Track, Inpaint, Resplat: Subject-driven 3D and 4D Generation with Progressive Texture Infilling

Shuhong Zheng[1,2] **Ashkan Mirzaei**[3*] **Igor Gilitschenski**[1,2]

[1]University of Toronto [2]Vector Institute [3]Snap Inc.

{shuhong, ashkan, gilitschenski}@cs.toronto.edu

Abstract

Current 3D/4D generation methods are usually optimized for photorealism, efficiency, and aesthetics. However, they often fail to preserve the semantic identity of the subject across different viewpoints. Adapting generation methods with one or few images of a specific subject (also known as *Personalization* or *Subject-driven* generation) allows generating visual content that aligns with the identity of the subject. However, personalized 3D/4D generation is still largely underexplored. In this work, we introduce TIRE *(Track, Inpaint, RE-splat)*, a novel method for subject-driven 3D/4D generation. It takes an initial 3D asset produced by an existing 3D generative model as input and uses video tracking to identify the regions that need to be modified. Then, we adopt a subject-driven 2D inpainting model for progressively infilling the identified regions. Finally, we resplat the modified 2D multi-view observations back to 3D while still maintaining consistency. Extensive experiments demonstrate that our approach significantly improves identity preservation in 3D/4D generation compared to state-of-the-art methods. Our project website is available at https://zsh2000.github.io/track-inpaint-resplat.github.io/.

1 Introduction

Improving on the personalization quality of 3D/4D generation is a core challenge to enable impact and improve user experience. Current generation methods, however, are mostly guided by text prompts [11, 46, 68, 79, 96, 107, 119] or single images/videos [51, 88, 123, 129] to determine the front-facing appearance of the generated assets. Although these methods provide users with a certain amount of control on the content of the generated scene, they fall short in delivering *identity-preserving* outputs that are desired for personalized generation. As illustrated in Fig. 1(a), the state-of-the-art 4D generation model L4GM [73] fails to preserve the identity for the side and back views in the generated 4D asset. In the given example, it results in a blueish tone on the originally occluded regions of the cat. These limitations highlight the need for methods to enable subject-driven, identity-preserving 3D/4D generation for personalized applications.

Identity preservation in 3D/4D generation, though tempting, is challenging to accomplish. In single image- or video-guided 3D/4D generation, the model has few cues to infer the appearance of unobserved viewpoints and is forced to hallucinate. One line of research adopts score distillation sampling [1, 5, 12, 26, 29, 59, 68, 91] to optimize the appearance of the novel viewpoints. However, the time-consuming optimization process prevents the paradigm from being widely used in real-world applications. Moreover, during optimization, the appearance and motions of the 3D/4D assets often become averaged out [2, 132], which can reduce the quality of the generated content. More recently,

*Work done while at University of Toronto and Vector Institute.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

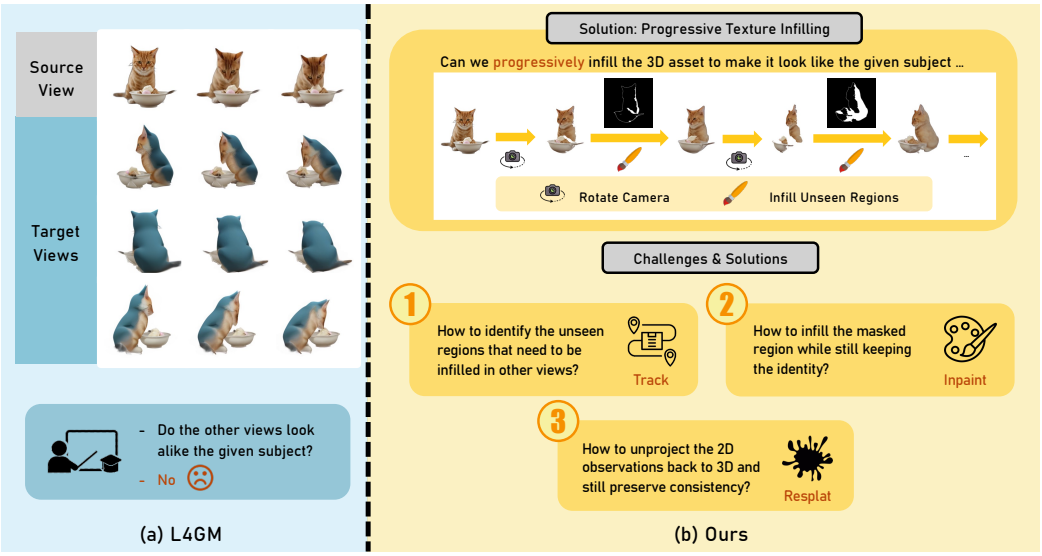

Figure 1: State-of-the-art 4D generation model L4GM [73] and our solution. **(a)** L4GM performs video-to-4D generation. The side and back views of the generated 4D asset does not look alike the subject in the given source view. **(b)** Our proposed solution TIRE *(**T**rack, **I**npaint, **RE**splat)* adopts the progressive texture infilling paradigm to inpaint the 3D asset to achieve subject-driven 3D/4D generation, which preserves the identity of the generated assets when observing from the novel views.

multi-view diffusion models are deployed in 3D/4D generation pipelines [54, 60, 78, 84, 92, 95, 106, 115] to hallucinate the appearance of a certain number of selected novel views. However, these models suffer from systematic errors in color and appearance for novel views. This is likely due to bias in the training data, which is difficult to address without careful dataset filtering. The most recent advancements [27, 42, 50, 102, 104, 110, 114, 124, 130] achieve superior efficiency with native 3D generation, *i.e.*, directly generating the 3D representation without per-scene optimization or multi-view observations as intermediate results. However, they still fail to produce results that can satisfactorily preserve the identity of the given reference as further demonstrated through the experimental analysis in Sec. 4.2 and App. F.

To effectively handle the challenges in 3D/4D generation for better personalization, we propose to perform subject-driven 3D/4D generation with progressive texture infilling, as shown in Fig. 1(b). Our proposed method, TIRE, is named after its three key components of the pipeline: ***T**rack, **I**npaint, and **RE**splat*. These three key components handle specific subtasks to achieve progressive texture infilling in a coordinated way: (1) *Track* identifies regions in other views that need infilling using long-video tracking. (2) *Inpaint* uses a customized 2D inpainting model to progressively infill unseen regions identified by *Track*, and ensures that the infilled content matches the subject's identity from the given source view. (3) *Resplat* unprojects the multi-view 2D infilled observations from *Inpaint* back to 3D while still maintaining consistency across multiple views. With these three components collaborating together in a cascaded manner, we achieve subject-driven 3D/4D generation while preserving the subject's identity.

We conduct extensive qualitative and quantitative evaluations, and find that our method serves as an effective general solution for enhancing the identity preservation in 3D/4D generation results upon the baseline methods. Moreover, our solution takes the exploration step in an *orthogonal direction* to current feed-forward approaches that utilize multi-view foundation models or native 3D/4D generation pipelines, which can be *complementary* to other advancements made in 3D/4D generation to collaboratively push forward the research field.

To summarize, our key contributions are threefold: First, we propose to solve subject-driven 3D/4D generation to enhance identity preservation of the generated 3D/4D assets for more personalized experience, which still remains a challenge for the most recent advancements on 3D/4D generation. It is an orthogonal while complementary effort towards high-quality generation from the foundation 3D/4D generation models. Second, to achieve the goal, we propose an innovative three-stage method,

TIRE: *Track, Inpaint, Resplat*. We first adopt video tracking for identifying regions that need infilling. Then, a customized 2D inpainting model is applied to infill the unseen regions. Afterwards, the 2D infilled observations are reprojected back to 3D to create the identity-preserving 3D/4D assets. Finally, comprehensive experimental results on our constructed DreamBooth-Dynamic benchmark and in-the-wild data showcase the superior performance of TIRE on subject-driven 3D/4D generation.

2 Related Works

3D Generation. Diffusion models [18, 80] have recently advanced and greatly improved 2D image/video generation [74]. To apply 2D diffusion model knowledge to 3D generation, DreamFusion [68] introduced score distillation sampling (SDS) to transfer knowledge to 3D. While some works [19, 25, 35, 39, 45, 69, 96, 99, 121] improved the quality and efficiency of optimization-based SDS, 3DiM [98], Zero-1-to-3 [51], and their successors [48, 49, 77] offered a different approach. They synthesized novel view images using diffusion models and then reconstructed 3D models from these synthetic views. Later works [7, 9, 10, 23, 28, 41, 52–55, 92, 97, 101] further enhanced the correctness and consistency of multi-view diffusion models. To accelerate generation, LRM [21] and concurrent works [40, 112] directly generate 3D representations using feed-forward networks. Subsequent works [86, 111, 126] applied this feed-forward approach to 3D Gaussians [36], a common 3D representation. Recent advancements [27, 100, 102, 110, 127] including MeshFormer [50], TRELLIS [104], and the Hunyuan3D series [114, 130] achieve fast native 3D generation that directly produce 3D representations after training on massive data. However, the goal of personalized customization is largely neglected as the field evolves, and even the most recent state-of-the-arts struggle to generate 3D content that satisfactorily preserve the identity.

4D Generation. Similar to 3D generation, early 4D generative methods [1, 2, 32, 72, 132] used a video version of SDS to transfer knowledge from video generation models to 3D space. Subsequently, multi-view video generation models [37, 108, 123, 125] were introduced. These models enabled 4D optimization using synthetic multi-view videos. Later research [31, 43, 44, 73, 82, 83, 94, 103, 109, 116, 118, 120, 122, 128] further improved the geometry, consistency, and efficiency of generated 4D assets. In contrast to efficiency-focused 3D/4D generation, we focus on subject-driven generation to preserve identity in generated assets, offering a more personalized option for 3D/4D content creation.

Subject-driven Generation. To personalize generated content for users, subject-driven generation became a prominent topic. Textual inversion [14] optimized a special text prompt token to represent a specific subject. DreamBooth [75] enabled text-to-image diffusion models to generate customized content for subjects by finetuning pre-trained models with a small number of example images. RealFill [89] extended this idea to subject-driven inpainting and adopted LoRA [22] for parameter-efficient finetuning. Subject-driven model personalization was also studied in multi-subject compositional generation [105] and videos [17, 30]. Beyond prior work in subject-driven 2D generation, DreamBooth3D [70] used image-to-image translation with personalized 2D models on multi-view observations to enhance identity in generated 3D objects. Customize-It-3D [24] proposed using a subject-specified prior to guide SDS optimization of generated assets. Make-Your-3D [47] finetuned a multi-view generation model with identity-aware optimization. Unlike previous works, we proposed leveraging the powerful 2D video tracking and inpainting tools to progressively infill the occluded regions of 3D/4D assets, preserving identity in the generated assets with the knowledge in 2D models.

3 Method

Given a single image or video of a certain subject, our goal is to generate a 3D (for images) or 4D (for videos) asset that faithfully represents the identity of this specific subject. Our proposed method, TIRE, consists of three stages: Track, Inpaint, Resplat. *Track* aims at providing the masks indicating the infilling regions from other viewpoints beyond the given source view (Sec. 3.2). *Inpaint* targets at progressively infilling the unobserved regions in other viewpoints with the infilled contents preserving the identity, while the regions are identified by the previous *Track* step (Sec. 3.3). *Resplat* is responsible for unprojecting the 2D infilled observations back to 3D (Sec. 3.4).

Our algorithm starts from a rough 3D/4D representation generated by existing models, as shown in the leftmost part in Fig. 2 about the setup stage. We render multi-view observations from the viewpoints within azimuth angle $\pm 180^\circ$ and elevation angle 0°, following the practice in [73, 86].

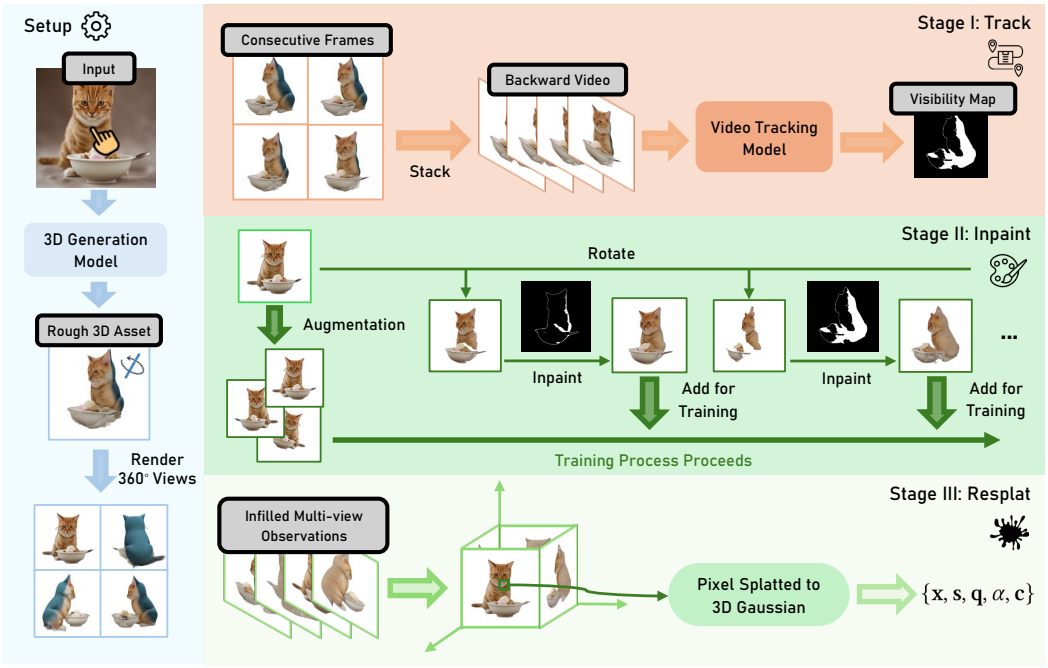

Figure 2: Pipeline of TIRE. TIRE starts from a rough 3D asset created by existing models and its rendered multi-view observations. Afterwards, the three stages *Track, Inpaint, Resplat* target at identifying the inpainting masks, infilling the occluded regions, and unprojecting back to 3D, respectively.

These initial rendered results will be used in our three-stage pipeline for infilling mask calculation, inpainting, and unprojection as discussed in the following subsections.

3.1 Preliminary

Diffusion Models [18] are generative models that learn a given data distribution by learning a denoising process that starts with random (typically Gaussian) noise and transforms this noise over multiple steps to the data distribution. In order to reduce the computational cost of learning distributions of image data, latent diffusion models [74] perform the diffusion and denoising processes in latent space. Pretrained autoencoders are used for the conversion between latent and image spaces.

In the diffusion process, we convert a clean latent z_0 to a noisy latent z_T of arbitrary timestep T as

$$z_T \sim q(z_T|z_0) = \mathcal{N}(z_T; \sqrt{\bar{\alpha}_T} z_0, (1-\bar{\alpha}_T)\mathbf{I}), \tag{1}$$

where the notation $\alpha_T = 1 - \beta_T$ and $\bar{\alpha}_T = \prod_{s=1}^{T} \alpha_s$ simplify the formulation with β_T representing the schedule of the strength of the noise added in timestep T. When $T \to \infty$, z_T is close to being equivalent to sampling from an isotropic Gaussian distribution.

The denoising process is the inversed operation to the diffusion process. The denoised latent at timestep $t-1$ can be estimated with the latent at timestep t by

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)), \tag{2}$$

where the parameters $\mu_\theta(z_t, t), \mathbf{\Sigma}_\theta(z_t, t)$ of the Gaussian distribution are estimated from the diffusion model. As revealed in [18], $\mathbf{\Sigma}_\theta(z_t, t)$ only has negligible contribution on the results from the experiments. Therefore, the main objective of the denoising framework is to estimate $\mu_\theta(z_t, t)$, which is reparameterized with

$$\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right), \tag{3}$$

where $\epsilon_\theta(z_t, t)$ is the denoising network to predict the added noise ϵ for z_t at timestep t.

The training objective for the denoising network is

$$\mathcal{L} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1),t} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]. \tag{4}$$

With the well-trained denoising network $\epsilon_\theta(z_t, t)$ and the deterministic sampling schedule in DDIM [81], the denoising process can be represented as

$$z_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t, t). \tag{5}$$

Large Reconstruction Models (LRM) [21] are foundation models for 3D reconstruction that predicts triplane representation in a feed-forward manner. Later, large Gaussian model (LGM) [86] achieves feed-forward prediction from multi-view observations to 3D Gaussians, inspired by the previous works [6, 85] that utilize a U-Net to splat each pixel into a 3D Gaussian in space. To be more concrete, taking 4 multi-view images as input, LGM functions as

$$f : \mathbb{R}^{4\times H\times W\times 3} \rightarrow \mathbb{R}^{\left(4\times \frac{H}{2}\times \frac{W}{2}\right)\times 14}, \tag{6}$$

where the output is a total of $\left(4 \times \frac{H}{2} \times \frac{W}{2}\right)$ 3D Gaussians, each has 14 parameters representing the center position $\mathbf{x} \in \mathbb{R}^3$, scaling $\mathbf{s} \in \mathbb{R}^3$, rotation quaternion $\mathbf{q} \in \mathbb{R}^4$, opacity $\alpha \in \mathbb{R}$, and color $\mathbf{c} \in \mathbb{R}^3$ of the 3D Gaussians. Follow-up work L4GM [73] extends the feed-forward 3D Gaussian generation to 4D with the additional time dimension, generating *a sequence of 3D Gaussians* that are consistent across both the spatial and time dimension.

3.2 First Stage: Track

Goal. As shown in Fig. 1(b), our proposed method adopts progressive texture infilling to achieve identity-preserving 3D/4D generation results. The first subtask we need to solve is to identify the regions that need to be infilled in viewpoints not observed in the original data. Obtaining a decent mask for infilling, however, poses non-trivial challenge.

Identifying Infilling Masks with Video Tracking. With the multi-view observations rendered from the setup stage, we can stack the consecutive frames together in the order of the camera movement to form a video. Then, we leverage the video tracking model CoTracker [34] to find the correspondence between the source view and the target views. The underlying principle is that, if the tracking result of a point on the target view is still within the valid mask of the target image, and with its visibility flag still active, we consider the point on the source view to be visible on the target view. An intuitive implementation is to start tracking from the given view, and see how the valid 2D pixels get propagated to the target views. However, as shown in the first row in Fig. 3, many small inpainting regions appear in the infilling mask, resulting in suboptimal grainy inpainting performance. Although this issue may be mitigated by performing other post-processing operations like dilations on the mask, it would require case-specific parameters to control the post-processing operations. Instead, we design a wiser approach to obtain the mask with *backward tracking*, which performs video tracking from the target views to the given source view. As displayed in the second row in Fig. 3, the masks obtained with backward tracking is more accurate and also more suitable for the following inpainting process. The insight behind backward tracking is that the given source view contains the richest information about the subject's identity. Therefore, we start tracking from the target views to establish as many correspondences as possible with the given view, to effectively leverage the subject information present in the given source view. Our proposed solution is a model-agnostic approach that is generally applicable to any types of 3D representations, and also leverages the power of 2D models that are trained on massive video data.

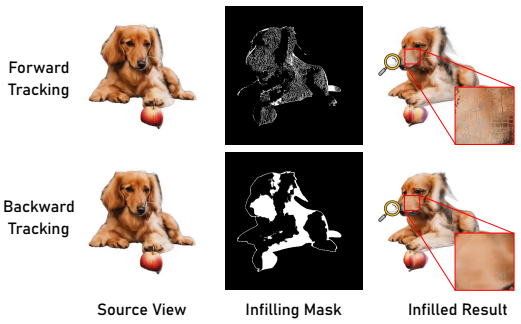

Figure 3: Comparison between forward tracking and backward tracking when identifying the inpainting mask. Forward tracking, which means that the tracking process starts from the given source view to the target views, though being more intuitive, leads to grainy inpainting results. In contrast, backward tracking produces more accurate masks in better shapes, which benefits the following inpainting process.

3.3 Second Stage: Inpaint

Challenges. After obtaining the infilling masks for novel views in the *Track* stage, we need to wisely inpaint these regions to maintain the identity of the subject. We are facing two substantial challenges: **(1)** How to faithfully preserve the identity in the infilling regions; **(2)** How to inpaint the viewpoints that are far away from the given source view, as the reference appearance in the source view may be very different and is unable to provide direct guidance. To address challenge **(1)**, inspired by RealFill [89], we propose to personalize the pretrained stable diffusion [74] inpainting model to be subject-driven, aiming to preserve the identity of the inpainted region. To handle challenge **(2)**, we perform the inpainting process in a *progressive* manner, as we first start inpainting from the viewpoints that are close to the given source view. As the training proceeds, the model progressively learn to inpaint the viewpoints that are farther away from the original source view.

Solution. With the pretrained stable diffusion inpainting model in hand, we inject LoRA [22] weights in the pretrained model and finetune with randomly generated binary mask m_i for inpainting. The loss calculation will only be conducted on the valid regions m_v as

$$\mathcal{L} = m_v \odot [\epsilon_\theta(x_t, t, p, m_i, (1 - m_i) \odot x) - \epsilon], \tag{7}$$

where $m_v \in \{0, 1\}^{H \times W}$ is the valid mask in which 1 indicates the foreground of the image, while 0 indicates the background of the image. The valid mask is obtained by the same background removal tool used in recent 3D/4D generation works [54, 73, 86]. p is a fixed language prompt *"A photo of sks"*. "$\odot$" denotes the element-wise multiplication and therefore $(1 - m_i) \odot x$ is the masked image. The other notations follow the same convention as Eq. 4.

At the beginning of the tuning process, since only a single image/video from the source view is available, we perform horizontal flipping and small-scale rotations within $15°$ on the original image for data augmentation. After training the inpainting task on the original image together with its augmented counterparts, we perform inpainting of azimuth angle $\theta = \pm 20°$ which we name it a *sweet spot*. This viewpoint serves as an *anchor viewpoint* for later processes when inpainting the viewpoints that are farther away from the source viewpoint. To be more concrete, when inpainting the farther viewpoints within $\pm 90°$, the similar operation of *backward tracking* in Sec. 3.2 will be applied to track from the queried viewpoints to this anchor viewpoint. It helps further reduce the area that needs to be infilled, therefore lowering the difficulty for the model to inpaint the far away viewpoints. The reasons for choosing $\pm 20°$ as the sweet spot for the anchor viewpoint in tracking are **(1)** compared to the source viewpoint at $0°$, it has certain exploration on unseen regions that can provide larger known regions for the farther viewpoints; **(2)** when inpainting the $\pm 20°$ viewpoint, since it does not have significant shift from the original training image, the inpainted results are relatively decent and reliable. Therefore, this *sweet spot* strikes a balance between *"exploration and exploitation"* of the given source view observation. After the $\pm 90°$ viewpoint is inpainted, it serves as the next anchor point for inpainting the rest of the viewpoints within $\pm 90° \sim \pm 180°$. As we do not want significant change on the original structure, we perform denoising with the first 30% of the denoising schedule following similar practice as [33].

3.4 Third Stage: Resplat

The goal of this stage is to unproject the inpainted 2D observations back to 3D. As the frames are infilled separately during the *inpaint* stage in Sec. 3.3, there may exist inconsistency across the inpainted frames. Thus, before lifting to 3D, we propose to use the multi-view diffusion model [92] to refine the consistency of the multi-view observations. More specifically, inspired by the previous work on mask-aware image editing [62], our multi-view denoising process only updates the latents on the unseen viewpoints as

$$z_{t-1} = \tilde{z}_{t-1} \odot M + \hat{z}_{t-1} \odot (1 - M), \tag{8}$$

where $z_{t-1}, \tilde{z}_{t-1}, \hat{z}_{t-1} \in \mathbb{R}^{(V+1) \times c \times h \times w}$. $\tilde{z}_{t-1}$ is the predicted latent at $T = t - 1$ during the denoising process obtained by Eq. 5. $\hat{z}_{t-1}$ is the noisy latent obtained by Eq. 1 with the forward diffusion process. $V = 4$ is the number of views, while c, h, w are the channel number, height and width of the latents. As the multi-view diffusion model is image-conditioned, there are $(V + 1)$ entries on the first dimension of the latents, as the additional one being the latent of the conditional image. $M \in \{0, 1\}^{(V+1) \times c \times h \times w}$ is the mask for refinement with value 0 for the source view and 1 elsewhere. Similar to the practice of the *Inpaint* stage, only the first 30% of the denoising schedule

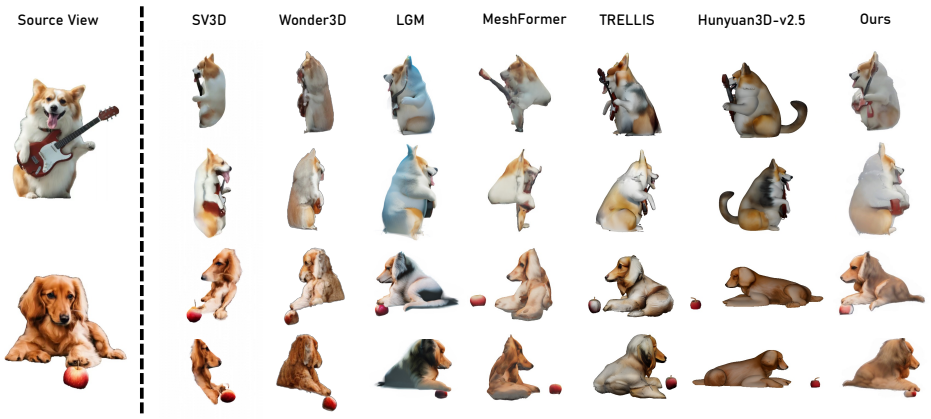

Figure 4: Qualitative comparison on image-to-3D generation with SV3D [90], Wonder3D [54], LGM [86], MeshFormer [50], TRELLIS [104], and Hunyuan3D-v2.5 [38]. Compared against other method in the image-to-3D setting, our method better preserves the identity of the reference image, and also reaches superior quality on geometry. It is noticeable that even for the most recent advancements in image-to-3D like TRELLIS and Hunyuan3D-v2.5, the challenge of producing identity-preserving 3D assets is still not well solved.

is applied. The mask-aware latent update strategy reinforces the identity preservation of the front view of the final 3D/4D assets. After refining the multi-view observations with Eq. 8, we *resplat* pixels in every viewpoint to Gaussians with [73, 86]. Note that this multi-view-to-Gaussian process is also adaptable to different large reconstruction models, and thus can be applied to different 3D representations beyond 3D Gaussians.

4 Experiments

4.1 Experimental Settings

Datasets. Starting from the original DreamBooth [75] dataset that focuses on subject-driven image generation, we construct a *DreamBooth-Dynamic* dataset. It is based on animatable subjects in the original DreamBooth dataset and will be used for subject-driven image-to-3D and video-to-4D generation. More details about the dataset are provided in App. B in the appendix. Besides the qualitative and quantitative evaluations on the constructed DreamBooth-Dynamic dataset, we also demonstrate that our method works for *in-the-wild data* that are displayed on the official project webpage of L4GM [73] in App. F in the appendix due to space issue.

Baseline Methods. We select Wonder3D [54], SV3D [90], LGM [86], MeshFormer [50], TRELLIS [104], and Hunyuan3D-v2.5 [38] for the comparison on the image-to-3D task. For video-to-4D, we select recent 4D generation methods STAG4D [123] and SV4D [108][1], along with L4GM [73]. As few existing works focus on subject-driven 3D/4D generation [24, 47, 70], we choose Customize-It-3D [24] for comparison, as it is the most recent open-source option.

4.2 Qualitative Evaluation

We show visualizations of our results and the compared methods in Fig. 4 for image-to-3D generation and Fig. 5 and Fig. 6 for video-to-4D generation, rendered from different viewpoints and timesteps. From the qualitative comparisons, we observed the following:

Identity-preserving Appearance. Fig. 4 and Fig. 5 demonstrate that the 3D/4D assets generated by our method achieves significantly better identity preservation compared to the baselines. Notably, even the most recent advancements in image-to-3D generation, such as TRELLIS [104] and Hunyuan3D-v2.5 [38], still face substantial challenges in producing identity-preserving 3D assets, demonstrating that the task of subject-driven 3D/4D generation is a valid and important problem to investigate, highlighting the necessity of addressing it to achieve better identity preservation for 3D/4D generation.

[1]Comparison included in App. F in the supplementary material due to different pose settings.

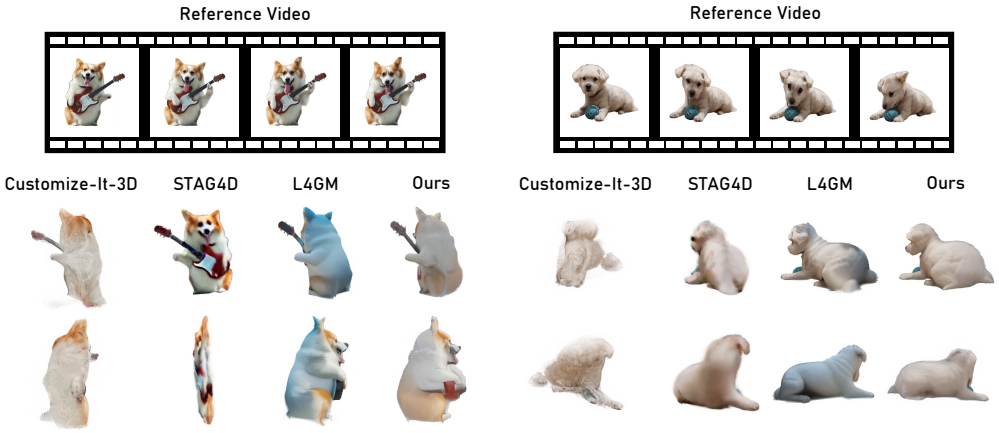

Figure 5: Comparison between our method and the baselines Customize-It-3D [24] (additional feed-forward operation from L4GM is applied after obtaining multi-view observations to allow it to generate dynamic 3D assets), STAG4D [123], and L4GM [73].

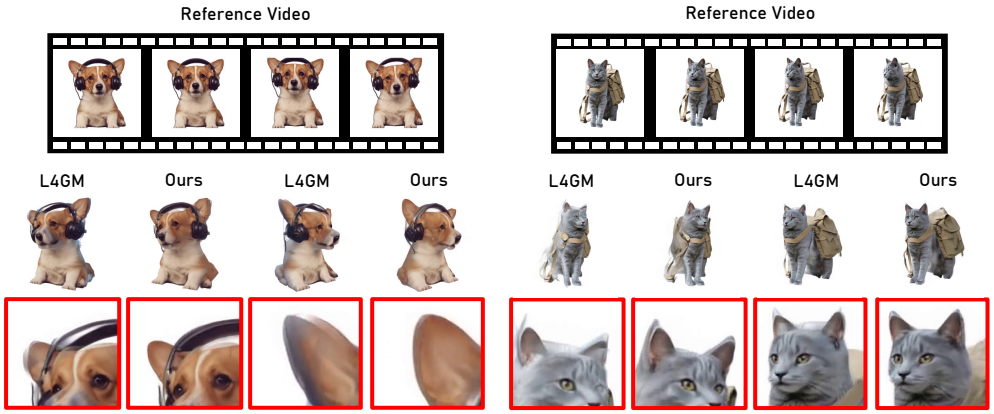

Figure 6: Comparison between our method and L4GM [73]. Although the main objective of our TIRE is to preserve the identity of the generated assets, the geometry of the generated assets also gets improved because the three stages in our pipeline collectively promote cross-view consistency.

Enhanced Quality on Geometry. Although TIRE mainly aims to improve subject-driven appearance modeling of the generated 3D/4D assets, it also demonstrates better geometry quality than the baseline methods. As shown in Fig. 6, our results outperform L4GM on the geometry of the generated assets. The rendered viewpoints from our method have fewer ghosting artifacts, which are caused by cross-view inconsistency. The reason that TIRE can fix the geometry is that the *Track* and *Inpaint* stages collaboratively propagate pixels seen in source views to the target views. Also, the *Resplat* stage includes a mask-aware halfway diffusion-denoising process in Eq. 8. These designs within TIRE implicitly refine the consistency of the multi-view observations and the generated 3D/4D assets.

General Solution to 3D/4D Generation Methods. As discussed in Sec. 3.4, the demonstrated results are generated with LGM [86] and L4GM [73] as the base model for our method. As our progressive texture infilling process only needs to perform on rendered 2D frames for tracking and inpainting, our method has the advantage of being generally applicable to all types of 3D/4D generation methods, no matter what representation the method adopts. To support this, we also include the results of how our TIRE can further enhance identity preservation for even the most advanced models like Hunyuan3D-v2.5 [38] in App. F.

4.3 Quantitative Evaluation

As there are no standard evaluation protocols for benchmarking the quality of subject-driven 3D/4D generation, we first follow [75] to adopt DINO [4] feature similarity to measure the subject fidelity of

the generated assets. More specifically, the source view is the reference image of the subject, and without loss of generality, 4D generation is selected as a case study.

For each generated asset, we render every 10 timesteps from 8 views around the assets, with a 45° interval in azimuth angle, to calculate the DINO similarity score with the reference image.

Table 1: Quantitative comparisons on the DINO feature similarity metrics from DreamBooth [75]. Best: **bold**, 2nd best: *italics*.

Method	DINO (ViT-S/16) (↑)	DINO (ViT-B/16) (↑)
Customize-It-3D [24]	**0.5773**	**0.6087**
SV4D [108]	0.5213	0.5426
STAG4D [123]	0.5287	0.5592
L4GM [73]	0.5506	0.5694
TIRE (Ours)	*0.5665*	*0.5815*

We report the results in Tab. 1, which is the average similarity score across all rendered images in all the scenes. Although TIRE outperforms L4GM and STAG4D in subject fidelity according to the DINO similarity metric, which is commonly used in the literature [13, 75], it is surprising that Customize-It-3D ranks highest. However, the qualitative comparisons in Sec. 4.2 clearly show that Customize-It-3D performs worse than other methods. This suggests that the original DINO similarity metric, while being a standard metric for subject fidelity, is not the most suitable quantitative measure for the nuances of the subject-driven 3D/4D generation task. We further discuss this limitation in App. C in the appendix.

Having recognized the challenges of conducting quantitative evaluations with purely vision-based models, we turn to vision-language models (VLMs) for more comprehensive assessment across multiple dimensions including identity preservation and visual quality. Specifically, we follow DreamBench++ [66] to adopt VLM-based evaluation on the identity preservation quality of the generated assets. Following the subject similarity evaluation option in DreamBench++, we ask the VLMs to give an overall score from 0-4 (0 is the worst, while 4 is the best) on subject consistency between the reference image and the generated images rendered from the assets, considering the following aspects: shape, color, texture, and facial features. Detailed implementations including the used prompts for VLMs are described in App. D. For the choice of VLMs, besides GPT-4o [64] which is adopted in DreamBench++, we also use five other VLMs (OpenAI o4-mini [65], Gemma 3 27B [15], Gemini 2.0 Flash [16], Qwen2.5-VL-7B [3], Mistral-Small-3.1-24B-Instruct [63]) with different architectures and sizes to foster the reliability of the evaluation. For each generated asset, we evenly render 8 views for each generated asset for evaluating the quality. The results in Tab. 2 show that our method achieves the best subject consistency when taking multiple aspects of consistency into account, demonstrating the effectiveness of our method for enhancing identity preservation for the generated 3D assets. Nevertheless, we can observe that the scores for all the methods still remain distant from the perfect score of 4, indicating that the task of subject-driven 3D/4D generation is still far from being solved.

Table 2: Quantitative comparisons on the VLM-based scores on the views rendered from the generated assets from different methods. Best is marked as **bold**.

Methods	**VLM-based scores (↑)**						
	GPT-4o	OpenAI o4-mini	Gemma 3 27B	Gemini 2.0 Flash	Qwen2.5-VL-7B	Mistral-Small-3.1-24B-Instruct	Average
TRELLIS	1.332	1.426	1.870	1.402	1.596	1.228	1.476
Hunyuan3D-v2.5	1.614	1.690	2.098	1.533	1.780	1.501	1.703
TIRE (Ours)	**1.777**	**1.834**	**2.103**	**1.793**	**1.880**	**1.739**	**1.854**

4.4 User Study

To further calibrate the identity preservation quality with human perception, we conduct user study on the 4D assets generated by Customize-It-3D [24], L4GM [73], and our method. We ask the volunteers to score the generated assets in a scale of 1-10 on the overall quality, where the participants are guided to focus on subject fidelity, cross-view consistency, *etc.* that are considered important factors for 3D/4D generation quality. We randomly select 10 samples from the DreamBooth-Dynamic dataset. Note that we *do not* explicitly tell the participants that we are working on improving subject-driven generation,

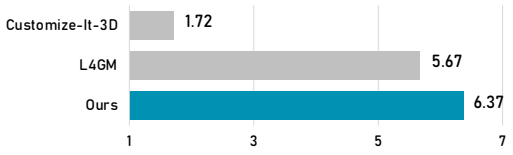

Figure 7: Results of our user study. Our method scores the highest in overall quality *without* explicitly informing the users that we are focusing on subject-driven generation.

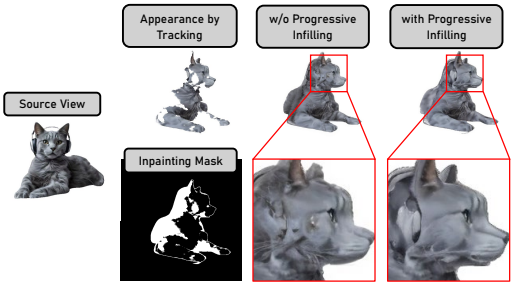

Figure 8: Ablation study on the progressive learning strategy in TIRE. Without adopting progressive learning, the model tends to consistently infill the appearance of the given source view *regardless of the current pose*, which results in wrongly infilling the textures on the side and back views.

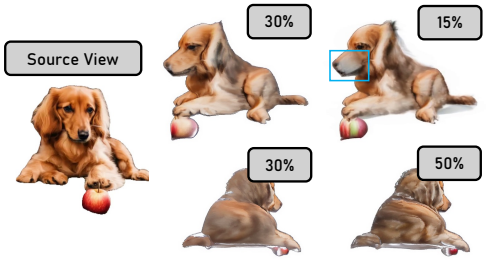

Figure 9: Ablation study on different denoising schedules for inpainting. A smaller schedule may leave some regions unchanged, while a larger one may overly distort the textures.

which fosters the fairness in user study and reduces the underlying bias of the subjective preference towards our method. Details on the instructions and interface of our user study can be referred in App. A in the appendix. With 18 volunteers participating in the user study, we collect a total of 540 scores from participants. The results of the user study are reported in Fig. 7, which shows that our method is more subjectively preferred from the user experience.

4.5 Ablation Study

Progressive Texture Infilling is a core idea of our pipeline to infill the texture of unseen regions in the generated assets. Therefore, to prove the validity of our progressive infilling strategy, we demonstrate the ablation results illustrated in Fig. 8. More specifically, for the variant "w/o Progressive Inpainting", we use the inpainting model finetuned only on the original image of the source view and its augmented counterparts. When the progressive infilling strategy is not applied, the model tends to inpaint the appearance corresponding to the given source view of the object. We can observe that the model fills in textures of a cat's face with a pair of cat whiskers in the target view, even if the target view is 60° off the source view and actually displays the side view of the cat.

Denoising Schedule for Inpainting. As discussed in Sec. 3.3, we empirically adopt the first 30% of the denoising schedule in the *Inpaint* stage. To provide a more intuitive demonstration for guiding real practice, we examine the impact of a larger or smaller denoising schedule in Fig. 9. We can observe that when we adopt a smaller denoising schedule of 15%, some of the regions where the textures need to be refined remain unaltered, like the regions in the blue bounding box. On the other hand, when we choose a larger denoising schedule of 50%, we may have the risk of having an overly intense change on the textures, such that the appearance becomes less realistic.

Due to the limit of space, more ablation study can be found in App. E in the appendix.

5 Conclusions

We present TIRE, which aims at subject-driven 3D/4D generation that progressively infill the textures of the occluded regions. TIRE contains three important steps: *Track, Inpaint, Resplat*. *Track* helps identify the regions that need to be infilled via video tracking tools. *Inpaint* aims at infilling the occluded regions while preserving the identity of the subject. *Resplat* unprojects the multi-view infilled observations back to 3D with cross-view consistency. Extensive experimental results demonstrate that the 3D/4D assets obtained from our method achieve superior performance on both *more identity-preserving appearance* and *refined quality on geometry*. We believe that our solution is an important exploration on the direction complementary to the current research on efficient feed-forward 3D/4D generation pipelines, and can serve as a useful tool for expressing the creativity and enhancing the personalization on subject-driven 3D/4D generation. More discussions can be referred in App. C in the appendix.

References

[1] S. Bahmani, X. Liu, W. Yifan, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein, A. Tagliasacchi, and D. B. Lindell. TC4D: Trajectory-conditioned text-to-4D generation. In *ECCV*, 2024. 1, 3

[2] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4D-fy: Text-to-4D generation using hybrid score distillation sampling. In *CVPR*, 2024. 1, 3

[3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 9

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 8

[5] I. Chachy, G. Yariv, and S. Benaim. RewardSDS: Aligning score distillation via reward-weighted sampling. *arXiv preprint arXiv:2503.09601*, 2025. 1

[6] D. Charatan, S. Li, A. Tagliasacchi, and V. Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, 2024. 5

[7] C. Chen, X. Yang, F. Yang, C. Feng, Z. Fu, C.-S. Foo, G. Lin, and F. Liu. Sculpt3D: Multi-view consistent text-to-3D generation with sparse 3D prior. In *CVPR*, 2024. 3

[8] D. Z. Chen, Y. Siddiqui, H.-Y. Lee, S. Tulyakov, and M. Nießner. Text2Tex: Text-driven texture synthesis via diffusion models. In *ICCV*, 2023. 26

[9] H. Chen, B. Shen, Y. Liu, R. Shi, L. Zhou, C. Z. Lin, J. Gu, H. Su, G. Wetzstein, and L. Guibas. 3D-Adapter: Geometry-consistent multi-view diffusion for high-quality 3D generation. *arXiv preprint arXiv:2410.18974*, 2024. 3

[10] H. Chen, R. Shi, Y. Liu, B. Shen, J. Gu, G. Wetzstein, H. Su, and L. Guibas. Generic 3D diffusion adapter using controlled multi-view editing. *arXiv preprint arXiv:2403.12032*, 2024. 3

[11] Z. Chen, F. Wang, Y. Wang, and H. Liu. Text-to-3D using gaussian splatting. In *CVPR*, 2024. 1

[12] K. Do and B.-S. Hua. Text-to-3D generation using Jensen-Shannon score distillation. *arXiv preprint arXiv:2503.10660*, 2025. 1

[13] Z. Fan, Z. Yin, G. Li, Y. Zhan, and H. Zheng. DreamBooth++: Boosting subject-driven generation via region-level references packing. In *ACM MM*, 2024. 9

[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3

[15] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 9

[16] Google. Introducing Gemini 2.0: our new AI model for the agentic era, 2025. 9

[17] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 3

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4

[19] S. Hong, D. Ahn, and S. Kim. Debiasing scores and prompts of 2D diffusion for view-consistent text-to-3D generation. In *NeurIPS*, 2023. 3

[20] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 24

[21] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. LRM: Large reconstruction model for single image to 3D. In *ICLR*, 2024. 3, 5, 25

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 6, 24

[23] H. Hu, T. Yin, F. Luan, Y. Hu, H. Tan, Z. Xu, S. Bi, S. Tulsiani, and K. Zhang. Turbo3D: Ultra-fast text-to-3D generation. In *CVPR*, 2025. 3, 25, 28

[24] N. Huang, T. Zhang, Y. Yuan, D. Chen, and S. Zhang. Customize-It-3D: High-quality 3D creation from a single image using subject-specific knowledge prior. *arXiv preprint arXiv:2312.11535*, 2023. 3, 7, 8, 9, 37

[25] T. Huang, Y. Zeng, Z. Zhang, W. Xu, H. Xu, S. Xu, R. W. H. Lau, and W. Zuo. DreamControl: Control-based text-to-3D generation with 3D self-prior. In *CVPR*, 2024. 3

[26] Y. Huang, J. Wang, Y. Shi, B. Tang, X. Qi, and L. Zhang. DreamTime: An improved optimization strategy for diffusion-guided 3D generation. In *ICLR*, 2024. 1

[27] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani. SPAR3D: Stable point-aware reconstruction of 3D objects from single images. In *CVPR*, 2025. 2, 3, 32, 35

[28] K.-H. Hui, A. Sanghi, A. Rampini, K. R. Malekshan, Z. Liu, H. Shayani, and C.-W. Fu. Make-A-Shape: a ten-million-scale 3D shape model. In *ICML*, 2024. 3

[29] C. Jiang, Y. Zeng, T. Hu, S. Xu, W. Zhang, W. Xu, and D.-Y. Yeung. JointDreamer: Ensuring geometry consistency and text congruence in text-to-3D generation via joint score distillation. In *ECCV*, 2024. 1

[30] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, and Z. Liu. VideoBooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024. 3

[31] Y. Jiang, C. Yu, C. Cao, F. Wang, W. Hu, and J. Gao. Animate3D: Animating any 3D model with multi-view video diffusion. In *NeurIPS*, 2024. 3

[32] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao. Consistent4D: Consistent 360° dynamic object generation from monocular video. In *ICLR*, 2024. 3

[33] Y. Kant, A. Siarohin, M. Vasilkovsky, R. A. Guler, J. Ren, S. Tulyakov, and I. Gilitschenski. iNVS: Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia*, 2023. 6, 28

[34] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. CoTracker: It is better to track together. In *ECCV*, 2024. 5

[35] O. Katzir, O. Patashnik, D. Cohen-Or, and D. Lischinski. Noise-free score distillation. In *ICLR*, 2024. 3

[36] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023. 3

[37] J.-G. Kwak, E. Dong, Y. Jin, H. Ko, S. Mahajan, and K. M. Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *CVPR*, 2024. 3

[38] Z. Lai, Y. Zhao, H. Liu, Z. Zhao, Q. Lin, H. Shi, X. Yang, M. Yang, S. Yang, Y. Feng, S. Zhang, X. Huang, D. Luo, F. Yang, F. Yang, L. Wang, S. Liu, Y. Tang, Y. Cai, Z. He, T. Liu, Y. Liu, J. Jiang, Linus, J. Huang, and C. Guo. Hunyuan3D 2.5: Towards high-fidelity 3D assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025. 7, 8, 31, 34

[39] K. Lee, K. Sohn, and J. Shin. DreamFlow: High-quality text-to-3D generation by approximating probability flow. In *ICLR*, 2024. 3

[40] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. In *ICLR*, 2024. 3

[41] P. Li, Y. Liu, X. Long, F. Zhang, C. Lin, M. Li, X. Qi, S. Zhang, W. Xue, W. Luo, P. Tan, W. Wang, Q. Liu, and Y. Guo. Era3D: High-resolution multiview diffusion using efficient row-wise attention. In *NeurIPS*, 2024. 3

[42] W. Li, J. Liu, H. Yan, R. Chen, Y. Liang, X. Chen, P. Tan, and X. Long. CraftsMan3D: High-fidelity mesh generation with 3D native generation and interactive geometry refiner. In *CVPR*, 2025. 2

[43] Z. Li, Y. Chen, and P. Liu. DreamMesh4D: Video-to-4D generation with sparse-controlled gaussian-mesh hybrid representation. In *NeurIPS*, 2024. 3

[44] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei. Diffusion4D: Fast spatial-temporal consistent 4D generation via video diffusion models. In *NeurIPS*, 2024. 3

[45] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3D: High-resolution text-to-3D content creation. In *CVPR*, 2023. 3

[46] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis. Align your gaussians: Text-to-4D with dynamic 3D gaussians and composed diffusion models. In *CVPR*, 2024. 1

[47] F. Liu, H. Wang, W. Chen, H. Sun, and Y. Duan. Make-Your-3D: Fast and consistent subject-driven 3D content generation. In *ECCV*, 2024. 3, 7

[48] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 3

[49] M. Liu, R. Shi, L. Chen, Z. Zhang, C. Xu, X. Wei, H. Chen, C. Zeng, J. Gu, and H. Su. One-2-3-45++: Fast single image to 3D objects with consistent multi-view generation and 3D diffusion. In *CVPR*, 2024. 3, 32, 35

[50] M. Liu, C. Zeng, X. Wei, R. Shi, L. Chen, C. Xu, M. Zhang, Z. Wang, X. Zhang, I. Liu, H. Wu, and H. Su. MeshFormer: High-quality mesh generation with 3D-guided reconstruction model. In *NeurIPS*, 2024. 2, 3, 7, 25, 32

[51] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, 2023. 1, 3

[52] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *ICLR*, 2024. 3

[53] Y. Liu, M. Xie, H. Liu, and T.-T. Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia*, 2024.

[54] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, and W. Wang. Wonder3D: Single image to 3D using cross-domain diffusion. In *CVPR*, 2024. 2, 6, 7, 28, 31, 32, 36

[55] Y. Lu, J. Zhang, S. Li, T. Fang, D. McKinnon, Y. Tsin, L. Quan, X. Cao, and Y. Yao. Direct2.5: Diverse text-to-3D generation via multi-view 2.5D diffusion. In *CVPR*, 2024. 3

[56] A. Lukoianov, H. S. de Ocáriz Borde, K. Greenewald, V. C. Guizilini, T. Bagautdinov, V. Sitzmann, and J. Solomon. Score distillation via reparametrized DDIM. In *NeurIPS*, 2024. 25

[57] J. Ma, J. Liang, C. Chen, and H. Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *SIGGRAPH*, 2024. 28

[58] Z. Ma, Y. Wei, Y. Zhang, X. Zhu, Z. Lei, and L. Zhang. ScaleDreamer: Scalable text-to-3D synthesis with asynchronous score distillation. In *ECCV*, 2024. 25

[59] D. McAllister, S. Ge, J.-B. Huang, D. W. Jacobs, A. A. Efros, A. Holynski, and A. Kanazawa. Rethinking score distillation as a bridge between image distributions. In *NeurIPS*, 2024. 1

[60] L. Melas-Kyriazi, I. Laina, C. Rupprecht, N. Neverova, A. Vedaldi, O. Gafni, and F. Kokkinos. IM-3D: Iterative multiview diffusion and reconstruction for high-quality 3D generation. In *ICML*, 2024. 2

[61] 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. In *ECCV*, 2020. 26

[62] A. Mirzaei, T. Aumentado-Armstrong, M. A. Brubaker, J. Kelly, A. Levinshtein, K. G. Derpanis, and I. Gilitschenski. Watch your steps: Local image and scene editing by text instructions. In *ECCV*, 2024. 6

[63] Mistral AI. Mistral small 3.1, 2025. 9

[64] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 9

[65] OpenAI. OpenAI o3 and o4-mini system card, 2025. 9

[66] Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia. DreamBench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025. 9, 29

[67] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 24

[68] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 1, 3

[69] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, and B. Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *ICLR*, 2024. 3

[70] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron, Y. Li, and V. Jampani. DreamBooth3D: Subject-driven text-to-3D generation. In *ICCV*, 2023. 3, 7

[71] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. 27

[72] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu. DreamGaussian4D: Generative 4D gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 3

[73] J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, and H. Ling. L4GM: Large 4D gaussian reconstruction model. In *NeurIPS*, 2024. 1, 2, 3, 5, 6, 7, 8, 9, 28, 31, 33, 37

[74] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 4, 6

[75] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3, 7, 8, 9, 24, 27

[76] J. Shi, W. Xiong, Z. Lin, and H. J. Jung. InstantBooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 28

[77] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[78] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang. MVDream: Multi-view diffusion for 3D generation. In *ICLR*, 2024. 2

[79] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson, and Y. Taigman. Text-to-4D dynamic scene generation. In *ICML*, 2023. 1

[80] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[81] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[82] Q. Sun, Z. Guo, Z. Wan, J. N. Yan, S. Yin, W. Zhou, J. Liao, and H. Li. EG4D: Explicit generation of 4D object without score distillation. In *ICLR*, 2025. 3

[83] W. Sun, S. Chen, F. Liu, Z. Chen, Y. Duan, J. Zhang, and Y. Wang. DimensionX: Create any 3D and 4D scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 3

[84] Z. Sun, T. Wu, P. Zhang, Y. Zang, X. Dong, Y. Xiong, D. Lin, and J. Wang. Bootstrap3D: Improving multi-view diffusion model with synthetic data. *arXiv preprint arXiv:2406.00093*, 2024. 2

[85] S. Szymanowicz, C. Rupprecht, and A. Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *CVPR*, 2024. 5

[86] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. LGM: Large multi-view gaussian model for high-resolution 3D content creation. In *ECCV*, 2024. 3, 5, 6, 7, 8, 31, 32, 36

[87] J. Tang, R. Lu, X. Chen, X. Wen, G. Zeng, and Z. Liu. InTeX: Interactive text-to-texture synthesis via unified depth-aware inpainting. *arXiv preprint arXiv:2403.11878*, 2024. 26

[88] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng. DreamGaussian: Generative gaussian splatting for efficient 3D content creation. In *ICLR*, 2024. 1

[89] L. Tang, N. Ruiz, Q. Chu, Y. Li, A. Holynski, D. E. Jacobs, B. Hariharan, Y. Pritch, N. Wadhwa, K. Aberman, and M. Rubinstein. RealFill: Reference-driven generation for authentic image completion. *ACM TOG*, 43(4), 2024. 3, 6, 28

[90] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *ECCV*, 2024. 7, 32, 36

[91] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *CVPR*, 2023. 1

[92] P. Wang and Y. Shi. ImageDream: Image-prompt multi-view diffusion for 3D generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 3, 6

[93] P. Wang, D. Xu, Z. Fan, D. Wang, S. Mohan, F. Iandola, R. Ranjan, Y. Li, Q. Liu, Z. Wang, and V. Chandra. Taming mode collapse in score distillation for text-to-3D generation. In *CVPR*, 2024. 25

[94] Y. Wang, X. Wang, Z. Chen, Z. Wang, F. Sun, and J. Zhu. Vidu4D: Single generated video to high-fidelity 4D reconstruction with dynamic gaussian surfels. In *NeurIPS*, 2024. 3

[95] Y. Wang, F. Hong, S. Yang, L. Jiang, W. Wu, and C. C. Loy. MEAT: Multiview diffusion model for human generation on megapixels with mesh attention. In *CVPR*, 2025. 2

[96] 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. In *NeurIPS*, 2023. 1, 3, 25

[97] Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu. CRM: Single image to 3D textured mesh with convolutional reconstruction model. In *ECCV*, 2024. 3

[98] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, and M. Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 3

[99] M. Wei, J. Zhou, J. Sun, and X. Zhang. Adversarial score distillation: When score distillation meets GAN. In *CVPR*, 2024. 3

[100] S.-T. Wei, R.-H. Wang, C.-Z. Zhou, B. Chen, and P.-S. Wang. OctGPT: Octree-based multiscale autoregressive models for 3D shape generation. In *SIGGRAPH*, 2025. 3

[101] K. Wu, F. Liu, Z. Cai, R. Yan, H. Wang, Y. Hu, Y. Duan, and K. Ma. Unique3D: High-quality and efficient 3D mesh generation from a single image. In *NeurIPS*, 2024. 3

[102] S. Wu, Y. Lin, Y. Zeng, F. Zhang, J. Xu, P. Torr, X. Cao, and Y. Yao. Direct3D: Scalable image-to-3D generation via 3D latent diffusion transformer. In *NeurIPS*, 2024. 2, 3, 25, 32, 35

[103] Z. Wu, C. Yu, Y. Jiang, C. Cao, F. Wang, and X. Bai. SC4D: Sparse-controlled video-to-4D generation and motion transfer. In *ECCV*, 2024. 3

[104] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang. Structured 3D latents for scalable and versatile 3D generation. In *CVPR*, 2025. 2, 3, 7, 25, 32

[105] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, 2024. 3

[106] D. Xie, J. Li, H. Tan, X. Sun, Z. Shu, Y. Zhou, S. Bi, S. Pirk, and A. E. Kaufman. Carve3D: Improving multi-view reconstruction consistency for diffusion models with RL finetuning. In *CVPR*, 2024. 2

[107] K. Xie, J. Lorraine, T. Cao, J. Gao, J. Lucas, A. Torralba, S. Fidler, and X. Zeng. LATTE3D: Large-scale amortized text-to-enhanced3D synthesis. In *ECCV*, 2024. 1

[108] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani. SV4D: Dynamic 3D content generation with multi-frame and multi-view consistency. In *ICLR*, 2025. 3, 7, 9, 32, 38

[109] D. Xu, H. Liang, N. P. Bhatt, H. Hu, H. Liang, K. N. Plataniotis, and Z. Wang. Comp4D: LLM-guided compositional 4D scene generation. *arXiv preprint arXiv:2403.16993*, 2024. 3

[110] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 25

[111] Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. In *ECCV*, 2024. 3

[112] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu, and K. Zhang. DMV3D: Denoising multi-view diffusion using 3D large reconstruction model. In *ICLR*, 2024. 3

[113] R. Yan, Y. Chen, and X. Wang. Consistent flow distillation for text-to-3D generation. In *ICLR*, 2025. 25

[114] X. Yang, H. Shi, B. Zhang, F. Yang, J. Wang, H. Zhao, X. Liu, X. Wang, Q. Lin, J. Yu, L. Wang, J. Xu, Z. He, Z. Chen, S. Liu, J. Wu, Y. Lian, S. Yang, Y. Liu, Y. Yang, D. Wang, J. Jiang, and C. Guo. Hunyuan3D 1.0: A unified framework for text-to-3D and image-to-3D generation. *arXiv preprint arXiv:2411.02293*, 2024. 2, 3, 25, 32, 35

[115] Y. Yang, Y. Huang, X. Wu, Y.-C. Guo, S.-H. Zhang, H. Zhao, T. He, and X. Liu. DreamComposer: Controllable 3D object generation via multi-view conditions. In *CVPR*, 2024. 2

[116] Z. Yang, Z. Pan, C. Gu, and L. Zhang. Diffusion^2: Dynamic 3D content generation via score composition of video and multi-view diffusion models. In *ICLR*, 2025. 3

[117] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, X. Gu, Y. Zhang, W. Wang, Y. Cheng, T. Liu, B. Xu, Y. Dong, and J. Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 24, 27

[118] C.-H. Yao, Y. Xie, V. Voleti, H. Jiang, and V. Jampani. SV4D 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4D generation. *arXiv preprint arXiv:2503.16396*, 2025. 3

[119] J. Ye, F. Liu, Q. Li, Z. Wang, Y. Wang, X. Wang, Y. Duan, and J. Zhu. DreamReward: Text-to-3D generation with human preference. In *ECCV*, 2024. 1

[120] H. Yu, C. Wang, P. Zhuang, W. Menapace, A. Siarohin, J. Cao, L. A. Jeni, S. Tulyakov, and H.-Y. Lee. 4Real: Towards photorealistic 4D scene generation via video diffusion models. In *NeurIPS*, 2024. 3, 25

[121] X. Yu, Y.-C. Guo, Y. Li, D. Liang, S.-H. Zhang, and X. Qi. Text-to-3D with classifier score distillation. In *ICLR*, 2024. 3

[122] B. Zeng, L. Yang, S. Li, J. Liu, Z. Zhang, J. Tian, K. Zhu, Y. Guo, F.-Y. Wang, M. Xu, S. Ermon, and W. Zhang. Trans4D: Realistic geometry-aware transition for compositional text-to-4D synthesis. *arXiv preprint arXiv:2410.07155*, 2024. 3

[123] Y. Zeng, Y. Jiang, S. Zhu, Y. Lu, Y. Lin, H. Zhu, W. Hu, X. Cao, and Y. Yao. STAG4D: Spatial-temporal anchored generative 4D gaussians. In *ECCV*, 2024. 1, 3, 7, 8, 9, 37

[124] B. Zhang, J. Tang, M. Nießner, and P. Wonka. 3DShape2VecSet: A 3D shape representation for neural fields and generative diffusion models. *ACM TOG*, 42(4), 2023. 2

[125] H. Zhang, X. Chen, Y. Wang, X. Liu, Y. Wang, and Y. Qiao. 4Diffusion: Multi-view video diffusion model for 4D generation. In *NeurIPS*, 2024. 3

[126] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu. GS-LRM: Large reconstruction model for 3D gaussian splatting. In *ECCV*, 2024. 3

[127] L. Zhang, Z. Wang, Q. Zhang, Q. Qiu, A. Pang, H. Jiang, W. Yang, L. Xu, and J. Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM TOG*, 43(4), 2024. 3, 25

[128] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee. Animate124: Animating one image to 4D dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. 3

[129] Z. Zhao, W. Liu, X. Chen, X. Zeng, R. Wang, P. Cheng, B. Fu, T. Chen, G. Yu, and S. Gao. Michelangelo: Conditional 3D shape generation based on shape-image-text aligned latent representation. In *NeurIPS*, 2023. 1

[130] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, H. Shi, S. Liu, J. Wu, Y. Lian, F. Yang, R. Tang, Z. He, X. Wang, J. Liu, X. Zuo, Z. Chen, B. Lei, H. Weng, J. Xu, Y. Zhu, X. Liu, L. Xu, C. Hu, S. Yang, S. Zhang, Y. Liu, T. Huang, L. Wang, J. Zhang, M. Chen, L. Dong, Y. Jia, Y. Cai, J. Yu, Y. Tang, H. Zhang, Z. Ye, P. He, R. Wu, C. Zhang, Y. Tan, J. Xiao, Y. Tao, J. Zhu, J. Xue, K. Liu, C. Zhao, X. Wu, Z. Hu, L. Qin, J. Peng, Z. Li, M. Chen, X. Zhang, L. Niu, P. Wang, Y. Wang, H. Kuang, Z. Fan, X. Zheng, W. Zhuang, Y. He, T. Liu, Y. Yang, D. Wang, Y. Liu, J. Jiang, J. Huang, and C. Guo. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 2, 3, 25, 32

[131] C. Zheng, Y. Lin, B. Liu, X. Xu, Y. Nie, and S. He. RecDreamer: Consistent text-to-3D generation via uniform score distillation. In *ICLR*, 2025. 25

[132] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, and S. D. Mello. A unified approach for text- and image-guided 4D scene generation. In *CVPR*, 2024. 1, 3

NeurIPS Paper Checklist

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 claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   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: The limitations of the work have been discussed in App. C in the appendix.

   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.
   - 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: The paper does not include theoretical results.

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 have discussed all the details to reproduce all experimental results in our paper.

   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.
     (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: [No]

Justification: While the current submission does not contain code implementation, we have included detailed instructions in the appendix to provide guidance for reproducing the main experimental results. Also, we promise that we will open-source the data and code after paper acceptance.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details have been specified in the main paper and the appendix, which are enough to understand the results.

   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?

   Answer: [No]

   Justification: The paper mainly focuses on qualitative comparisons.

   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: The compute resources required for reproduce the is specified in App. C in the appendix.

   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 conducted research conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - 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: Both positive and negative societal impacts have been discussed in App. G in the appendix.

    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: [NA]

    Justification: Our paper does not have such risks.

    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.
    - 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: We have properly credited and listed the license in App. B in the appendix. The assets are also properly used under the terms of use.

    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: At the submission time, we have not released the new assets used in the paper. We will well document our data and model at the time we release our data after acceptance.

    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: We have included the full text of instructions given to the participants in App. A in the appendix, as well as the screenshots of the user study.

    Guidelines:

    - 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: There is no potential risks in the crowdsourcing research conducted for this work.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Track, Inpaint, Resplat: Subject-driven 3D and 4D Generation with Progressive Texture Infilling

Technical Appendices and Supplementary Material

In the appendix, we first provide more details on the instructions and interface used in our user study in App. A to show how we foster the fairness and soundness of our user study. Next, in App. B, we illustrate the pipeline of how we construct our dataset for subject-driven 3D/4D generation. In App. C, we carry out more detailed discussions on how our proposed TIRE can *co-exist* and *co-develop* with other 3D/4D generation methods, the insights of our method design, together with the limitations of our approach. Then, more ablation study is presented in App. E. Afterwards, we provide additional qualitative results in App. F on in-the-wild data, more state-of-the-art advancements on 3D generation, *etc.* Moreover, in App. D, we elaborate on more implementation details to better support the reproducibility of the proposed method. Finally, the societal impact of our work is discussed in App. G.

A User Study

Below we present more details of our user study setup. We use the following instructions for participants at the start of the user study.

For each scene, we use three different methods to generate 4D assets, which are rendered in the forms of videos and placed in random orders. We also provide a reference image to show the appearance of the subject. We would like to invite you to give an overall score from 1 to 10 to measure the quality of the generated results.

There are a few points to consider when providing the scores:

- *Whether the videos are consistent across different views*
- *Whether different viewpoints of the generated results look alike the subject in the given reference image*
- *The visual quality of the videos (e.g., whether it has blurry artifacts, whether the rendered views look realistic)*

From the above instructions, note that we *do not* explicitly tell the participants that we are improving on subject-driven 4D generation. Instead, we ask the users to rate the generation results based on the overall quality. It fosters the fairness in user study and reduces the underlying bias of the subjective preference towards our method. We also include screenshots of our user study interface in Fig. A and Fig. B.

B More Details on Datasets

We show an overview of our dataset construction pipeline in Fig. C. More specifically, we train a customized text-to-image generation model on each subject following DreamBooth [75], with the backbone being an SDXL [67] model finetuned with LoRA [22]. The link of the dataset is `https://github.com/google/dreambooth`, and the license is Creative Commons Attribution 4.0 International (CC-BY-4.0). Then, we use manually created prompts to generate images with the subjects performing specific activities. Afterwards, we adopt the image-to-video model CogVideoX [20, 117] to create an animated video capturing the scene. Altogether, we curate a total of 23 animable subjects in the original DreamBooth dataset performing diverse activities.

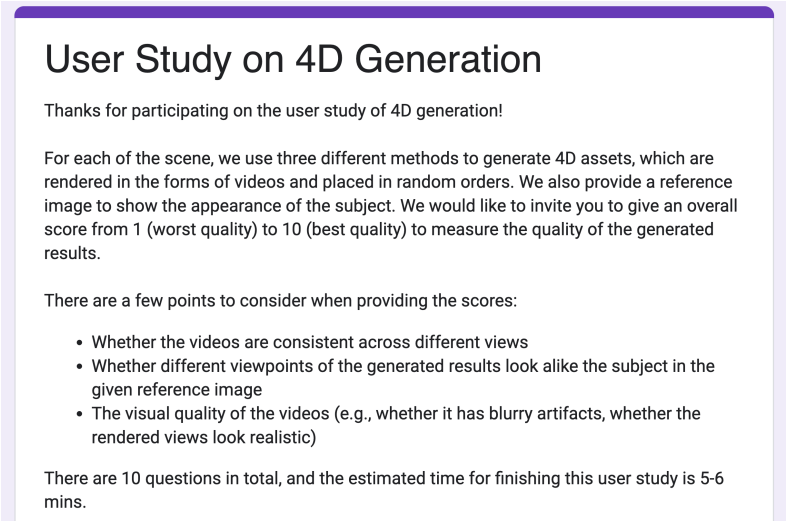

Figure A: Screenshot of the instructions of our user study. From the instructions, we *do not* explicitly tell the participants that our project is targeting at subject-driven 4D generation (*e.g.,* in the title we only mention "User Study on 4D Generation"). Instead, the users are asked to rate the generated results based on the overall quality, which enhances the fairness and reduces the underlying bias of the potential subjective preference towards our method.

C Discussions and Limitations

Position of our proposed TIRE relative to existing models. Existing state-of-the-art 3D/4D generation models [21, 23, 50, 58, 102, 104, 110, 114, 127, 130] are mostly feed-forward models with native 3D generation to achieve superior generation efficiency. Therefore, it is no denying that their feed-forward generation schema pioneers the "foundation models" for 3D/4D generation. Per-scene optimization-based methods [56, 93, 96, 113, 120, 131], do suffer from the limitation of efficiency as discussed in the training efficiency part in this section. However, we consider our method as an orthogonal effort towards subject-driven generation which is more customized. From the perspective of customers, it is valuable to have an option, supported by our method, to further preserve the identity and refine the generated results produced by the other feed-forward 3D/4D generation methods.

From a methodology perspective, some may argue that, to resolve the issues like the blueish shade shown in Fig. 1 in the main paper, we could heuristically design specific rules to filter out problematic data before training our model. However, it is infeasible to enumerate all the issues that may lead to data bias during the data filtering stage, as the issue shown in Fig. 1 is only one specific example of the many underlying issues that undermine the robustness of the feed-forward generation models. For example, if we place the rendered views of the generated assets from a specific model side by side, as shown in Fig. D, it becomes obvious to see that each model suffers from a specific bias pattern for side views and back views. For L4GM, the generated assets suffer from the unrealistic blueish or whitish color patterns similar to the case in Fig. 1. For SV3D, there are whitish tone and blurry texture at the generated viewpoints, which may originate from the imbalance object properties of the dataset used for training the model, or the bias from the network design. For Hunyuan3D-v2.5, although the shape and color tones looks generally reasonable, the appearance is over-smoothed and looks unrealistic. This could be still the consequence of using certain rules for data filtering and preprocessing, and the rest of the data gets biased towards the opposite direction. Therefore, it is extremely challenging and laborious to design data filtering principles to address the issue.

Insights and innovations of the proposed TIRE pipeline. Our solution is not a simple concatenation of existing methods, as seamlessly combining these works (some of them like video tracking methods are even seldom used in 3D generation) itself is already a non-trivial task. Our insight for devising the TIRE paradigm is that we can convert the problem of progressive inpainting for 3D assets to 2D problems, with *Track* and *Inpaint* identify and fill the occluded regions step by step with 2D

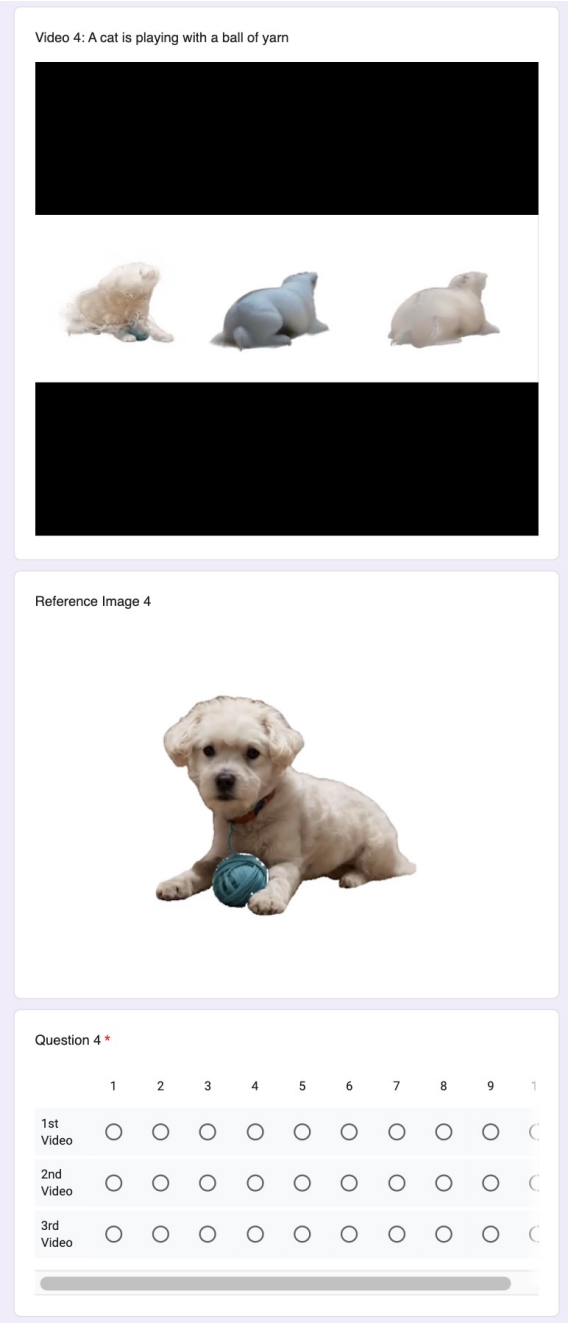

Figure B: Screenshot of an example question, containing videos generated by three methods placed side by side in random order, the reference front view, and the questions to score the three generated results.

models, followed by *Resplat* to fix the consistency in 3D. As a result, our pipeline is able to leverage the powerful 2D foundation models that are trained on substantially abundant and more diverse 2D data. Also, compared with other inpainting-based texture completion methods like Text2Tex [8] and InTeX [87], our pipeline can perform on any 3D representation including the implicit representations like neural radiance fields [61], while Text2Tex and InTeX require the adopted 3D representation to support surface modeling. Moreover, the three-stage pipeline is well orchestrated in a logical cascade. For example, the *Track* stage not only identifies the regions that need to be inpainted, but also

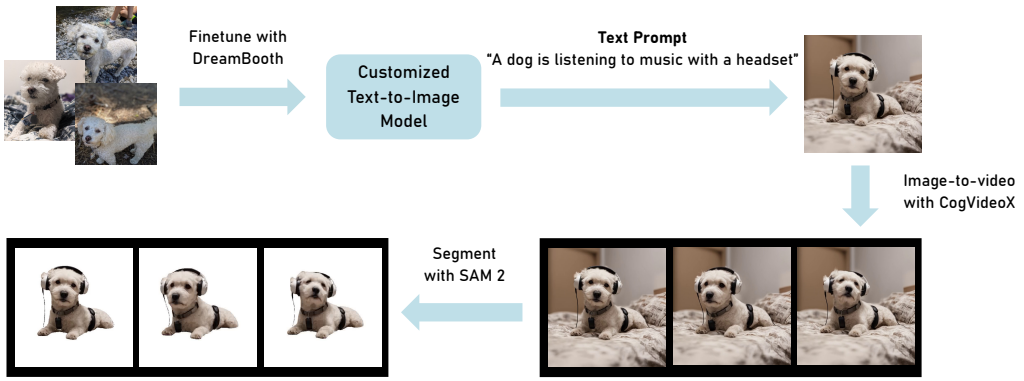

Figure C: Pipeline of constructing data for our *DreamBooth-Dynamic* dataset. First, a DreamBooth-finetuned [75] customized text-to-image model is trained on several casually collected images of a subject. Afterwards, we use manually created text prompts to generate images with the specific subjects. Afterwards, a powerful image-to-video model [117] is applied to animate the static images into videos. Finally, we use SAM 2 [71] to segment the foreground subjects to form the source view videos used in our evaluation process.

Figure D: Side and back viewpoints generated by specific models. It is obvious to see each model suffers from a bias pattern for generating the appearance of the occluded viewpoints.

propagate the corresponding pixels along the tracking path to novel views to mitigate the difficulty of the following *Inpaint* stage.

Quantitative metrics for 3D/4D subject-driven generation. As shown in Sec. 4.3 in the main paper, the quantitative metric adopted in DreamBooth leads to unreasonable results. We visualize an example in Fig. E to demonstrate the failure case of the DINO similarity metric. We can observe that the compared methods Customize-It-3D and STAG4D have obviously incorrect geometry for this viewpoint which is 135° away from the source view. However, as their shape or appearance looks similar to the reference which is the source view image, they score significantly higher than L4GM and our method, which at least correctly display the back view of the subject. Therefore, we believe that currently there are limitations in the quantitative evaluation on the subject fidelity of 3D/4D generation. A better evaluation metric for subject-driven 3D/4D generation needs to be proposed for properly benchmarking the performance of the methods. Potential solutions could be integrating the geometric evaluation with the appearance evaluation that is conducted by the current DINO feature similarity metric. The geometric assessment may pertain to the structural characteristics of the foreground of the generated images. We leave the design of proper evaluation metrics on subject-driven 3D/4D generation as future work.

Training efficiency. Our method takes around 100 mins on a single NVIDIA A100 GPU (when considering the memory consumption, an NVIDIA Quadro RTX 6000 or RTX 4090 with 24GB

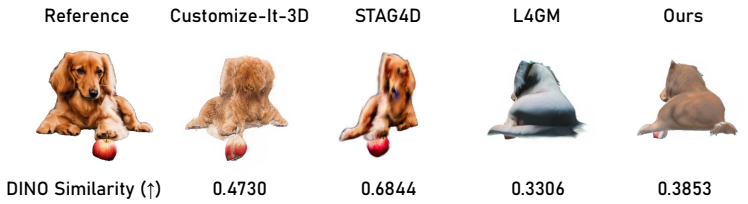

Figure E: An example of DINO (ViT-S/16) similarity scores of a rendered viewpoint which is 135° away from the source view across different methods. The baseline methods Customize-It-3D and STAG4D are obviously incorrect regarding geometry but get higher scores for DINO similarity.

will suffice), which is not very efficient. This issue is inherited from RealFill [89] which our method's pipeline is based on. However, since our solution focuses more on subject-driven 3D/4D generation, our work is an orthogonal effort to existing methods like L4GM [73], Turbo3D [23], *etc.* that work towards increasingly faster feed-forward 3D/4D generation pipelines. Moreover, personalization models are also in the trend of becoming feed-forward models [57, 76], without the need of finetuning on every subject like DreamBooth. Therefore, we believe that there will be more efficient subject-driven 2D inpainting models in the future that can greatly improve the efficiency of our approach.

Why there are no specific operations designed for 4D generation in the temporal dimension? Subject-driven generation, in the current context, mainly refers to the point that the appearance of the generated asset needs to conform to the reference image, which is often a static attribute. Therefore, temporal reasoning is not needed in most cases. Also, in our current implementation, the method L4GM [73] that we builds upon for 4D generation essentially generates a sequence of 3D Gaussians as mentioned in Sec. 3.1, which makes 4D generation, in the current situation, can be understood as the composition of a series of 3D generation processes. Nevertheless, we do acknowledge that there exist significant challenges specifically for 4D generation compared with 3D generation, and there are more complicated cases that customization in 4D generation also becomes important and especially challenging. For example, the reference objects shown in the given video may contain temporal patterns that change their appearance over time, and we want to generate 4D assets that faithfully preserve this property. We leave this as an interesting future direction to explore.

D Additional Implementation Details

D.1 Training Details

As mentioned in the main paper, during the inpainting process, as we do not want to significant change the original structure of the observations from novel views, we perform denoising in the first 30% of the denoising schedule following similar practice as [33]. Therefore, the original structure of the images will not be destroyed after the inpainting process. Since we set the denoising schedule to only 30%, the original color of the images from the other views cannot be too deviant, which sometimes may not be satisfied with the original multi-view observations obtained by ImageDream. Therefore, we leverage another multi-view diffusion model Wonder3D [54], which empirically has inferior geometry but superior color than ImageDream, to provide guidance for color. Specifically, we splat multi-view observations from both ImageDream and Wonder3D to 3D Gaussians with LGM, and concatenate color features from Wonder3D and other features from ImageDream, followed by another halfway diffusion-denoising process with ImageDream. For progressive inpainting, except for the sweet spot $\pm 20^\circ$ and the secondary anchor viewpoint $\pm 90^\circ$, the other viewpoints are $\{\pm 40^\circ, \pm 60^\circ, \pm 80^\circ, \pm 110^\circ, \pm 130^\circ, \pm 150^\circ, \pm 170^\circ\}$. Each training stage consists of 300 training steps before advancing to the next stage. When adding the newly inpainted sample to the training data, we perform the same augmentation strategy when there is only one single source view image in the training set. We follow the other training parameters in RealFill [89] for finetuning the stable diffusion inpainting model (*e.g.,* batch size is 16, LoRA rank is 8, learning rate is 2e-4 for U-Net and 4e-5 for the text encoder). For the halfway diffusion-denoising process used after fixing the color of the rough assets with Wonder3D and before resplatting the pixels back to 3D, we find that the same denoising schedule of 0.3 as the inpainting stage generally works well in both cases.

D.2 VLM-based Evaluation Details

We follow the concept preservation evaluation proposed in DreamBench++ [66] to prepare the prompts for VLMs to assign scores for the identity preservation quality between the reference image and the rendered view of the generated asset. The prompts are shown as follows.

Task Definition

You will be provided with an image generated based on reference image.

As an experienced evaluator, your task is to evaluate the semantic consistency between the subject of the generated image and the reference image, according to the scoring criteria.

Scoring Criteria

It is often compared whether two subjects are consistent based on four basic visual features:

1. *Shape: Evaluate whether the main body outline, structure, and proportions of the generated image match those of the reference image. This includes the geometric shape of the main body, clarity of edges, relative sizes, and spatial relationships between various parts composing the main body.*
2. *Color: Comparing the accuracy and consistency of the main colors generated in the image with those of the reference image. This includes saturation, hue, brightness, and whether the distribution of colors is similar to that of the subject in the reference image.*
3. *Texture: Focus on the local parts of the RGB image, whether the generated image effectively captures fine details without appearing blurry, and whether it possesses the required realism, clarity, and aesthetic appeal. Please note that unless specifically mentioned in the text prompt, excessive abstraction and formalization of texture are not necessary.*
4. *Facial Features: If the evaluation is of a person or animal, facial features will greatly affect the judgment of image consistency, and you also need to focus on judging whether the facial area looks very similar visually.*

Scoring Range

You need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from 0 to 4:

- *Very Poor (0): No resemblance. The generated image's subject has no relation to the reference.*
- *Poor (1): Minimal resemblance. The subject falls within the same broad category but differs significantly.*
- *Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.*
- *Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.*
- *Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the reference.*

Input Format

Every time you will receive two images, the first image is a reference image, and the second image is the generated image.

Please carefully review each image of the subject.

Output Format

Score: [Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

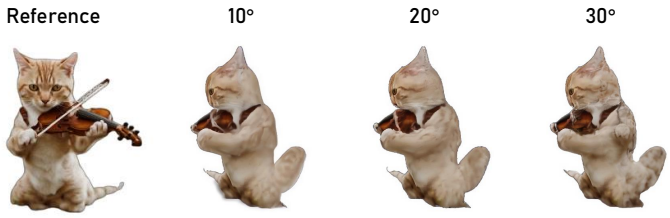

Figure F: Ablation study on the degree of progressiveness during the *Inpaint* stage. Our current choice of degree of progressiveness (20°) strikes a balance between inpainting quality and efficiency. Smaller degree of progressiveness (10°) yields decent results but slows down the whole inpainting process, while larger degree of progressiveness (30°) brings noticeable degradation in the inpainting quality.

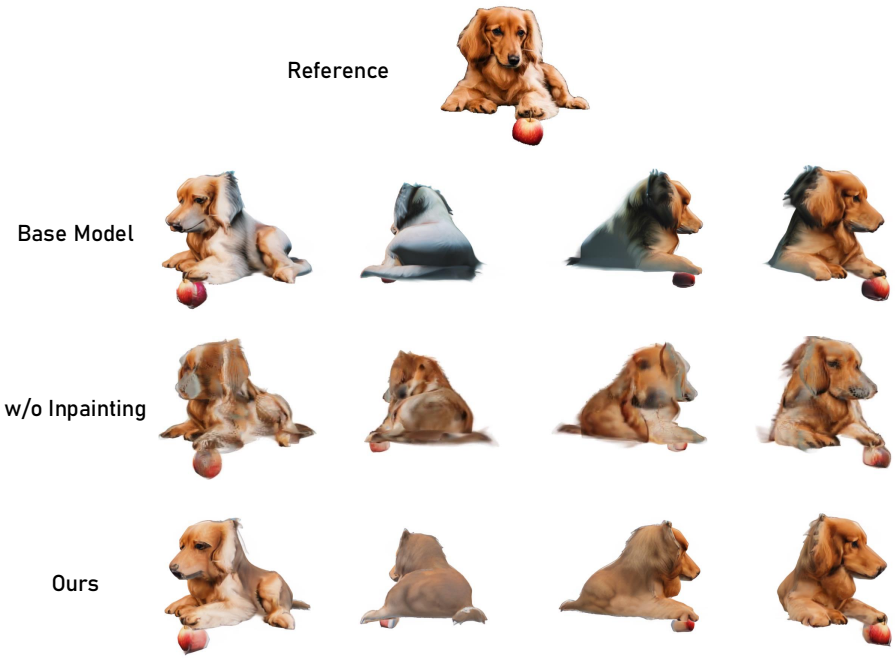

Figure G: Ablation study on the necessity of having the inpainting process. The color and texture of the assets without our inpainting process would be far from having satisfactory quality and decent identity preservation.

E More Ablation Study

E.1 Degree of Progressiveness for Inpainting

We empirically choose 20° as the degree of progressiveness in the inpainting stage as described in Sec. 3.3 for striking a balance between the inpainting quality and the algorithm efficiency. We display the ablation on selecting a smaller or larger degree of progressiveness (10° and 30°) in Fig. F to show the effect of different choices on progressiveness. Basically, for the degrees of progressiveness smaller than the current choice, the visual results are very similar to the current results, but we need additional steps to calculate the inpainting masks for the increased number of training stages. For the degrees of progressiveness larger than the current choice, we could observe that the regions (especially for those that are far away from the reference view, e.g., the regions that are close to the back view for the target views of $\pm 90°$) are inpainted worse than the current implementation. Our current choice of 20° of progressiveness balances between the number of frames that need for tracking, and the difficulty of inpainting the unseen regions that are far away from the source view.

Table A: Quantitative comparisons the assets generated by Hunyuan3D-v2.5 [38] and our enhancement built upon it. Best is marked as **bold**.

Methods	VLM-based scores (↑)						
	GPT-4o	OpenAI o4-mini	Gemma 3 27B	Gemini 2.0 Flash	Qwen2.5-VL-7B	Mistral-Small-3.1-24B-Instruct	Average
Hunyuan3D-v2.5	1.614	1.690	**2.098**	1.533	1.780	1.501	1.703
TIRE on Hunyuan3D-v2.5 (Ours)	**1.750**	**1.712**	2.092	**1.609**	**1.899**	**1.707**	**1.795**

Table B: Quantitative comparisons the assets generated by L4GM [73] and our enhancement built upon it on the in-the-wild data displayed on the L4GM official website. Best is marked as **bold**.

Methods	VLM-based scores (↑)						
	GPT-4o	OpenAI o4-mini	Gemma 3 27B	Gemini 2.0 Flash	Qwen2.5-VL-7B	Mistral-Small-3.1-24B-Instruct	Average
L4GM	**2.743**	2.441	2.640	2.022	2.104	2.125	2.346
TIRE on L4GM (Ours)	2.601	**2.478**	**2.684**	**2.110**	**2.191**	**2.163**	**2.371**

E.2 Necessity of the Inpainting Process

As mentioned in App. D, we leverage another multi-view diffusion model Wonder3D [54] which empirically has superior color than the base 3D generation model LGM [86] that we mainly use in the paper, to roughly fix the color first before applying our method. We would like to show the necessity of our method by demonstrating how much identity is restored with our inpainting process. In Fig. G, we can see that without inpainting, there are many obvious flaws from the results. For example, on the right view of the scene, the dog's nose has a large grayish patch. The dog's body is also covered with messy dark brown patterns. Therefore, it is still far away from a good 3D asset for decent identity preservation. After the inpainting is done, the large grayish patch on the dog's nose gets infilled with reasonable color, and the textures on the dog's body gets more realistic and coherent with the given source view.

F More Experimental Results

For an overview of this section, first we show the comparison on in-the-wild data to show the robustness of our method in Sec. F.1. Then, we illustrate that our method is generally applicable to all types of 3D/4D generation methods by showing its application on one of the most recent image-to-3D methods. Next, comparisons with more state-of-the-art advancements are demonstrated in Sec. F.3 to show that recent advancements are generally not robust enough to produce personalized 3D assets that well preserve the identity of the subjects. Afterwards, more comparisons on image-to-3D and video-to-4D generation are presented in Sec. F.4. Finally, the comparisons with SV4D is displayed in Sec. F.5 due to different pose settings.

F.1 Comparisons on In-the-wild Data

We additionally demonstrate the performance of our model on in-the-wild data, which are the examples displayed on the official webpage of L4GM [73], as their validation dataset has not been released. The comparison of our method against L4GM is shown in Fig. H and Tab. B. We can observe that our method more faithfully preserves the subject fidelity in the generated 4D assets and achieves superior overall quality compared with the baseline.

F.2 General Solution to 3D/4D Generation Methods

Our solution is a *plug-in solution* that is generally applicable to all types of 3D/4D representations and methods. This benefit is a natural outcome from our method design of fully leveraging 2D tools to solve this 3D task. To apply on any 3D/4D methods, it is as simple as replacing the process of generating the initial assets in our pipeline with the 3D/4D generation that we want to improve upon. We show the improvement of our method upon the current state-of-the-art advancement in 3D generation Hunyuan3D-v2.5 [38] in Fig. I and Tab. A. We can observe that our method makes refinement to the texture of the generated 3D assets of Hunyuan3D-v2.5, yielding results that better match the identity of the reference images.

F.3 Comparisons with More State-of-the-art Methods

We showcase the comparison with more state-of-the-art advancements in Fig. J, including Hunyuan3D-v1.0 [114], SPAR3D [27], and commercial models like Sudo AI[1] (built upon One-2-3-45++ [49]) and Neural4D[2] (built upon Direct3D [102]). Together with the results of MeshFormer [50], TRELLIS [104], and Hunyuan3D-v2.5 [130] in Sec. 4.2, we can observe that even the most recent advancements in 3D generation still struggle with producing well-rounded personalized 3D/4D contents that well preserve the identity of the subject.

F.4 More Qualitative Comparisons on 3D/4D Generation

We include more comparisons on image-to-3D generation with methods Wonder3D [54], SV3D [90], and LGM [86] in Fig. K. The comparisons demonstrate that our method outperforms existing image-to-3D approaches with better identity preservation and superior geometry accuracy. Additional visual comparisons on video-to-4D generation in Fig. L to demonstrate the *identity-preserving appearance* and the *enhanced quality on geometry* produced from our method, indicating the effectiveness of our approach.

F.5 Comparisons with SV4D

We display the comparison with SV4D [108] in Fig. M. The reason that we show the SV4D results separately in the supplementary material is that the official code of SV4D has only released the multi-view video generation part, without the following 4D optimization step. Therefore, only a certain number of selected views can be rendered with its official code. For simplicity, the viewpoints selected for our method are close viewpoints but not the identical ones. However, we can still observe from the qualitative comparisons that SV4D fails to produce decent results regarding both texture and geometry.

G Societal Impact

We expect our work to have a meaningful and positive impact on the society. We sincerely wish that our method can ignite the creativity inside people to design personalized 3D/4D assets that serve versatile purposes. Also, we hope that our work highlights a critical and often overlooked challenge during the evolution of 3D/4D generation: the task of generating identity-preserved 3D/4D assets for personalized applications still remains unsolved. By bringing attention to this gap, we aim to raise awareness within the research community and encourage further exploration in this direction.

Potential negative societal impact. Our work is likely the same as other research on data generation regarding potential negative societal impact with the risk of digital forgery. Also, if the technique is mistakenly used, copyright and ethical issues may occur.

[1] https://www.sudo.ai/image-to-3d

[2] https://www.neural4d.com/studio/image-to-3d

Figure H: Qualitative comparisons on the in-the-wild videos which are displayed on the official website of L4GM [73]. The visualizations demonstrate that our results more faithfully reflect the identity of the given subjects, yielding better overall quality.

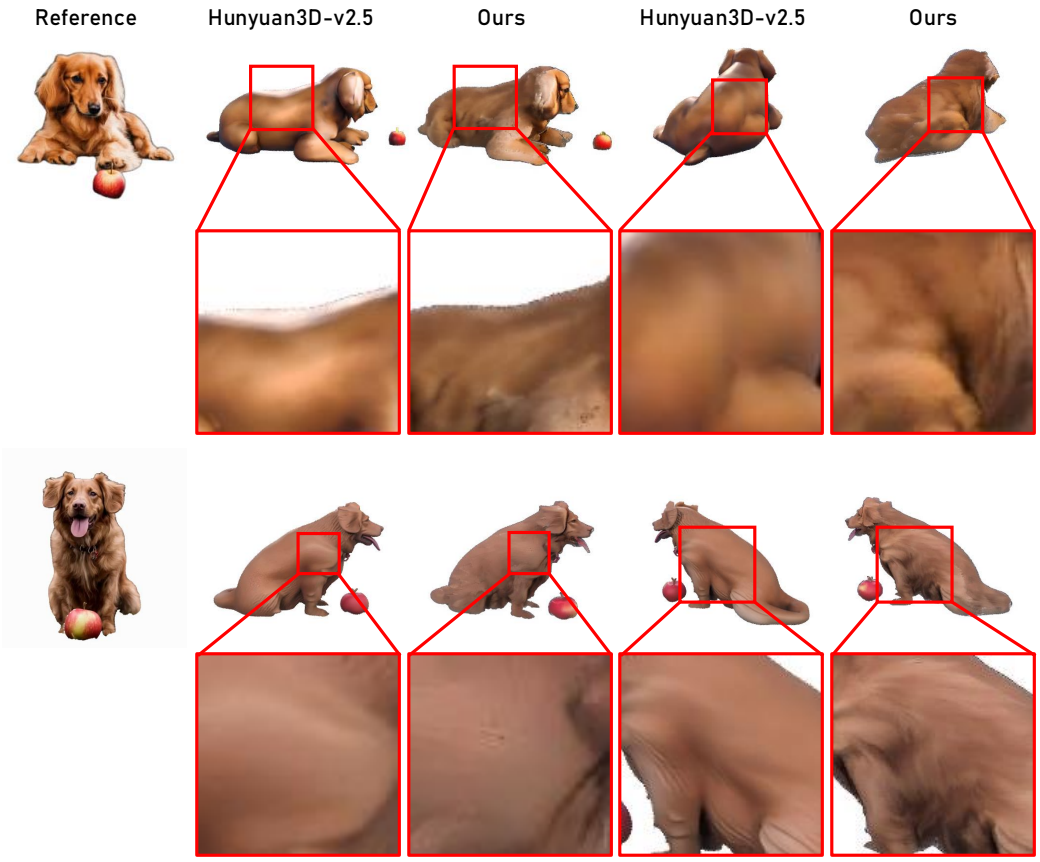

Figure I: Qualitative results showing the improvement of our method when applying on the most advanced image-to-3D method Hunyuan3D-v2.5 [38]. We can observe that our method produces superior texture of the generated assets which better match the identity of the reference image.

Figure J: Qualitative comparison with more advanced methods including Hunyuan3D-v1.0 [114], SPAR3D [27], Sudo AI (built upon One-2-3-45++ [49]), and Neural4D (built upon Direct3D [102]). Even the recent advancements in 3D generation cannot satisfactorily handle the personalized 3D generation challenge.

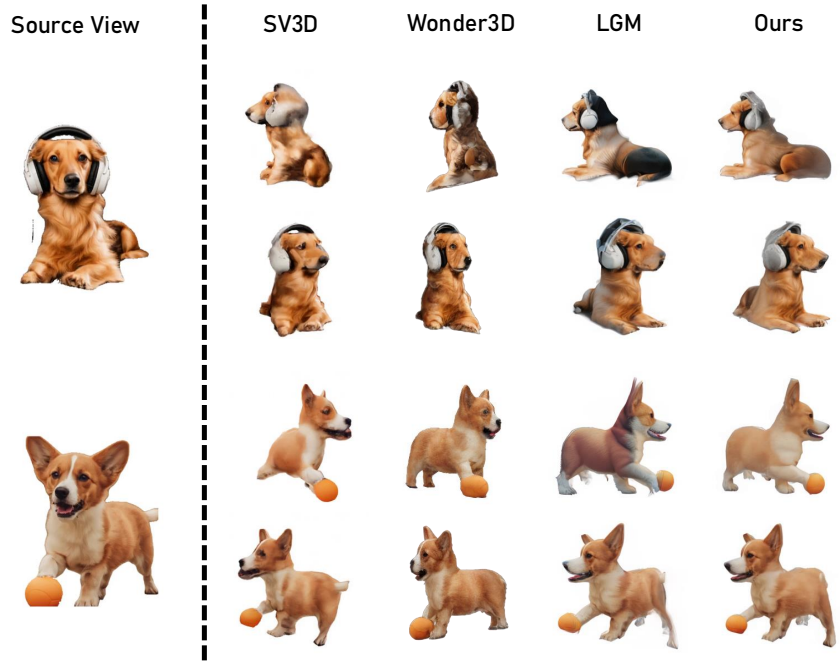

Figure K: Qualitative comparison with SV3D [90], Wonder3D [54], and LGM [86]. Compared against other method in the image-to-3D setting, our method achieves better preserves the identity of the reference image, and also reaches superior quality on geometry.

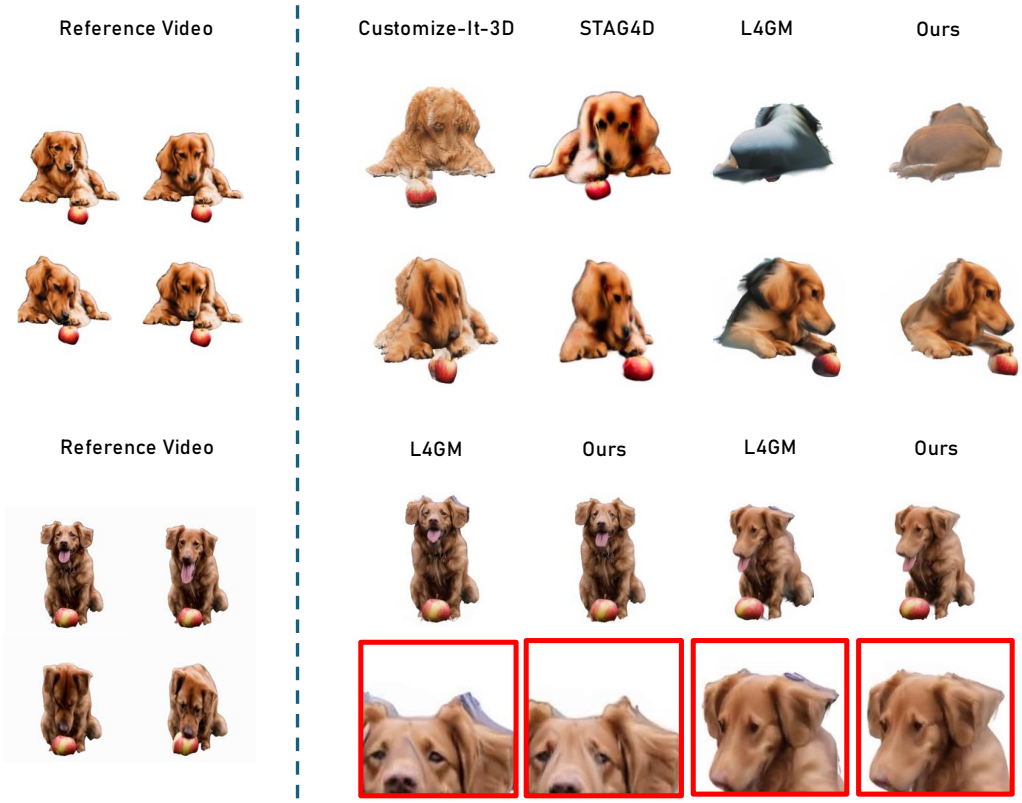

Figure L: Additional qualitative results of the comparison between our method and the baselines Customize-It-3D [24], STAG4D [123], and L4GM [73]. Compared with other methods, our solution achieves superior subject fidelity along with improved geometry on the generated assets, due to our design in our three-stage framework that fosters cross-view consistency.

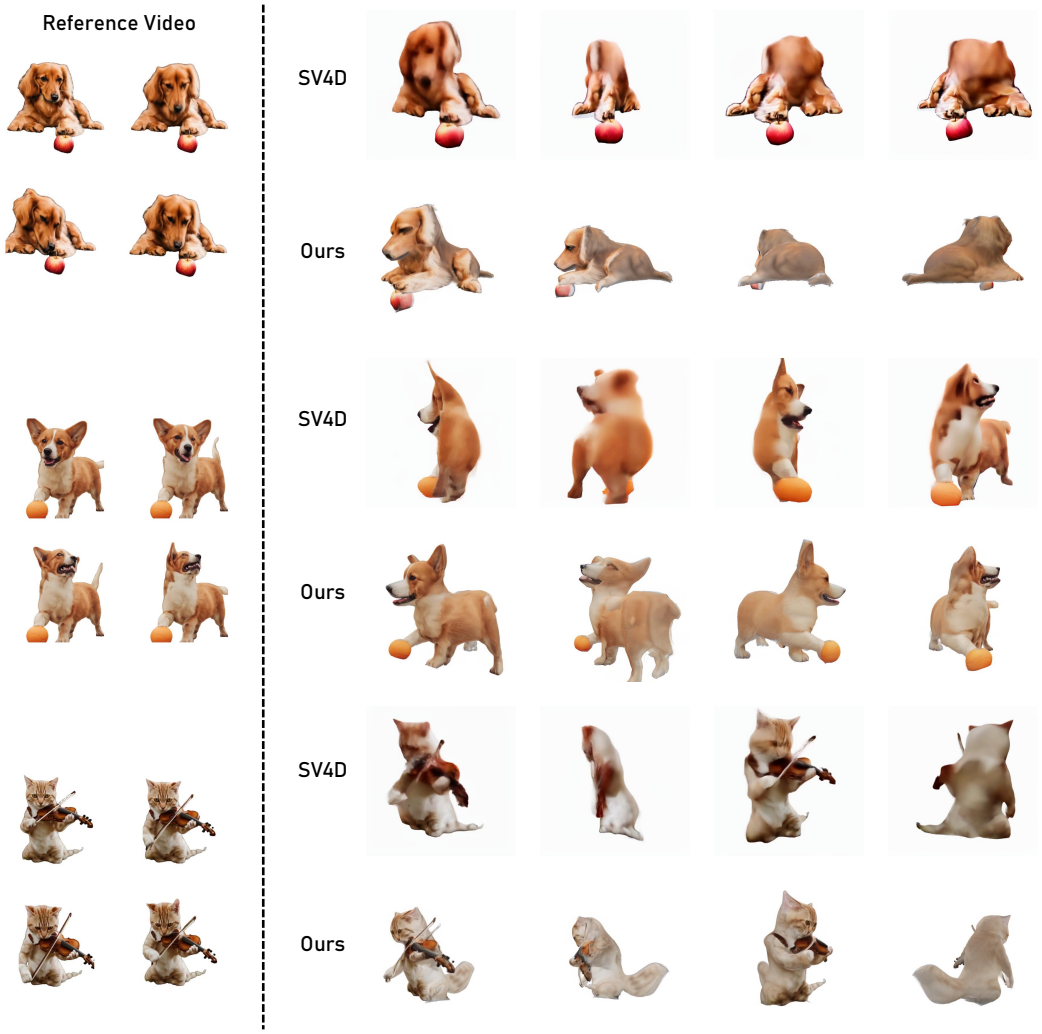

Figure M: Qualitative comparison with SV4D [108]. Since only a certain number of selected views can be rendered with the official code of SV4D, we choose the close viewpoints but not the identical ones for comparison for the sake of simplicity. Nevertheless, it is still obvious that SV4D has inferior performance compared with our method in the perspectives of both appearance and geometry.