TransPixeler: Advancing Text-to-Video Generation with Transparency

Luozhou Wang[1*] Yijun Li[3†] Zhifei Chen[1] Jui-Hsien Wang[3] Zhifei Zhang[3]
He Zhang[3] Zhe Lin[3] Ying-Cong Chen[1, 2‡]
[1] HKUST(GZ) [2] HKUST [3] Adobe Research

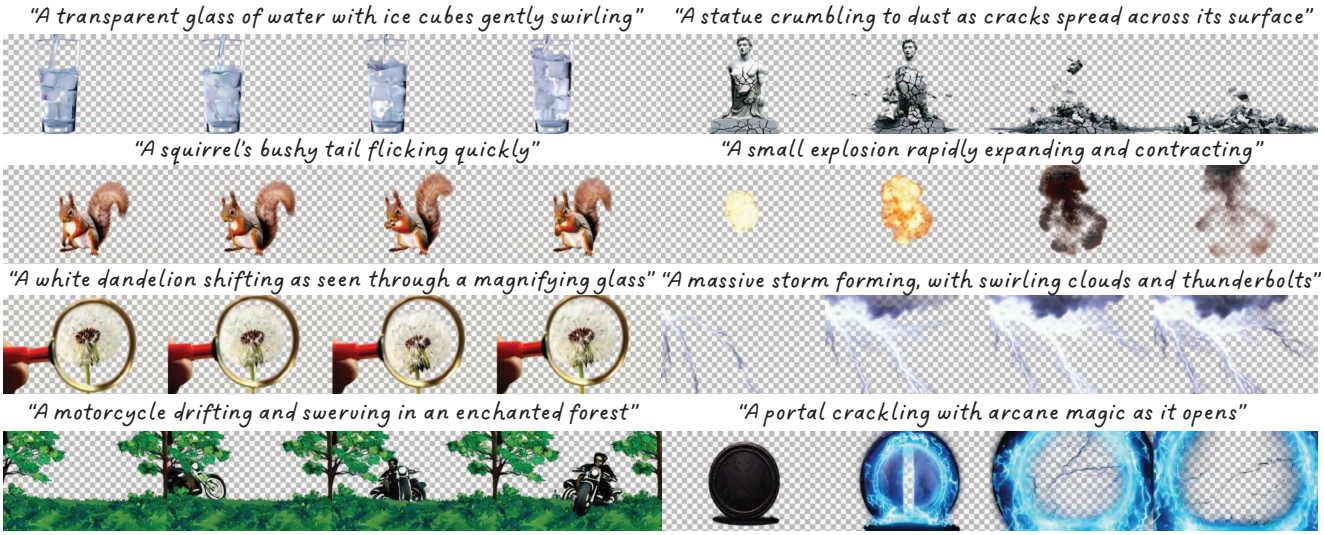

Figure 1. **RGBA Video Generation with TransPixeler.** By introducing LoRA layers into DiT-based text-to-video model with a novel alpha channel adaptive attention mechanism, our method enables RGBA video generation from text while preserving Text-to-Video quality.

Abstract

Text-to-video generative models have made significant strides, enabling diverse applications in entertainment, advertising, and education. However, generating RGBA video, which includes alpha channels for transparency, remains a challenge due to limited datasets and the difficulty of adapting existing models. Alpha channels are crucial for visual effects (VFX), allowing transparent elements like smoke and reflections to blend seamlessly into scenes. We introduce TransPixeler, a method to extend pretrained video models for RGBA generation while retaining the original RGB capabilities. TransPixeler leverages a diffusion transformer (DiT) architecture, incorporating alpha-specific tokens and using LoRA-based fine-tuning to jointly generate RGB and alpha channels with high consistency. By optimizing attention mechanisms, TransPixeler preserves the strengths of the original RGB model and achieves strong alignment between RGB and alpha channels despite limited training data. Our approach effectively generates diverse and consistent RGBA videos, advancing the possibilities for VFX and interactive content creation. The code is available at https://wileewang.github.io/TransPixeler/.

1. Introduction

Text-to-Video generative models have quickly advanced, achieving impressive results [7, 16, 20, 26, 47, 49, 56, 61, 64]. This progress has enabled various applications, such as video editing [11, 13, 32, 39, 52, 55], image animation [2, 14, 15, 38], and motion customization [18, 24, 31, 36, 48, 51, 59]. Diffusion Transformers (DiT) enhance these models by using self-attention to capture long-range dependencies [3, 26, 56, 64]. These models are now widely used in entertainment, advertising, and education, meeting the demand for customizable, dynamic content. Notably, Text-to-RGBA (A denotes Alpha channel) video generation is invaluable for VFX and creative industries. The inclusion of an alpha channel in RGBA formats allows for transparent effects, enabling seamless blending of elements like smoke

*This work was done during an internship at Adobe Research.
†Project Leader
‡Corresponding author

and reflections (see Fig. 1). This transparency creates realistic visuals that can integrate smoothly into scenes without modifying the background. Such flexibility is crucial in gaming, virtual reality (VR), and augmented reality (AR), where dynamic and interactive content is in high demand.

Currently, no direct solutions exist for RGBA video generation, which remains a challenging task due to the scarcity of RGBA video data, with only around 484 videos available in [29]. This scarcity will significantly limit the diversity of generated content, resulting in a constrained set of object types and motion patterns. One feasible solution is to use video matting [28, 30, 40] to obtain alpha channels from generated videos. However, these methods are still limited by the scarcity of RGBA video data and struggle to generalize to a wider range of objects, as shown in Fig. 2 (b). Other video segmentation methods, such as SAM-2 [41], may generalize well to different tasks. However, they cannot generate alpha channels and are therefore unsuitable for direct compositing. There have been attempts to generate RGBA at the image level, such as LayerDiffusion [63]. However, adapting its concept directly to a temporal VAE used in video generative models remains challenging.

In this paper, we explore how to extend pretrained video models to generate corresponding alpha channels while retaining the original capabilities of pretrained models. Our goal is to generate content beyond the limitations of the current RGBA training set. Existing works such as Lotus [19] and Marigold [25] demonstrate that leveraging pretrained generation model weights significantly enhances out-of-distribution in dense prediction, hinting at the potential for predicting alpha channels. However, in the context of RGBA video generation, these approaches typically require generating RGB channels first, followed by separate alpha channel prediction. Consequently, information flows unidirectionally from RGB to alpha, keeping the two processes largely disconnected. Given the limited availability of RGBA video data, this imbalance results in insufficient alpha prediction when challenging objects are generated, as shown in Fig. 2 (c).

In this work, we propose **TransPixeler**, which effectively adapts the pretrained RGB video models to generate RGB channels and the alpha channel simultaneously. We leverage state-of-the-art DiT-like video generation models [26, 56] , and additionally introduce new tokens appended after text and RGB tokens for generating the alpha channels. To facilitate convergence, we reinitialize the positional embeddings for the alpha tokens and introduce a zero-initialized, learnable domain embedding to distinguish alpha tokens from RGB tokens. Furthermore, we employ a LoRA-based fine-tuning scheme [23], applied exclusively to project alpha tokens into the qkv space, thereby maintaining RGB generation quality. With the proposed approach, we extend the modality while preserving the original input-output structure and relying on the existing attention mechanism through LoRA adaptation.

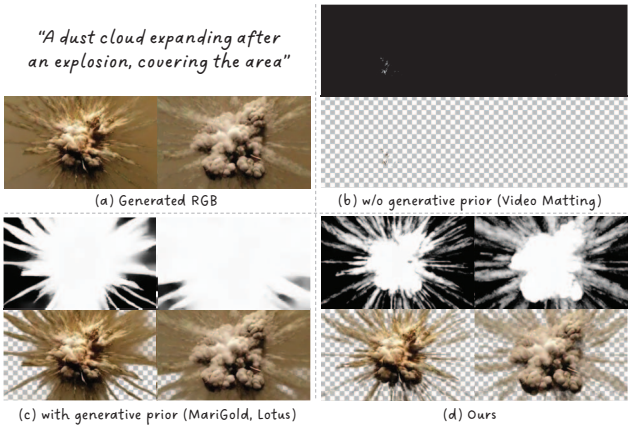

Figure 2. Comparison between **Generation-Then-Prediction** and our **Joint Generation** approach. Given the generated RGB in (a), (b) and (c) show the predicted alpha (top) and the composited result (bottom). In (d), the top shows the jointly generated alpha.

The extended sequence contains text, RGB, and alpha tokens, with self-attention divided into a 3x3 grouped attention matrix involving interactions like **Text-attend-to-RGB** (Text as query, RGB as key) and others. We also systematically analyze the attention mechanisms for RGBA generation: 1) **Text-attend-to-RGB** and **RGB-attend-to-Text**. The interaction between text and RGB tokens represents original model's generation capabilities. Minimizing the impact on text and RGB tokens during these attention computation processes can better retain the original model's performance; 2) **RGB-attend-to-Alpha**. We reveals a fundamental limitation in conventional methods is the lack of **RGB-attend-to-Alpha** attention. This attention is necessary to refine RGB tokens based on alpha information, improving RGB-alpha alignment; 3) **Text-attend-to-Alpha**. We remove this attention mechanism to reduce the risk caused by limited training data, which could degrade the model's performance. This removal also enhances the retention of the model's original capabilities.

By integrating these techniques, our method achieves diverse RGBA generation with limited training data while maintaining strong RGB-alpha alignment. To summarize, our contributions are as follows:

- We propose an RGBA video generation framework using DiT models that requires limited data and training parameters, achieving diverse generation with strong alignment.
- We analyze the role of each attention component in the generation process, optimize their interactions, and introduce necessary modifications to improve RGBA generation quality.
- Our method is validated through extensive experiments, demonstrating its effectiveness across a variety of challenging scenarios.

2. Related Work

Text-to-Video Generation. Early video generation models were primarily based on Unet-based latent diffusion models (LDMs) extended from text-to-image models like Stable Diffusion [42]. For example, AnimateDiff [16] introduced a temporal attention module to improve temporal consistency across frames. Subsequent video generation models [5, 7, 8, 47, 61, 62] adopted an alternating approach with 2D spatial and 1D temporal attention, including works like ModelScope, VideoCrafter, Moonshot, and Show-1.

With advancements in large language models (LLMs) and the introduction of Sora [3], attention shifted from Unet architectures to transformer-based architectures (DiT). DiT-based video generation models, such as Latte [37] and OpenSora [64], extended the DiT text-to-image (T2I) model [9] and maintained the 2D and 1D alternating attention approach, achieving promising results. Recently, DiT-based video generation has rapidly progressed, achieving further improvements in quality. Several methods [26, 44, 56] have moved away from the 2D and 1D alternating approach, instead treating video frames as a single long sequence with 3D positional embeddings for encoding. These approaches also prepend text tokens—processed through a text encoder—to the video sequence, creating a streamlined network that relies solely on full self-attention and feed-forward layers. Our method builds upon these recent open-source transformer-based video generation models.

Video Matting. A straightforward approach for RGBA video generation is to extract the alpha channel from generated RGB content, as done with traditional green screen keying or learning-based video matting expert models [28–30]. OmnimatteRF [28] introduces a video matting method that combines dynamic 2D foreground layers with a 3D background model, enabling more realistic scene reconstruction for real-world videos. Robust Video Matting (RVM) [30] proposes a real-time, high-quality human video matting method with a recurrent architecture for improved temporal coherence, achieving state-of-the-art results without auxiliary inputs. Another work presents a high-speed, high-resolution background replacement technique with precise alpha matte extraction, supported by the VideoMatte240K and PhotoMatte13K/85 datasets [29]. Additionally, many image matting methods [4, 6, 27, 50, 57] can be applied for frame-by-frame matting.

Further, several works [19, 25, 53] in image depth estimation adapt pretrained generation models for prediction tasks, achieving strong performance that often surpasses traditional, scratch-trained expert models. Marigold [25] modifies architectures to create image-conditioned generation models, while Lotus [19] explores the role of the diffusion process in this context. Although there is currently no dedicated approach for video matting within video generation models, we replicate and extend these methods to evaluate their performance, allowing us to highlight the limitations of prediction-based pipelines for RGBA generation.

Generation beyond RGB. Another category of methods [1, 17, 34, 35, 54, 60, 63] explores expanding generation models to simultaneously generate additional channels, though they are not specifically designed for RGBA video generation. For instance, LayerDiffusion [63] modifies the VAE in latent diffusion models to decode alpha channels. However, VAEs typically lack the semantic understanding required for precise alpha generation, limiting their effectiveness in complex visual scenarios where texture and contour details are critical. In contrast, other approaches [1, 34, 35, 60] modify the denoising model directly to enable joint generation. Wonder3D [34] uses a domain embedding to control the model's generation modality, while methods like IntrinsicDiffusion [35] and RGB↔X [60] adapt the UNet's input and output layers to jointly produce intrinsic modalities. However, all these methods are designed for image tasks and rely on UNet architectures. When applied to video generation, they face limitations in quality and diversity due to the scarcity of RGBA video data.

3. Method

3.1. Preliminary

We first introduce the open-sourced state-of-the-art DiT-based video generation models [44, 56]. The core components of DiT-based video models are attention modules, and there are two primary distinctions between these models and previous approaches. On one hand, unlike previous models that alternate between 1D temporal attention and 2D spatial attention [5, 7, 8, 64], current methods typically employ 3D spatio-temporal attention, allowing them to capture spatio-temporal dependencies more effectively. On the other hand, instead of using cross-attention for text conditioning, these models concatenate text tokens $\mathbf{x}_{\text{text}}$ with visual tokens $\mathbf{x}_{\text{video}}$ into a single long sequence. The shape of video tokens and text tokens are $B \times L \times D$ and $B \times L_{\text{text}} \times D$, wher B equals to batch size, L_{text} equals to the length of text tokens, L equals to the length of video tokens and D equals to the latent dimension of transformer. Full self-attention is then applied across the combined sequence:

$$\begin{aligned} &\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad \text{where} \\ &\mathbf{Z} : \mathbf{Z} \in \{\mathbf{Q},\mathbf{K},\mathbf{V}\} \\ &= [\mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{text}}); \mathbf{f}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}})] \end{aligned} \tag{1}$$

Here $\mathbf{W}_t$ (for $t \in \{q,k,v\}$) represents the projection matrixs in the transformer model, and $\mathbf{f}_t$ (for $t \in \{q,k,v\}$) represents a combined operation that incorporates both the projection and positional encoding for visual tokens. There

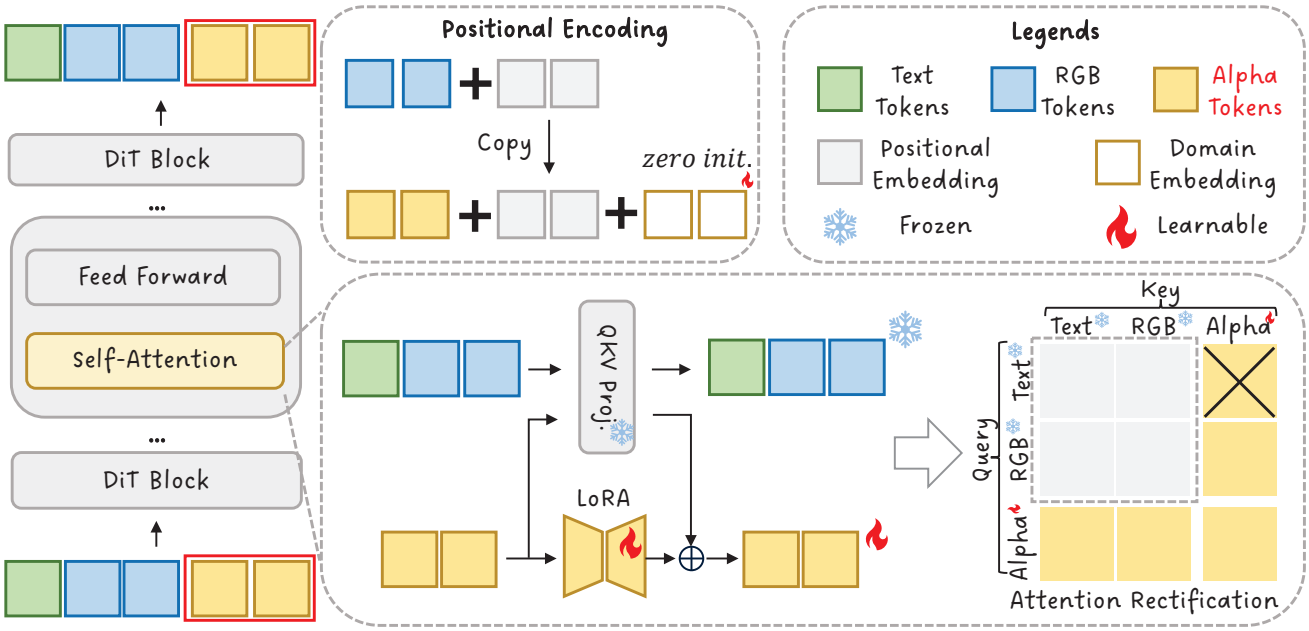

Figure 3. **Pipeline of TransPixeler.** Our method is organized as follows: (1) **Left**: we extend the input of DiT to include new alpha tokens; (2) **Top Center**: we initialize alpha tokens with our positional encoding; (3) **Bottom Center**: we insert a partial LoRA and adjust attention computation during training and inference.

are two commonly used types of positional encoding. One is absolute positional encoding formulated as follows:

$$\mathbf{f}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}}) := \mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^m), \quad (2)$$

where $\mathbf{p}$ is the positional embedding (e.g., a sinusoidal function) and m denotes the position of each RGB video token. Another approach is the Rotary Position Embedding (RoPE) [43], often used by [44, 56]. This is expressed as

$$\mathbf{f}_{z:z\in\{q,k\}}(\mathbf{x}_{\text{video}}) := \mathbf{W}_{z:z\in\{q,k\}}(\mathbf{x}^m_{\text{video}}) \circ e^{im\theta}, \quad (3)$$

where m is the positional index, i is the imaginary unit for rotation, and θ is the rotation angle.

3.2. Our Approach

To jointly generate RGB and alpha videos, we adapt a pre-trained RGB video generation model through several modifications. The whole pipeline is visualized in Fig. 3.

Firstly, we double the sequence length of noisy input tokens to enable the model to generate videos of double length, from $\mathbf{x}^{1:L}_{\text{video}}$ to $\mathbf{x}^{1:2*L}_{\text{video}}$. Here, $\mathbf{x}^{1:L}_{\text{video}}$ will be decoded into the RGB video, while $\mathbf{x}^{L+1:2*L}_{\text{video}}$ will be decoded into the corresponding alpha video. The Query(Q), Key(K), Value(V) representations are formulated as:

$$\begin{aligned}&\mathbf{Z} : \mathbf{Z} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} \\ &= [\mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{text}}); \mathbf{f}_{z:z\in\{q,k,v\}}(\mathbf{x}^{1:2*L}_{\text{video}})]\end{aligned} \quad (4)$$

In addition to sequence doubling, we explored increasing batch size or latent dimensions and splitting output into two domains; however, these approaches showed limited effectiveness under constrained datasets, which we discuss later.

Secondly, we modify the positional encoding function $\mathbf{f}_{t:t\in\{q,k,v\}}(\cdot)$, as shown in Fig. 4. Instead of continuously numbering indices, we allow RGB and alpha tokens to share the same positional encoding. Taking absolute positional encoding as an example:

$$\begin{aligned}&\mathbf{f}^*_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}}) \\ &:= \begin{cases}\mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^m), & \text{if } m \le L, \\ \mathbf{W}^*_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^{m-L} + d), & \text{if } m > L.\end{cases}\end{aligned} \quad (5)$$

Here we introduce a domain embedding d, initialized to zero. We make it learnable to help the model adaptively differentiate between RGB ($m \le L$) and alpha tokens ($m > L$). The motivation behind this design is we observe that with same postional encoding, even initializing with different noises, the tokens from two domains tend to generate same results. It minimizes spatial-temporal alignment challenges at the very beginning of training and thus accelerates convergence.

Next we propose a fine-tuning scheme using LoRA [23], in which the LoRA layer is applied only to alpha domain tokens:

$$\begin{aligned}&\mathbf{W}^*_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^{m-L} + d) \\ &= \mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^{m-L} + d) \\ &+ \gamma \cdot \text{LoRA}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^{m-L} + d), \quad \text{if } m > L,\end{aligned} \quad (6)$$

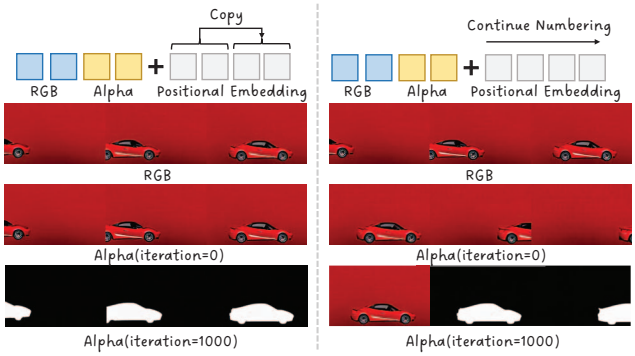

Figure 4. **Positional Encoding Design for RGBA Generation.** Assigning alpha tokens the same positional encoding as RGB yields similar results, resulting in faster convergence after 1000 iterations compared to standard encoding strategies.

where γ controls the residual strength. Additionally, we design an attention mask to block unwanted attention computation. Given a text-video token sequence length $L_{\text{text}} + 2L$, where L_{text} represents text token length, the mask is defined as:

$$\mathbf{M}^*_{mn} = \begin{cases} -\infty, & \text{if } m \leq L_{\text{text}} \text{ and } n > L_{\text{text}} + L, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Combining these modifications, inference with our method is expressed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}^*\right)\mathbf{V}, \quad \text{where}$$
$$\mathbf{Z} : \mathbf{Z} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$$
$$= [\mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{text}}); \mathbf{f}^*_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}})] \tag{8}$$

Training is carried out using flow matching [33] or a traditional diffusion process [21].

3.3. Analysis

Given our goal of maximizing the inherited capabilities of the pretrained video model, enabling it to generate beyond the existing RGBA training set, we analyze the most critical component within our current 3D full attention DiT video generation model: the attention mechanism. The attention matrix, $\mathbf{Q}\mathbf{K}^T$, has dimensions $(L_{\text{text}}+2*L)\times(L_{\text{text}}+2*L)$, which we simplify by organizing it into a 3x3 grouped attention matrix—including **Text-attend-to-RGB**, **RGB-attend-to-Text**, and so forth, as illustrated in Fig. 3.

Text-Attend-to-RGB and RGB-Attend-to-Text. These represent the upper-left 2x2 section of and are computations that exist solely in the original RGB generation model. If we ensure that this part of the computation remains unaffected, we can replicate the original RGB generation performance. Therefore, we limit the scope of LoRA's influence, as defined in Eq. (7), by retaining the original QKV values for both text and RGB tokens, thus preserving the pretrained model's behavior in these domains.

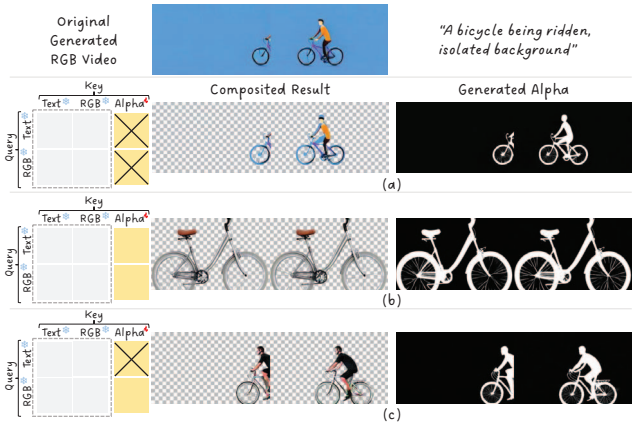

Figure 5. **Attention Rectification.** (a) Eliminating all attention from alpha as a key preserves 100% RGB generation but leads to poor alignment. (b) Retaining all attention significantly degrades quality, causing a lack of motion in bicycles. (c) Our method achieves an effective balance.

Besides the partial LoRA, the added alpha tokens requires the text and RGB tokens to also act as queries and interact with the alpha tokens as keys, which alters the computation in this 2x2 attention matrix. Therefore, we further analyze two additional attention computations that impact RGB generation, as shown in Fig. 5.

Text-Attend-to-Alpha. We find that this attention is detrimental to the generation quality. Since the model was originally trained with text and RGB data, introducing attention from text to alpha causes interference due to the domain gap between alpha and RGB. Specifically, the alpha modality provides only contour information and lacks the rich texture, color, and semantic details associated with the text prompt, thereby degrading generation quality. To mitigate this, we design the attention mask (Eq. (7)) that blocks this computation.

RGB-Attend-to-Alpha. In contrast, we identify **RGB-to-Alpha** as essential for successful joint generation. This attention allows the model to refine RGB tokens by considering alpha information, facilitating alignment between generated RGB and alpha channels. This refinement process is a critical component missing in previous generation-then-prediction pipelines, which lacked a feedback mechanism for RGB refinement based on alpha guidance.

4. Experiment

Training Dataset. We utilize the public VideoMatte240K dataset [29], a comprehensive collection of 484 high-resolution green screen videos consists of 240,709 unique frames of alpha mattes and foregrounds. These frames provide a diverse range of human subjects, clothing styles, and

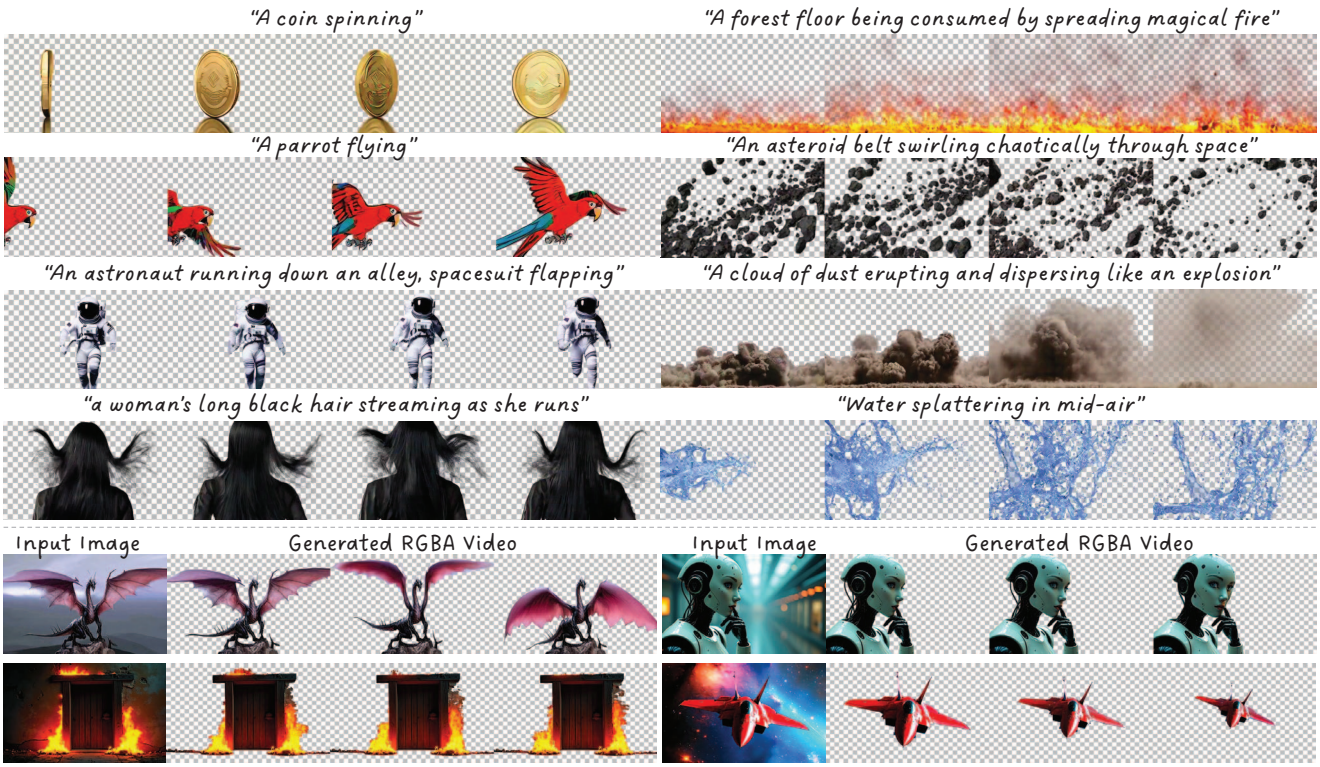

Figure 6. **Applications. Top**: Text-to-Video with Transparency. **Bottom**: Image-to-Video generation with transparency. .

poses. We apply fundamental preprocessing steps for them, including color decontamination and background blurring. Prompts are extracted using ShareGPT4V [10].

Model. Our RGBA video diffusion models are developed by fine-tuning pre-trained diffusion models. Specifically, we employ two models based on the diffusion transformer architecture: the open-source model CogVideoX [56] and a modified variant of CogVideoX denoted as J. CogVideoX generates RGB videos in 480x720 with 49 frames at 8 FPS, using 50 sampling steps. In contrast, the modified version produces videos in 176x320 with 64 frames at 24 FPS, while also using 50 sampling steps. Additionally, we integrate our method with CogVideoX-I2V (Image-to-Video) to support image-to-video generation with transparency. We set the LoRA rank to 128. For domain embedding, we initialize it with an original shape of $1 \times D$ and zero values, then expand it to $L \times D$ through repetition during training. We train these parameters over 5,000 iterations with a batch size of 8 in total, utilizing 8 NVIDIA A100 GPUs.

4.1. Applications

We mainly demonstrate two applications shown in Fig. 6:

Text-to-Video with Transparency. Our method is capable of generating moving objects with various types of motion, such as spinning, running, and flying, while also handling transparent properties of bottles and glasses. Additionally, it can produce complex visual effects, including fire, explosions, cracking, and lightning, as well as creative examples.

Image-to-Video with Transparency. Our method can also be integrated with an I2V video generation model-CogVideoX-I2V. Users can provide a single image along with an alpha channel (optional), and then we generate subsequent frames with dynamic effects and automatically propagate or generate alpha channels for these frames.

4.2. Comparisons

Generation-then-Prediction Pipeline. As shown in Fig. 2, video matting methods [29, 40, 58] struggle with matting non-human objects (see supplementary materials for additional results). Therefore, we selected Lotus [19] and SAM-2 [41] as baselines due to their stronger generalization: Lotus uses pretrained generative models, and SAM-2 is trained on large datasets. Since Lotus was originally designed for single-image depth estimation, we extended it for RGBA videos, denoted as Lotus + RGBA in our comparisons. Qualitative results are shown in Fig. 7. Since ground-truth alpha channels are not available for generated videos, we focus on qualitative comparison.

Joint Generation Pipeline. Since there are currently no existing RGBA video generation models, we integrate AnimateDiff [16] with LayerDiffusion [63] to generate RGBA videos. We use the open-source video generation model

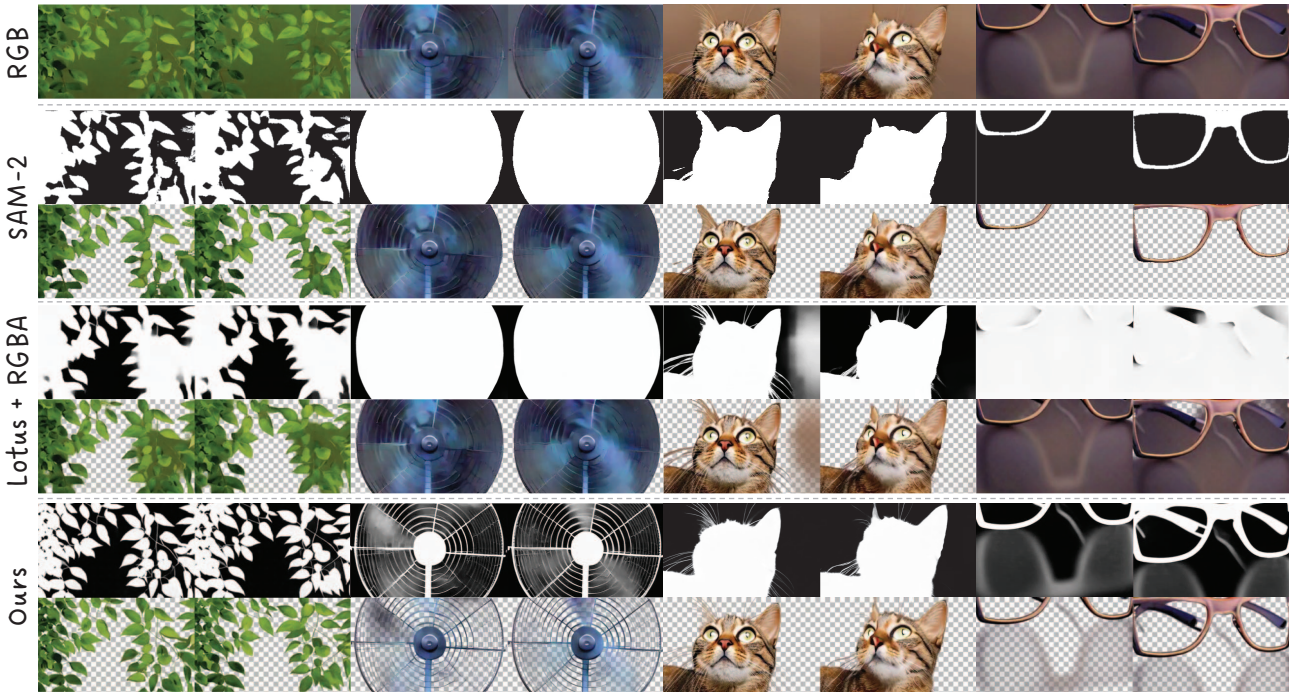

Figure 7. **Comparison with Generation-then-Prediction Pipelines.** Our method demonstrates superior alignment.

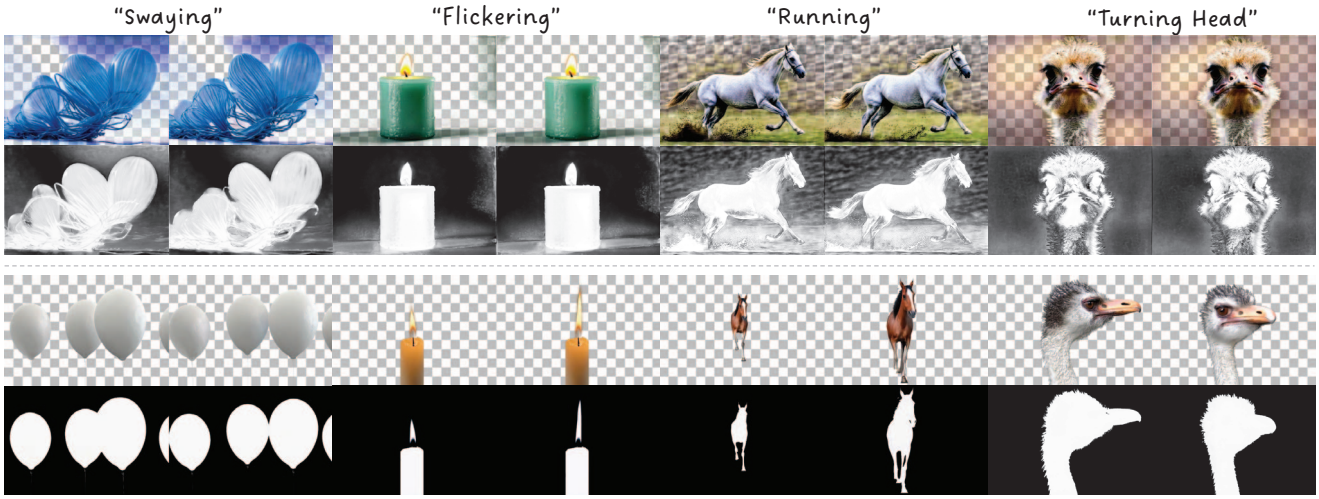

Figure 8. **Comparison with Joint Generation Pipelines. Top**: LayerDiffusion + AnimateDiff; **Bottom**: Ours. Our method achieves better alignment and generates corresponding motion described by prompts.

CogVideoX [56] as the base model for fair comparison. The qualitative results are illustrated in Fig. 8.

User Study. We also conduct a user study with Amazon Mechanical Turk to compare two joint generation methods, as shown in Table. 1. Participants are asked to evaluate two key aspects: 1) whether the RGB and alpha align correctly; and 2) whether the motion in the generated video matches the corresponding text description. A total of 30 videos are generated from distinct text prompts, and 87 users participated in the evaluation. The study shows that our method is obviously favored more by users with higher votes.

Table 1. **User Study.**

	RGBA Alignment	Motion Quality
AnimateDiff [16]+LayerDiff [63]	6.7%	21.7%
Ours + CogVideoX [56]	**93.3%**	**78.3%**

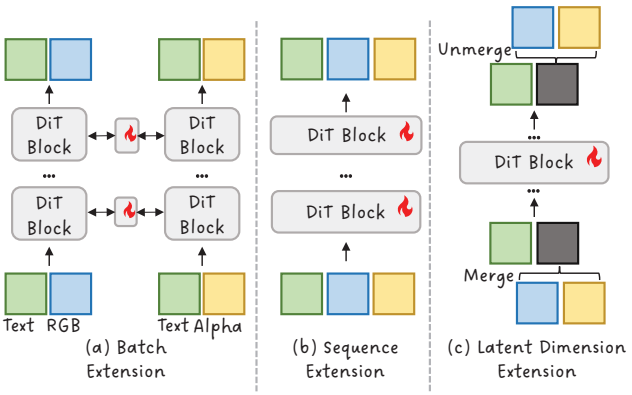

Figure 9. **Alternative Designs for Joint Generation with DiT**. Sequence extension (b) represents our method.

4.3. Ablation Study

As shown in Fig. 10, we conduct the ablation study across two dimensions: attention rectification and network design.

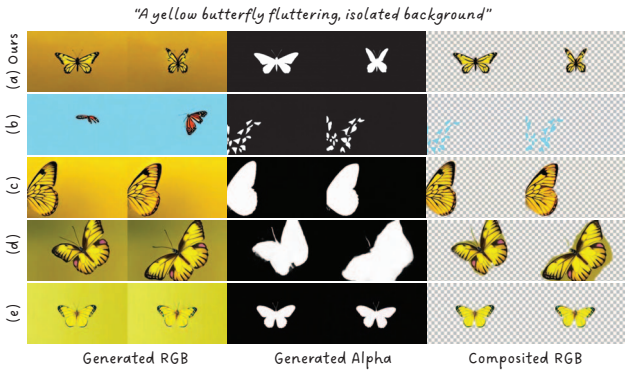

Figure 10. **Ablation Study.** (a) Ours; (b) Ours without RGB-attend-to-Alpha; (c) Ours with Text-attend-to-alpha; (d) Batch Extension Strategy; (e) Latent Dimension Extension Strategy. Our method maintains high-quality motion generation (e.g., butterflies waving their wings) while achieving good alignment.

Attention Rectification. By blocking RGB-to-Alpha attention, we first validate the importance of RGB-to-Alpha attention for aligning RGB and alpha channels, a feature lacking in most prediction-based methods. We also examine the effect of removing unnecessary attention to preserve the model's generative capacity, by learning Text-to-Alpha attention only. Without RGB-to-Alpha attention, the alpha channel misaligns with RGB content and the RGB output loses motion quality (e.g., reverse rocket).

Alternative Designs For Joint Generation. Given the transformer's input dimensions $B \times L \times D$, we extend the sequence dimension L to produce RGB and alpha channels, but alternative extensions are possible at the Batch B or Latent Dimension D levels (see Fig. 9). In the **Batch Extension** approach, a new module enables inter-batch communication, similar to the technique in [46]. For **Latent Dimension Extension**, we merge video and alpha tokens, project them into the DiT model's latent space, and unmerge post-generation, using learnable linear layers with fine-tuning. Batch Extension shows weaker RGB-alpha alignment, while Latent Dimension Extension, though akin to training from scratch, significantly reduces diversity.

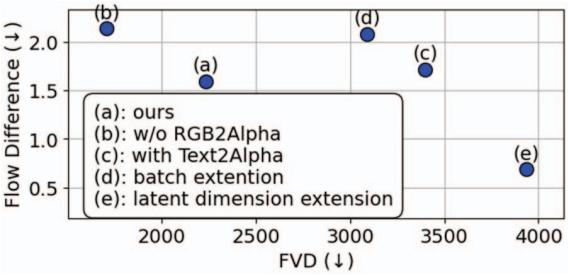

Figure 11. **Quantitative Evaluation**. Our approach achieves a good balance between alignment (low flow difference) and preserving generative quality (low FVD).

Evaluation. In addition to the qualitative comparisons shown in Fig. 10, we also generated a total of 80 videos, each consisting of 64 frames, and evaluated them using two primary metrics: **Flow Difference.** To measure alignment between the generated RGB and Alpha videos, we use optical flow [22] to focus on motion consistency while ignoring appearance. Specifically, we calculate optical flow with Farneback method [12] and compute the flow difference as the average Euclidean distance between RGB and Alpha flow fields. **Frechét Video Distance (FVD).** We use FVD [45] to compare the RGB videos generated by each RGBA method against those from the original RGB model, evaluating how well each method preserves the model's original generative quality. A lower FVD indicates that the generated results are closer to the original RGB model in terms of motion coherence and diversity, thus demonstrating a high fidelity to the model's intended generative quality. Results are shown in Fig. 11.

5. Conclusion

We introduce a novel Text-to-RGBA video generator that extends RGB diffusion models to produce RGBA output with minimal modifications and retained fidelity. Leveraging DiT backbones and RGBA-aware attention, the method preserves RGB quality while accurately inferring alpha channels. Targeted adjustments—adding alpha tokens, reinitializing positional embeddings, and applying selective LoRA fine-tuning—deliver rich, high-quality RGBA videos even from limited training data. Extensive quantitative and qualitative experiments validate the framework's versatility and robustness across challenging animation, complex compositing, and diverse real-world scenes.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 3

[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. 1

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 3

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22595–22604, 2024. 3

[5] cerspense. zeroscope_v2. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. Accessed: 2023-02-03. 3

[6] Guowei Chen, Yi Liu, Jian Wang, Juncai Peng, Yuying Hao, Lutao Chu, Shiyu Tang, Zewu Wu, Zeyu Chen, Zhiliang Yu, et al. Pp-matting: high-accuracy natural image matting. *arXiv preprint arXiv:2204.09433*, 2022. 3

[7] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 3

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024. 3

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[10] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 6

[11] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 1

[12] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, pages 363–370. Springer, 2003. 8

[13] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1

[14] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1

[15] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023. 1

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 3, 6, 7

[17] Hao He, Yixun Liang, Luozhou Wang, Yuanhao Cai, Xinli Xu, Hao-Xiang Guo, Xiang Wen, and Yingcong Chen. Lucidfusion: Generating 3d gaussians with arbitrary unposed images, 2024. 3

[18] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 1

[19] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 2, 3, 6

[20] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 1

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[22] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 8

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 4

[24] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similarity score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024. 1

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 3

[26] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 1, 2, 3

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

[28] Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, and Ayush Saraf. Omnimatterf: Robust omnimatte with 3d background modeling. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 23471–23480, 2023. 2, 3

[29] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 2, 3, 5, 6

[30] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 238–247, 2022. 2, 3

[31] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 1

[32] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 1

[33] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[34] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3

[35] Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Wang. Intrinsicdiffusion: Joint intrinsic layers from latent diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[36] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 1

[37] 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. 3

[38] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 1

[39] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 1

[40] Haotong Qin, Lei Ke, Xudong Ma, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Xianglong Liu, and Fisher Yu. Bimatting: Efficient video matting via binarization. *Advances in Neural Information Processing Systems*, 36: 43307–43321, 2023. 2, 6

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 6

[42] 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

[43] 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

[44] Genmo Team. Mochi, 2024. 3, 4

[45] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 8

[46] Shimon Vainer, Mark Boss, Mathias Parger, Konstantin Kutsy, Dante De Nigris, Ciara Rowles, Nicolas Perony, and Simon Donné. Collaborative control for geometry-conditioned pbr image generation. *arXiv preprint arXiv:2402.05919*, 2024. 8

[47] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 3

[48] Luozhou Wang, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024. 1

[49] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[50] Zhixiang Wang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang, and Shin'Ichi Satoh. Matting by generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[51] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[52] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3

[54] Shuai Yang, Zhifei Chen, Pengguang Chen, Xi Fang, Shu Liu, and Yingcong Chen. Defect spectrum: A granular look of large-scale defect datasets with rich semantics, 2023. 3

[55] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 1

[56] 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. 1, 2, 3, 4, 6, 7

[57] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 3

[58] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147: 105067, 2024. 6

[59] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1

[60] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[61] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 1, 3

[62] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024. 3

[63] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 2, 3, 6, 7

[64] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 3