Sketch-2-4D: Sketch Driven Dynamic 3D Scene Generation

Guo-Wei Yang
Tsinghua University
ygw19@mails.tsinghua.edu.cn

Dong-Yu Chen
Tsinghua University
dy-chen20@mails.tsinghua.edu

Tai-Jiang Mu
Tsinghua University
taijiang@tsinghua.edu.cn

Abstract

Sketch-based content generation offers flexible controllability, making it a promising narrative avenue in film production. Directors often visualize their imagination by crafting storyboards using sketches and textual descriptions for each shot. However, current video generation methods suffer from three-dimensional inconsistencies, with notably artifacts during large motion or camera pans around scenes. A suitable solution is to directly generate 4D scene, enabling consistent dynamic three-dimensional scenes generation. We define the Sketch-2-4D problem, aiming to enhance controllability and consistency in this context. We propose a noval Control Score Distillation Sampling (SDS-C) for sketch-based 4D scene generation, providing precise control over scene dynamics. We further design Spatial Consistency Modules and Temporal Consistency Modules to tackle the temporal and spatial inconsistencies introduced by sketch-based control, respectively. Extensive experiments have demonstrated the effectiveness of our approach.

***Keywords:** sketch driven generation, 4D generation, dynamic 3D scene, diffusion model.*

1. Introduction

In recent years, the rapid development of diffusion-based generative techniques has revolutionized the quality of image generation, reshaping the landscape of the art and design industries. Beyond static imagery, the spotlight has shifted towards video generation, with research efforts such as [7, 26, 13] and commercial software like GEN-2 [1] achieving photo-realistic quality. However, a challenging problem plaguing current video generation methods is their inability to maintain spatial consistency. This limitation becomes apparent when large movements or camera pans occur, resulting in noticeable artifacts within the generated videos.

To address this critical issue and pave the way for high-quality video generation, incorporating 3D priors into the generation is essential. One promising avenue is the direct creation of dynamic 3D scenes, referred to as 4D scenes, which can render high-quality videos with appropriate camera paths. In the realm of filmmaking, directors often start their creative journey by crafting storyboard scripts. These scripts are instrumental in governing the content and framing of each shot in a film. The storyboard typically consists of both visual sketches and accompanying textual descriptions for every scene or shot (see Fig. 2). This creative practice serves as the foundation upon which the entire cinematic narrative is meticulously planned and executed, ensuring that the director's vision is delivered and realized on screen effectively.

Inspired by the practise from film industry, this paper introduces the novel concept of "sketch-2-4D," pioneering a path towards the realization of dynamic 3D scenes driven by sketches and textual descriptions. A straightforward approach to supervise the generation of 4D scenes through sketches is to directly supervise the 4D scene using the corresponding mask derived from the sketch. However, this method falls short in providing detailed control over the internal intricacies of the 4D model. Recent advances, such as ControlNet [30], have proposed innovative strategies to extend the potential of Stable Diffusion. By introducing additional inputs like Canny edges, hand-drawn sketch, human key points, and segmentation maps, ControlNet significantly enhances the controllability of the m odel. In the context of 4D scene generation, we aim to harness this newfound control capability. It's worth noting that directly supervising the rendered image using ControlNet's outputs often leads to severe inconsistencies and overfitting. To achieve more precise control over 4D scene generation, inspired by the concepts introduced in DreamFusion [20], we present a novel approach known as Control Score Distillation Sampling (SDS-C), which serves as an extension of SDS specifically tailored for pretrained ControlNet.

The inherent challenge in sketch-based 4D scene generation lies in the fact that the 4D scenes conditioned by sketches may be outside the conventional domain of 4D generation, potentially leading to temporal and spatial inconsistencies. Considering that the Text-to-Video (T2V) models [7] trained on vast video datasets, are capable of learning three-dimensional priors from object motions, rotations, and camera movements, we propose a T2V-based Spatial Consistency Module to enhance the spatial coher-

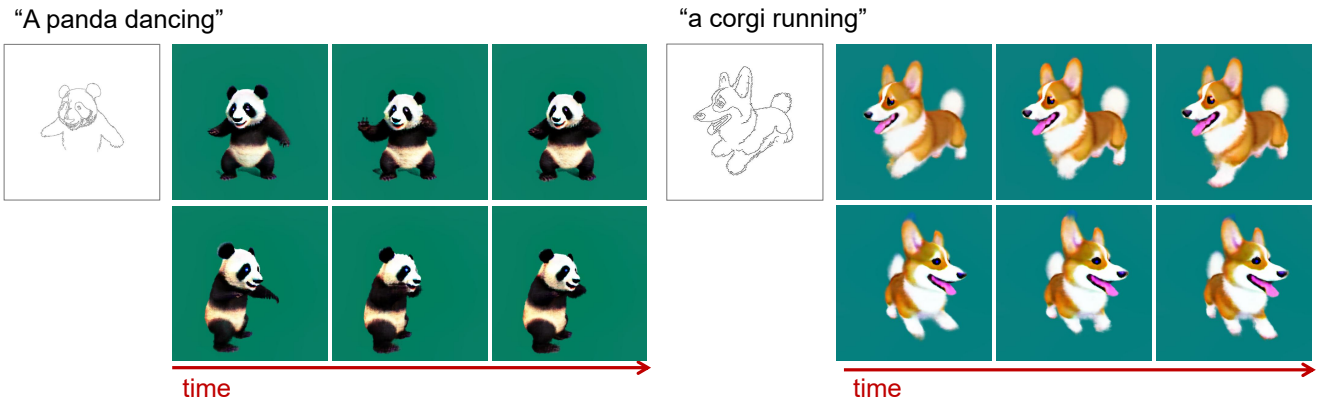

Figure 1. We present Sketch-2-4D, a method for generating dynamic 3D scenes (4D) based on sketch and text. The figure shows two examples of 4D scenes generated by our method, with different columns of the rendered results indicating different times and different rows indicating different viewpoints.

ence of the generated 4D scenes. In addition to spatial consistency, the temporal consistency ensures that the generated 4D scenes flow smoothly through time. We further design a Temporal Consistency Module, which ensures both geometric and semantic consistency over time by imposing margined dice loss on geometry and feature loss on rendered image, respectively.

In summary, our work makes the following main contributions:

- We pioneered the conceptualization and definition of the sketch-to-4D problem and successfully harnessed Control Score Distillation Sampling (SDS-C) to exert precise control over 4D scene generation by drawing sketches.
- We designed the Spatial Consistency Module and the Temporal Consistency Module to ensure the temporal and spatial consistency for the sketch-to-4D task.
- We propose staged training for consistent high-quality completion of sketch-2-4D tasks and experimentally demonstrate the effectiveness of our approach.

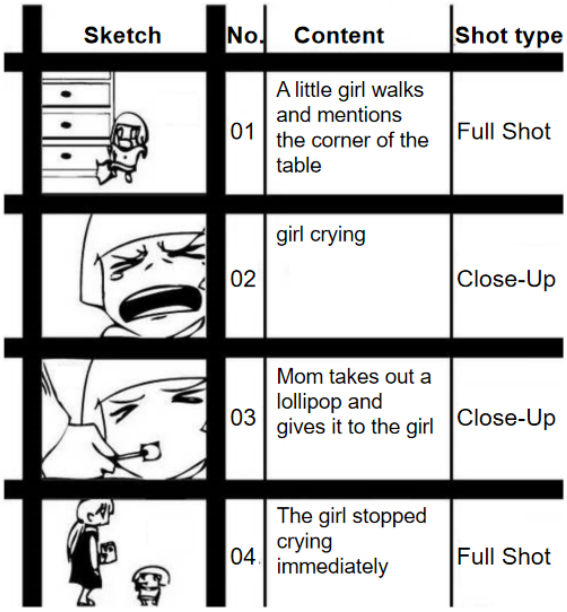

Figure 2. An example of storyboards for a lollipop advertisement from internet. Each line represents a shot, and the information of each shot includes sketch, textual descriptions, shot type, etc.

2. Related work

2.1. Diffusion-based Image and Video Generation

Many recent works on Diffusion- based image generation have produced high-quality and creative results. [4] proposes to use diffusion model to generate images, surpassing previous GAN-based methods [5, 9]. Latent Diffusion Models [23] uses VAE to encode the image into the hidden space and performs diffusion generation in the hidden space to improve the quality of the generation. Stable Diffusion [28] open-sources a pre-trained model which was trained on a huge amount of image data. ControlNet [30] uese Canny edges, human key points, segmentation maps, etc., as additional inputs to Stable Diffusion to extend the potential of Stable Diffusion and significantly improves the controllability.

Video generation based on Diffusion has also emerged [7, 26]. [12] combines diffusion model and transformer to capture temporal correlation. VideoFusion [13] proposes to capture information across video frames by sharing inter-frame noise, thereby enhancing

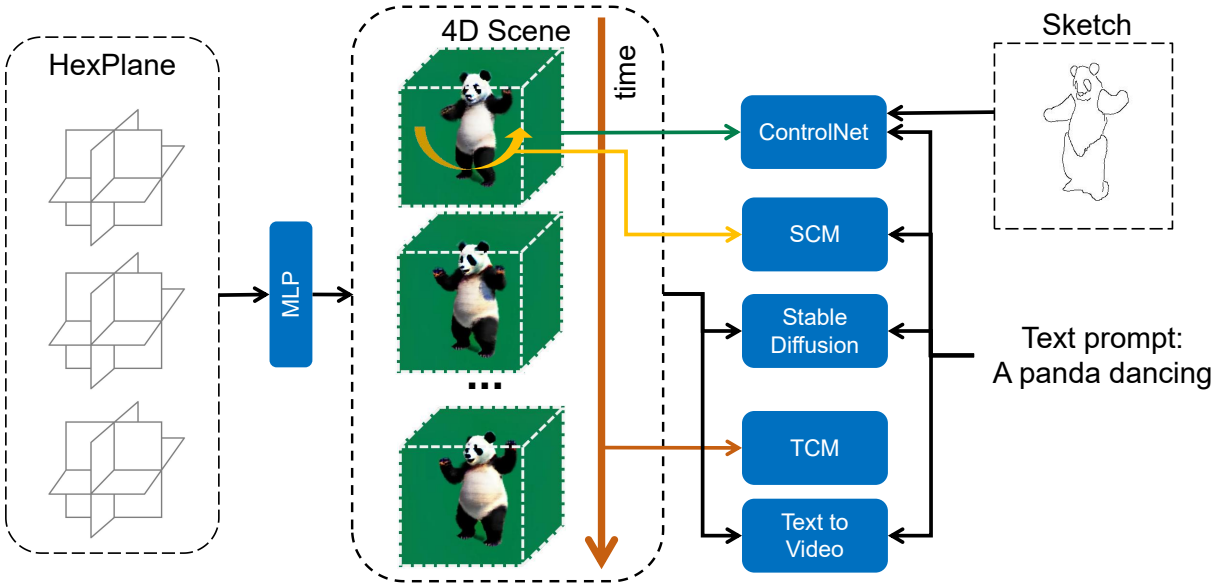

Figure 3. Pipeline of sketch-2-4D. We generate 4D scenes based on HexPlane [2] and add sketch constraints to the 4D scenes by ControlNet-based SDS-C loss. Spatial Consistency Module (SCM) and Temporal Consistency Module (TCM) are used to ensure spatial and temporal consistency.

the temporal consistency of the generated video. Some commercial software [1] have achieved good video generation effects by training on massive video data. However, existing video generation methods suffer from three-dimensional inconsistency. When objects in the scene have large movements or there are mirrors rotating around the objects, obvious artifacts will be observed. Scene generation based on 4D can solve this problem.

2.2. Static 3D Scene Generation

Recently, with the development of generation techniques, many 3D generation works for static scenes have been proposed. Such as three-plane representation and GAN-based generation and editing work [3, 10]. GAN-based static 3D generation and editing work are usually designed for single-category object generation such as faces, vehicles, etc., due to its unstable training and lack of 3D data.

Due to the versatile and creative generation effect of Stable Diffusion [28], the application of Diffusion techniques to static 3D scene generation has attracted widespread attention. Some of the work based on Diffusion directly learn and generate 3D scenes [8]. [25, 6] generate a scene by representing the 3D scene as a three-plane and learn the three-plane representation directly from a large amount of 3D data. [17] direct generates 3D scenes through point cloud diffusion.

However, due to the lack of 3D data and the huge computation for 3D generation training, it is difficult to achieve diverse and fine-grained results for methods directly based on 3D diffusion. DreamFusion [20] proposes to obtain a 2D prior from pre-trained image Stable Diffusion via score distillation sampling (SDS) loss and supervises the 3D scene generation by rendering viewpoints. Zero-1-to-3 [11] proposes single-image 3D reconstruction based on SDS supervision and Stable Diffusion with viewpoint condition. ProlificDreamer [29] proposes variational score distillation (VSD) loss to solve the problems of oversaturation, oversmoothing and low diversity based on SDS loss.

2.3. Dynamic 3D Scene Reconstruction and Generation

Neural implicit rendering methods based on NeRF [14, 16] have achieved photorealistic results in novel view synthesis for static scenes. For dynamic scene view synthesis, methods such as [21, 18] learn 3D deformation transformations from canonical space. HexPlane [2] represents dynamic 3D scene features by projecting them onto six feature planes. [19] proposes dynamic 3D scene representation by interpolate feature vectors in the time axis.

MAV3D [27] proposes to represent the 3D scene by Multi-Resolution HexPlane and supervises the dynamic scene 3D generation by frozening T2V model based on temporal Score Distillation Sampling (SDS-T). However, this method cannot control the generated dynamic 3D scene flexibly through inputs such as sketch. Control4D [24] proposes a 4D portrait editing method, which uses ControlNet as image editor and Dataset Update (DU) as the training. The method relies on the reconstructed dynamic 3D scene

for editing, and it is difficult to modify and control the motion.

3. Method

In this section, we first formulate the problem (Sec.3.1) for sketch driven 4D generation. Then we introduce our method of constraining the sketch for 4D scenes using control Score Distillation Sampling loss (Sec.3.2). Unlike text-to-4D, the 4D scene corresponding to sketch may be outside the domain of conventional 4D generation, leading to 3D inconsistency and temporal inconsistency. We thus propose to enhance spatial consistency by Spatial Consistency Module (Sec.3.3) and temporal consistency by Temporal Consistency Module (Sec.3.4). Finally, we describe our overall process for phased training(Sec.3.5).

3.1. Formulation

We will generate a 4D scene representation F_θ from the supplied prompt c_t, the desired scene sketch $\hat{S}$, and the sketch's corresponding mask $\hat{M}$.

$$F_\theta : (\mathbf{x}, \mathbf{d}, t) \rightarrow (c, \sigma) \tag{1}$$

where c and σ are color and density of rendering sample points with coordinates $\mathbf{x}$ and sampling direction $\mathbf{d}$ at time t.

Following the rendering approach R of NeRF [14], an image $I_{\theta,C,t} = R(F_\theta, C, t)$ can be rendered based on a 4D scene F_θ at time t with camera parameter C. For the sketch-2-4D task, we want the rendered image I_{θ,C_S,t_S} at time t_S from viewpoint C_S to be conformed to the conditional controls of sketch $\hat{S}$, mask $\hat{M}$ and prompt.

3.2. Sketch Driven 4D Generation

A straightforward way to generate 4D scene conditioned on sketches is to directly supervise the rendered mask M_{F_θ,C_S,t_S} of the 4D scene F_θ from previous text-to-4D approach [27] by the sketch counterpart $\hat{M}$ using the binary cross-entropy (BCE) loss:

$$L_{\text{mask}} = BCE(\hat{M}, M_{F_\theta,C_S,t_S}) \tag{2}$$

However this approach fails to control the internal details of the 4D model through sketch(see Fig.4). Recently, ControlNet [30] proposes to use Canny edges, hand-drawn sketch, human key points, segmentation maps as additional conditions to Stable Diffusion, in order to extend the potential of Stable Diffusion, achieving more flexible controllability. We expect to exploit this control capability in 4D generation, and a straightforward way to do so is to supervise the rendered image I_{θ,C_S,t_S} via L2Loss.

$$L_{\text{image}} = L2(I_{\theta,C_S,t_S}, Ctl(\hat{S})) \tag{3}$$

where $Ctl(\hat{S})$ is the generated image obtained by sketching $\hat{S}$ into ControlNet. Whereas supervising directly through ControlNet output images leads to serious inconsistencies and overfit(see Fig.4). Inspired by DreamFusion [20], we propose control Score Distillation Sampling (SDS-C), which is an extension of SDS for pretrained ControlNet.

To achieve sketch-based controllability, we use a pretrained ControlNet conditioned on Canny edge. It is noteworthy that both Canny edge and hand-drawn sketch are types of sketches. ControlNet demonstrates equally effective generative results when utilizing either of these inputs as conditions. Here we opt to employ Canny edge as our chosen condition. Specifically, we adopt ControlNet's image encoder E to extract the feature image $I' = E(I)$ of image I, and use ControlNet's denoiser U-Net, with the denoised feature image $O(I\theta, \tau, \epsilon)$ and sketch $\hat{S}$ as inputs, to predict the noise $\hat{\epsilon}(O(I\theta, \tau, \epsilon), \tau, \hat{S})$. Here, ϵ is the feature image after noise addition, conforming to a normal distribution and τ is the diffusion time step; $O(I'_\theta, \tau, \epsilon)$ is the denoised feature image of I' and ϵ at diffusion time step τ. From this we can define the gradient of the SDS-C loss $L_{\text{SDS-C}}$ of I_{θ,C_S,t_S} and optimise the 4D scene parameters θ by $L_{\text{SDS-C}}$.

$$\bigtriangledown L_{\text{SDS-C}} = \mathbb{E}_{\tau,\epsilon}[\omega(\tau)(\hat{\epsilon}(O(I'_\theta, \tau, \epsilon), \tau, \hat{S}) - \epsilon)\frac{\partial I'}{\partial \theta}] \tag{4}$$

where ω is a weighting function as the definition of [20]. By using $L_{\text{SDS-C}}$ on I_{θ,C_S,t_S} we can optimise the 4D scene parameters θ based on sketch control.

3.3. Spatial Consistency Module

Since the 4D scene conditioned on sketch may be outside the domain of the 4D generation space, it can easily lead to temporal and spatial inconsistencies. The Diffusion-based text to video (T2V) generation method is trained with a huge amount of video and several frames are generated simultaneously to compose the video. The 3D prior can be learned from the motion of the object, in the rotation, or the motion of the lens in the video of the training data. Therefore we propose the T2V-based space consistency module to enhance the spatial consistency of 4D scenes.

Let T be a sequence of sampling times, and P be a camera path consisting of a camera pose sequence,

$$T = (t_0, t_1, ..., t_{N_T-1}) \tag{5}$$

$$P = (C_0, C_1, ..., C_{N_T-1}) \tag{6}$$

where N_T is the number of time samplings. Then we can render a image sequence $V_{\theta,P,T}$ from the 4D scene F_θ given the camera path P, time sequence T:

$$V_{\theta,P,T} = (I_{\theta,C_0,t_0}, I_{\theta,C_1,t_1}, ..., I_{\theta,C_{N_T-1},t_{N_T-1}}) \tag{7}$$

To ensure the consistency of the frames where the sketch is located, we specifically define the sketch corresponding to the time sequence $T_S = (t_S, t_S, ..., t_S)$ as a stationary time sequence of length N_T, and $P_0 = (C_S, C_1, ..., C_{N_T-1})$ is a uniformly moving camera path starting from the camera position C_S. Referring to $L_{\text{SDS-T}}$ of [27], we define the gradient of the spatial consistency loss L_{SC} to be

$$\bigtriangledown L_{SC} = \mathbb{E}_{\tau,\epsilon}[\omega(\tau)(\hat{\epsilon}(V_{\theta,P_0,T_S}, \tau) - \epsilon)\frac{\partial V_{\theta,P_0,T_S}}{\partial \theta}] \quad (8)$$

And the total Spatial Consistency Module supervises the generated results by weighting L_{mask} and L_{SC}.

$$L_S = \beta_{\text{SC}} * L_{\text{SC}} + \beta_{mask} * L_{mask} \quad (9)$$

where β_{SC} and β_{mask} are the weights corresponding to L_{SC} and L_{mask}.

3.4. Temporal Consistency Module

SDS-C based sketch control may lead to inconsistency of the 4D scene at the time t_S where the sketch is located with other times. To enhance the temporal consistency of the generated 4D scene, we design two ways to ensure the temporal consistency of the 4D scene in terms of geometry and texture, respectively.

We impose the geometric temporal consistency of the scene by considering the density of sampling points between two consecutive frames. Let $P_S = (C_S, C_S, ..., C_S)$ be the camera path of N_T camera poses that are all C_S, i.e., the camera paths fixed to the corresponding camera poses in the sketch, and $\sigma_{C,t,k}$ be the density of the k-th sampling point of the view at camera pose C and time t. Since we use uniform sampling for rendering, the coordinates of the sampling points corresponding to $\sigma_{C_S,t_i,k}$ and $\sigma_{C_S,t_{i+1},k}$ are the same. Inspired by dice loss [15] for semantic segmentation, we propose 3D dice loss for measuring the geometric similarity of two 3D scenes. Specifically, we compute dice loss using the density of sampling points at the same location in two adjacent frames:

$$D(t_i, t_{i+1}) = \frac{2 * \sum_k M(\sigma_{C_S,t_i,k}) * \sigma_{C_S,t_{i+1},k}}{\sum_k M(\sigma_{C_S,t_i,k}) + \sum_k \sigma_{C_S,t_{i+1},k}} \quad (10)$$

$$M(\sigma) = \begin{cases} 0, & \sigma < \alpha \\ 1, & \sigma \geq \alpha \end{cases} \quad (11)$$

where $M(\sigma)$ is the density mask defined by thresholding α. To avoid that 3D dice loss restricts the motion of the scene, we further propose margined 3D dice loss L_D by setting a margin β to allow for a small range of density variations:

$$L_D = \frac{1}{N_T - 1} \sum_{i=0}^{N_T-2} max(D(t_i, t_{i+1}) - \beta, 0) \quad (12)$$

We also use the density of neighbouring frames for supervision to ensure its local temporal consistency

$$L_N = \frac{1}{(N_{\text{T}} - 1) * N_k} \sum_{i=0}^{N_T-2} \sum_k (\sigma_{C_S,t_{i+1},k} - \sigma_{C_S,t_{i+1},k})^2 \quad (13)$$

where N_k is the number of sampling points.

To ensure the consistency between two frames without affecting the magnitude of the motion, we impose semantic consistency on images rendered from the same viewpoint between consecutive frames by the similarity of CLIP [22] features. The image I is fed into a pre-trained CLIP to obtain the feature vector $CL(I)$, and the semantic consistency is computed by the L2 loss of features of consecutive frames:

$$L_{clip} = \frac{\sum_{i=0}^{N_T-2} L2(CL(I_{\theta,C_S,i}), CL(I_{\theta,C_S,i+1}))}{N_T - 1} \quad (14)$$

Finally, we sum L_D and L_{clip} to get the temporal consistency loss L_T

$$L_T = \beta_D * L_D + \beta_N * L_N + \beta_{clip} * L_{clip} \quad (15)$$

where β_D, β_N and β_{clip} are the weights corresponding to L_D, L_N and L_{clip}.

3.5. Training Strategy

In the first stage of training, we start with static scenes. We use the L_{SDS} proposed in [20] to supervise the random viewpoint rendering of the image $I_{\theta,C,0}$ and use $L_{\text{SDS-C}}$ to supervise the image $I_{\theta,C_S,0}$ rendered under the camera pose C_S of sketch.

$$L_{stage1} = \alpha_{\text{SDS}} * L_{\text{SDS}} + \alpha_{\text{SDS-C}} * L_{\text{SDS-C}} \quad (16)$$

where α_{SDS} and $\alpha_{\text{SDS-C}}$ are the weights. To reduce the computational consumption, we do not apply $L_{\text{SDS-C}}$ for every iteration and $I_{\theta,C_S,0}$ is rendered with some probability.

In the second stage of training, we introduce the Spatial Consistency Module for static scenes to enhance the spatial consistency of the scene. The reason why we do not apply the Spatial Consistency Module in the first stage is that using the Spatial Consistency Module when the scene does not have the initial contours will lead to unstable training.

$$L_{stage2} = \alpha_S * L_S + \alpha_{\text{SDS-C}} * L_{\text{SDS-C}} \quad (17)$$

where α_S is the weight of the loss L_S.

In the last stage we introduce temporal Score Distillation Sampling loss $L_{\text{SDS-T}}$ proposed by [27] and Temporal Consistency Module L_T to train the complete 4D scene. In

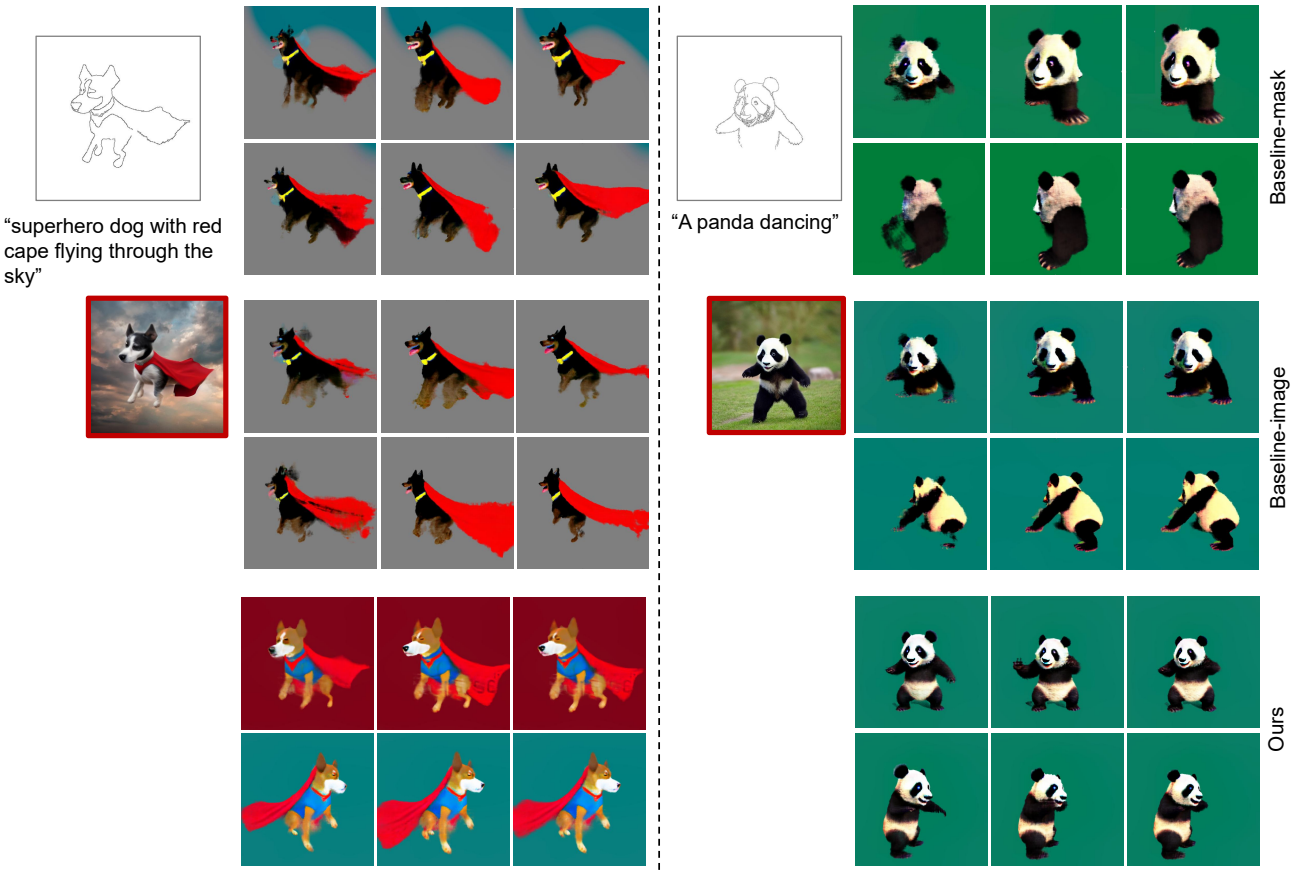

Figure 4. Comparison of the results with the baseline method.Top: Baseline-mask's result. Middle: Baseline-image's result. Bottom: our result. The image in the red section of the figure is the image generated using ControlNet via sketch for the Baseline-image supervision.

order to maintain sketch control and spatial consistency we still use $L_{\text{SDS-C}}$ and L_S

$$L_{stage3} = \alpha_S * L_S + \alpha_{\text{SDS-C}} * L_{\text{SDS-C}} + \alpha_{\text{SDS-T}} * L_{\text{SDS-T}} + \alpha_T * L_T \quad (18)$$

where $\alpha_{\text{SDS-T}}$ and α_{ST} are the weights of loss $L_{\text{SDS-T}}$ and L_T. Similarly, to reduce the computational consumption, we randomly apply L_S, $L_{\text{SDS-C}}$ and L_T in each iteration. To enhance the overall temporal and spatial consistency of the 4D scene, we supervise not only C_S and t_S, but also the Spatial Consistency Module and Temporal Consistency Module at other camera positions and time steps.

4. Experiments

4.1. Implementation Details

We used pre-trained Stable Diffusion [28], ControlNet [30], and a text-to-video model provided by VideoFusion [13] at huggingface. We trained 5k iterations in the first stage, 3k iterations in the second stage, and 10k to 20k iterations in the third stage (depending on whether the model converged or not). Our batch size during training is 1. Images with a resolution of 64×64 pixels are rendered for the first two stages of training and the first 3.5k iterations of the third stage, and images with a resolution of 128×128 pixels are used for the third stage after the 3.5k iterations. In the latter two stages, we render 16 frames of video for training. Fig. 6 shows the generation results of our method, and it can be seen that our method is able to generate 4D scene results that match sketch with high quality.

4.2. Baseline Comparison

We compared our mehtod with two baselines. Baseline-mask uses Eq. 2 on top of the text-to-4D approach for mask supervision. Baseline-image uses ControlNet generated images on top of the Baseline-mask to supervise the sketch viewpoint with Eq. 3.

Fig. 4 shows the results of our method compared to Baseline-mask and Baseline-image. In Fig. 4, "A panda dancing", it can be seen that the Baseline-mask method generates the whole face in the place of head and body due to the supervision of the mask only and no internal details. The

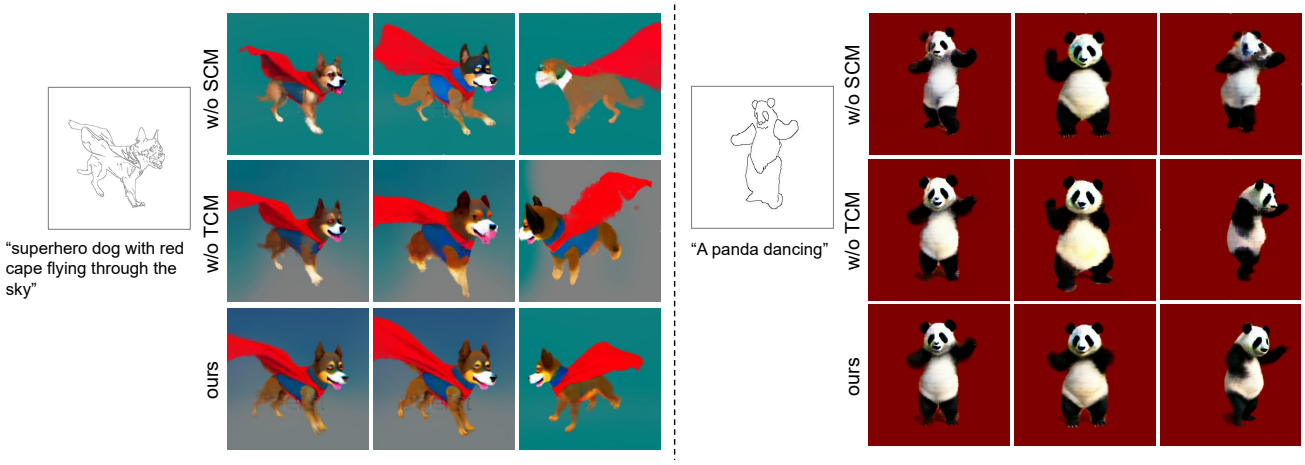

Figure 5. Comparison of results between our method and the ablation method.

Method	Quality	Match
Baseline-mask	4.343	3.936
Baseline-image	4.007	4.136
Ours	5.657	5.814

Table 1. Results of user ratings on the quality of our methods and baseline methods and how well they match sketch.

Method	Percentage
w/o SCM	20.71%
Ours	79.29%
w/o TCM	26.43%
Ours	73.57%

Table 2. Results of user experiments of our method with ablation methods.

Baseline-image method is able to generate the face in the correct position because it introduces the supervision of the image, but because the image is fixed and cannot be dynamically adjusted during the 4D generation process, the result of the 4D generation has strong 3D inconsistency in order to adapt to the image. In the case of "superhero dog", both Baseline-mask and Baseline-image are not able to control the content of the generated image effectively, which leads to the result not conforming to the sketch. Compared with the baseline method, our method achieve high-quality and conforming to the sketch.

We compare the quality of our method with that of the baseline methods by means of a user study. Users were asked to watch the rendering results of different methods given different prompts and different sketch inputs, and score the generation quality and How well it matches the sketches on a scale of 1 to 7, with higher scores indicating higher generation quality or matching. Table 1 shows the results of the user experiment, and it can be seen that our method is significantly better than the baseline methods in terms of both generation quality and match with sketch.

4.3. Ablation Study

We conducted ablation experiments on our proposed Spatial Consistency Module (SCM) and Temporal Consistency Module (TCM) to verify that they can enhance the temporal consistency and spatial consistency of the generated scenes.

Fig. 5 shows the comparison between our method and the post-ablation method. It can be seen that the spatial consistency of the generated scenes deteriorates significantly when SCM is not used. As in the example of "A panda dancing" in Fig. 5, although the rendered image from the sketch viewpoint matches the sketch, serious artifacts can be seen when the camera is moved to a similar viewpoint. When TCM is not used, although the first frame matches the sketch, all the objects in the subsequent scenes will have obvious inconsistency with the first frame. And our method can effectively avoid the above problem.

We also conducted a user experiment to verify the effectiveness of our method. We asked users to compare the 3D consistency of our results with that of w/o SCM, and select the results with better 3D consistency. We asked users to compare the temporal consistency of our results with those of w/o TCM, i.e., whether the overall generated 4D scene is similar to that of the corresponding sketch frames, and users are required to select the result with better temporal consistency.

Table. 2 shows the results of the user study, which shows that more than 70% of the users agree that our method has better temporal consistency and better temporal consistency than the ablation method.

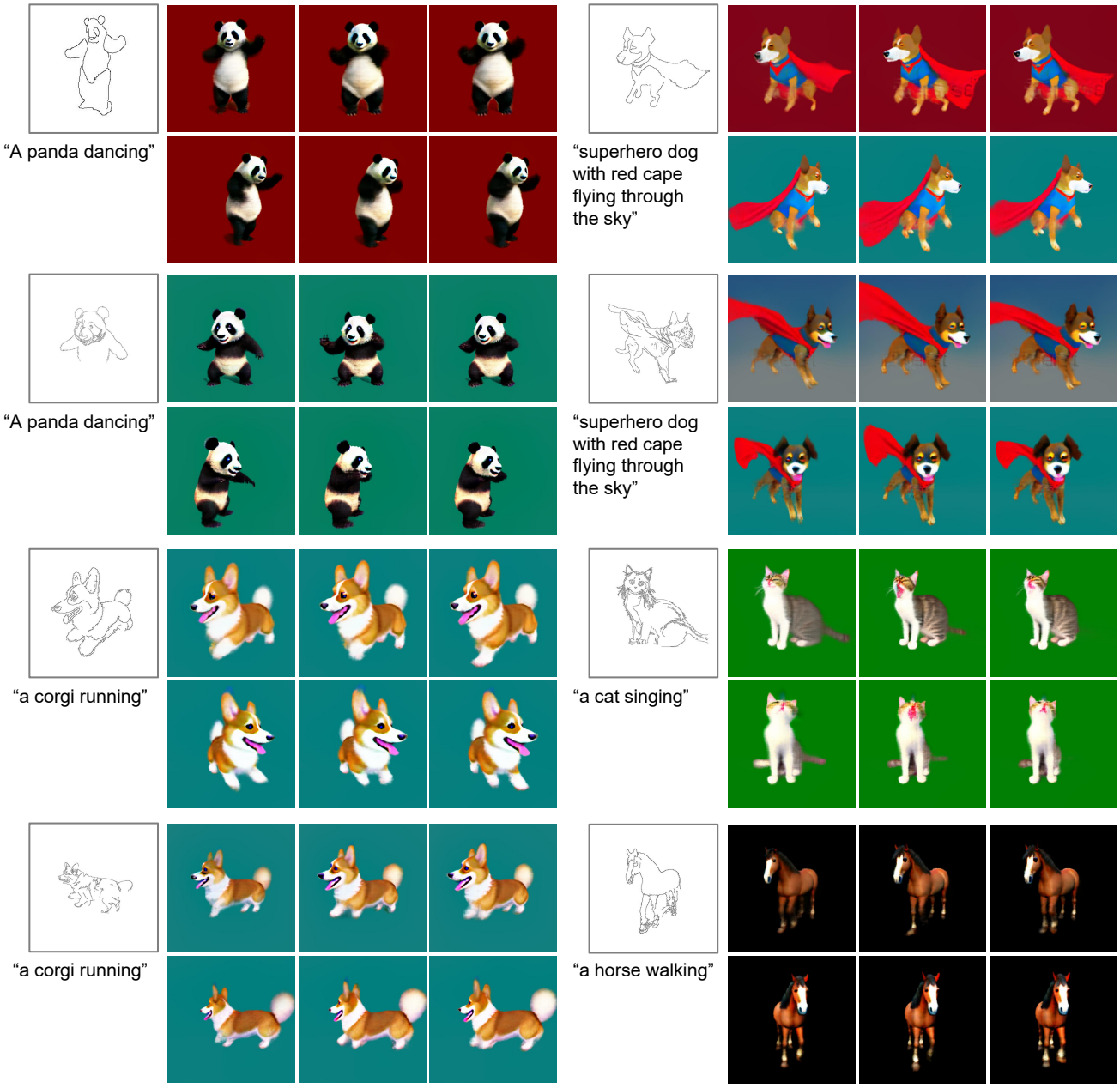

Figure 6. Results generated by our method. Our method generates rich and high quality 4D results based on sketch and text.

5. Conclusion

In this paper, we are the first to propose a sketch-driven 4D scene generation method, which performs sketch's constraints on 4D scenes via SDS-C. At the same time, SCM and TCM are proposed to solve the problem of spatial and temporal inconsistency brought by too strong control of sketch at a specific time and a specific viewpoint. Our experiments have demonstrated the effectiveness of our approach and shown that, compared to baseline, our approach can produce results in and of higher quality and more compatible with sketch.

The current state of 4D generation exhibits limitations, such as small motion amplitudes in generated results and slow generation speeds. The restricted motion range may result from insufficiently diverse training data. To address this, future work should focus on incorporating a more extensive dataset with larger motion amplitudes. In addition, the speed of generation may be improved by exploring

methods for direct 4D generation based on diffusion models, although this would require addressing the challenge of scarcity of 4D data. Overcoming these limitations will enhance the realism and efficiency of 4D content generation.

6. Acknowledgements

This work was supported by National Key Research and Development Program of China (Grant Number: 2021ZD0112902), National Natural Science Foundation of China (Grant Number: 62220106003), Research Grant of Beijing Higher Institution Engineering Research Center and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] R. AI. Gen-2: The next step forward for generative ai. 2023. https://research.runwayml.com/gen2. 1, 3

[2] A. Cao and J. Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3

[3] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[4] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[6] A. Gupta, W. Xiong, Y. Nie, I. Jones, and B. Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 3

[7] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models, 2022. 1, 2

[8] H. Jun and A. Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[9] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[10] G. Lin, L. Feng-Lin, C. Shu-Yu, J. Kaiwen, L. Chunpeng, Y. Lai, and F. Hongbo. Sketchfacenerf: Sketch-based facial generation and editing in neural radiance fields. *ACM Transactions on Graphics*, 2023. 3

[11] 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*, pages 9298–9309, 2023. 3

[12] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv preprint arXiv:2305.13311*, 2023. 2

[13] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 1, 2, 6

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3, 4

[15] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[16] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[17] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[18] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 3

[19] S. Park, M. Son, S. Jang, Y. C. Ahn, J.-Y. Kim, and N. Kang. Temporal interpolation is all you need for dynamic neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4212–4221, 2023. 3

[20] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 3, 4, 5

[21] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[22] 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*, pages 8748–8763. PMLR, 2021. 5

[23] 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*, pages 10684–10695, 2022. 2

[24] R. Shao, J. Sun, C. Peng, Z. Zheng, B. Zhou, H. Zhang, and Y. Liu. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2023. 3

[25] J. R. Shue, E. R. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, pages 20875–20886, 2023. 3

[26] 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. 1, 2

[27] 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. 3, 4, 5

[28] Stability. Stable diffusion v1.5 model card. 2022. https://huggingface.co/runwayml/stable-diffusion-v1-5. 2, 3, 6

[29] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[30] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 4, 6