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/ICASSP49660.2025.10887785

Generating Customized 4D Motions from Text Inputs Using Spatial-Temporal Slicing Approaches

Zhichao Zhang,Hui Chen,Ming Xu
College of Computer Science and Technology
National University of Defense Technology
Changsha China

Jinsheng Deng,Xingshen Song
College of Advanced Interdisciplinary Studies
National University of Defense Technology
Changsha China
jsdeng@nudt.edu.cn

***Abstract*—Text-guided diffusion models have revolutionized static 3D generation, which significantly accelerated progress in 4D content creation. However, applying diffusion models to 4D content creation poses huge challenges due to the complexity and diversity of motion. The task of text to 4D customized generation requires a large amount of guide data, and it is challenging to integrate diverse knowledge from multiple diffusion models. To handle these challenges, we present Motion4D, a novel framework focusing on motion customization in 4D creation tasks, adopting a spatial-temporal slicing strategy towards the generation process. Firstly, the initialized 4D Gaussian field (XYZ-T) is temporally sliced into 3D scenes corresponding to discrete time points along the time axis. Secondly, for spatial dimension, 3D objects are further decomposed into orthogonal multi-view images to capture geometric and appearance features from various perspectives. This spatial-temporal slicing enables a comprehensive representation of object motion and variation across both temporal and spatial dimensions, facilitating customized 4D modeling. Extensive experiments demonstrate that our method surpasses prior state-of-the-art methods in terms of generation efficiency and motion consistency across various prompts.**

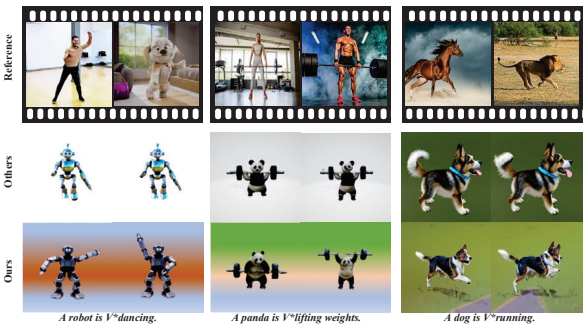

Fig. 1: **Comparisons of text to 4D generation with reference motions(first row).**The results produced by existing methods and our proposed method are shown in the second and third rows, respectively. Our approach allows for the customization of both subject identity and motion patterns, enabling the generation of desired 4D objects based on contextual descriptions.

***Index Terms*—Text-to-4D Generation, Decoupled Training Pipeline, Motion Pattern Optimization, Spatial-Temporal Slicing**

I. Introduction

Generative models have recently achieved remarkable advancements, bringing transformative changes to the fields of image, video, and 3D generation [1], [2]. Building on these advancements, text-to-4D generation, which aims to create four-dimensional (3D space + time) dynamic scenes or objects from input prompts, has shown huge potential to study. Leveraging advanced diffusion models, current methods for 4D generation have demonstrated impressive efficacy. This breakthrough can potentially revolutionize dynamic scene simulations, animation, and the creation of entire virtual worlds.

Universal methods [3]–[5] typically aim for generalized 4D generation but often struggle with limited motion diversity and lack of customization as Fig. 1 shows. A key challenge lies in generating customized 4D content due to the limited availability of 4D datasets and the inherent complexity of modeling both temporal and spatial dimensions for specific objects. On the one hand, creating a comprehensive 4D customization dataset requires considering individual variations, which is both resource-intensive and time-consuming. On the other hand, dynamic 3D scenes involve diverse spatial content coupled with intricate temporal dynamics. Effective space-time modeling must simultaneously capture detailed spatial information (such as geometry and surface texture) and temporal changes (such as object movements and deformations), further complicating the modeling process.

Current existing methods blend gradient updates from multiple pre-trained diffusion models and synthesize 4D scenes. For example, the pioneer work, MAV3D [6], leverages text-to-image, text-to-video, and 3D-aware text-to-image, generating customized static 3D object first and then introducing time dimension for dynamic scenes(3D space + time) step by step. Furthermore, 4d-fy [7] noticed that this direct combination strategy shows opposing weakness. It develops a three-way trade-off method for introducing a hybrid SDS, aiming to synthesize 4D scenes using the best qualities of each diffusion model. However, incorporating motion into a static 3D scene using SDS with a text-to-video model typically degrades the 3D structure relative to static scenes generated by text-to-3D models. Therefore, this weakness leads us to a question: how can we harness an existing text-to-3D model's knowledge about 3D structure and appearance while augmenting them with new, custom motions?

We propose a novel 4D content customization model, Motion4D, to fully capture both the spatial and temporal features for consistent and diversity dynamic generation. In contrast to previous cascaded methods, we designed a spatial-temporal slicing strategy, achieving motion customization across frames and geometry consistency of space dimension simultaneously rather than generating a static 3D object first and then animating it progressively. Firstly, we take dual slicing steps both in time and space dimensions for the initialized 4D Gaussian field, where time slicing refers to dividing a dynamic scene into multiple static 3D scene frames along time dimensions and spatial slicing represents generating multi-view images from different perspectives to represent the appearance and geometric features of the object. Subsequently, some real motion videos are selected for motion customizing tuning by optimizing cross-attention layers parameters of UNet blocks, aiming to generate motion patterns assisted with target pattern. This dual-slicing multi-view framework is efficient in capturing object features and motion features with shared-weighted

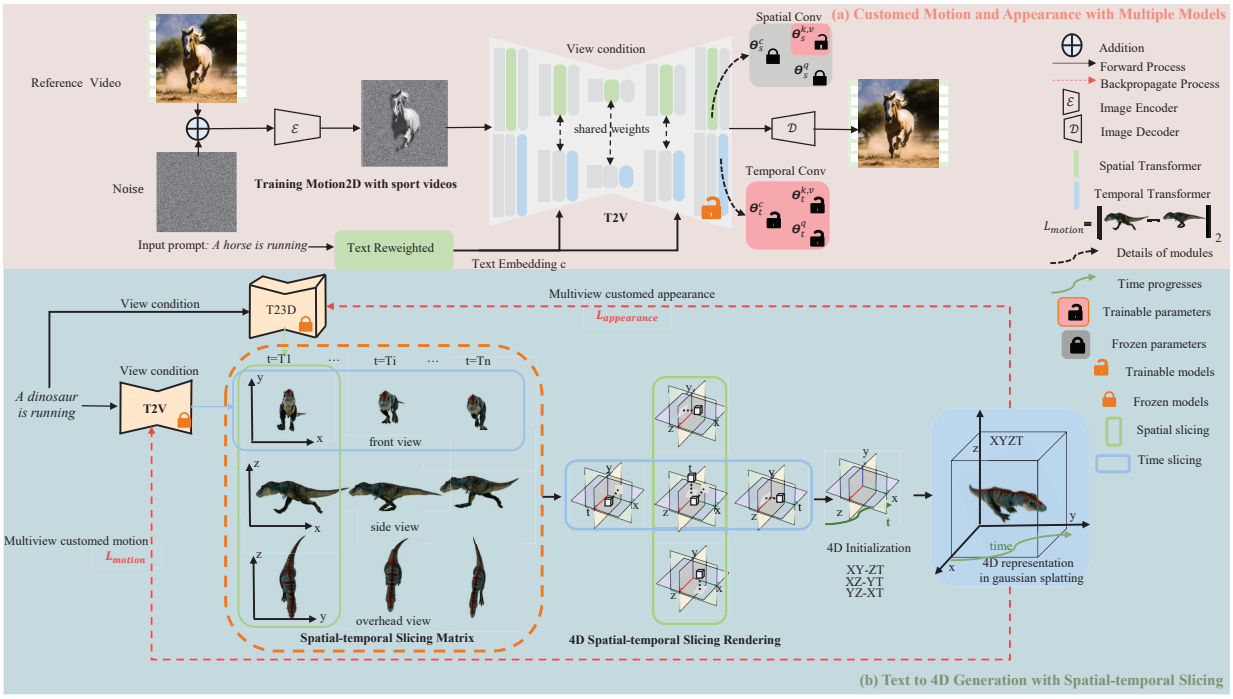

Fig. 2: **The diagram illustrates our proposed framework for 4D customized motion and appearance from input text using spatial-temporal slicing techniques.**

parameters. To summarize, our main contributions are as follows: (1) We propose Motion4D, a novel 4D content customization model, able to maintain the variety of motion generation in case of generating a specific 4D scene. (2) We present a spatial-temporal slicing training strategy that could accurately capture the movement trajectory and deformation of objects at each time point while maintaining 3D geometry and appearance consistency. (3) We conduct extensive qualitative and quantitative experiments, demonstrating the superiority of Motion4D over the existing state-of-the-art methods. Our project is displayed on https://zhangzhichao19020123.github.io/motion4d1/

II. Methods

A. Multiview Appearance Customization

Current methods for 3D appearance customization mainly focus on converting 2D images into 3D models using pretrained models. However, the generated 3D models often exhibit geometric inconsistencies. To address this, we incorporate fine-tuning the T2I model directly into 3D diffusion models to achieve multiview customizable appearance. Specifically, we fine-tune the self-attention layers of the Transformer in the UNet structure of MVDream [8] to customize the appearance. MVDream's pseudo-3D structure naturally integrates four orthogonal views (front, back, left, right) by merging 2D images from each view. This allows for spatial layout learning directly across these views, enabling the extraction, local editing, and optimization of spatial features and attributes, resulting in highly consistent multiview customizable 3D appearances.

Text-to-appearance consistency is foundational. Initially, text embedding handles the task of averaging word attribute representations. For customized appearances or to emphasize specific attributes, text reweighting is essential to highlight particular characteristics. This leads to the introduction of a weight adjustment strategy, as outlined in Equation 1. In the process of three-dimensional customization, viewpoint consistency must also be addressed to prevent issues such as multi-head artifacts or color shifts. Fine-tuning techniques are applied across all three planes to customize the appearance, and a viewpoint-sensitive 3D diffusion model is used to ensure consistency across different perspectives.

$$\begin{aligned} &\hat{W} = \arg\min_{W} ||WC_{\text{reg}_i}^{\top} - W_0 C_{\text{reg}_i}^{\top}||_F \\ &\text{s.t.} WC^{\top} = P, \text{ where } C = [\mathbf{c}_1 \cdots \mathbf{c}_N]^{\top} \\ &\quad \text{and } P = [W_1\mathbf{c}_1^{\top} \cdots W_N\mathbf{c}_N^{\top}]^{\top}, i = z, x, y. \end{aligned} \tag{1}$$

B. Multiview Motion Customization

Current video generation approaches [9], [10] primarily utilize large-scale text-video pairs to train models, ensuring text-motion consistency. However, this often results in randomly selected views, limited motion diversity, and suboptimal video quality [11]. To overcome this, we apply a multiview generation method. Using specific action videos, we fine-tune the UNet structure of Zeroscope by applying cross-attention layer fine-tuning in two stages to customize both the motion and the camera view (front, back, left, and right views bound to text input). Then, we generate customized videos from the three orthogonal views produced by the T23D model, ensuring strong geometric consistency across views in the generated customizable motions.

Adapting Equation 2 for time slicing, we apply a strategy of fine-tuning motion models across three views—front, side, and overhead—each corresponding to a unique orthogonal projection. These views are defined as:

$$\begin{cases} P_z(u(t)) = (u_x, u_y)^T + (t - u_t)\frac{V}{W} \\ P_x(u(t)) = (u_y, u_z)^T + (t - u_t)\frac{V}{W} \\ P_y(u(t)) = (u_x, u_z)^T + (t - u_t)\frac{V}{W} \end{cases} \tag{2}$$

Here, $P_z(u(t))$, $P_x(u(t))$, and $P_y(u(t))$ represent the front, side, and overhead projections of the object's motion, respectively. The term $u(t)$ is the object's positional vector at time t, and the matrix V/W defines the relationship between spatial and temporal dimensions.

The objective of this step is to minimize the difference between the predicted noise ϵ^{pr} and the initial noise ϵ^{gt} during the fine-tuning process. The multi-view fine-tuning motion loss $\mathcal{L}$ is formulated as:

$$\begin{aligned}\mathcal{L}_i = \mathbb{E}_{Z_t,\epsilon_t^{gt}\sim\mathcal{N}(0,1),\hat{t}}\Big[\left\|\epsilon_t^{gt}-\epsilon_t^{pr}\right\|_2^2\Big],\\ i = P_z(u(t)),\ P_x(u(t)),\ P_y(u(t))\end{aligned} \tag{3}$$

This loss function ensures that the generated motion matches the ground truth motion at each timestep for all three views. By calculating the L_2 norm between the predicted noise and ground truth, the model is optimized to produce accurate motion dynamics across all views. By adopting this strategy, the model learns to generalize motion patterns across multiple views, while still retaining the ability to generate high-quality, fine-tuned dynamic videos. This ensures that both motion and appearance are preserved, enabling high-fidelity motion customization in 4D video generation.

During the training phase, the model primarily focuses on capturing common motion patterns in D^m, without heavily focusing on appearance or the specific subject presenting the motion. This is achieved by leveraging the denoising process inherent in diffusion models. In these models, Gaussian noise is sampled at various timesteps, and progressively removed to recover the underlying motion. Early denoising steps significantly impact the dynamic structure of the video, while later steps add finer detail. To prioritize learning dynamic motion patterns over visual details such as background or subject appearance, we define a timestep sampling strategy that biases the training process towards earlier denoising steps. Unlike traditional approaches that uniformly sample timesteps for denoising, we define a probability distribution over the timesteps to emphasize earlier stages:

$$f_\alpha(t) = \frac{1}{T}(1-\alpha\cos(\frac{\pi t}{T})) \tag{4}$$

This function $f_\alpha(t)$ defines a cosine distribution over timesteps, where α controls the weighting towards earlier steps in the denoising process. This approach ensures that the model focuses on the overall dynamic structure of the video, as represented in the early stages of denoising, rather than the fine details that are recovered later.

C. 4D Spatial Temporal Slicing

In text-to-4D space generation, existing methods either use NeRF rendering [12]–[16] (with 4D representations encoded using hash functions, resulting in long backpropagation times, slow training, limited motion range, or scene distortion) or Gaussian deformation fields [17]–[20](which are faster for object formation but fail to render 4D dynamic objects correctly under large motions and drastic camera view changes, often leading to severe structural distortions). To resolve these issues, we introduce temporal slicing in 4DGS (4D Gaussian Splatting), which allows rendering in 2D slices across the 4D space, enabling highly customizable large-scale 4D space expression.

However, a new challenge arises: there are no highly consistent spatio-temporal slices available for rendering. Therefore, we develop a spatial-temporal slicing matrix to serve as a suitable representation for 4DGS slicing. Additionally, we design a 4D consistency loss to optimize and adjust the 4D space. Specifically, we initialize the columns of the matrix as spatial slices using the orthogonal view images from the T23D model and the rows as temporal slices using video frames obtained from fine-tuned views. The combination of temporal and spatial slices forms the spatial-temporal slicing matrix, which is aligned with the dimensions of 4DGS. The matrix is then fed into a 4D spatial-temporal feature extraction network for iterative training, ultimately producing a consistent and accurate 4D dynamic scene representation.

In the process of converting the 4D representation into 3D projections, we formalize the slicing operation using Gaussian splatting. Starting from the 4D covariance matrix Σ_{4D}, we define the following relationship:

$$\Sigma_{4D} = \begin{pmatrix}\mathbf{U} & \mathbf{V}\\ \mathbf{V}^T & \mathbf{W}\end{pmatrix} \text{ and } \Sigma_{4D}^{-1} = \begin{pmatrix}\mathbf{A} & \mathbf{M}\\ \mathbf{M}^T & \mathbf{Z}\end{pmatrix}, \tag{5}$$

Here, $\mathbf{U}$ and $\mathbf{A}$ are 3×3 matrices that describe the spatial variance, while $\mathbf{V}$ captures the interaction between spatial and temporal dimensions. For a given time t, the projected 3D Gaussian is computed as:

$$G_{3D}(\mathbf{x},t) = \mathrm{e}^{-\frac{1}{2}\lambda(t-\mu_t)^2}\mathrm{e}^{-\frac{1}{2}[\mathbf{x}-\mu(t)]^T\Sigma_{3D}^{-1}[\mathbf{x}-\mu(t)]}, \tag{6}$$

where the spatial and temporal components are decoupled, and the time evolution is controlled by the temporal decay term $\mathrm{e}^{-\frac{1}{2}\lambda(t-\mu_t)^2}$. The 3D covariance matrix Σ_{3D} and the time-dependent mean $\mu(t)$ are given by:

$$\begin{aligned}\lambda = \mathbf{W}^{-1}, \Sigma_{3D} = \mathbf{A}^{-1} = \mathbf{U} - \frac{\mathbf{V}\mathbf{V}^T}{\mathbf{W}},\\ \mu(t) = (\mu_x,\mu_y,\mu_z)^T + (t-\mu_t)\frac{\mathbf{V}}{\mathbf{W}}.\end{aligned} \tag{7}$$

Compared to the original 3D Gaussian Splatting (3DGS) method, the sliced 3D Gaussian in Equation 7 includes a temporal decay term $\mathrm{e}^{-\frac{1}{2}\lambda(t-\mu_t)^2}$. As time t progresses, a Gaussian point becomes visible when t is near its temporal position μ_t, gradually increasing in opacity until reaching its peak at $t=\mu_t$. It then decreases in density, vanishing when t is sufficiently far from μ_t. Controlling the temporal position and scaling factor allows a 4D Gaussian to model complex dynamics effectively, such as motions that appear or disappear suddenly. During rendering, temporally distant points are filtered out, with the visibility threshold $\lambda(t-\mu_t)^2$ empirically set to 16.

III. Experiment

We implement Motion4D under the 4D Gaussian Splatting framework. For motion slices finetuning, we regard the ZeroScope T2V diffusion model [21] as pretrained model, with multi-view resolutions for the resolutions of 512 $\times$ 512, using a batch size of 16 on a single Nvidia A100 80GB GPU. For spatial slices, we optimize the partial parameters of UNets for Zero-1-to-3-XL [22], optimizing the model for an additional 5,000 iterations. The Adam optimizer, with a learning rate of 0.001, was used throughout all stages.

To evaluate our approach, we used Fréchet Video Distance (FVD) [23], T3Bench [24]. We also use four qualitative metrics by asking human raters their preferences based on:(1)3D appearance(3D-A), (2)3D text alignment(3D-T), (3)motion text alignment(MT), and (4)motion realism(MR). A total of 200 questionnaires were distributed to gather comprehensive feedback on the generated 4D objects.

A. Comparative Experimental Results

We compared our method against state-of-the-art models using the same input prompt, as illustrated in Fig. 3. The results demonstrate significant differences in how each method responds to the prompt "a dinosaur is running." The outputs from Dreamgaussion4D and

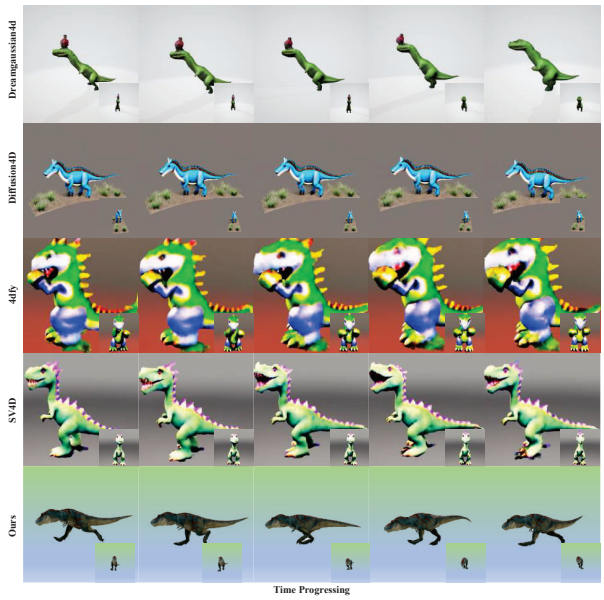

Fig. 3: **Visual comparisons of 4D generation methods with the input prompt "A dinosaur is running".**

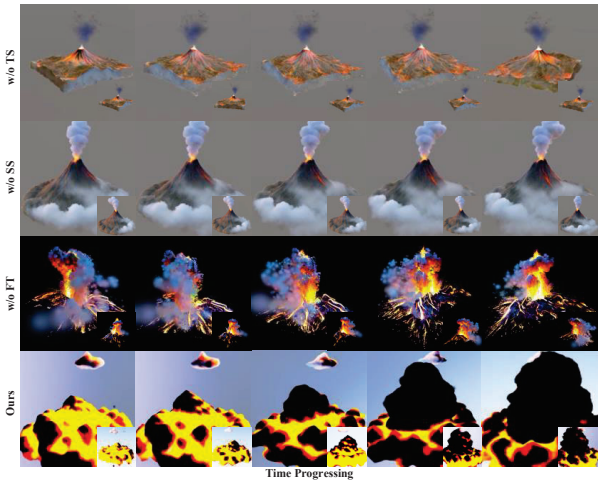

Fig. 4: **Ablation experiment results for the 4D generation of "Cloud rolls and volcano erupts"**

Diffusion4D exhibit simple, repetitive motions with a lack of fluidity, which hinders their ability to achieve high levels of customization. While 4dfy offers a distinctive stylization, the generated motions suffer from low detail resolution, distortion, and insufficient continuity. SV4D shows improvements in both smoothness and motion diversity, suggesting some potential for customization. In contrast, our proposed method outperforms the others, producing smooth, natural dinosaur movements with finely detailed and realistic actions. Its robust customization capabilities make it particularly well-suited for applications requiring high-precision dynamic representations. As shown in Table 1, our method also achieves the highest scores, particularly in MT and MR metrics, further highlighting its strength in generating both natural and customizable motion.

TABLE I: Quantitative comparison with 4D generation methods.

Combined Method	FVD↓	Text to 3D		User Preference	
		3D-A↑	3D-T↑	MT↑	MR↑
DreamGaussian4D [25]	71.25	56.74	47.36	23.90	26.57
Diffusion4D [26]	61.07	32.50	42.59	32.60	35.89
4dfy [7]	54.41	26.05	34.95	43.31	57.41
SV4D [27]	60.15	51.79	42.47	51.40	56.34
Ours	**40.12**	**57.13**	**59.28**	**63.99**	**62.58**

B. Ablation Experimental Results

TABLE II: Ablation study of various rendering methods with/without motion control.

Method	FVD↓	Text to 3D		User Preference	
		3D-A↑	3D-T↑	MT↑	MR↑
w/o TS	48.54	46.78	54.79	55.94	55.02
w/o SS	52.07	50.76	52.31	53.57	58.12
w/o FT	44.62	55.54	55.88	60.84	61.26
Ours	**42.82**	**56.25**	**58.24**	**61.21**	**65.72**

We provide an in-depth analysis of our temporal-spatial slicing training strategy through an ablation study, removing each component individually. The results of this study are presented in Fig. 4 and Table 2. To evaluate the contribution of each component, we conduct experiments by removing temporal slicing (w/o TS), spatial slicing (w/o SS), and fine-tuning (w/o FT), and compare the results to our full method. Fig. 4 shows the effect of omitting each component in the 4D generation process using the input prompt "cloud rolls and volcano eruptions."

Without the TS module, we observed less coherent motion trajectories and reduced fluidity in the first row compared to our full model. Additionally, removing the spatial slicing module notably diminished the level of detail, especially in the smoke dispersion and volcanic eruptions. When the fine-tuning strategy is excluded from the training process, the detail in the volcanic eruptions and smoke generation slightly decreased in both overall motion and 3D appearance. Overall, our method demonstrates superior performance in enhancing the dynamic realism and complexity of 4D scene generation. This is further supported by the quantitative results in Table 2, where the superior performance of our approach is clearly verified.

IV. Conclusion

We propose Motion4D, a novel framework for generating customized text-to-4D outputs. To achieve greater precision in generation, we employ temporal-spatial slicing techniques to accurately capture motion features of objects at various time slots, while utilizing multi-view images to capture the geometric structure and appearance features of objects. This approach enhances the specificity of both motion and appearance adjustments. By splitting 4D scenes into smaller, more manageable temporal and spatial slices, we reduce the dependence on large amounts of 4D datasets, improve processing efficiency, and offer better control over object behavior in both time and space. This ultimately leads to more precise fine-tuning of motion and appearance.

V. Acknowledgments

This research was supported by the Hunan Province Graduate Student Innovation Project. We gratefully acknowledge the funding provided by Project XJQY2024040,Project XJZH2024038 and QL20220009.

REFERENCES

[1] Z. Zhang, H. Chen, X. Yin, and J. Deng, "Eawnet: An edge attention-wise objector for real-time visual internet of things," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 7258649, 2021.

[2] Z. Zhang, H. Chen, X. Yin, J. Deng, and W. Li, "Dynamic selection of proper kernels for image deblurring: a multistrategy design," *The Visual Computer*, vol. 39, no. 4, pp. 1375–1390, 2023.

[3] Z. Yang, Z. Pan, C. Gu, and L. Zhang, "Diffusion2: Dynamic 3d content generation via score composition of orthogonal diffusion models," *arXiv preprint arXiv:2404.02148*, 2024.

[4] Q. Sun, Z. Guo, Z. Wan, J. N. Yan, S. Yin, W. Zhou, J. Liao, and H. Li, "Eg4d: Explicit generation of 4d object without score distillation," *arXiv preprint arXiv:2405.18132*, 2024.

[5] H. Zhang, X. Chen, Y. Wang, X. Liu, Y. Wang, and Y. Qiao, "4diffusion: Multi-view video diffusion model for 4d generation," *arXiv preprint arXiv:2405.20674*, 2024.

[6] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson *et al.*, "Text-to-4d dynamic scene generation," *arXiv preprint arXiv:2301.11280*, 2023.

[7] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell, "4d-fy: Text-to-4d generation using hybrid score distillation sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7996–8006.

[8] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "MVDream: Multi-view diffusion for 3d generation," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=FUgrjq2pbB

[9] X. Shi, Z. Huang, F.-Y. Wang, W. Bian, D. Li, Y. Zhang, M. Zhang, K. C. Cheung, S. See, H. Qin *et al.*, "Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[10] Q. Zuo, X. Gu, L. Qiu, Y. Dong, Z. Zhao, W. Yuan, R. Peng, S. Zhu, Z. Dong, L. Bo *et al.*, "Videomv: Consistent multi-view generation based on large video generative model," *arXiv preprint arXiv:2403.12010*, 2024.

[11] Z. Zhang, J. Liao, M. Li, L. Qin, and W. Wang, "Tora: Trajectory-oriented diffusion transformer for video generation," *arXiv preprint arXiv:2407.21705*, 2024.

[12] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen, "Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6517–6526.

[13] S. Huang, S. Sun, Z. Wang, X. Qin, Y. Xiong, Y. Zhang, P. Wan, D. Zhang, and J. Jia, "Placiddreamer: Advancing harmony in text-to-3d generation," *arXiv preprint arXiv:2407.13976*, 2024.

[14] K. Xie, J. Lorraine, T. Cao, J. Gao, J. Lucas, A. Torralba, S. Fidler, and X. Zeng, "Latte3d: Large-scale amortized text-to-enhanced3d synthesis," *arXiv preprint arXiv:2403.15385*, 2024.

[15] L. Jiang and L. Wang, "Brightdreamer: Generic 3d gaussian generative framework for fast text-to-3d synthesis," *arXiv preprint arXiv:2403.11273*, 2024.

[16] F. Liu, H. Wang, W. Chen, H. Sun, and Y. Duan, "Make-your-3d: Fast and consistent subject-driven 3d content generation," *arXiv preprint arXiv:2403.09625*, 2024.

[17] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen, "4d-rotor gaussian splatting: Towards efficient novel view synthesis for dynamic scenes," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[18] Y. Wang, X. Wang, Z. Chen, Z. Wang, F. Sun, and J. Zhu, "Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels," *arXiv preprint arXiv:2405.16822*, 2024.

[19] F. Li, H. Zhang, and N. Ahuja, "Self-calibrating 4d novel view synthesis from monocular videos using gaussian splatting," *arXiv preprint arXiv:2406.01042*, 2024.

[20] D. Li, S.-S. Huang, Z. Lu, X. Duan, and H. Huang, "St-4dgs: Spatial-temporally consistent 4d gaussian splatting for efficient dynamic scene rendering," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[21] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[22] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309.

[23] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Fvd: A new metric for video generation," *arXiv preprint arXiv:2106.09685*, 2019.

[24] Y. He, Y. Bai, M. Lin, W. Zhao, Y. Hu, J. Sheng, R. Yi, J. Li, and Y.-J. Liu, "T3bench: Benchmarking current progress in text-to-3d generation," *arXiv preprint arXiv:2310.02977*, 2023.

[25] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu, "Dreamgaussian4d: Generative 4d gaussian splatting," *arXiv preprint arXiv:2312.17142*, 2023.

[26] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei, "Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models," *arXiv preprint arXiv:2405.16645*, 2024.

[27] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani, "Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency," *arXiv preprint arXiv:2407.17470*, 2024.