1. Introduction

Since the impressive success of diffusion models [10, 27–29], extensive exploration [12, 31, 41] of transparent static-layer image generation has prompted more controllable and harmonious compositions for digital art and advertising posters [22, 35]. While the effectiveness of layer image generation has been demonstrated in interactive synthesis for creators, these approaches remain fundamentally static

*Equal Contribution. ✉ Corresponding author.

in composition, limited to a single and flattened canvas. Imagine a more captivating visual effect, such as the magical newspapers from the world of Harry Potter, where figures within a photograph move and interact, transforming the image into a living, dynamic, and layered scene. This animated vision of layer element storytelling in a dynamic poster for advertising is more eye-catching than its static counterpart. Therefore, in this paper, we pursue effective solutions for static-to-dynamic layer generation, achieving more vision-attractive effects, as shown in Figure 1.

Transparent dynamic layers primarily represent object motion across the temporal dimension and can be defined as a unique form of video sequence. However, transparent dynamic layer generation presents challenges fundamentally distinct from those of conventional video synthesis, primarily due to two factors: 1) dynamic layer objects possess varying resolution proportions, and 2) the RGBA format of animated layers, in which the background is transparent, while the foreground remains opaque. In particular, typical image2video generation models synthesize holistic sequences conditioned on a fixed-resolution image containing many elements with natural proportions. Consequently, issues such as shape deformation arise during dynamic layer synthesis, as illustrated in Figure 1 (b), since objects in layers share small resolution proportions that are not compatible with the existing I2V mode. Moreover, existing video generation specializes in operating on standard RGB video streams, making it challenging to directly synthesize editable elements (RGBA format) within a scene (as depicted in Figure 1 (a)). Therefore, these intrinsic differences suggest that naively applying existing video generation models would be suboptimal. It is crucial to construct a specialized dataset and then develop a tailored generative model.

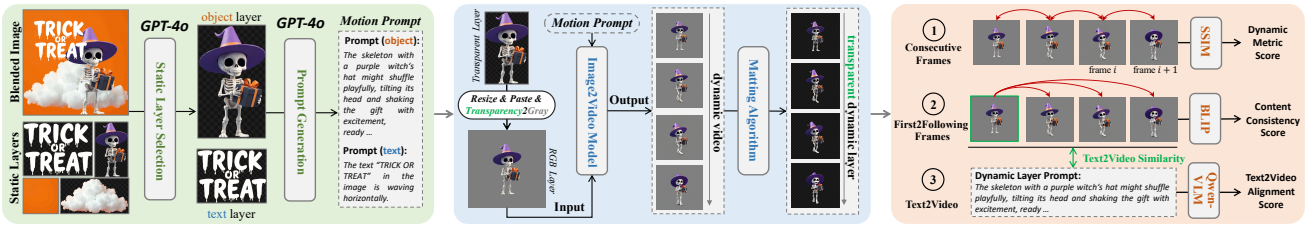
In this paper, as shown in Figure 2, we design a training-free transparent dynamic layers generation pipeline, constructing a dataset that comprises 100k transparent dynamic layers of varying resolutions with labeled text and motion region layout, categorized into two main types: salient objects motion and textual special effects. Moreover, we propose Diffusive Multiple Dynamic Layers (DMDL), a novel generative framework designed to support the simultaneous generation of multiple dynamic layers or a single dynamic layer with varying resolutions. Our framework comprises two key components: a latent multiple dynamic-layer diffusion model and a specialized transparent animated layer autoencoder. The diffusion model adopts a static-to-dynamic generation paradigm (akin to image-to-video synthesis) conditioned on static-layer latent representations, associated motion prompts, and motion-region layouts. To enable joint generation while preserving the uniqueness of each dynamic layer, we design a crucial layout conditional 4D-RoPE strategy, accommodating the extensional layer dimension. This mechanism provides cross-communication of visual token information between different layers in the shared layer-spatio-temporal embedding space, allowing more harmonious group action synthesis of multiple layers (as shown in Figure 1 (c)). For the final synthesis, our autoencoder is accompanied by a ViT-based transparent video decoder to reconstruct the transparent dynamic layers from their latent counterparts. A two-stage layout-conditional 3D RoPE mechanism is introduced to effectively capture spatio-temporal positional relationships among visual tokens for each temporal upscaling stage.

In summary, our contribution is three-fold: (i) We design a training-free transparent dynamic layer generation scheme, producing 100k transparent dynamic layers with various resolutions for the corresponding investigation. (ii) We propose a latent multiple dynamic-layer diffusion model featuring a novel layout conditional 4D-RoPE strategy, safeguarding cross-interaction of visual tokens across layer-spatial-temporal dimensions. (iii) We present a transparent dynamic layer autoencoder incorporating a custom ViT-based decoder. This decoder integrates a two-stage, layout-conditioned 3D RoPE mechanism that accommodates distinct phases of temporal upscaling.

2. Related Work

Transparent Layer Generation. Recent advances [12, 22, 31, 41] have been made in generating transparent image layers based on text-to-image technologies [7, 17, 21, 25], which are crucial for editable content creation. Existing works can be categorized by the number of layers generated. Single-layer generation methods, such as LayerDiffuse [38] and Text2Layer [41], focus on producing one high-quality foreground layer, sometimes conditioned on or generated alongside a background. Other methods like LayeringDiff [13] achieve this through a post-hoc separation of a generated composite image. Concurrently, multi-layer generation approaches like LayerDiff [12] and ART [22] have successfully synthesized multiple, coherent transparent layers from a single prompt in a "bottom-up" fashion, distinct from "top-down" image decomposition techniques [31]. Despite their significant contributions to compositional flexibility, these methods are fundamentally limited to generating static assets. The layers, while editable in space, are frozen in time. To make generated content significantly more captivating, motion is an essential ingredient. Therefore, instead of generating still images, we focus on the novel and challenging task of synthesizing transparent dynamic layers, aiming to imbue generative assets with life and motion, thereby making them more attractive.

Video Generation. Video generation approaches [2, 3, 8, 11, 18, 26, 34] based on the diffusion model have achieved great success in creating high-quality content, largely following text-to-video (T2V) and image-to-video (I2V) paradigms. And these generation models can be



Stage I: ***Layer Selection and Annotation*** Stage II: ***Transparent Dynamic Layer Generation*** Stage III: ***Dynamic Layer Filtering Metric***

Figure 2. **Data Construction Pipeline**. Stage I: MLLM selects interesting static image layers and produces reasonable motion prompts. Stage II: Dynamic layers with gray backgrounds are generated by the I2V model and following the matting operation to obtain the transparent counterparts. Stage III: SSIM (temporal dynamic score), BLIP-visual score encoder (content semantic consistency), and Qwen-VLM (text2video alignment score) are utilized to filter out bad cases.

mainly divided into two types: U-Net [24] based generation models and diffusion transformer (DiT) [20] based synthesis models. Models, such as SVD [2] and LVDM [3], extend the established U-Net architecture from image diffusion models [23] by inserting temporal convolution layers to model temporal dynamics. More recent approaches, employed by models like Latte [18], CogvideoX [37], and Hunyuan [14], adopt a DiT architecture since it has greater potential for scalability and capturing complex video dynamics compared to its U-Net-based counterparts. However, these methods are fundamentally designed for holistic scene generation, producing content that typically features multiple elements within a fixed-resolution canvas. Dynamic layers are isolated, transparent-subject-centric assets, each potentially having a unique, variable resolution. Based on these differences, we aim to propose a novel static-to-dynamic layer model to create transparent, variable-resolution animated layers.

3. Data Construction

Overview. We propose a training-free pipeline that generates animated, transparent dynamic layers, built on the existing multi-layer image dataset PrismLayers [5]. Here, each instance is composed of N_s static RGBA image layers, formulated as $I_s = \{(L_i^s, B_i^s)_{i=1}^{N_s}, C_s\}$, where $L_i^s \in R^{H_i \times W_i \times 4}$ is a RGBA layer and $B_i^s = [x_i^s, y_i^s, H_i, W_i]$ is the corresponding layout box, and C_s is a text description for the all layers. Here, we select N interesting static layers from instance I_s to produce the corresponding transparent dynamic layer L_i assigned with proper text motion prompts C_i and motion region layout B_i similar to B_i^s locating dispaly area, the constructed data is defined as $I_d = \{(L_i, B_i, C_i)_{i=1}^{N}\}$. Note that we also construct the harmony dynamic multi-layer data to investigate group animation generation, and the corresponding data is formulated as $I_d = \{(L_i, B_i)_{i=1}^{N}, C_g\}$, C_g is the group action text prompt for multiple layers. The constructed dynamic layers are categorized into two types: object-centric dynamic layer (e.g., characters, animals, flowers) and textual dynamic layer, and they are diverse in resolution, reflecting the varied nature of real-world visual assets. Dynamic layer construction follows a meticulous three-stage pipeline, as illustrated in Figure 2, and we will detail it below.

Static Layer Selection and Annotation. Since a dynamic layer can be conceptualized as an animated representation of a static one, the first step is to select suitable source static layers. Here, we employ Multimodal Large Language Models (MLLM) [19] to perform suitable layer selection, given the static RGBA layers and their corresponding spatial information (layout box). The selection process is guided by a carefully designed prompt that instructs the MLLM. Upon selection, the layers are classified as either text layers or subject layers. For the object layer, MLLM is utilized to generate an open-vocabulary prompt C_i describing a plausible and engaging motion based on the layer's visual content, and MLLM produces the group motion prompt C_g by understanding the merged image with multiple layers. For text layers, owing to their distinct structural properties (e.g., text cannot "smile"), we restrict motion effects to a predefined set of closed-set animations.

Transparent Dynamic Layer Synthesis. Here, we aim to apply the advanced I2V generation model [39] to synthesize the dynamic effects based on the selected static layers and corresponding motion prompts. Firstly, we standardize the input by proportionally scaling each variable-resolution static layer and placing it onto a fixed-size transparent canvas. Then, the RGBA layer is converted to RGB by compositing it onto a neutral gray background to fit the input format of the I2V model. After that, the transferred RGB layer, paired with its motion prompt, is input to the I2V model to generate an RGB video sequence depicting the object executing the prompted motion against the gray background. Finally, this matting model [4] is leveraged to precisely separate the animated foreground object from the gray background for generating final transparent dynamic layers, and the dynamic region box B_i can be obtained by discerning the alpha channel of transparent dynamic layers, like Art. For group dynamic generation, similar to the above steps, we input the merged image of well-positioned multiple object layers into the I2V model to produce a har-

mony group-action animation of multiple layers. After the matting process, we split the generated multi-object group animation layer into unique object dynamic layers by using pixel-level K-means [9] clustering algorithm operated on alpha channel.

Dynamic Layer Filtering. We introduce a sample filtering metric to obtain high-quality and reliable transparent dynamic layers in three respects: 1) temporal dynamic, 2) semantic consistency, and 3) text2video alignment. *Temporal dynamic*: Structural Similarity (SSIM) [36] is used to measure the temporal consistency among consecutive frames at the pixel level, discarding samples exhibiting negligible frame-to-frame changes (lack meaningful dynamics) and removing instances indicating excessively large changes (temporal inconsistency or flickering artifacts). *Semantic consistency*: We use BLIP-visual encoder [15] to assess semantic consistency scores between the first frame and subsequent frames, discarding instances with low scores to preserve high semantic fidelity to the given static layer. *Text2video alignment*: Qwen-VL2.5 [1] is leveraged to score the alignment between generated dynamic layer and its motion prompt, and samples with alignment scores below a predefined threshold are filtered out.

4. Approach

Distinct from prevailing I2V models [14, 37, 39] that produce a single fixed-resolution video, we propose Diffusive Multiple Dynamic Layers (DMDL), allowing controllable generation of multiple or single transparent dynamic layers with various resolutions, conditioned by static layers, text prompts, and motion region layout provided by LLM or user. As shown in Fig. 3, DMDL consists of two core components: 1) Transparent Dynamic Layer Autoencoder for encoding and decoding transparent dynamic layers and the corresponding latent representation; 2) latent Multiple Dynamic-Layer Diffusion model that jointly generates latent representations for multiple animated layers located within the motion region layout. Additionally, we design a Dynamic Region Planner with LLM that can automatically predict the animated region for a dynamic layer given its static region layout and motion prompt.

4.1. Transparent Dynamic-Layer Autoencoder

Transparent Dynamic Layer Encoder. Following the success of ART [22], which represents the alpha channel of RGBA images as a gray background for encoding RGBA layers, we transfer the transparent dynamic layer to an RGB animated image sequence $\bar{L}_i$ with a gray background. Then, we employ the video VAE encoder E_v [14] to extract a compact spatio-temporal latent representation of the dynamic layer. Concretely, for the gray-background dynamic layer $\bar{L}_i$, E_v compresses the spatial dimensions by a factor of 8×8 and the temporal dimension to $\frac{(T-1)}{4}+1$, while maintaining a feature channel dimension of 16. Following the encoding, we crop the resulting latent feature according to the motion region B_i and flatten it along the spatio-temporal dimensions to obtain the final latent representation:

$$\mathbf{z}_i = \mathrm{Flatten}(\mathrm{Crop}(\mathrm{E}_v(\bar{L}_i), B_i)), \tag{1}$$

where $\mathbf{z}_i \in R^{K\times 16}$, K is the number of tokens after flattening. The latent code $\mathbf{z}_i$ is then decoded as the transparent dynamic layer.

Transparent Dynamic Layer Decoder. Since the success of ViT-based architecture [6] in compatible with diverse token lengths, we design our transparent dynamic layer decoder based on ViT to provide the flexibility to accommodate different spatial resolutions inherent in our dynamic layers. To upscale the temporal dimension of $\mathbf{z}_i$ compressed in the encoding stage, we design two temporal scale blocks TS to recover it to the original temporal length of $\bar{L}_i$. The two TS are inserted in the start and middle parts of ViT. Moreover, we use a dynamic two-stage 3D-RoPE [30] strategy to obtain the temporal-spatial positional information $\mathbf{P}_1$ and $\mathbf{P}_2$ during the decoding process. This decode-forward process can be formulated as:

$$\hat{\mathbf{z}}_i = \mathrm{ViT}_1(\mathrm{TS}(\mathrm{Linear}_{\mathrm{in}}(\mathbf{z}_i)), \mathbf{P}_1), \tag{2}$$

$$\bar{\mathbf{z}}_i = \mathrm{Linear}_{out}(\mathrm{ViT}_2(\mathrm{TS}(\hat{\mathbf{z}}_i), \mathbf{P}_2)), \tag{3}$$

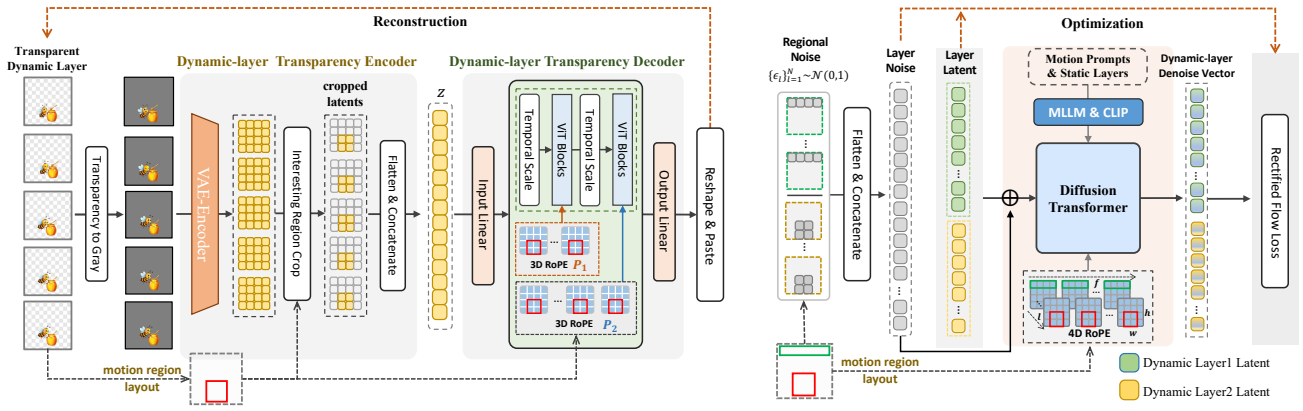
where $\mathrm{ViT}_*(\cdot)$ means cascaded ViT blocks, Linear_{in} is a linear layer to map the channel dimension of $\mathbf{z}_i$ from 16 to the ViT's internal token dimension (e.g., 768). $\hat{\mathbf{z}}_i$ is the middle-state latent representation, and $\bar{\mathbf{z}}_i$ means the final latent that is projected by Linear_{out}, transferring dimension 768 to 256, and each token of $\bar{\mathbf{z}}_i$ will be reshaped to $8\times8\times4$ RGBA patch size, recovering to transparent dynamic layer.

Since TS will upscale the latent temporal dimension, and each dynamic layer owns a unique spatial resolution, our dynamic two-stage 3D-RoPE strategy can produce corresponding positional embedding $\mathbf{P}_1$ and $\mathbf{P}_2$ changing following different instances, and the process progress of them can be formulated as:

$$\mathbf{P}_* = \mathrm{PE}(t_*, H, W, B_i), \tag{4}$$

where H, W are default spatial resolution, $t_* \in \{t_1, t_2\}$ are latent temporal lengths corresponding to the two stages, and PE is the positional embedding function. With the given layout B_i and temporal size t_*, PE can locate precise position and then produce positional embedding for each dynamic layer, safeguarding the decoding process.

During training, we use L_1 regularization loss as the reconstruction constraint to enforce pixel-level accuracy, the parameters of the pretrained encoder E_v are frozen, and only the decoder is optimized. We discuss the different structure design of our decoder in the experiment. The development of this RGBA dynamic layer decoder significantly improves generation efficiency, as it bypasses the need for a two-step process of decoding an RGB video and



(a) Transparent Dynamic Layer Autoencoder (b) Latent Multiple Dynamic Layer Diffusion Model

Figure 3. (a) Transparent dynamic layer autoencoder first encodes the transparency-to-gray-background layer as latent, then the ViT-based decoder reconstructs the motion-region-cropped latent to its original state. (b) Latent multiple dynamic layer diffusion model jointly denoises the multi-layer noisy latents owning distinct resolutions located with motion layout.

then applying a separate matting model.

4.2. Latent Multiple Dynamic-Layer Diffusion

Different from the conventional Image-to-Video (I2V) paradigm, which typically generates a single video from a source image within fixed resolution, we aim to introduce a method for the simultaneous synthesis of multiple dynamic layers from their corresponding static counterparts. Based on motion region layout, this new framework supports two distinct modes of animation: 1) individual layers can be jointly transformed into dynamic sequences guided by their own specific motion prompts, and 2) the harmony group-action animation of multiple layers can be generated under the guidance of a shared dynamic prompt. As depicted in Fig. 3 (b), we design a Multiple Dynamic-Layer Diffusion model with parameters ϕ based on HunyuanI2V [14] to build such a diffusion generation process, and the likelihood formula of this generation paradigm is expressed as:

$$\begin{aligned}\log P_\phi(Z \mid L^s, B, C) &= \log P_\phi(\mathbf{z}_1, \ldots, \mathbf{z}_N \mid L^s, B, C) \\ &= \sum_{i=1}^{N} \log P_\phi(\mathbf{z}_i \mid L^s, B, C),\end{aligned} \tag{5}$$

where $\mathbf{z}_i$ is the layer latent, $C = \text{Concat}[C_1, \ldots, C_N]$ to drive each layer to follow their independent prompts or $C = C_g$ to guide multiple layers leading to the group action, B is the motion region layout set of multiple dynamic layers, and L^s means the selected static layers.

Layout Conditional 4D RoPE. A key architectural challenge in our framework is encoding token positions across a 4D space (layer, temporal, spatial). While 3D Rotary Position Embedding (3D-RoPE) [30] effectively handles spatiotemporal coordinates in image2video models, it cannot account for the additional layer dimension integral to our approach. To resolve this, we design the Layout-Conditional 4D-RoPE, which extends RoPE's principles to four dimensions. This extension provides comprehensive positional information, allowing the model to accurately discern the location of any visual token within the complete layer-time-space continuum. Here, we introduce a global positional space $S \in R^{d_l \times d_f \times d_h \times d_w}$ to accommodate multiple dynamic layers with various spatial-temporal resolutions. The angular frequency vector $\boldsymbol{\theta}^k, k \in \{l, f, h, w\}$ for 4D can be calculated as:

$$\boldsymbol{\theta}^k = \{\theta_j^k = 10000^{-2(j-1)/d_k}, \quad j \in [1, \ldots, d_k/2]\}, \tag{6}$$

and $\Theta = [\boldsymbol{\theta}^l, \boldsymbol{\theta}^f, \boldsymbol{\theta}^h, \boldsymbol{\theta}^w]$ is the angular frequency vector for 4D-RoPE. With the provided motion layout B, we can obtain the u-th visual token's positional indices $\boldsymbol{P}_u = [p_u^l, p_u^f, p_u^h, p_u^w]$ in space S, p_u^* means position index of layer-spatial-temporal dimensions. The positional information $\boldsymbol{P}_u$ is then transferred to positional embedding with our 4D-RoPE angular frequency Θ, formulated as:

$$\boldsymbol{P}_u \cdot \Theta = [p_u^l \boldsymbol{\theta}^l, p_u^f \boldsymbol{\theta}^f, p_u^h \boldsymbol{\theta}^h, p_u^w \boldsymbol{\theta}^w]. \tag{7}$$

Given u-th query embed $\boldsymbol{Q}_u$ and v-th key embed $\boldsymbol{K}_v$, the rotary embedding operated in attention mechanism [33] for our 4D information is represented as:

$$\boldsymbol{A}(u, v) = Re[\boldsymbol{Q}_u (\boldsymbol{K}_v)^* e^{i(\boldsymbol{P}_u \cdot \Theta - \boldsymbol{P}_v \cdot \Theta)}], \tag{8}$$

where $\boldsymbol{A}(u, v)$ is component unit of attention matrix, $Re[\cdot]$ is real part extracting operation, $(\cdot)^*$ produces the conjugate complex number.

Inter-Layer Isolation. Additionally, we consider another strategy to accommodate multiple dynamic layers to the original 3D positional embedding space by revising the representation modeling among visual tokens from different layers. We find that dynamic layers, sharing overlapping spatial-temporal positions, lead to visual artifacts since the model is unable to distinguish their identity difference. Therefore, we introduce the Layer-Isolation Attention Mask

to explicitly prohibit the visual token information exchange among distinct layers, ensuring their representational integrity. This operation is formulated as:

$$\mathrm{ATT}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{M}_I) = \mathrm{SoftMax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}} + \boldsymbol{M}_I)\boldsymbol{V}, \quad (9)$$

where $\boldsymbol{M}_I$ is the independence binary mask matrix. For any query token $\boldsymbol{Q}_i$ belonging to layer L_i and key token $\boldsymbol{K}_j$ belonging to another layer L_j, the corresponding entry in $\boldsymbol{M}_I$ is set to $-\infty$. Although this mechanism guarantees visual token information isolation among layers and prevents visual artifacts in overlapping regions, it hinders the cooperation between layers to generate more harmony-consistent effects (shown in the experiment). Hence, we only integrate our proposed 4D-RoPE into the generation model.

Training loss. We follow the rectified flow loss [16] to optimize our model for multiple dynamic layer generation:

$$\mathcal{L} = \sum_{i=1}^{N} \left\| \boldsymbol{v}_\phi(\mathbf{z}_i^t, t, L^s, C) - (\mathbf{z}_i - \epsilon) \right\|_2^2, \quad (10)$$

where $\mathbf{z}_i^t$ is the rectified flow trajectory latent, $\boldsymbol{v}_\phi(\cdot)$ is predicted flow vector, $t \in [0, 1]$, and ϵ is Gaussian noise.

4.3. Dynamic Layer Region Planer

To automate the estimation of the motion region layout for dynamic layer generation, we propose a Dynamic Region Planner. This component utilizes GPT-4o [19] to predict a suitable dynamic region based on the static layer's initial layout and its intended motion. Specifically, for each layer, we input its static bounding box and the textual motion prompt into the LLM, and it reasons about the necessary spatial expansion and outputs the coordinates of the predicted motion region box.

5. Experiments

Implement Details. Our multiple dynamic-layer diffusion model is developed by enhancing the existing image-to-video model, Hunyuan-I2V [14]. During the training phase, the model was fine-tuned on four A100 GPUs with a global batch size of 4. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4}. The transparent dynamic layer decoder is constructed based on the Vision Transformer (ViT) architecture and is further augmented with two Temporal Scale (TS) blocks. The ViT backbone consists of 12 attention blocks, with a token channel dimension of 768. Each TS block adopts a residual structure, which includes a temporal upsampling layer, followed by a temporal convolution with a $3 \times 1 \times 1$ kernel and a spatial convolution with a $1 \times 3 \times 3$ kernel. Further architectural details are provided in the supplementary materials.

Experiment Setting. For our experiments, the TDL dataset is divided into two parts: a primary dataset for individual layer animation containing 100k training and 1k validation samples, where each layer is paired with a specific motion prompt, and another smaller dataset containing 4k group-action multi-layer samples under a shared motion prompt for training and about 800 samples for validation. The resolution of all layers is within 512×512 pixels. To assess the dynamic-layer generation quality of diffusion models, we employ standard metrics including Fréchet Video Distance (FVD) [32], Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM) [36], and LPIP [40], which are computed between the predicted animated layers and the validation samples. Concurrently, the reconstruction fidelity of our transparent autoencoder is measured by the PSNR and SSIM between its reconstructed output and the original input layers.

5.1. Comparisons

The comparison against other methods is confined to the generation of the object dynamic layer to ensure fairness.

Individual-Layer Animation. In this section, we conduct a comparative analysis of dynamic layer controlled generation against leading Image-to-Video (I2V) methods, including CogvideoX [37], Hunyuan-I2V [14], and FramePack [39]. To ensure compatibility with these baseline models, which typically process RGB inputs, we adapted our static RGBA image layers by converting their transparent channels into a gray background, thereby producing standard RGB-formatted input images. Performance was evaluated using a suite of standard metrics: FVD, PSNR, SSIM, and LPIP. The evaluation was structured across three distinct settings to comprehensively validate the proposed method, as shown in Table 1. First, at the layer-by-layer generation setting, our method achieves great performance for dynamic layer text-controlled synthesis. Furthermore, when challenged with the more complex task of multiple dynamic-layer generation simultaneously, our approach again demonstrates superior efficacy over competing methods. Finally, we also evaluated the I2V models at the poster level to control the generation of a specific layer within it, and this mode is more closely aligned with the conventional use cases of I2V models compared to using isolated layers on a gray background. We paste our dynamic layers to the poster for comparison. The quantitative evaluation across these metrics consistently shows that our method outperforms I2V methods. These comprehensive results consistently validate the superior performance and robustness of our proposed method.

Group Animation of Multi-layers. We once again benchmarked our method against the results generated by CogvideoX, Hunyuan-I2V, and FramePack, in the task setting of generating coherent group actions across multiple layers jointly, our model achieves optimal performance, as evidenced by the results in Table 2.

Computing Cost Comparison. We benchmarked the com-

Method	Layer by Layer				Joint Layers				Poster			
	FVD↓	PSNR↑	SSIM↑	LPIP↓	FVD↓	PSNR↑	SSIM↑	LPIP↓	FVD↓	PSNR↑	SSIM↑	LPIP↓
CogVideoX [37]	484.82	14.97	0.71	0.33	463.30	15.07	0.72	0.33	524.20	13.19	0.59	0.42
HunyuanI2V [14]	310.73	13.61	0.69	0.42	267.65	14.17	0.69	0.39	243.84	13.67	0.57	0.39
FramePack [39]	151.42	16.74	0.77	0.23	142.91	17.10	0.79	0.21	133.79	16.45	0.67	0.16
DMDL (**ours**)	**87.17**	**20.65**	**0.85**	**0.12**	**85.07**	**20.66**	**0.86**	**0.12**	**72.56**	**22.71**	**0.89**	**0.08**

Table 1. Quantitative results for dynamic layer generation on three settings.

Method	FVD ↓	PSNR ↑	SSIM ↑	LPIP ↓
CogVideoX [37]	767.87	15.2	0.73	0.31
HunyuanI2V [14]	465.17	13.40	0.67	0.44
FramePack [39]	126.50	18.43	0.83	0.13
DMDL (**ours**)	**63.25**	**23.79**	**0.91**	**0.06**

Table 2. Quantitative results of group action for multiple dynamic layer generation.

Method	Time (s)	Memory (MB)	Parameters (B)
CogvideoX [37]	88	28526	5.57
HunyuanI2V [14]	298	67558	12.82
Framepack [39]	264	69126	12.87
DMDL (**ours**)	40	54422	12.82

Table 3. Comparison about computing cost.

putational efficiency of our method in the multi-layer group-action mode on a 50-sample test set. As shown in Table 3, our approach achieves the fastest inference speed while remaining competitive in GPU memory usage and parameter count. Notably, it is about 6× faster than Hunyuan-I2V, demonstrating a clear efficiency advantage.

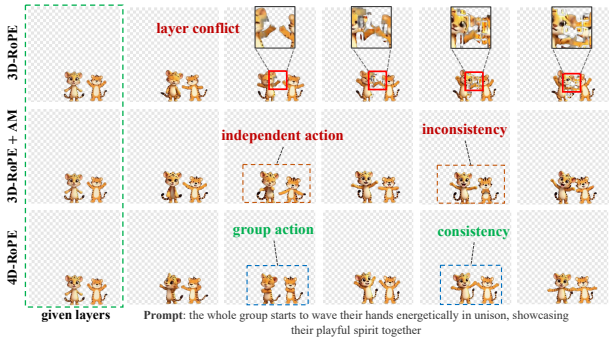


Figure 4. Visualization for different positional embeddings.

5.2. Ablation Study

Positional Embedding. We conducted an ablation study to compare the performance of different positional embedding methods in the multi-layer group action generation setting. As shown in Table 4, the results reveal that our proposed layout-based 4D-RoPE achieves superior performance compoared with 3D-RoPE and 3D-RoPE + attention mask (AM). As illustrated in Figure 4, the naive application of 3D-RoPE results in positional conflicts between layers, manifesting as conspicuous visual artifacts in overlapping regions. While augmenting 3D-RoPE with AM resolves these artifacts by preventing direct visual information exchange, this strict isolation consequently inhibits the model's ability to generate harmonious motions between layers. In contrast, our proposed layout-conditional 4D-RoPE effectively overcomes both of these limitations.

Components	FVD ↓	PSNR ↑	SSIM ↑	LPIP ↓
3D-RoPE	86.79	23.11	0.88	0.07
3D-RoPE+AM	68.22	23.62	0.90	0.06
4D-RoPE	63.25	23.79	0.91	0.06

Table 4. Evaluation of different positional embedding manners.

Components	PSNR ↑	SSIM ↑
ViT	26.73	0.996
ViT+EE	26.42	0.996
ViT+SE	28.49	0.997
ViT+SM	29.83	0.998

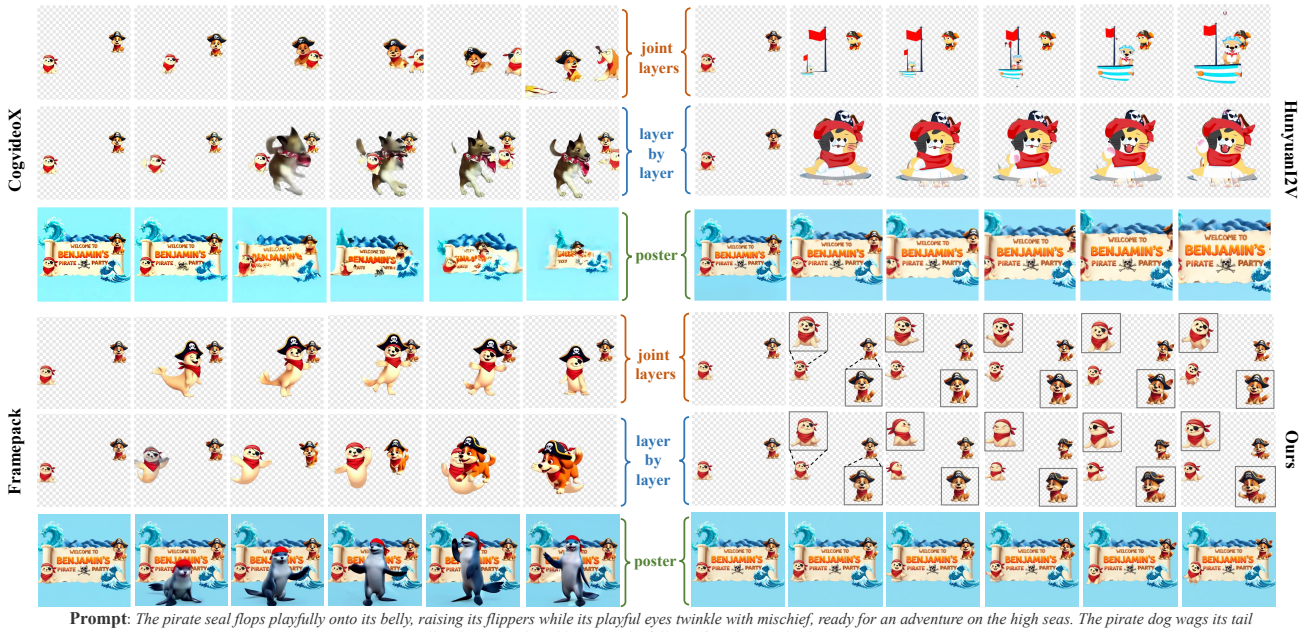
Table 5. Ablation study on decoder structures.

Transparent Decoder Backbone. We explored three architectural configurations for integrating the Temporal Scale (TS) blocks within the ViT backbone: (1) appending two TS blocks after the ViT (EE); (2) placing one TS block before and one after the ViT (SE); and (3) inserting one block at the beginning and another in the middle of the ViT (SM). Table 5 shows that the third configuration performs best, as it enables the attention mechanism to model temporal dependencies among visual features at two distinct stages of varying lengths, achieving superior performance.

5.3. Qualitative Result

For visualization purposes, we composite the multiple layers onto a single background.

Qualitative Comparison. Figure 5 provides a visual comparison of our method against CogvideoX, HunyuanI2V, and FramePack for text-controlled generation of dynamic layers across three paradigms: layer-by-layer, joint multi-layer, and poster-level synthesis. In the layer-by-layer generation mode, it is evident that the baseline I2V methods struggle to preserve object identity. As shown in the figure, objects in subsequently generated dynamic layers often undergo significant deformations (e.g., the seal's body becomes disproportionately larger), failing to maintain content consistency with the provided static. For joint multi-layer generation, competing methods exhibit a failure to align all layers with their corresponding textual prompts.



Prompt: *The pirate seal flops playfully onto its belly, raising its flippers while its playful eyes twinkle with mischief, ready for an adventure on the high seas. The pirate dog wags its tail excitedly, tilts its head, and bounces playfully as if ready to embark on a treasure hunt.*

Figure 5. Visualization of multiple dynamic layers under joint layers, layer by layer, and poster-level settings.

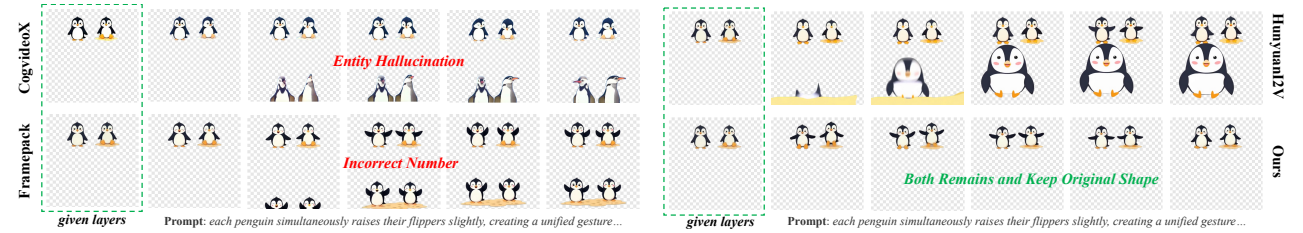


Figure 6. Visualization of multiple dynamic layers for group action.

Figure 7. Visualization of textual effect.

The visualizations reveal that only the first layer (the seal) is animated in response to its prompt, while the other layer fails to generate the intended dynamic effect. In the poster-level synthesis setting, where the goal is to animate specific layers within a complete poster, the other methods again fail to correctly map the textual instruction to the corresponding object. For example, FramePack hallucinates an entirely new seal rather than animating the existing one. Figure 6 further illustrates the qualitative outcomes when generating group action dynamic layers across multiple layers using text control. Compared with our method, the existing approaches often hallucinate and generate additional objects not present in the original layers, incorrectly altering the number of entities within the scene. These visual results underscore the necessity and superiority of our proposed approach for generating dynamic transparent layers.

Textual Effects. Figure 7 presents qualitative text-effect generations produced by our method:(i) vertical up–down oscillation and (ii) horizontal wave-like rotation.

6. Conclusion

In this paper, we addressed the novel and challenging task of transparent dynamic layer generation by proposing a comprehensive framework, Diffusive Multiple Dynamic Layers (DMDL). To facilitate research in this new domain, we first constructed a dataset comprising 100k dynamic layers via a training-free scheme. DMDL owns two pivotal innovations: (i) a latent multiple dynamic-layer diffusion model that leverages a novel layout-conditional 4D-RoPE to identify visual tokens across layer-spatial-temporal dimensions; and (ii) a transparent dynamic-layer autoencoder featuring a specialized ViT-based decoder with a two-stage 3D-RoPE, ensuring robust layer reconstruction across distinct temporal upscaling phases. Extensive experiments demonstrate that DMDL significantly outperforms leading I2V models both qualitatively and quantitatively, establishing a strong baseline for this new generation paradigm.

Acknowledgments

This work was supported by Grants of NSFC 62302246, ZJNSFC LQ23F010008, Ningbo 2023Z237 & 2024Z284 & 2024Z289 & 2023CX050011 & 2025Z038 & 2025Z059, and supported by High Performance Computing Center at Eastern Instituteof Technology and Ningbo Institute of Digital Twin.

References

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

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2, 3

[4] BRIA AI. Bria background removal v2.0 (rmbg-2.0). https://huggingface.co/briaai/RMBG-2.0, 2024. 3

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

[6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[8] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2

[9] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[12] Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024. 1, 2

[13] Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025. 2

[14] W Kong, Q Tian, Z Zhang, R Min, Z Dai, J Zhou, J Xiong, X Li, B Wu, J Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2025. *URL https://arxiv. org/abs/2412.03603*. 3, 4, 5, 6, 7

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 4

[16] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 6

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2

[18] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2, 3

[19] OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. 3, 6

[20] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[22] Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7952–7962, 2025. 1, 2, 4

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

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

tation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[25] 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. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[26] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 1

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

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4, 5

[31] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22413–22422, 2024. 1, 2

[32] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[34] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 2

[35] Yunnan Wang, Ziqiang Li, Wenyao Zhang, Lexiang Lv, Zequn Zhang, Xiaoyu Shen, Xin Jin, and Wenjun Zeng. Canvas: Compositional generation for art painting with seamless subject-driven infusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 6

[37] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3, 4, 6, 7

[38] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2

[39] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 1, 3, 4, 6, 7

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[41] 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