Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models

Hanwen Liang[1]*, **Yuyang Yin**[2]*, **Dejia Xu**[3], **Hanxue Liang**[4],
Zhangyang Wang[3], **Konstantinos N. Plataniotis**[1], **Yao Zhao**[2], **Yunchao Wei**[2]†
[1]University of Toronto, [2]Beijing Jiaotong University,
[3]University of Texas at Austin, [4]University of Cambridge

Abstract

The availability of large-scale multimodal datasets and advancements in diffusion models have significantly accelerated progress in 4D content generation. Most prior approaches rely on multiple images or video diffusion models, utilizing score distillation sampling for optimization or generating pseudo novel views for direct supervision. However, these methods are hindered by slow optimization speeds and multi-view inconsistency issues. Spatial and temporal consistency in 4D geometry has been extensively explored respectively in 3D-aware diffusion models and traditional monocular video diffusion models. Building on this foundation, we propose a strategy to migrate the temporal consistency in video diffusion models to the spatial-temporal consistency required for 4D generation. Specifically, we present a novel framework, **Diffusion4D**, for efficient and scalable 4D content generation. Leveraging a meticulously curated dynamic 3D dataset, we develop a 4D-aware video diffusion model capable of synthesizing orbital views of dynamic 3D assets. To control the dynamic strength of these assets, we introduce a 3D-to-4D motion magnitude metric as guidance. Additionally, we propose a novel motion magnitude reconstruction loss and 3D-aware classifier-free guidance to refine the learning and generation of motion dynamics. After obtaining orbital views of the 4D asset, we perform explicit 4D construction with Gaussian splatting in a coarse-to-fine manner. Extensive experiments demonstrate that our method surpasses prior state-of-the-art techniques in terms of generation efficiency and 4D geometry consistency across various prompt modalities. Our project page is https://vita-group.github.io/Diffusion4D.

1 Introduction

The availability of internet-scale image-text-video datasets, coupled with progress in diffusion model techniques [37, 32, 31], has propelled significant advancements in generating diverse and high-quality visual content, including images, videos, and 3D assets [23, 13, 19, 34, 6, 48, 2]. These advancements have further fostered much progress in the realm of 4D content generation [47, 16, 51, 1, 11, 30, 36, 7, 52], which has gained widespread attention across both research and industrial domains. The ability to generate high-quality 4D content is key to various applications, ranging from the artistic realms of animation and film production to the dynamic worlds of augmented reality.

However, generating high-quality 4D content **efficiently** and **consistently** remains a significant challenge. ❶ Due to the scarcity of large-scale multi-view consistent 4D datasets, earlier 4D synthesis works [1, 57, 36] borrow appearance and motion priors from pretrained image- or video-diffusion models, and leverage score distillation sampling (SDS) [28] for optimization. This strategy is

---

*Equal contribution.

†Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).
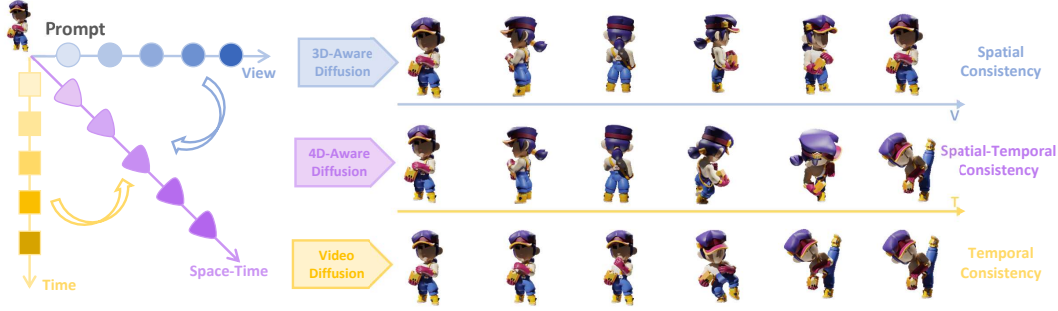
Figure 1: Decomposition of spatial-temporal consistency in 4D generation. The proposed Diffusion4D embeds geometrical consistency and temporal coherence into a single network.

time-consuming and computationally inefficient due to the heavy supervision back-propagation, limiting its widespread applicability. [51, 54] explored video-conditioned 4D generation and utilized 3D-aware image diffusion models to acquire pseudo multi-view data for photometric supervision. Yet the issue of slow optimization speed persists. ❷ Another primary challenge in 4D synthesis is to guarantee 4D geometry consistency, which can be decomposed into two components: spatial consistency and temporal consistency. As illustrated in Fig. 1, spatial consistency implies that the object maintains consistent 3D geometry at each distinct timestamp, while temporal consistency indicates that the object's appearance and movement exhibit coherence and smoothness across timestamps. These two components have been thoroughly explored separately in static multi-view 3D synthesis [19, 39, 34] and monocular video generation [3, 35, 4]. Multi-view 3D diffusion models possess robust knowledge of geometrical consistency, whereas video diffusion models encapsulate temporal appearance coherence and smooth motion priors. Recent approaches [27, 49] tried combining these two generative models to enhance efficiency and consistency in 4D generation. However, their underlying assumption of spatial-temporal conditional independence, the utilization of multi-model supervision, or inference at distinct timestamps inherently leads to 4D geometry inconsistencies. This raises the question: **can we integrate spatial and temporal consistency into a single network, and obtain multi-timestamp cross-view supervision in one shot?** Inspired by recent works in static 3D generation [41, 58] that repurposed temporal consistency in video generation to spatial consistency in 3D generation, we design a strategy that achieves spatial-temporal 4D consistency with a singular 4D-aware video diffusion model.

To this end, we propose a novel framework, **Diffusion4D**, to achieve efficient and consistent generation of 4D content. First, to overcome the data scarcity issue, we meticulously curate a large-scale, high-quality 4D dataset from a vast corpus of 3D datasets [10, 9]. Utilizing the curated dataset, we develop a *4D-aware video diffusion model* capable of generating orbital videos of dynamic 3D objects, imitating the photographing process of 4D assets. To the best of our knowledge, this is the first endeavor to adapt a video diffusion model and train on a 4D dataset for explicit novel view synthesis of 4D assets. To explicitly control the dynamic strength of the assets, we propose a novel 3D-to-4D motion magnitude metric and embed it as conditions into the diffusion network. We also incorporate the motion magnitude reconstruction loss to encourage the model to directly learn the object's 3D-to-4D dynamics. During the sampling stage, we propose a novel 3D-aware classifier-free guidance to further enhance the dynamics of 3D assets. Leveraging the flexibility of the video diffusion model architecture, our framework can seamlessly accommodate various prompts, including text, single image, and static 3D content. It demonstrates the capability to generate highly spatial-temporal consistent orbital views around 4D assets, thus providing comprehensive information for the construction of 4D assets. Based on the generated orbital views of 4D assets, we perform explicit *4D construction* by optimizing Gaussian splatting representations with off-the-shell 4D construction pipelines [46, 15]. We develop a coarse-to-fine optimization strategy with photometric losses. The synthesized multi-view consistent 4D image set enables us to swiftly generate high-fidelity and diverse 4D assets. In summary, our contributions can be summarized into three folds:

• We present a novel 4D content generation framework that, for the first time, adapts video diffusion models for explicit synthesis of spatial-temporal consistent novel views of 4D assets. The 4D-aware video diffusion model can seamlessly integrate with the off-the-shelf modern 4D construction pipelines to efficiently create 4D content.

2

- We introduce the 3D-to-4D motion magnitude metric to enable explicit control over the dynamic strength and propose the motion magnitude reconstruction loss and 3D-aware classifer-free guidance to refine the dynamics learning and generation.

- Extensive experiments demonstrate that the proposed framework outperforms previous approaches in terms of generation efficiency and 4D geometry consistency, establishing a new benchmark for 4D generation under different types of prompts, including text, single image, and static 3D content.

2 Background

4D generation aims at synthesizing dynamic 3D content from various inputs, such as text, single images, monocular videos, or static 3D assets. Extending 3D generation into the space-time domain, this task requires not only consistent geometry prediction but also the generation of temporally consistent dynamics. There are many works dedicated to this task. MAV3D[36] deals with text-conditioned 4D generation by utilizing SDS derived from video diffusion models to optimize a dynamic NeRF representation. 4DFY [1] tackles this problem by combining supervision signals from image, video, and 3D-aware diffusion models. Animate124 [57] leverages both text and image prompts to improve the appearance of the dynamic 3D assets. DreamGaussian4D [30] adopts the deformable 3D Gaussians for 4D representations and tackles the image-conditioned generation with a driving video. Similarly, 4DGen [51] proposes utilizing spatial-temporal pseudo labels and obtains anchor frames with a multi-view diffusion model. Consistent4D [16] leverages the object-level 3D-aware image diffusion model for the primary supervision and proposes cascade DyNeRF to facilitate stable training. More recent work STAG4D [54] uses a multi-view diffusion model to initialize multi-view images anchored on input video frames and introduces a fusion strategy to improve the temporal consistency. However, the use of SDS loss in these works results in slow optimization speed and limits their applicability. To improve the efficiency, Efficient4D [27] generates multi-view captures of dynamic 3D objects through SyncDreamer-T, while Diffusion^2[49] leverages orthogonal diffusion models to sample dense views. Despite the high efficiency, the assumptions of spatial-temporal conditional independence, multiple model utilization, and distinct-timestamp inference design inherently lead to 4D geometry inconsistencies. Unlike previous methods, our framework is the first to integrate 4D spatial-temporal consistency into a singular video diffusion model. More discussions about related works are in App. A.1.

3 Method

In this section, we provide a detailed illustration of the proposed framework, **Diffusion4D**, designed for the efficient and consistent generation of 4D content. We initiate by formulating the problem and outlining the key objectives in 3.1. Then, we introduce our data curation strategy pivotal in acquiring a large-scale, high-quality 4D dataset in 3.2. Subsequently, we delve into the methodology of applying the curated dataset to develop **4D-aware video diffusion models**, capable of synthesizing orbital views of dynamic 3D assets conditioning on various forms of prompts in 3.3. Finally, we introduce the explicit construction of 4D assets using 4D models based on the Gaussian Splitting representations in 3.4.

3.1 Problem Setting and Key Objectives

Formally, given a prompt y, we aim to generate an orbital video $\mathcal{V} = \{I_i \in R^{H \times W \times 3}\}_{i=1}^{T}$ around a dynamic 3D asset. This video comprises T multi-view images captured at T consecutive timestamps $\mathcal{T} = \{\tau_i\}_{i=1}^{T}$ along a predefined camera pose trajectory, where H and W are the height and width dimensions of images. To simplify the problem, we put constraints on the camera pose trajectory, assuming that the camera always looks at the center of an object (origin of the world coordinates) with a fixed elevation angle and camera distance. Thus, the viewpoint can be uniquely specified by the azimuth angle. The azimuth angle uniformly increases from 0 to 360 degrees along T timestamps, constructing a complete orbital video. We aim to generate this orbital video with a 4D-aware video diffusion model that can iteratively denoise samples from the learned conditional distribution $p(\mathcal{V}|y)$, where y can be text, single images, or static 3D content. At the end, explicit 4D construction is performed with generated orbital video $\mathcal{V}$ as supervisions.

3
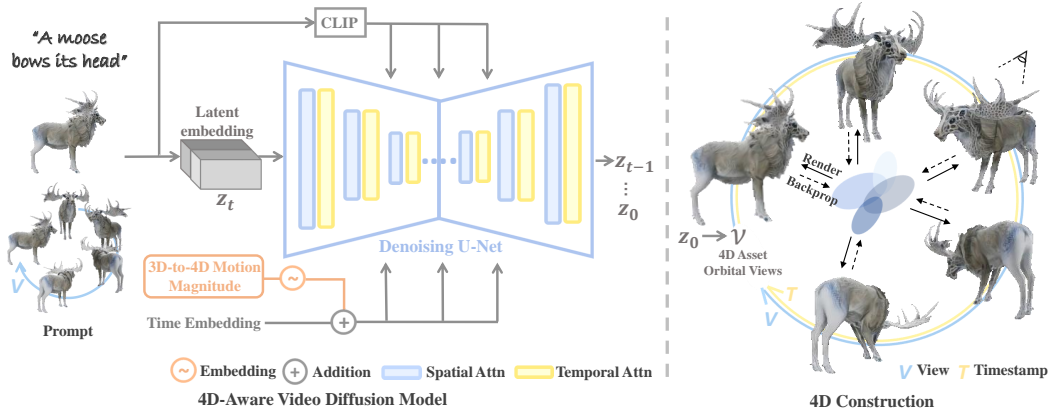
Figure 2: Our proposed Diffusion4D consists of a 4D-aware video diffusion model and explicit 4D construction, capable of synthesizing 4D assets conditioned on text, single images, or static 3D assets.

3.2 Data Curation

The development of the proposed 4D-aware video diffusion model requires a substantial amount of high-quality 4D assets. We curate a large-scale, high-quality dynamic 3D dataset sourced from the vast 3D data corpus of Objaverse-1.0 [10] and Objaverse-XL [9]. Since the original Objaverse dataset primarily consists of static 3D assets and many of them are of low quality (e.g., partial scans, missing textures) [38], we applied a series of empirical rules to filter the dataset. The curation process includes an initial selection with *dynamic* label, removal of assets with subtle or overly dramatic motion, and out-of-boundary detection. In the initial step, we select 3D assets labeled as "dynamic". However, upon closer examination, we observe that many assets exhibit subtle or imperceptible movement, limiting their utility in learning dynamic properties in the proposed task. To mitigate this problem, we employ Structural Similarity Index Measure (SSIM) to evaluate the temporal dynamics of the assets. Specifically, for each 4D asset, we fix the camera pose at the front view and render three images at three distinct timestamps ($\tau_0, \tau_{T/4}$, and $\tau_{T/2}$), then compute two SSIM scores between them. Assets with both SSIM scores higher than a predetermined threshold $s_{high} = 0.95$, suggesting high resemblance and little movement, are discarded. We also observe that many samples exhibit significant geometric distortion or drastic appearance changes over time. Therefore, we empirically set an SSIM score of small value $s_{low} = 0.4$ as the lower bound to filter out cases of poor quality. Subsequently, for each remaining 4D asset, we render $T = 24$ multi-view images following the camera pose trajectory and timeline settings as we defined in Sec. 3.1. In the final step, we handle cases where assets exhibit over-dramatic movements that extend out of the boundaries of the scene. We employ alpha maps to identify and remove such cases and ensure that only appropriately positioned dynamic assets are included in the curated dataset, enhancing the overall quality and coherence of the generated content. Finally, we also conduct a manual inspection over Objaverse-1.0 [10] to remove low-quality cases that do not make sense semantically. This comprehensive strategy results in a total of 81K high-quality animated assets. We released the IDs of the dynamic 3D assets in our curated dataset.

3.3 4D-Aware Video Generation

Utilizing the curated dataset, we can render a vast number of orbital videos of dynamic 3D assets to train a 4D-aware video diffusion model. For pretrained model selection, given that conventional video diffusion models designed for monocular video generation lack 3D geometry priors, we resort to the recent works of 3D-aware video generation models [41, 58]. In the following, we first introduce the pretrained 3D-aware video diffusion model. Then, we describe how to adapt the model to our 4D-aware orbital video generation task. At this point, we focus on explicitly specifying the motion magnitude guidance, directly learning the 3D-to-4D dynamics in the training stage, and further augmenting the 3D object dynamics in the inference stage. Finally, we introduce how we utilize the curated dataset and customize the model architecture to accommodate various condition modalities.

Pretrained 3D-aware video diffusion models. The main idea of 3D-aware video diffusion models [41, 58] is to repurpose the temporal consistency in video generation models for the spatial 3D

4

consistency of static 3D objects. Capitalizing on pretrained video diffusion models [3, 42, 56], these models are finetuned on large-scale datasets of high-quality 3D assets. It facilitates them to generate smooth and consistent orbital views of static 3D objects with user-specified camera pose conditions. In our method, we inherit the 3D geometry consistency in these models and extend it to modeling the spatial-temporal consistency in orbital views of 4D objects.

Vanilla 4D-aware video diffusion models. Modern video diffusion models typically carry out diffusion and denoising processes in latent space [4, 53]. In our specific task, given an orbital video $\mathcal{V}$ around a 4D asset, which is rendered from curated dataset following the camera position trajectory introduced in Sec. 3.1, we first use a pretrained encoder to encode images into a compressed latent $z_0 \in R^{T \times h \times w \times c}$. The h, w, c respectively denote the height, width, and channel dimension of the latent representation. The diffusion forward process samples a time step t and adds noise ϵ_t to source input and obtain z_t. A denoising network ϵ_θ, parameterized by θ, is trained to predict the added noise conditioned on y, with a noise prediction loss $\mathcal{L}_{ldm} = \|\epsilon_t - \epsilon_\theta(z_t, y, t)\|_2^2$. In the inference stage, as shown in Fig. 2, given an initial random gaussian noise and prompt condition y, the denoising network predicts the added noise to iteratively denoise the latent embedding. A denoised latent $\hat{z}_0$ is finally obtained, which is decoded via a pretrained decoder to recover a high-fidelity orbital video.

3D-to-4D motion magnitude guidance and reconstruction. For the 4D-aware video generation, one of the most interesting attributes is the dynamic strength of the 3D assets. In an effort to enhance the control over the 3D-to-4D dynamic strength, we introduce the motion magnitude guidance into the diffusion model. To begin with, we need to quantitatively measure the motion magnitude of the moving objects. In contrast to the previous video models [8] that measure the motion magnitude by computing the inter-frame differences, the appearance differences across frames in $\mathcal{V}$ originate from both camera-view changes and object movement. To remove the influence from the camera-view changes, for each dynamic 3D object, we also render an orbital video $\bar{\mathcal{V}} = \{\bar{I}_i \in R^{H \times W \times 3}\}_{i=1}^T$ around the static 3D asset, consisting of T multi-view images captured at the same camera poses as $\mathcal{V}$ at timestamp τ_0. Consequently, we propose a metric named **3D-to-4D motion magnitude** m measuring the dynamic strength:

$$m(\mathcal{V}) = \frac{1}{T} \sum_{i=1}^{T} ||I_i - \bar{I}_i||_2^2. \tag{1}$$

Motivated by previous works [58], we use a two-layer multi-layer perception (MLP) to extract a motion magnitude embedding, which is combined with the time embedding and injected into the denosing network. The noise prediction function by the denoising network accordingly changes to $\epsilon_\theta(z_t, y, m, t)$. This design allows the user to explicitly control the motion magnitude in inference stage. The impact of the motion magnitude guidance on the generation results is illustrated in Fig. 5.

In the training phase, to encourage the denoising network to learn the 3D-to-4D dynamics, we incorporate the motion magnitude reconstruction loss. This loss facilitates the direct learning of dynamic strength with explicit supervision over motion magnitude from 3D to 4D dynamics. For the sake of computation cost, we apply the reconstruction loss in the latent space and it is formulated as:

$$\mathcal{L}_{mr} = ||m(z_0) - m(\hat{z}_0)||_2^2, \ m(z_0) = \frac{1}{T}||z_0 - \bar{z}_0||_2^2. \tag{2}$$

, where $\hat{z}_0$ denotes the estimated clean video latent during training, the $\bar{z}_0$ denotes the compressed latent of $\bar{\mathcal{V}}$ by the pretrained encoder. Our diffusion model training loss is formulated by combining latent diffusion loss and motion magnitude reconstruction loss with a weight ω: $\mathcal{L} = \mathcal{L}_{ldm} + \omega \mathcal{L}_{mr}$.

3D-aware classifier-free guidance. To further augment the dynamic strength of 3D objects, we draw inspiration from classifier-free guidance [14] and propose a 3D-aware classifier-free guidance. It uses the pretrained 3D-aware video diffusion model, formulated as $\bar{\epsilon}_\theta$, to provide classifier-free guidance during the inference stage. Combining the prompt condition y, motion magnitude guidance m, the unconditional noise prediction $\epsilon_\theta(z_t, t)$ and 3D-aware noise prediction $\bar{\epsilon}_\theta(z_t, y, t)$, the denoising step is reformulated as:

$$\hat{\epsilon}_\theta(z_t, y, m, t) = \epsilon_\theta(z_t, y, m, t) + w_1(\epsilon_\theta(z_t, y, m, t) - \epsilon_\theta(z_t, t)) + w_2(\epsilon_\theta(z_t, y, m, t) - \bar{\epsilon}_\theta(z_t, y, t))$$
$$\tag{3}$$

, where $\hat{\epsilon}_\theta(z_t, y, m, t)$ is joint noise prediction, and w_1 and w_2 are the classifier-free guidance scalers.

5

Generation with various condition modalities. In the above formulations, we use y as a general prompt condition. Thanks to the versatility of our high-quality 4D dataset and the flexibility of the 3D-aware video diffusion model architecture, our framework can readily accommodate diverse prompt modalities, including text, single images, or static 3D contents. **For the text condition**, we obtain the text description of each dynamic 3D asset from the prior work [22], and the text embedding extracted by CLIP model is fed into the diffusion model via the cross-attention mechanism. You can refer to [42] for more details. **For the image condition**, we use the first view image I_0 in $\mathcal{V}$ captured at timestamp τ_0 as the reference image. The image condition is injected into the diffusion model with both cross-attention mechanism and feature concatenation. Please refer to [56, 3] for more details. **For static 3D content condition**, we use $\bar{\mathcal{V}}$ as the reference video. Similar to the image condition, the video features are extracted by pretrained encoder and and fed into the diffusion model via feature concatenation. This versatility allows our framework to effectively respond to different condition modalities, facilitating seamless integration into a wide range of applications and scenarios.

3.4 Coarse-to-Fine 4D Construction

The spatial-temporal consistent multi-view videos generated by our 4D-aware video diffusion model offer comprehensive information about the 3D geometry and motions of dynamic 3D assets. At this stage, we explicitly model the 4D assets with Gaussian splatting(GS) [18] owing to its explicit representation, powerful fitting capabilities, and efficient optimization with dense-view supervision. We resort to the state-of-the-art 4D construction pipelines, i.e. 4DGS [46] and SC-GS [15]. When we directly train the GS on the generated multi-view video $\mathcal{V}$, we find that the model does not perform well in capturing the 3D geometry details. The visualization can be found in Fig. 5(e). Therefore, we augment the supervision signals and propose a coarse-to-fine construction strategy. In the coarse stage, given the first-view image I_0 in $\mathcal{V}$, we use the pretrained 3D-aware video diffusion model to generate an orbital-view video $\bar{\mathcal{V}}'$ of the static 3D object. As our 4D-aware video diffusion model is finetuned on this model, we observe high 3D geometry consistency between $\bar{\mathcal{V}}'$ and $\mathcal{V}$. We merge them together to train the GS in the coarse training stage. Note that the two sets of images ($\mathcal{V}$ and $\bar{\mathcal{V}}'$) are of the same orbital views at different timestamps, $\mathcal{V}$ ranging from 0 to 1, and $\bar{\mathcal{V}}'$ at 0 only. Actually, for static 3D content conditioned generation, we can readily replace $\bar{\mathcal{V}}'$ with conditional signal $\bar{\mathcal{V}}$ into the image set for coarse training. In the fine training stage, we use $\mathcal{V}$ only to further tune the GS to improve the spatial-temporal consistency. Thanks to the 4D consistency in our generated videos, we can achieve precise pixel-level matching across different views and timestamps. We follow [46] and use L_1 and L_{lpips} [55] losses for optimization. To enforce geometry smoothness, we also involve depth smoothness loss as regularizer [5, 25]. For 4DGS-based reconstruction [46], the total loss is formulated as $\mathcal{L}_{gs} = \lambda_{l1}\mathcal{L}_1 + \lambda_{lpips}\mathcal{L}_{lpips} + \lambda_{depth}\mathcal{L}_{depth}$. For SC-GS based reconstruction [15], $\lambda_{arap}\mathcal{L}_{arap}$ is also involved. The λ_{l1}, λ_{lpips}, λ_{arap}, and λ_{depth} are losses weights.

4 Experimentation

4.1 Experiment Setup

We employ Diffusion4D to generate 4D assets conditioned on multiple modalities of prompts, i.e. text, single image, and static 3D content. Leveraging the curated dataset, for each dynamic 3D asset, we render two 360-degree orbital videos $\mathcal{V}$'s with $T = 24$, one starting at the front view and the other starting at a random azimuth angle. We fix the elevation angle as 0 and the camera distance as 2-meter. The image resolution is chosen at 256×256 across the experiments. In preparation for static 3D content-conditioned generation, following the camera poses of each rendered video $\mathcal{V}$, we also render two 360-degree orbital video $\bar{\mathcal{V}}'$s of each static 3D asset fixed at initial timestamp τ_0. We use VideoMV [58] as the pretrained 3D-aware video diffusion model, which is trained on a large scale of static 3D assets of superior quality. Specifically, for text-conditioned 4D-aware video generation, we adopt the text-to-3D video generator [58] that is based on the ModelScopeT2V [42] model architecture. For image and video conditions, we adopt the image-to-3D video generator [58] that is finetuned from I2VGen-XL [56]. We train our 4D-aware video diffusion model for 6k iterations with a constant learning rate of 3×10^{-5}. We use a valid batch size of 128 and train on 8 NVIDIA A100 GPUs. During the sampling stage, we use text, front-view images, front-view started orbital videos of static 3D assets as conditions. we use DDIM [37] sampling with sampling step 50, and $w_1 = 7.0$ and $w_2 = 0.5$ in classifier-free guidance. In the 4D construction stage, for 4DGS, we

6

Figure 3: Qualitative comparisons between Diffusion4D and other baselines in Text-to-4D (upper) and Image-to-4D (lower) generation. For our method, we show five views from consecutive timestamps. (* results from 4D-aware video diffusion model).7
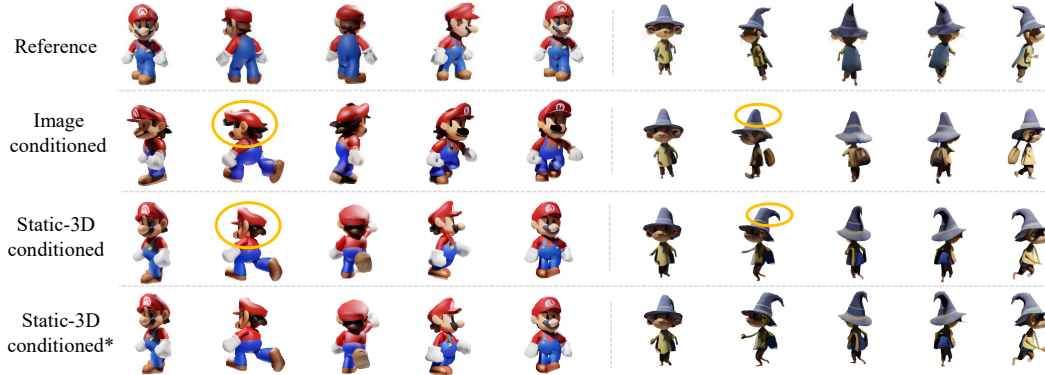
Figure 4: Visualization of Static-3D conditioned Diffusion4D. The first row shows the conditions, and the rest shows the results. (* results from 4D-aware video diffusion model.)

optimized the GS representation for 5k iterations in the coarse stage and 2k iterations in the fine stage. For SC-GS, we optimized the GS for 16k iterations in the coarse stage and 4k iterations in the fine stage. Following prior works [30, 51], by default, we use reconstructions from 4DGS for quantitative and qualitative comparisons with the baselines.

4.2 Metric

From the curated dataset, we leave out 20 cases as the test set for evaluation. For quantitative assessment, we first use the CLIP [29] score to evaluate the semantic correlation between the prompts (reference) and synthesized images (target). In each case, depending on the condition modality, the text description or front-view image serves as the reference. For the diffusion generation, we use all 24 images in orbital videos as targets, and for 4D construction, we uniformly render 36 orbital views from constructed 4D assets as targets (CLIP-O). We also measure using only front-view images as targets(CLIP-F). To evaluate the appearance and texture quality and the spatial-temporal consistency of the synthesized images, we also use the following metrics, i.e. LPIPS [55], PSNR, SSIM [45], and FVD [40], to help assess image- and static 3D content-conditioned generation. For these evaluations, we use ground truth images rendered from the dynamic 3D assets as references. Images generated by the diffusion model or rendered from constructed 4D assets are used as targets. Metrics are computed between ground truth images and synthesized images from the same camera pose. The same procedure is applied to calculate the scores for baseline methods. We also conduct qualitative comparisons through a user study involving 30 human evaluators from diverse backgrounds. Participants are asked to specify their preference for rendered orbital videos around 4D assets based on four properties, following the approach in [1]: 3D geometry consistency (3DC), appearance quality(AQ), motion fidelity(MF), and text alignment(TA). We report the percentage of users who prefer each method overall and based on each of the four properties.

4.3 Main Results

Text-conditioned and image-conditioned 4D generation. We take 4DFY [1] and Animate124 [57] as baselines for text-conditioned generation, and 4DGen[51] and STAG4D[54] as baselines for image-conditioned generation, due to their remarkable performance and adaptability to the task settings. In Fig. 3 we provide a detailed comparison between our method and baselines with various prompts. For the baselines, we provide two views at starting and ending timestamps. For our method, we visualize the spatial-temporal renderings from 4D constructions at five consecutive timestamps in multiple views. We also show the generated multi-view images from our video diffusion models (denoted with *). Our efficient and elegant pipeline is capable of generating diverse and realistic 4D assets with satisfactory geometrical and temporal consistency. While the baselines also synthesize 4D assets, their results exhibit very limited or even invisible motion. In contrast, our results show apparent motions of great fidelity. As highlighted by the contours in Fig. 3, our animations include cartoon characters and humans stepping forward, running or raising arms, birds flapping wings, and lights changing. Our method also provides more detailed appearances. The quantitative results in

8

Table 1: Quantitative comparison between our method with other baselines under different task settings. The human study includes 3D geometry consistency(3DC), appearance quality(AQ), motion fidelity(MF), text alignment(TA), and overall score. By default, Diffusion4D suggests results from 4D construction, and Diffusion4D* suggests results from our diffusion models.

Method	CLIP-F↑	CLIP-O↑	Generation time↓	SSIM↑	PSNR↑	LPIPS↓	FVD↓	3DC	AQ	MF	TA	Overall
			Metrics							*Human Preference*		
Text-to-4D												
4DFY [1]	0.78	0.61	23h			—		26%	34%	25%	37%	29%
Animate124 [57]	0.75	0.58	9h			—		22%	28%	19%	23%	22%
Diffusion4D	**0.81**	**0.65**	**8m**			—		**52%**	**38%**	**56%**	**40%**	**49%**
Diffusion4D*	**0.82**	**0.69**	–			—				—		
Image-to-4D												
4DGen [51]	0.84	0.71	1h30m	0.69	14.4	0.31	736.6	18%	22%	18%	29%	22%
STAG4D [54]	0.86	0.72	2h30m	0.76	15.2	0.27	675.4	15%	25%	26%	33%	24%
Diffusion4D	**0.89**	**0.75**	**8m**	**0.83**	**16.7**	0.21	560.8	**67%**	**53%**	**56%**	**38%**	**54%**
Diffusion4D*	**0.90**	**0.80**	–	0.82	**16.8**	**0.19**	**490.2**			—		
Static 3D content-to-4D												
Diffusion4D	0.88	0.77	8m	0.82	16.8	0.19	544.7			—		
Diffusion4D*	**0.91**	**0.81**	–	**0.83**	**17.2**	**0.18**	**482.4**			—		
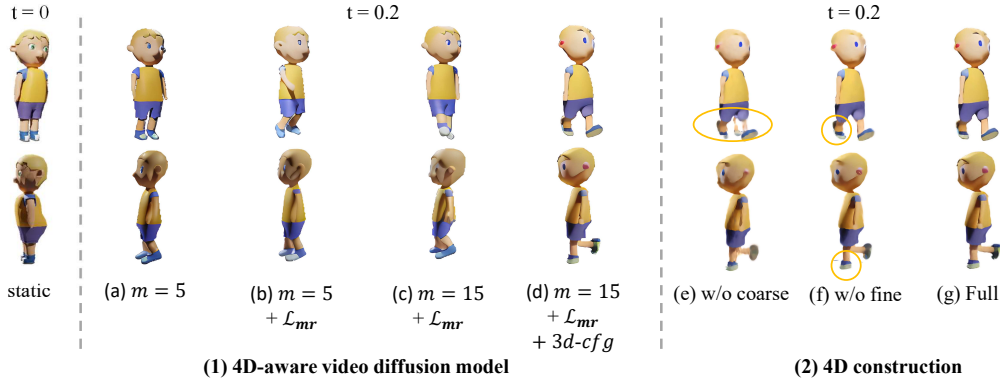
Figure 5: Ablation study of different components of our proposed framework.

Tab. 1 demonstrate that our method outperforms previous approaches across all metrics and user preferences. Compared with the state-of-the-art SDS-based methods, which involve sophisticated and time-consuming optimization, our method is much more efficient and produces more favorable results. Users showed a strong preference for Diffusion4D over other baselines, especially in 3D geometry consistency and motion fidelity. Also, as 4D construction is based on the images from the diffusion models, we can observe that diffusion outputs perform slightly better than rendered outputs.

Static 3D content-conditioned 4D generation. We visualized the results of static 3D content-conditioned generation in Fig. 4. Our framework successfully activates static 3D assets and generates spatial-temporal consistent 4D assets. Due to the lack of direct baselines, we compare our image-conditioned 4D generation with 3D content-conditioned generation. The difference between these two settings is that the former relies only on a single image, whereas the latter has access to dense views of the 3D assets. While both can generate realistic dynamic 3D assets, the latter can generate more 3D geometry-consistent objects. As shown in the last two rows in Fig. 4, Mario and anthropomorphic mouse closely follows the appearance and geometry of the prompts, particularly from the side and back views. The quantitative results in Tab. 1 indicate that the static 3D content-conditioned setting achieves the best performance among all tasks, thanks to the access to dense views of source assets.

4.4 Ablation Study

We conduct an analysis on the effect of various design components in our framework with the typical image-to-4D generation task, and the results are shown in Fig. 5 and Tab. 2. In the 4D-aware diffusion model, we incorporated multiple features to enhance the dynamics of 3D assets. We added each feature incrementally to demonstrate their impact. In Fig. 5 (1), we can observe that training without motion magnitude reconstruction loss results in nearly invisible movement (a), while incorporating it introduces subtle motion (b). Increasing the motion magnitude guidance (m) augments the dynamics of the kid (c). The involvement of 3D-aware classifier-free guidance significantly augments the

9

Table 2: Ablation study on the effect of proposed components on 4D construction. * denotes results from diffusion models.

Image-to-4D	CLIP-F↑	CLIP-O↑	SSIM↑	PSNR↑	LPIPS↓	FVD↓
w/o $\mathcal{L}_{mr}$	0.84	0.73	0.78	15.4	0.26	602.5
w/o $\mathcal{L}_{mr}$*	0.86	0.77	0.79	15.8	0.23	524.6
w/o coarse stage	0.79	0.70	0.72	14.5	0.32	651.2
w/o fine stage	0.86	0.74	0.77	15.2	0.25	581.4
Full model	**0.89**	**0.75**	**0.83**	**16.7**	**0.21**	**560.8**
Full model*	**0.90**	**0.80**	**0.82**	**16.8**	**0.19**	**490.2**

dynamics of the kid (d). In the 4D construction stage, as you can observe in Fig. 5 (2), training without the coarse stage leads to incomplete 3D geometry (e). Comparing (f) and (g), we can observe that adding the fine stage enhances the appearance and texture details of the generated 4D assets. Results in Tab. 2 indicate that the removal of each proposed component leads to a noticeable drop in metrics. In summary, the full model delivers the best results both quantitatively and qualitatively.

5 Limitation and Future work

While our proposed Diffusion4D framework demonstrates significant advancements in terms of efficiency and consistency for 4D content generation, we acknowledge the following limitations to provide a foundation for future work. First, we use a video size of $24 \times 256 \times 256$. While this resolution provides a good balance between quality and computational feasibility, higher resolution and longer temporal sequences could further enhance the detail and realism of 4D content generation. Image resolution could be improved using off-the-shelf super-resolution models. Second, though we curated a substantial amount of high-quality data, the dataset still lacks diversity. The asset textures are simple, with actions restricted to basic movements such as walking, jumping, and stretching arms. Although there is a large number of characters, each character appears only a few times, performing a limited range of actions. Last, in the 4D reconstruction stage, the performance is subject to the capacity of off-the-shelf 4D construction pipelines. In our settings, we empirically find that both of the chosen 4D construction pipelines work well in constructing consistent 4D assets from orbital views. Comparatively, SC-GS performs better than 4DGS from more novel views. Due to the limited views for supervision, the constructed 4D assets may meet floaters and blurry problems when the objects have large motions. In future research, we will try to expand the diversity and quality of the dynamic 3D dataset. We will focus on exploring 4D reconstruction to improve the model's ability to interpret and leverage orbital-view supervisions, which will be beneficial for generating 4D assets with better consistency across novel views. We will put more effort into generating longer and higher-resolution 4D orbital videos to support more detailed 4D construction, potentially broadening the applications for Diffusion4D.

6 Conclusion

In this work, we introduced Diffusion4D, an efficient and scalable framework for spatial-temporal consistent 4D generation. Motivated by the prior explorations using video diffusion model to generate orbital views of static 3D assets, we migrate the temporal consistency in the video diffusion model to spatial-temporal consistency in 4D generation. Leveraging a meticulously curated dynamic 3D dataset, we developed a 4D-aware video diffusion model capable of synthesizing orbital views of dynamic 3D assets. We incorporate 3D-to-4D motion magnitude guidance and the motion magnitude reconstruction loss to enhance the dynamics learning and control. 3D-aware classifier-free guidance is introduced to further augment the dynamic strength. We fulfill explicit 4D construction with a coarse-to-fine learning strategy. Extensive experiments demonstrated that our method surpasses prior state-of-the-art techniques in terms of generation efficiency and 4D geometry consistency across various prompt modalities. This innovative approach not only addresses the limitations of previous methods but also sets a new benchmark for 4D content generation.

7 Acknowledgment

This research was funded by the Fundamental Research Funds for the Central Universities (2024XKRC082).

10

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023.

[2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

[3] 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.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, pages 168–172. IEEE, 1994.

[6] Zhixiang Chi, Li Gu, Tao Zhong, Huan Liu, Yuanhao Yu, Konstantinos N Plataniotis, and Yang Wang. Adapting to distribution shift by visual domain prompt generation. *arXiv preprint arXiv:2405.02797*, 2024.

[7] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024.

[8] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023.

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[11] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024.

[12] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4220–4230, 2024.

11

[16] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

[17] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl 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.

[20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[21] 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. *arXiv preprint arXiv:2310.15008*, 2023.

[22] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[24] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[25] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[26] Zijie Pan, Jiachen Lu, Xiatian Zhu, and Li Zhang. Enhancing high-resolution 3d generation through pixel-wise gradient clipping. *arXiv preprint arXiv:2310.12474*, 2023.

[27] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024.

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[31] 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.

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023.

12

[33] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[34] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

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

[36] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023.

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

[38] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024.

[39] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *arXiv preprint arXiv:2402.12712*, 2024.

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

[41] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

[42] 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.

[43] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

[44] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[46] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023.

[47] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024.

[48] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

[49] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion^2: Dynamic 3d content generation via score composition of orthogonal diffusion models. *arXiv preprint arXiv:2404.02148*, 2024.

13

[50] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023.

[51] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

[52] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *arXiv preprint arXiv:2406.07472*, 2024.

[53] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023.

[54] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024.

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

[56] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[57] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

[58] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024.

14

A Appendices

A.1 Related Work

A.1.1 3D Generation

3D generation aims to create static 3D content from text or images. DreamFusion [28], first introduces the score distillation sampling (SDS) loss to optimize NeRF with diffusion models. However, The original form of SDS encounters challenges such as multi-face Janus issues and slow optimization. Subsequent works [19, 20, 43, 44, 26, 33, 34, 50] have attempted to refine the mechanism to address these issues. Another research direction [19, 21, 20, 39] focus on generating dense multi-view images with consistent 3D properties directly. They achieve this by training or fine-tuning 2D diffusion models on 3D datasets to better suit 3D generation tasks. Zero123 [19] generates a 2D image from an unseen view based on a single image. SyncDreamer [20] can generate multiview-consistent images from a single input image. Some later works [33, 34] explicitly generate fixed multi-view images in one diffusion pass. The generated images can be used for reconstructing 3D content. Some other works like Point-E [24] and Shap-E [17] train models to directly generate 3D point clouds or meshes.

A.1.2 Video Generation

Video generation has obtained significant advancements. Diffusion has been employed for video prediction tasks in recent years, which have achieved great levels of realism, diversity, and controllability. Among them, video LDM [4] applies the latent diffusion framework for video generation. SVD [3] follows the same architecture and inserts temporal convolution and attention layers after every spatial convolution and attention layer, achieving effective performance improvement. W.A.L.T [12] introduces window attention architecture for joint spatial and spatio-temporal generative modeling. Some recent works [41, 58] have tried to leverage temporal consistency [?] from video generation but are limited to static 3D generation tasks. Inspired by their success, our work focuses on utilizing spatial consistency from video diffusion models for 4D generation tasks.

A.1.3 4D Generation

Compared to 3D generation, 4D generation requires not only predicting consistent geometry but also generating temporal-consistent dynamics. There are a few recent works dedicated to this task. One line of research works predict 4D representations conditioned on a single image and text description. MAV3D [36] deals with a text-to-4D problem by utilizing score distillation sampling derived from video diffusion models to optimize a dynamic NeRF based on textual prompts. Animate124 [57] leverages a dynamic NeRF-based representation to tackle this problem. 4DFY [1] addresses text-to-4D synthesis by combining supervision signals from image, video and 3D-aware diffusion models. DreamGaussian4D [30] adopts the deformable 3D Gaussian as 4D representations. 4DGen [51] proposes a novel pipeline where they utilize spatial-temporal pseudo labels into anchor frames with a multi-view diffusion model. Another line of work predicts dynamic objects from single-view videos. Consistent4D [16] leverages the object-level 3D-aware image diffusion model as the primary supervision signal for training Dynamic Neural Radiance Fields, and proposing cascade DyNeRF to facilitate stable training. More recent work STAG4D [54] utilizes a multi-view diffusion model to initialize multi-view images anchoring on the input video frames and introduce a fusion strategy to ensure temporal consistency. Efficient4D [27] generates multi-frame multi-view captures of dynamic 3D objects through SyncDreamer-T and reconstructs 4D representations with them.

A.2 Experimental Details about Baselines implementation

For the baselines, we followed the instructions from their public websites to do experiments. 4DFY [1] is designed for text-to-4D generation. We fed in the text prompts directly to get the generation results. Animate124 [57] requires a pair of text and image prompts for generation. Therefore, we use the first front-view image in our generated video as the image prompt for generation. In the image-to-4D generation, both 4DGen [51] and STAG4D [54] propose to use pretrained video diffusion models to generate monocular videos as the prompt conditions. To boost their performance, we directly render ground-truth images at the front view from the source dynamic 3D datasets and obtain the monocular dynamic videos as prompts.

15

NeurIPS Paper Checklist

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction both outline the challenges in 4D content generation. It then introduces our novel framework, Diffusion4D, and highlights key innovations. These contributions are accurately detailed in the main body of the paper, where each component is thoroughly explained, implemented, and evaluated.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed about the limitations and future works of this submission in APP. 5. This includes the need for higher image resolution and longer temporal sequences to enhance visualization and detail in 4D construction.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

16

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not include theoretical results in our submission.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe our method in great details in Sec.4.1, experimental details in Sec.4.1 and App. A.2, the comprehensive evaluation metrics in Sec. 4.2, and the data curation strategies in Sec. 3.2 and App. **??**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

17

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code and the dynamic 3D assets idx of the dataset for reproduction of the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the details in Sec. 4.1 and App. A.2. All training details such as model parameters, dataset used, and metrics are discussed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

18

Answer: [No] ,

Justification: In this particular literature, error bars and confidence intervals are not typically reported. It is not clear what assumptions should be made on error distributions. So we do not report error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information about compute resources in Table 1. Our 4D construction model is trained on an A100 in 8 minutes to get the results. We also provide the details about cost of training diffusion model in App. A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The paper demonstrates a high level of research integrity by providing detailed methodologies, transparent data curation processes, and thorough analyses. We also emphasize reproducibility by fully disclosing the information needed to replicate the experiments. The research adheres to all relevant legal and ethical standards regarding the use of data and research practices.

Guidelines:

19

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We disucss our positive societal impacts in Sec. 1. Our work has a positive effect on the generating area, helping artists and designers to create high-quality 4D content quickly. We do not find that there are direct negative societal impacts of this work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: By providing detailed information about the curation process and the safeguards implemented, the paper promotes transparency. This allows other researchers and practitioners to understand the steps taken to ensure ethical use and to replicate these safeguards in their own work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

20

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

   Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

   Answer: [Yes]

   Justification: The creators or original owners of the assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected. The paper clearly credits the original source of the 3D data corpus, and provides citations for all the models, methods, and algorithms it builds upon or compares against.

   Guidelines:

   - The answer NA means that the paper does not use existing assets.
   - The authors should cite the original paper that produced the code package or dataset.
   - The authors should state which version of the asset is used and, if possible, include a URL.
   - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
   - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
   - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
   - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
   - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

   Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

   Answer: [NA]

   Justification: We did not provide any new assets alongside this paper. The paper describes in detail the curation process of a high-quality dynamic 3D dataset sourced from the Objaverse dataset.

   Guidelines:

   - The answer NA means that the paper does not release new assets.
   - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
   - The paper should discuss whether and how consent was obtained from people whose asset is used.
   - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

   Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

   Answer: [Yes]

   Justification: For the user study, we provide the details of this part in Sec. 4.3.

   Guidelines:

21

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We did not study on human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

22