Layer-Animate for Transparent Video Generation

Jingqi Bai [*†] Jingkai Zhou [§] Benzhi Wang [*†]
Weihua Chen [§‡] Yang Yang [*] Zhen Lei [*¶‡] Fan Wang [§]
[*]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[†]School of Artificial Intelligence, University of Chinese Academy of Sciences
[‡]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation,
Chinese Academy of Sciences
[§]Alibaba Group

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887999

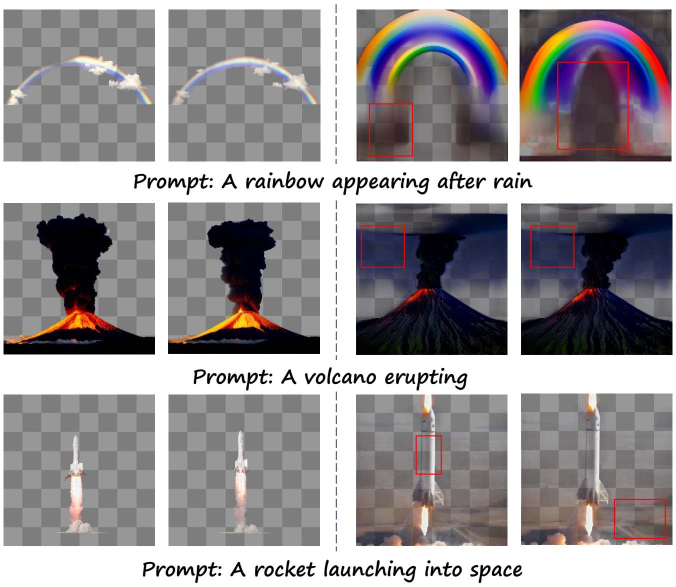
***Abstract*—Transparent videos with alpha channels play a crucial role in film production, advertising, and augmented reality fields. However, there is currently no available method for producing transparent videos. Traditional methods are time-consuming and labor-intensive, and employing alternative approaches for this task will result in inaccurate transparent regions, constrained motion, and artifacts. To address these challenges, we propose Layer-Animate, the first method capable of generating transparent videos. Our method comprises two stages: in the first stage, transparent images are generated as the base images to provide content and transparency information for the next stage. In the second stage, Inter-Frame Attention is applied to decouple content from motion, enabling the motion module to focus better on action. Layer-Animate is the first method used to generate transparent videos with accurate transparent regions, sufficient motion, and no artifacts, as demonstrated by notable improvements in qualitative and quantitative metrics.**

***Index Terms*—diffusion models, transparent videos, deep generative models, video generation.**

I. Introduction

Green screen technology is widely employed in film, television, and video production. However, the current mainstream method for obtaining green screen videos is shooting green screen videos in professional studios, which is often costly and time-consuming. In contrast, a transparent video preserves alpha channel transparency information within its frames. This alpha data allows for direct video layering, achieving the same effects as green screen technology.

Although generative technologies have advanced significantly in recent years [1], current models are typically only capable of generating standard videos without transparent information. Research into transparent video technology remains surprisingly limited. Zhang et al. proposed LayerDiffuse [2] to generate transparent images, which encodes the transparent alpha channel into the latent distribution of Stable Diffusion. The initial approach for extending transparent image generation to transparent video generation was to integrate the pre-trained motion module [3] into the layer-diffusion model. However, since the pre-trained layer-diffusion model was trained on transparent images, while the motion module was trained on standard video data, a mismatch in data distribution arises between the two. This inconsistency leads to problems, as shown in the red box of Fig.1, such as errors in transparent regions and significant distortion when directly applying the pre-trained motion module. Meanwhile, due to the cosiderable model size, training with a small amount of data prevents the model from genuinely learning the distribution of transparent data, leading to poor results [4], [5]. However, training with a large dataset is often costly and impossible. To address these challenges, we propose Layer-Animate, the first method capable of generating transparent videos with accurate transparency and sufficient motion amplitude while minimizing artifacts. This method achieves these results using only a small amount of training data. Our method consists of two stages: In the first stage, our objective is to generate an transparency image. In the second stage, the transparency image generated in the previous stage will be used as the base image, providing prior information through Inter-Frame Attention for the current stage. This allows the model to conduct a foundational understanding of the layout between the foreground and the background, enabling it to focus more on generating the motion.


Prompt: A rainbow appearing after rain

Prompt: A volcano erupting

Prompt: A rocket launching into space

Fig. 1: Illustration of our transparent video results with given prompts. The right side of the figure illustrates the significant problems resulting from applying alternative methods, such as Still-Moving. These issues include artifacts, inaccurate rendering of transparent regions, and insufficient motion. The left side displays our results.

Our contributions are summarized as follows:

- We propose a novel framework, Layer-Animate, which is the first to address the task of transparent video generation, effectively mitigating the issue of inaccurate transparent regions, insufficient motion, and artifacts.
- We propose Inter-Frame Attention, which utilizes prior information from images to decouple motion from content, enabling the generation of high-quality transparent videos with minimal data requirements.
- The Layer-Animate method substantially improves the quality of transparent video generation, as evidenced by extensive experiments demonstrating enhancements in qualitative and quantitative metrics.

II. Related Work

A. Diffusion Model

Recently, diffusion models [6]–[9] have garnered significant attention due to their powerful generative capabilities, emerging as a prominent research focus in computer vision. These models have demonstrated superior performance, surpassing traditional techniques [10] through their inherent ability to generate high-quality and diverse outputs. However, the high dimensionality of images poses substantial computational challenges. To address this, the Latent Diffusion Model (LDM) [11] was introduced, performing denoising in a lower-dimensional latent space through a pre-trained autoencoder [12]. In text-to-image generation models, it is common practice to employ language models like CLIP [13] and T5 [14] as text encoders, integrating them via cross-attention mechanisms [15] to enhance text-image alignment. In addition to natural language inputs, incorporating additional image conditions to guide the layout of generated images [16], [17] has also emerged as an active area of research.

B. Text to Video Generation

The success of diffusion-based models in the text-to-image (T2I) [18]–[21] field underscores their potential in text-to-video (T2V) [22]–[25] generation. T2V generation surpasses the capabilities of T2I generation by producing a sequence of continuous frames in which diffusion models play a pivotal role. In contrast to T2I, T2V requires synthesizing temporally coherent and visually consistent video content from textual descriptions, which presents distinct challenges in maintaining continuity and realism over time [26]. Tune-A-Video [27] cleverly incorporates trainable temporal layers while retaining the core structure of existing T2I frameworks. This strategy capitalizes on the effectiveness of T2I models, extending their capability to manage the temporal dimension, thereby facilitating the generation of seamless video sequences. However, these methods are suitable only for conventional video generation; their direct application to transparent videos leads to significant artifacts and inaccuracies in transparent regions.

III. METHOD

The overall framework pipeline is depicted in Fig.2. Essentially, the LayerDiffuse [2] model can be regarded as a customized variant of the text-to-image model. We aim to advance LayerDiffuse to generate transparent videos with precise transparency, adequate motion amplitude, and minimal artifacts. We divide the pipeline into two stages. In the first stage, we utilize a large-scale, custom-trained text-to-image model to generate high-quality transparency images. In the second stage, the transparency image serves as the base image, providing content information and latent transparency to generate transparent videos.

A. Prelimiaries

Latent Diffusion Models (LDMs) are a set of efficient diffusion models that perform the denoising process in the compressed latent space rather than in the pixel space. Specifically, LDMs leverage the VAE encoder [28] to compress images into latent space, learning the data distribution by applying forward and reverse diffusion processes within this latent space [29]. Herein, the VAE and the diffusion model should share the same latent distribution, as any major mismatch can significantly degrade the inference, training, or fine-tuning [30] of the latent diffusion framework.

To incorporate transparency information into the diffusion model, LayerDiffuse [2] employs a direct approach: it verifies whether the image decoded from the latent x with the added offset x_ϵ remains consistent with the original image and free from artifacts. The absence of artifacts suggests that the introduced offset does not adversely affect the latent representation. Specially, LayerDiffuse make use of the latent offset x_ϵ to establish "latent transparency" for encoding and decoding transparent images, which trained from scratch a latent transparency encoder $\varepsilon(\cdot,\cdot)$ that takes the RGB channels I_c and alpha channel I_α as input to convert pixel-space transparency into a latent offset, which is formulated as follows:

$$x_\epsilon = \varepsilon(I_c, I_\alpha). \tag{1}$$

LayerDiffuse then train from scratch another latent transparency decoder $\mathcal{D}(\cdot,\cdot)$ that takes the adjusted latent $x_a = x + x_\varepsilon$ and the RGB reconstruction $\hat{I} = \mathcal{D}^*_{sd}(x_a)$ to extract the transparent image from the adjusted latent space :

$$[\hat{I}_c \ \hat{I}_\alpha] = \mathcal{D}(\hat{I}, x_a), \tag{2}$$

where $\hat{I}_c$, $\hat{I}_\alpha$ are the reconstructed color and alpha channels and $\mathcal{D}^*_{sd}(\cdot)$ dentoes the pretrained and frozen Stable Diffusion latent decoder. Through a specially designed loss function, the latent space containing both alpha channel and RGB information is aligned with the latent space of normal images. This latent space maintains consistency with the distribution in the standard LDM model while exhibiting latent transparency.

B. Network Architecture

Transpaernt image generation. We employ the LayerDiffuse as our pre-trained text-to-image method during the first stage. Through LayerDiffuse, we generate a transparent image as the base image for the second stage. This base image can provide high-quality content priors for the second stage, including layout, id and etc., and contains necessary latent transparency. We decompose the transparent image into an

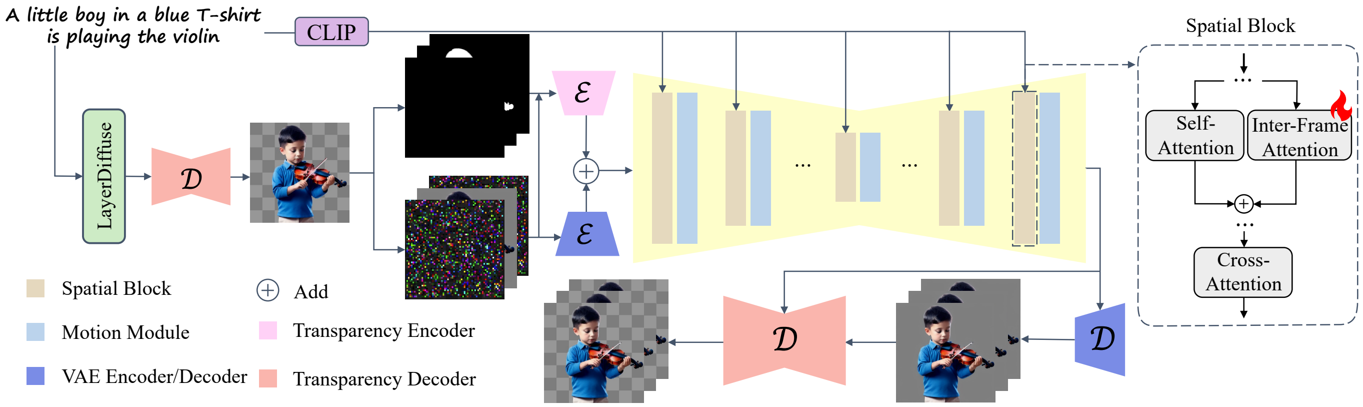


Fig. 2: The overall framework of our method. Our method separates the transparent video generation process into two distinct stages: 1) a transparent image generation stage and 2) a transparent video generation stage. During training, noise is applied to the video frames, excluding the middle frame, and the loss is subsequently calculated for these frames. The middle frame serves as a base image, and the newly introduced trainable Inter-Frame Attention facilitates the decoupling of content and motion. Consequently, the motion module can focus exclusively on motion without being influenced by content, leading to more realistic, transparent videos. It is important to note that only the Inter-Frame Attention component is trainable in our model.

RGB image and an alpha image. Then, we encode the transparent information separately using the Transparency Encoder and VAE Encoder and sum the results to obtain the adjusted latent. This adjusted latent not only conforms to normal data distribution but also incorporates transparent information.

Intermediate frame information injection. In the second stage, we employ the text and the transpaernt images produced in the first stage as conditions. Previous studies [31], [32] have shown that the self-attention mechanism can substantially improve the preservation of detail in reference images. Inspired by these findings, we introduce an Inter-Frame Attention mechanism that facilitates the transmission of detailed information across frames. By providing content and transparent information as priors, this method allows the motion module to be decoupled from the content and transparency, enabling it to focus more effectively on motion. This approach helps mitigate issues such as artifacts and errors in transparent regions arising from distribution mismatches between the LayerDiffuse and the motion module.

During the training process, we keep the intermediate frame, such as the eighth frame, noise-free while adding noise to the other frames. By this way, we add the image features obtained from the intermediate frame to the Inter-Frame Attention features. Previous approaches [33] frequently employed the first frame as the reference image. During the inference phase, we use the VAE Decoder to decode a grayscale background image and input it along with the adjusted latent into the Transparency Decoder to generate video frames with transparency information. Formally, let $\mathcal{V} = \{f^i | i = 1, ..., l\}$ represent a video consisting of l frames, given the key and value features of the intermediate frame (m-th frame), the Inter-Frame Attention can be computed as follows:

$$\begin{aligned}\text{Out}^{'} &= \text{Attention}\left((Q^i)^{'}, K^m, V^m\right) W_O^{'} \\ &= \left(\text{softmax}\left(\frac{(Q^i)^{'}(K^m)^T}{\sqrt{d}}\right) V^m\right) W_O^{'}, \qquad (3)\end{aligned}$$

where $(Q^i)^{'} = X^i W_Q^{'}$, $K^m = X^m W_K$ and $V^m = X^m W_V$. It is crucial to note that only the weight $W_Q^{'}$ and the output matrix $W_O^{'}$ serve as trainable parameters in the entire network. The weights W_K and W_V are directly inherited from the parallel self-attention mechanism. Finally, the output of self-attention is added to that of Inter-Frame Attention.

IV. Experiments

A. Dataset and Evaluation Protocol

We have compiled a dataset of 20,000 high-quality videos with a resolution of 720p or higher for fine-tuning. Each video incorporates pixel-level alpha channel data and each has an average duration of approximately 18 seconds. Each video clip is accompanied by a detailed textual description of its content.

B. Results and Comparisons

Using our existing data, we reproduced two methods for comparison: Still-Moving [34] and direct training of the motion module. Still-Moving introduces a Motion Adapters to control the dynamic levels in the video and incorporates a Spatial Adapters to map the customized model's output to the distribution of standard videos.

Qualitative Analysis: To evaluate the effectiveness of our method in transparent video generation, we present several visual examples of our method and two baselines in Fig. 3. This figure presents a comprehensive comparison between our method and the other two methods from two critical perspectives. **(a) Effectiveness in Artifact Prevention:** Our method demonstrates significant effectiveness in preventing artifact generation and effectively mitigates the issue of data distribution inconsistency between the transparent image model and the motion module. **(b) Sufficiency of Motion Amplitude:** Compared to other methods, our approach decouples motion from content through the integration of Inter-Frame Attention, enabling the motion module to concentrate more effectively on movement, enhancing the motion dynamics in the video. **(c) Ensuring Correct Transparency:** Our method, which

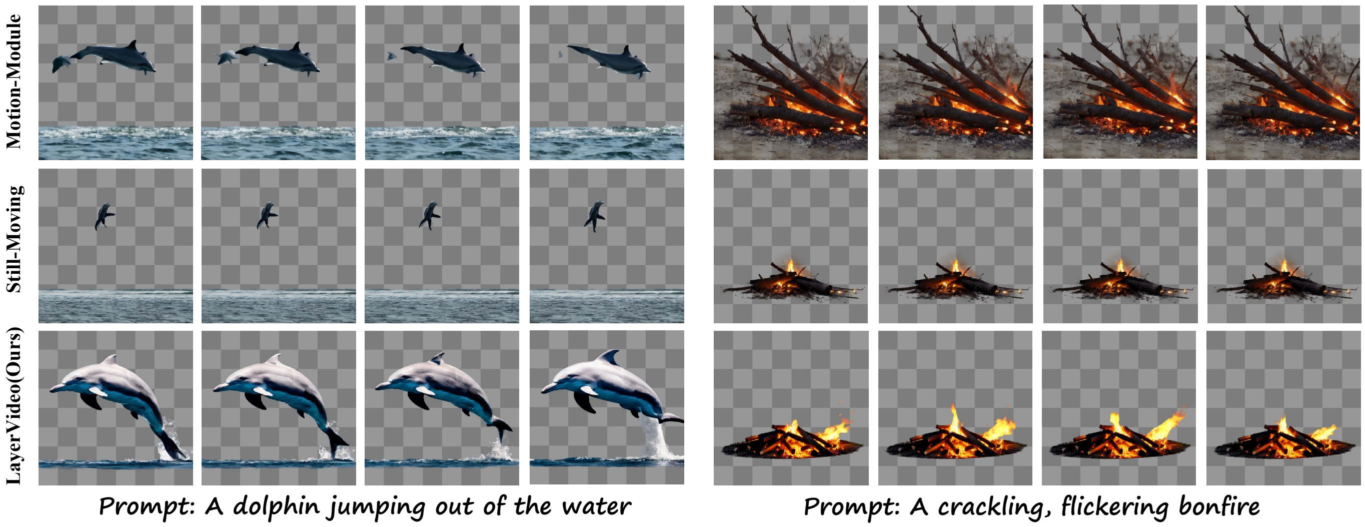

Motion-Module
Still-Moving
LayerVideo(Ours)
Prompt: A dolphin jumping out of the water
Prompt: A crackling, flickering bonfire

Fig. 3: Comparative results of transparent video generation across different methods.

utilizes high-quality transparent image as the base image, can effectively mitigate the issue of incorrect transparent regions within the transparent videos.

TABLE I: Quantitative comparison with evaluated other methods.

Candidate	User Study ↑	Automatic metrics	
		Clip Score ↑	Dynamic Degree ↑
Motion Module	10.34	27.72	3.96
Still-Moving	6.91	28.96	0.01
Layer-Animate (Ours)	**82.75**	**30.63**	**4.43**

Quantitative Analysis: We compare our proposed method against existing approaches through automatic metrics and user study, with the results shown in Table I. **(a) User study.** Participants compared the transparent videos generated by our method with those produced by other methods. Specifically, we selected ten volunteer to participate in the study. We randomly sampled 100 results generated from three different methods for evaluation. Participants were required to assess the videos from three perspectives: motion dynamics, the precision of transparent regions, and the authenticity of content, and rank each video set. We then computed the proportion of instances where each method was ranked the highest across all video sets. **(b) Automatic metrics.** To measure textual faithfulness, we compute the average CLIP [13] score between all frames of output videos and corresponding prompts. The degree of motion is a critical metric for assessing video quality [35]. While we employed the RAFT [36] method, the uniformly gray backgrounds in our videos significantly skew this metric, leading to markedly lower overall scores when compared to standard videos. Our results indicate that our method outperforms the other two approaches in content authenticity and motion sufficiency. However, due to the lack of a precise metric to assess the accuracy of transparent regions, we conducted a user survey to compare the outcomes of the different methods.

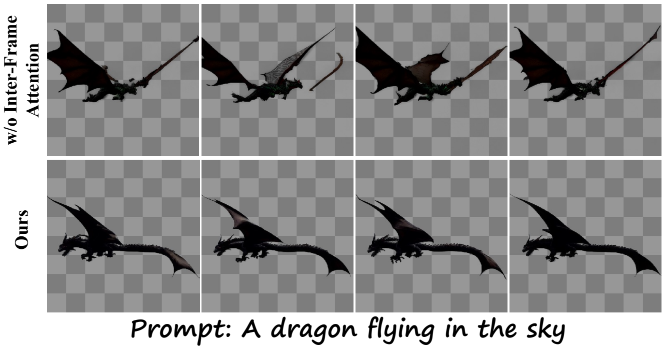


Fig. 4: Qualitative results of ablation study on model design.

Ablation Study: To validate the effectiveness of the key components in our method, we compared the proposed model as the baseline with the method that removes Inter Frame Attention. The corresponding results are shown in Fig. 4. The result demonstrates that the absence of Inter Frame Attention leads to severe artifacts, significantly affecting video quality.

V. CONCLUSION

In this paper, we propose a novel method for transparent video generation, marking it as the first approach specifically designed for this purpose. Our method enhances the quality of transparent video generation by partitioning the process into two distinct stages, effectively mitigating artifacts errors in transparent regions, and insufficient motion. Comprehensive experiments demonstrate significant improvements in both qualitative and quantitative measures, validating the effectiveness and robustness of our approach.

ACKNOWLEDGEMENTS

This work was supported in part by the Chinese National Natural Science Foundation Project 62276254, 62206276, U23B2054, 62106264 and the InnoHK program.

References

[1] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[2] L. Zhang and M. Agrawala, "Transparent image layer diffusion using latent transparency," *arXiv preprint arXiv:2402.17113*, 2024.

[3] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[5] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli *et al.*, "Lumiere: A space-time diffusion model for video generation," *arXiv preprint arXiv:2401.12945*, 2024.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[9] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[12] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[15] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[16] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," *arXiv preprint arXiv:2302.09778*, 2023.

[17] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou, "I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models," *arXiv preprint arXiv:2311.04145*, 2023.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[19] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[21] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[22] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. 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*, 2023, pp. 22 563–22 575.

[23] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[24] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[25] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[26] L. Gong, Y. Zhu, W. Li, X. Kang, B. Wang, T. Ge, and B. Zheng, "Atomovideo: High fidelity image-to-video generation," *arXiv preprint arXiv:2403.01800*, 2024.

[27] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. 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*, 2023, pp. 7623–7633.

[28] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[30] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, S. Huang, and W. Chen, "Consisti2v: Enhancing visual consistency for image-to-video generation," *arXiv preprint arXiv:2402.04324*, 2024.

[31] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 560–22 570.

[32] Y. Zhang, Z. Xing, Y. Zeng, Y. Fang, and K. Chen, "Pia: Your personalized image animator via plug-and-play modules in text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7747–7756.

[33] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, and X. Zhang, "Lamp: Learn a motion pattern for few-shot-based video generation," *arXiv preprint arXiv:2310.10769*, 2023.

[34] H. Chefer, S. Zada, R. Paiss, A. Ephrat, O. Tov, M. Rubinstein, L. Wolf, T. Dekel, T. Michaeli, and I. Mosseri, "Still-moving: Customized video generation without customized video data," *arXiv preprint arXiv:2407.08674*, 2024.

[35] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 807–21 818.

[36] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.