DreamControl: Control-Based Text-to-3D Generation with 3D Self-Prior

Tianyu Huang[1,3] Yihan Zeng[2] Zhilu Zhang[1] Wan Xu[1] Hang Xu[2]
Songcen Xu[2] Rynson W. H. Lau[3] Wangmeng Zuo[1†]
[1]Harbin Institute of Technology [2]Huawei Noah's Ark Lab [3]City University of Hong Kong

Abstract

3D generation has raised great attention in recent years. With the success of text-to-image diffusion models, the 2D-lifting technique becomes a promising route to controllable 3D generation. However, these methods tend to present inconsistent geometry, which is also known as the Janus problem. We observe that the problem is caused mainly by two aspects, i.e., viewpoint bias in 2D diffusion models and overfitting of the optimization objective. To address it, we propose a two-stage 2D-lifting framework, namely DreamControl, which optimizes coarse NeRF scenes as 3D self-prior and then generates fine-grained objects with control-based score distillation. Specifically, adaptive viewpoint sampling and boundary integrity metric are proposed to ensure the consistency of generated priors. The priors are then regarded as input conditions to maintain reasonable geometries, in which conditional LoRA and weighted score are further proposed to optimize detailed textures. DreamControl can generate high-quality 3D content in terms of both geometry consistency and texture fidelity. Moreover, our control-based optimization guidance is applicable to more downstream tasks, including user-guided generation and 3D animation. The project page is available at https://github.com/tyhuang0428/DreamControl.

1. Introduction

Digital 3D content plays a crucial role in various fields, including medicine, education, entertainment, *etc*. Generating 3D content in an automatic system is thus raising more and more attention. Recently, with the success of score distillation sampling [26] (SDS), 2D-lifting technique [2, 16, 20, 26, 33] becomes a promising route to 3D generation, where the pre-trained 2D diffusion models [24, 27, 29] are utilized to optimize 3D representations. Compared with methods [5, 10, 12, 23, 34] that are supervised with 3D assets, 2D-lifting technique is capable of generating high-fidelity 3D textures in open-world scenarios.

Nonetheless, the results generated by 2D-lifting methods tend to present inconsistent 3D geometry, *e.g.*, multi-face, also known as the Janus problem. Some recent works [14, 17–19, 31, 41] attribute this problem primarily to the lack of view awareness in 2D diffusion models. They propose to incorporate 3D prior knowledge into 2D diffusion models, thus learning to perceive view-dependent conditions. Albeit the improvement of geometry consistency, it compromises texture fidelity and fine-grained details, as the used 3D contents are mostly created manually in a cartoonish style. Moreover, the limited scale of available 3D assets affects the generalizability of these methods, making it challenging to acquire a comprehensive 3D prior.

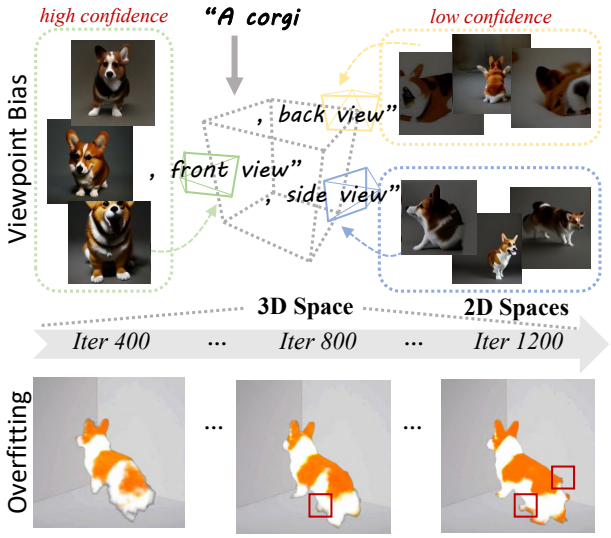


Figure 1. Main causes of inconsistent 3D generation. Images sampled by 2D diffusion models are biased in viewpoint distribution. The generation confidence decreases as the viewpoint turns from front to back. 3D representations are thus gradually overfitted to the highest probability image during the optimization, generating artifacts as shown in red b-boxes.

Rethinking the optimization target of SDS, *i.e.*, *creating 3D models that look like good images when rendered from random angles*, we observe that the causes of 3D inconsistency can be further divided into two terms: (1) viewpoint

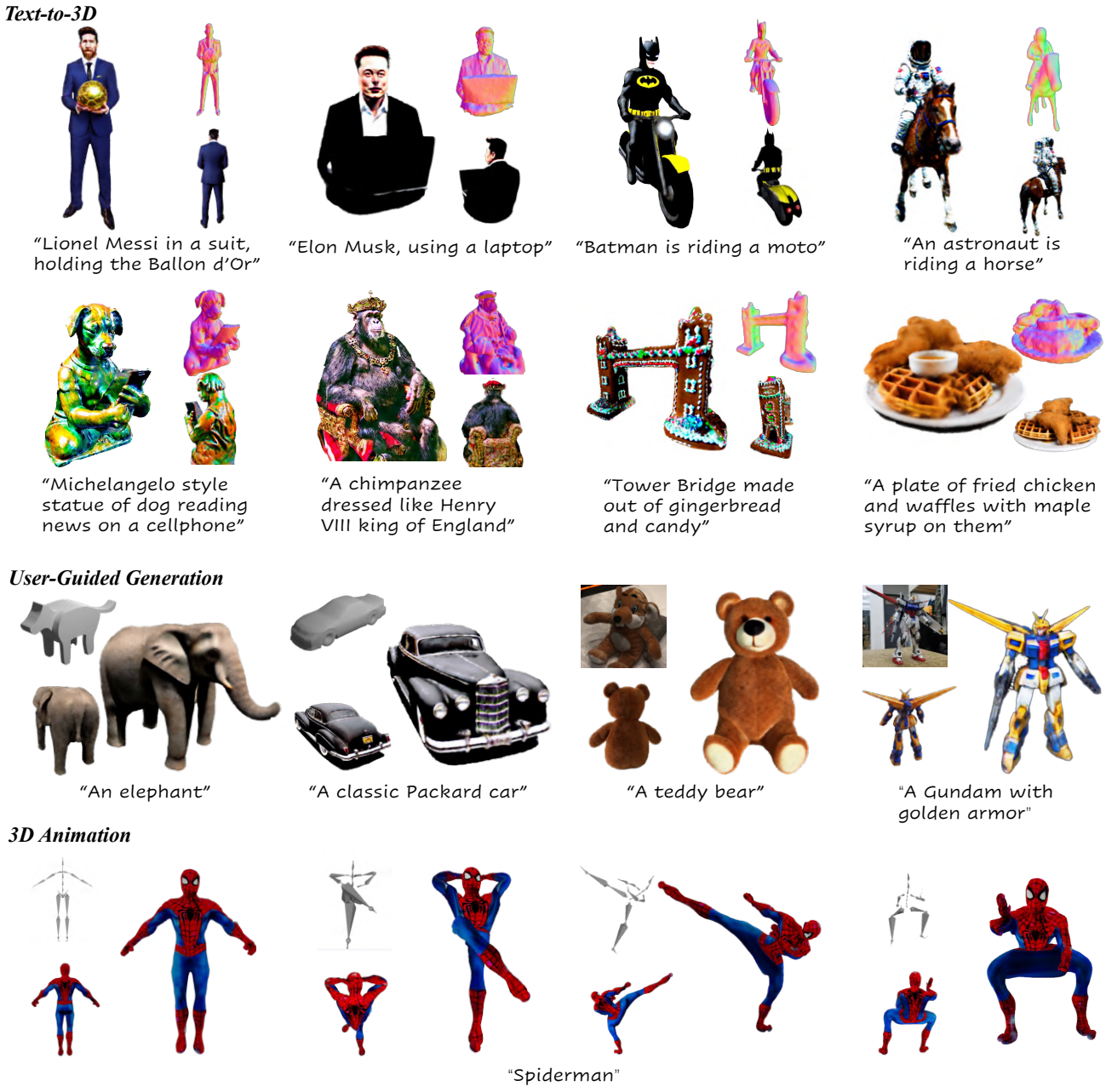


Figure 2. DreamControl can generate diverse 3D content with high-consistency geometries and high-fidelity textures. Beyond text-to-3D generation, our control-based guidance is applicable to controllable generation tasks, including user-guided generation and 3D animation.

bias in 2D diffusion models; (2) overfitting of the optimization objective. Take neural radiance fields [21] (NeRF) in SDS as an example: Points in NeRF scenes are supervised by the casting rays from uniformly sampled viewpoints, while the viewpoint distribution of 2D diffusion models is biased, as shown in Figure 1. Under the optimization of SDS, all the rendered images may overfit to the highest-probability image generated by the diffusion model. In other words, all the views of the 3D model look similar to one specific image, giving rise to the Janus problem.

Based on this observation, we aim to leverage the 3D representation before overfitting as a self-generated 3D prior, namely 3D self-prior. Accordingly, we propose DreamControl, a two-stage 2D-lifting framework that maintains self-priors by control-based distillation. Specifically, we optimize coarse NeRF scenes as 3D self-prior and then

generate fine-grained objects with prior-based control. In the first stage, we adopt SDS to construct a coarse shape that keeps good geometry consistency. For alleviating possible 3D artifacts, an adaptive viewpoint sampling is proposed to adjust the viewpoint distribution of diffusion models, and a boundary integrity metric is proposed to avoid the overfitting of optimization. In the second stage, we regard the 3D prior as a conditional input and deploy ControlNet [39] to supervise the generation, thus obtaining a detailed texture while maintaining the geometry of the prior. Considering the diversity of ControlNet can be easily constrained by fixed conditions, we propose control-based score distillation, in which a conditional LoRA and a weighted score are presented to stabilize the optimization process.

Extensive experiments on text-to-3D generation demonstrate that DreamControl can obtain high-quality 3D content regarding geometry consistency and texture fidelity. Benefiting from the control-based guidance, our framework can be further applied to more downstream tasks, including user-guided generation and 3D animation.

Our contributions can be summarized as:

- We propose to optimize NeRF as 3D self-prior, where adaptive viewpoint sampling and boundary integrity metric are suggested to alleviate inconsistent generation.
- We propose a control-based score distillation to maintain geometries in self-prior, where conditional LoRA and weighted score are presented to stabilize the optimization.
- Our two-stage framework, namely DreamControl, can generate high-quality 3D content in text-to-3D generation, and the control-based guidance can be further applicable to more downstream tasks.

2. Related Work

Text-to-3D Generation. Text-to-3D generation has witnessed rapid progress in recent years, in which methods can generally be split into two categories, *i.e.*, 3D supervised and 2D lifting. 3D supervised methods [5, 10, 12, 23, 34, 37] train generators with text-3D data. Albeit the efficiency to generate solid 3D content, these methods lack generalizability due to the limited scale of available 3D data. In contrast, 2D lifting methods [2, 16, 20, 26, 33] take advantage of 2D diffusion models, distilling 3D representations to 2D priors. Although presenting photorealistic generation, these methods can easily fall into 3D inconsistency issues, also known as the famous Janus problem. To address the problem, recent works [14, 17–19, 31, 41] attempt to incorporate 3D prior into 2D diffusion models. Since the prior is trained with limited 3D data, these methods still suffer from the lack of generalizability. Moreover, their generation may lose high-fidelity texture due to the cartoonish style of 3D data. In this work, we propose to optimize a coarse NeRF representation as a training-free 3D prior, enhancing generation consistency while keeping texture fidelity.

Controllable Generation. Text input is a flexible control in 3D tasks [9], while other conditions like image [28, 35], video [3, 40], and 3D sketch [20] are also available for guiding generation. ControlNet [39] supports text-to-image synthesis with additional conditions, *e.g.*, edge, normal, *etc*. It allows 3D generators to create 3D content with the guidance of 2D sketch [4], depth [38], and even video [30]. In this work, we use 2D conditions to maintain geometry consistency, which are rendered from our 3D prior.

3. Preliminaries

NeRF [21] (Neural Radiance Fields) is a widely-used 3D representation, which combines neural networks with graphical principles. A multilayer perceptron (MLP) θ is trained to predict the color *RGB* and density $\boldsymbol{\sigma}$ of sampling points in the 3D space, supervised by the total squared error between the rendered and ground-truth pixel colors. Given a camera pose c, images x are rendered by the density summation of colored points. We represent the rendering process as $x = \mathbf{g}(\theta, c)$, in which $\mathbf{g}$ denotes the renderer.

SDS [26] (Score Distillation Sampling) distills the parameters of 3D representation NeRF (θ) to a pre-trained 2D diffusion models (ϕ). Given a text prompt y, it optimizes the rendered image x_t with the predicted noise in the timestep t. The gradient can be formulated as,

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\omega(t) \left(\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon} \right) \frac{\partial \boldsymbol{x}}{\partial \theta} \right], \quad (1)$$

where $\hat{\boldsymbol{\epsilon}}_\phi$ is the noise predicted by ϕ.

VSD [33] (Variational Score Distillation) supposes the corresponding 3D scene given a textual prompt as a distribution range, rather than a single point as in SDS, significantly improving quality and diversity of 3D generation. It proposes a particle-based update strategy via the Wasserstein gradient flow to optimize a 3D distribution,

$$\mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\omega(t) \left(\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t, t, y) - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, c, y) \right) \frac{\partial \boldsymbol{x}}{\partial \theta} \right], \quad (2)$$

where $\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, c, y)$ is the noise predicted by rendered images. In practice, it can be regarded as LoRA [8] (Low-Rank Adaption) conditioned with camera poses c, which is supervised by a standard diffusion loss as,

$$\mathbb{E}_{t\sim\mathcal{U}(0,1),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)} \left[\| \hat{\boldsymbol{\epsilon}}_\theta(\alpha_t \boldsymbol{x}_t + \sigma_t \boldsymbol{\epsilon}, t, c, y) - \boldsymbol{\epsilon} \|_2^2 \right]. \quad (3)$$

Nonetheless, neither SDS nor VSD considers 3D consistency issues. Their optimization objectives may force NeRF to present a "most-likely" diffusion-generated image in any camera view, resulting in the Janus problem.

4. DreamControl

In this section, we introduce a two-stage 2D-lifting framework, DreamControl. As shown in Figure 3, we first gener-

ate a coarse NeRF with SDS, which is regarded as a 3D self-prior (Section 4.1). Then, we propose a control-based score distillation, generating high-quality texture while maintaining the geometry from previous priors (Section 4.2).

4.1. 3D Self-Prior Generation

In SDS optimization, the gradient direction is towards minimizing the distribution gap between rendered images in NeRF and generated images by 2D diffusion models,

$$\min_{\theta\in\Theta}\mathcal{L}_{\text{SDS}}(\theta) \iff \min_{\theta\in\Theta} D_{\text{KL}}(q_t^\theta(\boldsymbol{x}_t|y,c)||p_t(\boldsymbol{x}_t|y,c)), \tag{4}$$

where q_t^θ and p_t the distribution probability of the NeRF scene θ and a pre-trained 2D diffusion model, respectively. However, there exists a viewpoint bias in the 2D model, which means it may not be able to acquire an accurate distribution $p_t(\boldsymbol{x}_t|y,c)$ for a given camera pose c. When overfitting to the biased distribution p_t, optimization target in Eq. (4) can degenerate to $D_{\text{KL}}(q_t^\theta(\boldsymbol{x}_t|y)||p_t(\boldsymbol{x}_t|y,\widetilde{c}))$, where $\widetilde{c}$ is the most frequent viewpoint. As a result, generated results may look similar in any camera view, giving rise to the Janus problem.

In this optimization process, two crucial aspects are the main cause of the problem, *i.e.*, viewpoint bias and overfitted distribution. Accordingly, an adaptive viewpoint sampling and a boundary integrity metric are proposed in the following to alleviate potential inconsistent generation.

Adaptive Viewpoint Sampling. 2D diffusion models inherit viewpoint bias from the Internet training data, *e.g.*, they tend to generate a person's front face rather than its back. To obtain a satisfied image, previous works [1, 26] attach view-dependent information to text prompts like "*front view*", but 2D networks cannot explicitly encode these view prompts. Some recent works [14, 17–19, 31, 41] attempt to train a view-aware generative model with 3D data. However, texture fidelity and content generalizability are limited by available 3D training data.

Considering incorporating 2D diffusion models with 3D information is challenging, we present a solution with a new perspective, *i.e.*, aligning 3D camera sampling with the viewpoint distribution of 2D diffusion. Specifically, we modify the camera pose sampling $p(c)$ in SDS to fit the distribution in 2D. Given a text prompt "$*$", we make three view-dependent prompts, *i.e.*, "$*$, *front view*" (y_1), "$*$, *side view*" (y_2), and "$*$, *back view*" (y_3), and then take them to generate corresponding 2D images $\phi(y_i,t)$ with the diffusion model ϕ in time step t. Thus, the expected probability distribution of each view range p^* can be calculated as,

$$p^* = \text{softmax}([s_1,s_2,s_3]), s_i = \frac{1}{|T|}\sum_{t\in T} s_{\text{CLIP}}(y_i,\phi(y_i,t)), \tag{5}$$

where T is a set of timesteps. s_{CLIP} denotes the CLIP similarity between text and image. In each view range, the viewpoint probability $p^*(c)$ follows a uniform distribution. In this way, we have a NeRF representation $q_t^\theta(\boldsymbol{x}_t|y) = \int q_t^\theta(\boldsymbol{x}_t|y,c)p^*(c)dc$, which can get closer to $p_t(\boldsymbol{x}_t|y)$ in terms of camera pose c.

Boundary Integrity Metric. Albeit NeRF is modeled based on graphical principles, its reconstruction result highly depends on the quality of training data. Each ground-truth image and its own camera pose are both required, so that the prediction of color and density can be supervised by casting rays $\mathbf{r}$. However, it is difficult to instruct a 2D generative model to generate images that accurately match viewpoints. Even though our sampling strategy can alleviate the effects of viewpoint bias, 3D artifacts are still unavoidable in the overfitting circumstance. Previous works [2, 16, 20, 26, 33] hardly considered to check the occurrence of overfit automatically. In practice, they usually optimize 3D representations for a fixed number of iterations or manually terminate the optimization.

In this work, we propose a geometric terminated metric to avoid the possible overfit. Specifically, we observe that NeRF generally can form a solid object without overfitting when the density between foreground and background starts to show a clear boundary. Thus, we can detect the situation by calculating the difference between the density $\boldsymbol{\sigma}$ of all valid pixels and boundary pixels, *i.e.*,

$$\Delta_{\mathbf{r}} = \frac{1}{|\mathcal{R}_v|}\sum_{\mathbf{r}\in\mathcal{R}_v}\boldsymbol{\sigma}(\mathbf{r}) - \frac{1}{|\mathcal{R}_b|}\sum_{\mathbf{r}\in\mathcal{R}_b}\boldsymbol{\sigma}(\mathbf{r}), \tag{6}$$

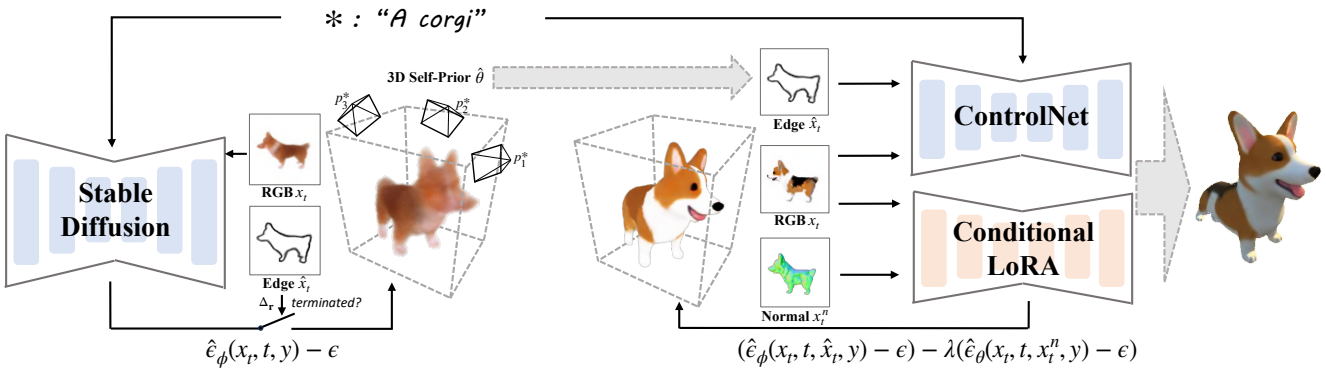
where $\mathcal{R}_v$ and $\mathcal{R}_b$ are the set of valid rays and boundary rays. And we terminate the optimization process when $\Delta_{\mathbf{r}}$ falls below our threshold $\delta_{\mathbf{r}}$.

With the adaptive viewpoint sampling and boundary integrity metric, a coarse shape $\hat{\theta}$ that keeps a reasonable geometry can be generated based on NeRF. We regard it as a 3D self-prior for the following control-based generation.

4.2. Control-Based Score Distillation

The 3D prior $\hat{\theta}$ cannot capture fine-grained texture in a short optimization period. Previous multi-stage methods [2, 16, 33] continually optimize the generation based on coarse shape, which contradicts our intention of early termination in Section 4.1 and may still lead to overfitting. To generate high-quality texture while maintaining a reasonable geometry, we propose to treat $\hat{\theta}$ as a conditional input and adopt ControlNet [39] as the optimization guidance.

Specifically, for one optimization step, we render an RGB image $\boldsymbol{x}$ from θ and a conditional image $\hat{\boldsymbol{x}}$ from $\hat{\theta}$ in the same camera pose. Then, ControlNet can supervise the generation by the predicted noise $\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t,t,\hat{\boldsymbol{x}}_t,y)$ according to Eq. (1), which is demonstrated effective in previous works [30, 38]. However, these methods generally require high-quality conditions for the generation in ControlNet, *e.g.*, depth maps or DSLR photos. Strong condi-



Stage 1: **3D Self-Prior Generation** *Stage 2:* **Control-Based Score Distillation**

Figure 3. Overview of DreamControl. In the first stage, a coarse NeRF is optimized as a 3D self-prior $\hat{\theta}$, in which an adaptive viewpoint sampling p^* and a boundary integrity metric $\Delta_\mathbf{r}$ are proposed to alleviate inconsistent generation. The prior $\hat{\theta}$ is then sent to the second stage as an input edge condition $\hat{x}_t$, in which a control-based score distillation can generate fine-grained textures and maintain geometries in the prior. A Conditional LoRA and a weighted score are further proposed to stabilize the optimization process.

tions such as depth may restrict the diversity of generated content. Meanwhile, our 3D prior $\hat{\theta}$ can only exhibit poor texture information, making it hard to extract photo-realistic images for generative control.

As for the issue of diversity, we take VSD [33] into consideration, which expands the generation range of a given prompt. Since photo-realistic images are not available from our 3D priors, we use the boundary mask from $\hat{\theta}$ as the condition $\hat{x}$, further weakening the control restriction. However, simply replacing the pre-trained diffusion model with ControlNet doesn't work well, as the pre-trained term $\hat{\boldsymbol{\epsilon}}_\phi$ and the LoRA term $\hat{\boldsymbol{\epsilon}}_\theta$ vary a lot, making it hard for the convergence of Eq. (3). To stabilize the optimization, we propose a conditional LoRA and a weighted score.

Conditional LoRA. In VSD, a LoRA is adopted to predict the noise of current NeRF scene $\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, c, y)$, in which a camera pose c is additionally embedded. We are concerned that camera pose is a high-level semantic concept, which is hard to tokenize and imbibe through LoRA. The LoRA term $\hat{\boldsymbol{\epsilon}}_\theta$ may predict a noise unrelated to c, enlarging the gap with the pre-trained term $\hat{\boldsymbol{\epsilon}}_\phi$. Instead, we replace camera pose c with the normal map $\boldsymbol{x}_t^n$ rendered by itself. Accordingly, a lightweight control LoRA is adopted to predict noise conditioned by $\boldsymbol{x}_t^n$. The training objective is similar to Eq. (3).

Weighted Score. The prediction of the LoRA term $\hat{\boldsymbol{\epsilon}}_\theta$ in the early stage lacks practical significance, as NeRF has not yet generated meaningful objects. To mitigate the possible disruption caused by LoRA at those early steps, we propose to incorporate a coefficient λ into its loss term, which is changed along with training steps. Specifically, we rewrite $\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t, t, y) - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, c, y)$ in Eq. (2) as two terms,

$$(\hat{\boldsymbol{\epsilon}}_\phi(\boldsymbol{x}_t, t, \hat{\boldsymbol{x}}_t, y) - \boldsymbol{\epsilon}) - \lambda(\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, \boldsymbol{x}_t^n, y) - \boldsymbol{\epsilon}). \quad (7)$$

When $\lambda = 0$, the loss function degenerates to SDS format, which can be regarded as a special case of VSD. We gradually increase λ as the training step increases.

Following the training strategy of ProlificDreamer [33], we alternately update the gradient of Eq. (2) and Eq. (3).

4.3. Implementation Details

We implement DreamControl based on threestudio [6]. In the first stage, we adopt DeepFloyd [32] guidance in the SDS optimization. In the second stage, we use the scribble version of ControlNet [39] v1.1 as the pre-trained term and the stable diffusion [27] v1.5 injected with Contraol-LoRA [7] as the LoRA term. 3D prior is the rendered density mask from the NeRF result in the first stage, further processed by HEDdetector [36] implemented in ControlNet. The LoRA condition is the rendered normal map from the current NeRF. The classifier-free guidance scale (CFG) is set as 7.5 and 1.0 in pre-trained and LoRA terms, respectively. We increase the loss coefficient λ from 0.5 to 0.75 linearly in the first 5,000 iterations. Please refer to the Suppl. for more details.

5. Experiments

In this section, extensive experiments are conducted to evaluate the generation quality of our proposed method DreamControl. We first show our text-to-3D generation results in Section 5.1, in which several state-of-the-art methods are compared in terms of geometry consistency and texture fidelity. In Section 5.2, we present controllable 3D generation tasks with our control-based optimization guidance, including user-guided generation and 3D animation. Finally, we conduct ablation studies in Section 5.3, demonstrating the effectiveness of our newly proposed designs.

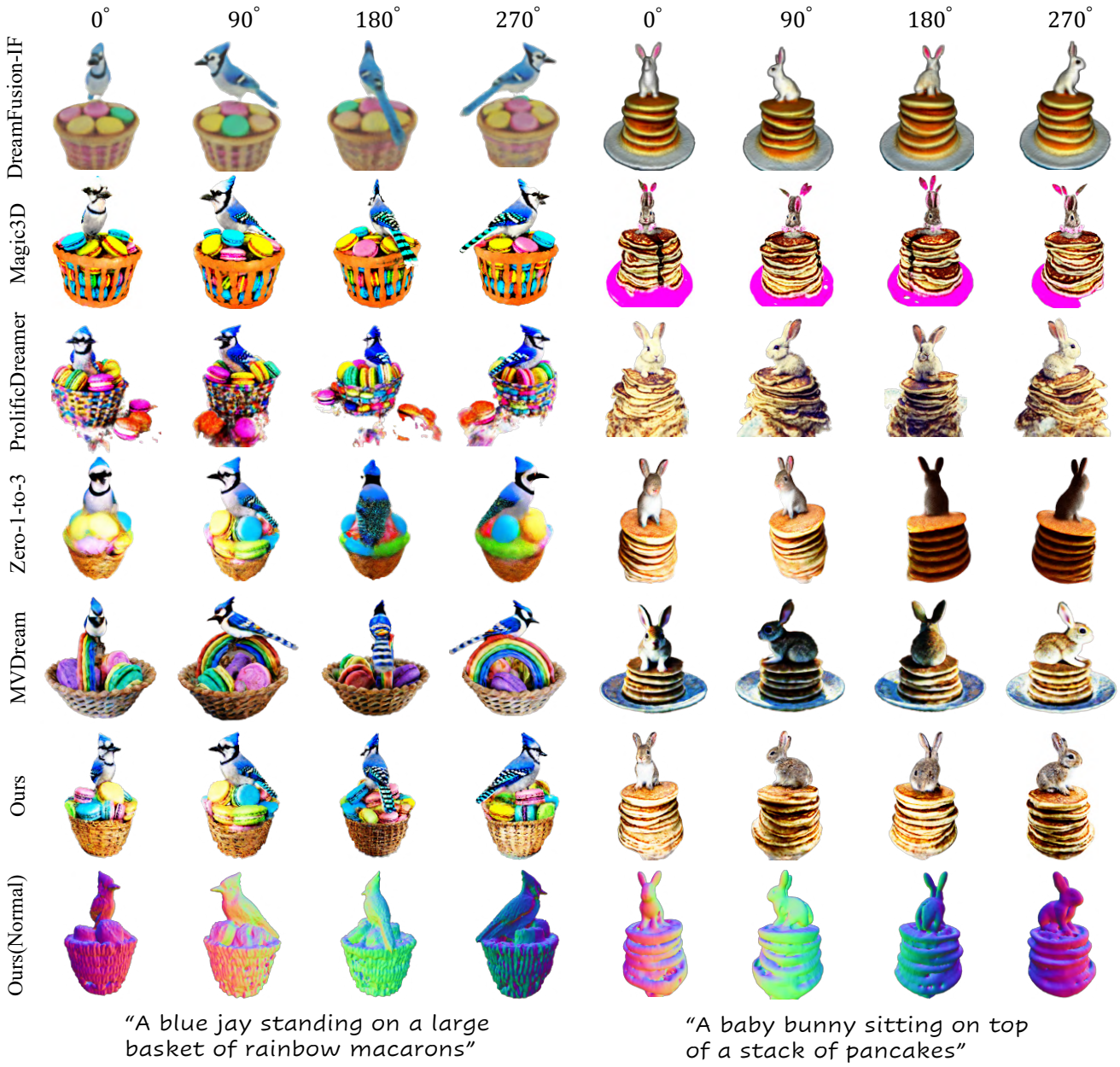


Figure 4. Qualitative results. Compared with other methods, DreamControl enjoys high-consistency geometry and high-fidelity texture.

5.1. Text-to-3D Generation

We compare DreamControl with five 3D generation methods: (1) DreamFusion [26], the early work in 2D-lifting methods, (2) Magic3D [16], the first two-stage optimization method, (3) ProlificDreamer [33], a high-fidelity optimization method, (4) Zero-1-to-3 [17], a view-conditional diffusion model, and (5) MVDream [31], a multi-view diffusion model. Since most of these methods haven't released official implementation, we use the reproduction in threestudio [6]. Note that the reproduction can be different from the original implementation. For example, DreamFusion adopts an unreleased diffusion model Imagen [29], which is replaced with DeepFloyd [32] in our comparison. We provide quantitative and qualitative results in the following.

Quantitative Results. We select 30 text prompts from the galleries of DreamFusion, Magic3D, and ProlificDreamer for quantitative experiments. To evaluate the consistency of 3D geometries, we count the number of inconsistent 3D content generated in each method, regarded as the occurrence rate of the Janus problem (JR). To evaluate the fidelity

Table 1. Quantitative results. DreamControl surpasses the competing methods in all the evaluation metrics.

Method	JR(%)↓	PS(%)↑	CS(%)↑
DreamFusion-IF [26]	36.67	10.01	26.36
Magic3D [16]	53.33	16.13	26.59
ProlificDreamer [33]	56.67	20.81	26.69
Zero-1-to-3 [17]	16.67	7.89	21.25
MVDream [31]	**10.00**	17.70	26.17
DreamControl (ours)	**10.00**	**27.46**	**28.14**

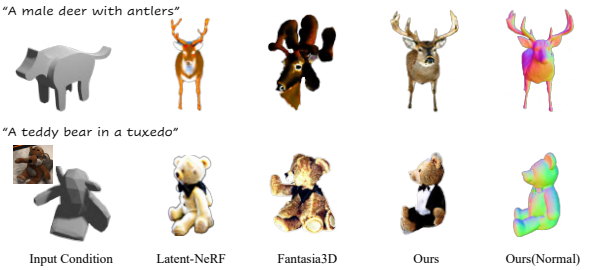


Figure 5. User-guided generation. DreamControl is flexible to loose input conditions, generating fine-grained content with a 3D sketch or even a coarse layout.

Figure 6. 3D Animation. DreamControl can generate 3D content conditioned by a template skeleton, which is naturally bound with the skeleton after generation.

of textures, we render multi-view images and introduce PickScore (PS) [13] for comparison, which is an aesthetic metric used to measure the quality of 2D content. Moreover, we adopt CLIP-Score (CS), verifying the text consistency of generation. See the Suppl. for the details of evaluation metrics. As shown in Table 1, our DreamControl outperforms other methods in all the metrics. The first three methods face a serious Janus problem. JR grows as PS increases, which indicates that the overfitting of optimization could exacerbate 3D inconsistency. Zero-1-to-3 and MVDream incorporate 3D prior knowledge into 2D diffusion, capable of generating high-consistent geometries. However, they suffer from low-quality texture and text-irrelevant generation, according to PickScore and CLIP-Score. We have analyzed that 3D training data tend to exhibit a cartoonish style and lack generalizability due to the limited scale. The results further demonstrate this point. In comparison, DreamControl adopts a coarse NeRF as a 3D self-prior, enjoying high-quality 3D generation in both geometry and texture, as well as the consistency of text input.

Qualitative Results. To compare the generation results qualitatively, we select two classic text prompts from quantitative experiments and show the corresponding multi-view rendered images in Figure 4. Textures in DreamFusion-IF are over-smoothed. The rabbit's ears are still red in the back view, implying the risk of the multi-face issue. Yet DreamFusion basically generates reasonable geometries, demonstrating our assumption that NeRF can be used as a self-generated 3D prior. The two-stage method Magic3D presents more detailed textures but 3D inconsistent problems emerge, *e.g.*, two beaks in the generated jay and three ears in the generated rabbit. ProlificDreamer can generate high-fidelity textures, while the Janus problem still exists. In particular, the other face is generated in the back view of the rabbit. Zero-1-to-3 and MVDream create highly consistent geometries. However, the generated textures are quite rough, *e.g.*, the macarons and pancakes. Moreover, MVDream generates an unrelated rainbow in the first prompt, which means 3D prior is not general enough to handle detailed text descriptions. In contrast, our DreamControl can generate 3D content with high-consistency geometries and high-fidelity textures. The visualization is consistent with quantitative results.

5.2. Controllable 3D Generation

Thanks to our control-based optimization guidance, DreamControl can also be applied to controllable 3D generation tasks. We present user-guided generation in Figure 5 and 3D animation in Figure 6.

User-Guided Generation. User-guided generation is to create 3D content based on a text prompt, as well as a spatial sketch. Compared with previous works Latent-NeRF [20] and Fantasia3D [2], our method is flexible to looser input conditions. In the first row of Figure 5, we use a simple animal shape as a template, successfully generating a male deer with antlers. Latent-NeRF adopts a sketch loss to minimize the geometric distance between generation and the input condition. Compared with it, our result is much more detailed in textures. Fantasia3D initializes the 3D representation with input conditions and then gradually refines an expected geometry. However, it fails to attach antlers to the animal body, generating a weird deer head instead. In the second row, we present the generation task based on a harder condition, *i.e.*, multi-view images. We use a classic multi-view image set in DTU MVS dataset [11] and reconstruct a coarse layout with [22]. Conditioned on that coarse shape, our method can generate a teddy bear in a fine-grained tuxedo, while the other two create teddy bears with multiple faces and legs. The results exhibit the effectiveness of our guidance on flexible conditions.

3D Animation. To animate a 3D object, the common practice is to bind the object with a template skeleton, also known as rigging, which is a time-consuming work that de-

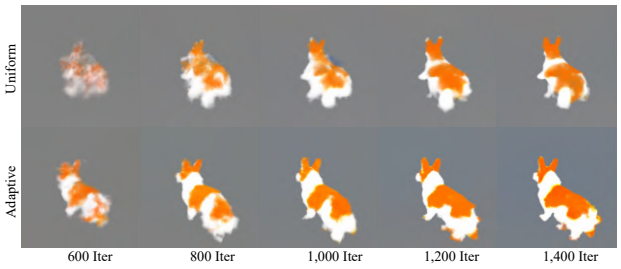


Figure 7. Ablation study on viewpoint sampling and termination metric. Our optimization can avoid the generation of extra legs.

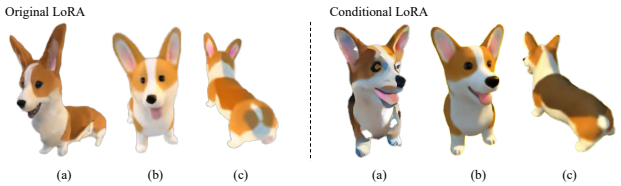


Figure 8. Ablation study on conditional LoRA. (a) is the LoRA sampling result in the front view. (b) and (c) are 3D generation results in the front view and the back view. Our conditional LoRA can precisely present the current representation from a correct viewpoint, improving the generation quality.

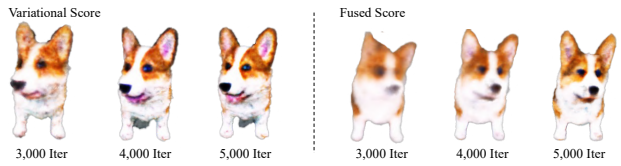


Figure 9. Ablation study on fused score distillation. Compared with VSD loss, our loss stabilizes the generation in the early stage.

mands related technical expertise. As a result, 3D animation is an expensive production currently. Instead, our method can generate 3D objects conditioned by a given skeleton, which can be directly bound with the skeleton after generation. Different from previous works TADA [15] that is based on a human-template SMPL-X [25], DreamControl can adapt to all kinds of templates like animal, machine, *etc*. As shown in Figure 6, we can easily animate the generated Woody and horse.

5.3. Ablation Studies

To further verify the effectiveness of our components, we take the generation of a corgi as an example, comparing results with or without our designs in Figures 7, 8, and 9.

Viewpoint Sampling and Optimization Termination. We compare our adaptive viewpoint sampling with uniform sampling in Figure 7. In the uniform sampling, the corgi's extra legs grow up evenly with normal legs, making it hard to find an appropriate timestep for terminating the optimization. Differently, the extra legs do not appear until 1,000 iterations with our sampling strategy. However, the multi-leg issue still exists, indicating that termination is necessary.

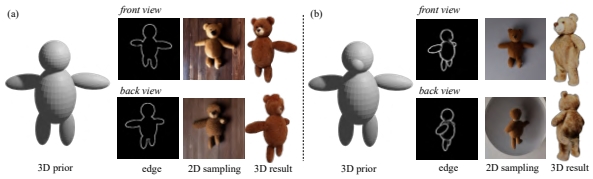


Figure 10. Failure case. (a) When edge conditions from different views are similar, ControlNet may fail to provide correct guidance. (b) A possible solution is to add more details in one side.

Conditional LoRA. We compare our conditional LoRA with ProlificDreamer's original LoRA in Figure 8. (a) is the sampling result of LoRA in the front view. With a camera pose c, the original LoRA can hardly generate 2D content in an accurate viewpoint. In contrast, our conditional LoRA successfully learns the current scene with a normal map $\mathbf{n}$. Due to the estimation bias of the original LoRA, the generated corgi is incompatible with ours. Especially in the back view, its ears are colored pink, and its body and bottom are disproportionately scaled.

Fused Score. We compare our fused score with the variational score of VSD in Figure 9. VSD optimization formulates a clear outline of a corgi after 3,000 iterations, which seems a stronger supervision than ours. However, the corgi contains too much noise, especially near the generation boundary. By restricting LoRA in the early stage, our optimization is much more stable than VSD.

6. Conclusion

In this work, we present a two-stage framework DreamControl to improve the geometry consistency of 3D generation. By optimizing coarse NeRF, our method is able to obtain 3D self-prior for the following generation. A control-based score distillation is further proposed, generating fine-grained texture while maintaining the prior geometry. DreamControl can create high-quality 3D content with high-consistency geometries and high-fidelity textures. The control-based guidance can also be applied to controllable tasks including user-guided generation and 3D animation.

Limitation. Although DreamControl maintains the geometry consistency with prior-conditional guidance, it may fail in some extreme cases where priors look similar in different views. For example, the left prior in Figure 10 looks exactly the same in opposite views, and ControlNet thus provides incorrect guidance in back views, leading to multi-face generation. Fortunately, this problem is avoidable by enlarging the view differences, *e.g.*, adding more details to the front view of 3D priors like the right prior in Figure 10.

Acknowledgements

This work was supported by National Key RD Program of China under Grant No. 2021ZD0112100, and the National Natural Science Foundation of China (NSFC) under Grant No. U19A2073.

References

[1] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 4

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 1, 3, 4, 7

[3] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[4] Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1148–1156, 2023. 3

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 1, 3

[6] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 5, 6

[7] Wu Hecong. ControlLoRA: A Lightweight Neural Network To Control Stable Diffusion Spatial Information, 2023. 5

[8] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 3

[9] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22157–22167, 2023. 3

[10] Tianyu Huang, Yihan Zeng, Bowen Dong, Hang Xu, Songcen Xu, Rynson WH Lau, and Wangmeng Zuo. Textfield3d: Towards enhancing open-vocabulary 3d generation with noisy text fields. *arXiv preprint arXiv:2309.17175*, 2023. 1, 3

[11] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 7

[12] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 3

[13] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 7

[14] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 1, 3, 4

[15] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 8

[16] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1, 3, 4, 6, 7

[17] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 3, 4, 6, 7

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

[19] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 1, 3, 4

[20] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 1, 3, 4, 7

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[22] Tom Monnier, Jake Austin, Angjoo Kanazawa, Alexei A Efros, and Mathieu Aubry. Differentiable blocks world: Qualitative 3d decomposition by rendering primitives. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 7

[23] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 1, 3

[24] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and

Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 8

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3, 4, 6, 7

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 5

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 6

[30] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082*, 2023. 3, 4

[31] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 3, 4, 6, 7

[32] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd. https://huggingface.co/DeepFloyd, 2023. 5, 6

[33] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 1, 3, 4, 5, 6, 7

[34] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16815, 2023. 1, 3

[35] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3

[36] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *Proceedings of IEEE International Conference on Computer Vision*, 2015. 5

[37] Cuican Yu, Guansong Lu, Yihan Zeng, Jian Sun, Xiaodan Liang, Huibin Li, Zongben Xu, Songcen Xu, Wei Zhang, and Hang Xu. Towards high-fidelity text-guided 3d face generation and manipulation using only images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15326–15337, 2023. 3

[38] Wangbo Yu, Li Yuan, Yan-Pei Cao, Xiangjun Gao, Xiaoyu Li, Long Quan, Ying Shan, and Yonghong Tian. Hifi-123: Towards high-fidelity one image to 3d content generation. *arXiv preprint arXiv:2310.06744*, 2023. 3, 4

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4, 5

[40] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[41] Minda Zhao, Chaoyi Zhao, Xinyue Liang, Lincheng Li, Zeng Zhao, Zhipeng Hu, Changjie Fan, and Xin Yu. Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior. *arXiv preprint arXiv:2308.13223*, 2023. 1, 3, 4