ControlDreamer: Blending Geometry and Style in Text-to-3D

Yeongtak Oh[*1]
dualism9306@snu.ac.kr
Jooyoung Choi[*1]
jy_choi@snu.ac.kr
Yongsung Kim[2]
libary753@snu.ac.kr
Minjun Park[2]
minjunpark@snu.ac.kr
Chaehun Shin[1]
chaehuny@snu.ac.kr
Sungroh Yoon[†1,2]
sryoon@snu.ac.kr

[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, South Korea

Abstract

Recent advancements in text-to-3D generation have significantly contributed to the automation and democratization of 3D content creation. Building upon these developments, we aim to address the limitations of current methods in blending geometries and styles in text-to-3D generation. We introduce multi-view ControlNet, a novel depth-aware multi-view diffusion model trained on generated datasets from a carefully curated text corpus. Our multi-view ControlNet is then integrated into our two-stage pipeline, ControlDreamer, enabling text-guided generation of stylized 3D models. Additionally, we present a comprehensive benchmark for 3D style editing, encompassing a broad range of subjects, including objects, animals, and characters, to further facilitate research on diverse 3D generation. Our comparative analysis reveals that this new pipeline outperforms existing text-to-3D methods as evidenced by human evaluations and CLIP score metrics. Project page: https://controldreamer.github.io

1 Introduction

3D content creation has recently been gathering attention, particularly in the realms of virtual reality and game development. This evolving landscape has been profoundly impacted by the recent advancement of 2D lifting methods. Dreamfusion's [29] introduction of score distillation sampling (SDS) has revolutionized and democratized the field, facilitating the generation of 3D models by leveraging large-scale text-to-image diffusion models [1, 33, 35]. This breakthrough has enabled more intuitive and imaginative 3D content creation from

[*] These authors contributed equally to this work
[†] Corresponding author

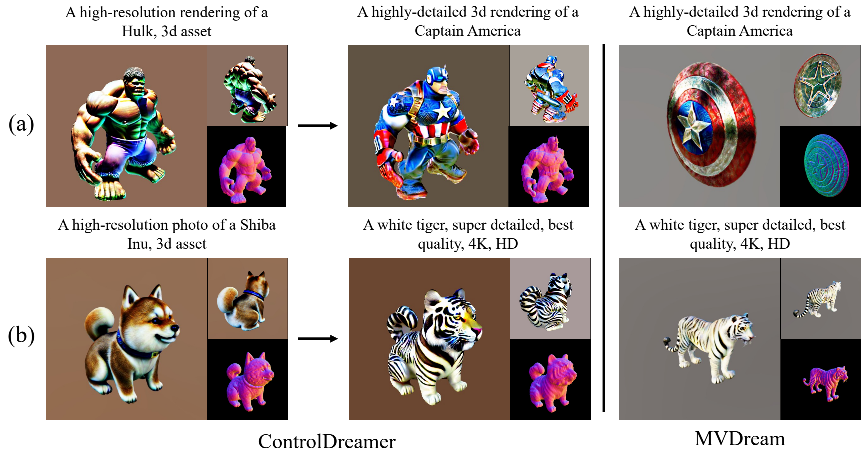

Figure 1: Comparing text-to-3D pipeline of our ControlDreamer and MVDream. On the right, MVDream's output reveals a vulnerability to pre-training geometry biases, often producing (a) unintended results such as shields or (b) stereotypical geometries related to prompts. On the left, ControlDreamer overcomes these biases, enabling unique combinations of geometry and style, even facilitating counterfactual generation in 3D models.

users' textual descriptions. Furthering this evolution, various works show high-resolution text-to-3D generation, employing multiple 3D representations [4, 5, 19] or variants of score distillation sampling [10, 40]. One such approach is MVDream [38], a model that extends the capabilities of Stable Diffusion [33], a well-known text-to-image diffusion model, by adapting it on the Objaverse dataset [8], which consists of multi-view images of 3D objects.

However, we observe that when attempting to generate imaginative 3D models using MVDream, it fails to seamlessly blend attributes such as geometry and style. For instance, when given a prompt like 'Captain America,' the model tends to produce a 3D model of a shield rather than the character itself. This issue likely arises due to dataset biases and the mode-seeking nature of score distillation sampling [29], which causes the model to generate geometries that reflect dominant modes in the dataset, rather than customized or more nuanced 3D models, such as a bulky version of Captain America (Fig. 1).

These challenges motivated us to decompose the text-to-3D pipeline into a two-stage approach given two text prompts. Initially, we generate a neural radiance field (NeRF) [22, 23] from a *geometry prompt* to sculpt geometric structures, followed by converting it to a mesh using DMTet algorithm [37] and refining both geometry and texture using a *style prompt*. To address this, we present *MV-ControlNet*, a depth-aware multi-view diffusion model. We first generate a paired dataset of text and multi-view images using GPT-4 [27] filtered texts and MVDream, followed by training a depth conditioner for the frozen MVDream, according to the principle of ControlNet [42].

In evaluations on our new benchmark, which consists of pairs of geometry and style prompts, ControlDreamer shows superior performance in generating styles on 3D models compared to existing two-stage pipelines. By using CLIP [30] scores and human assessments, we find that our depth-aware MV-ControlNet outperforms normal and edge-aware variants in rendering detailed 3D models. These findings confirm ControlDreamer's ability to produce high-quality 3D models that closely match the provided textual descriptions.

Table 1: Comparison to previous text-to-3D generation methods. Our method is the first two-stage pipeline designed for multi-view consistent 3D generation. Notably, it offers convenience by eliminating the need for annotations like bounding boxes and masks.

Method	3D Rep.	Multi-Stage	Annot.-Free	Multi-View
ProlificDreamer [40]	NeRF	✗	✓	✗
Fantasia3D [5]	DMTet	✓	✓	✗
Magic3D [19]	NeRF&DMTet	✓	✓	✗
MVDream [38]	NeRF	✗	✓	✓
Ours	NeRF&DMTet	✓	✓	✓

2 Related Works

2.1 Adding Conditions to Text-to-Image Models

Diffusion models like Stable Diffusion [33], trained on billions of image-text pairs [36], have shown remarkable capabilities in generating high-quality images from text or in removing harmful degradations from images [25]. ControlNet [42] introduces a method for adding conditions to Stable Diffusion by using a trainable copy alongside the original frozen model to perform specific tasks, such as depth-to-image conversion.

2.2 Text-to-3D Generation

While public datasets of 2D images are abundant, dataset of 3D assets are relatively scarce (∼80K in Objaverse [8]). Consequently, leveraging pre-trained 2D diffusion models to create diverse, user-customized 3D models continues to be a demanding task. To overcome such difficulties, DreamFusion [29] introduces score distillation sampling (SDS) to enable the optimization of an implicit 3D model [22, 23] without any 3D dataset, utilizing a pre-trained 2D diffusion model as a prior. Despite these advantages, DreamFusion suffers from slow scene generation and lower-quality results. To overcome these limitations, several methods are proposed involving primary optimization for NeRF and secondary optimization for the extracted mesh [5, 19], along with improvements for SDS [3, 14, 21, 40]. In contrast to prior 2D lifting-based methodologies, MVDream [38] fine-tunes a text-to-image diffusion model [33] through joint training on large-scale 2D and 3D datasets. Consequently, MVDream circumvents the *Janus Problem* and the *Content Drifting Problem*, consistently generating 3D models that align accurately with the given texts. Concurrent works have developed large-scale 3D generative models [2, 39]. However, these models primarily rely on image inputs, making them unsuitable for our objective of controlling both the geometry and style of 3D models through text descriptions.

2.3 Text-Guided 3D Editing

In text-to-3D generation, various innovative methods have been proposed. Magic3D [19] offers a coarse-to-fine pipeline for high-resolution 3D models, starting with NeRF and converting to deformable meshes. Fantasia3D [5] divides its process into geometry and appearance stages, beginning with a 3D ellipsoid and refining with DMTet. TEXTure [32] handles a similar task of generating textures on top of geometry, but unlike our approach, it does not allow for changes in the geometry. IT3D [6] refines initial rendered views using a 2D ControlNet [42] and ensures multi-view consistency through a GAN loss, whereas we use a single multi-view diffusion model. In Table 1, we outline a key gap in existing research: a

Figure 2: Main results. Our generation process begins by generating a coarse-grained geometry, followed by creating a fine-grained stylized 3D model using a style prompt.

lack of methods combining multi-stage, annotation-free, and multi-view capabilities for 3D generation. We found that current methods, especially in DMTet texture refinement, work well only when geometry and style prompts match. Furthermore, methods using 2D models struggle with multi-view consistency. Our ControlDreamer generates multi-view consistent 3D models by combining geometry and style specified by text prompts.

3 Method

Our ControlDreamer pipeline involves a two-stage process: first, training a coarse-grained NeRF [22, 23] with a geometry prompt and then creating a fine-grained stylized textured mesh [37] with a style prompt. We illustrate our pipeline in Fig. 3 and provide details in Sec. 3.1. Then, we emphasize a depth-aware, multi-view approach in the second stage and elaborate on the Multi-View (MV) ControlNet's training scheme in Sec. 3.2.

3.1 ControlDreamer: Two-Stage 3D Generation

3.1.1 Geometry Stage

Here, we generate a NeRF from a given geometry prompt. The style of this NeRF is subsequently modified in the second stage. We optimize the NeRF by distilling probability density [26] from the pre-trained MVDream [38]. This optimization is achieved through score

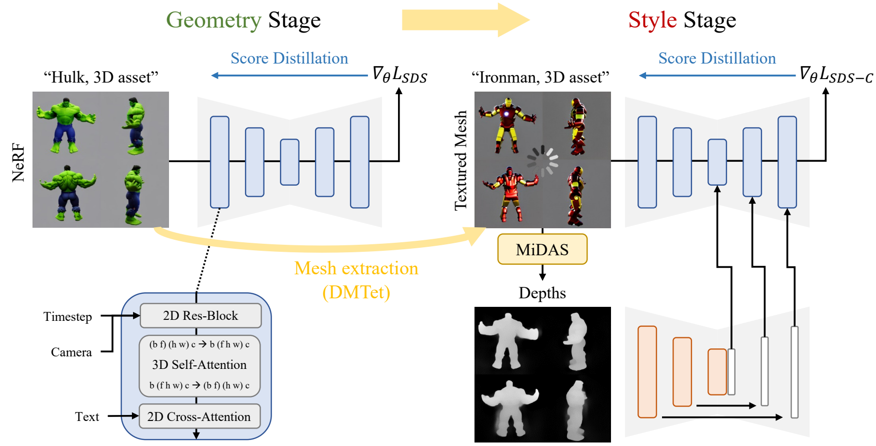

Figure 3: Illustration of ControlDreamer. (Left) Starting with a geometry prompt, we use MVDream to generate a NeRF, ensuring consistency through 3D self-attention. (Right) The NeRF is converted into a mesh via DMTet, followed by style generation through our MV-ControlNet, which integrates a trainable copy (red) and employs zero-initialized convolutions (white). MV-ControlNet is designed to understand geometry using multi-view depth.

distillation sampling (SDS) [29], where the diffusion model directly computes the loss gradient as shown in the following equation:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} \triangleq \mathbb{E}_{t,\varepsilon}\left[w(t)\left(\hat{\varepsilon}_\phi(z_t;y_{\text{geo}},c,t)-\varepsilon\right)\frac{\partial z}{\partial x}\frac{\partial x}{\partial \theta}\right], \tag{1}$$

where $t \sim U(t_{\min}, t_{\max})$, $w(t)$ denotes the weighting function, $\hat{\varepsilon}_\phi(z_t;y,t)$ represents a guided epsilon prediction [12], y_{geo} is the geometry text prompt, c represents the camera parameters, z refers to the VAE latents [16, 33] of rendered multi-view images x, and θ is the parameter of 3D model. Note that there is no backpropagation through the diffusion model.

To leverage the multi-view consistency of MVDream, we render four views using randomly selected camera parameters c in each iteration of optimization. These parameters are uniformly sampled from a range of view angles, each maintaining consistent elevation. In further detail, we incorporate point lighting [29], orientation loss [29], and soft shading [19]. The timestep range $[t_{\min}, t_{\max}]$ is linearly annealed during the optimization process. To prevent color saturation, we apply the x_0-reconstruction loss, which is the epsilon reconstruction loss weighted by the signal-to-noise ratio (SNR) [15], thus $w(t) = \text{SNR}(t)$.

3.1.2 Style Stage

Our second stage is motivated by the coarse-to-fine optimization of [19], which is utilized for generating high-resolution meshes. In this process, they optimize the textured mesh using DMTet [37], which uses a deformable tetrahedral grid and employs a differentiable rasterizer [17, 24] for efficient rendering. However, unlike the goal of [19], which is centered on high-resolution mesh refinement, our approach leverages the DMTet algorithm for generating *style* on the previously generated geometry.

In each iteration of optimization, we employ pre-trained MiDAS [31] to obtain predicted depth maps, using the same depth estimator that is used for training our MV-ControlNet in

Sec. 3.2. Subsequently, we optimize the mesh with the gradient computed by MV-ControlNet as shown in the following equation:

$$\nabla_\theta \mathcal{L}_{\text{SDS-C}} \triangleq \mathbb{E}_{t,\varepsilon}\left[w(t)\left(\hat{\varepsilon}_\phi(z_t;\mathcal{D}(x),y_{\text{style}},c,t)-\varepsilon\right)\frac{\partial z}{\partial x}\frac{\partial x}{\partial \theta}\right], \tag{2}$$

where y_{style} is the style text prompt, which is used differently from y_{geo}. $\mathcal{D}$ is our MV-ControlNet, which we later explain in Sec. 3.2. This gradient is coupled with regularizations to refine both geometry and texture meticulously. In contrast to recent methods [5, 19, 40] that necessitate narrow $[t_{\min}, t_{\max}]$ during the refinement phase, our ControlNet allows a large range of $[t_{\min}, t_{\max}]$ with scheduled timestep annealing. This approach ensures the depth information remains intact maintaining original geometry throughout the style optimization, even with significant timestep variations.

3.1.3 Training Details

We use the normal map consistency loss [19] and the Laplacian smoothing regularization loss [40] to eliminate pixelated artifacts in DMTet. These allow us to generate photorealistic 3D models. Utilizing MV-ControlNet, textures are effectively modified while preserving the initial geometry, even with significant variations in text descriptions. Furthermore, this effectiveness persists in scenarios where t is sampled from $t \sim U(0.02, 0.98)$. In the style stage, we sample t from this range for the initial $4/5$ of the total iterations, akin to the geometry stage. Subsequently, for the remaining iterations, t is sampled from $t \sim U(0.02, 0.5)$. These techniques differentiate MV-ControlNet from previous methods [5, 19] that rely on smaller timesteps for mesh refinement.

3.1.4 Theoretical Discussion

The key to the success of score distillation in text-to-3D generation lies in the guidance mechanism [12]. The guided score of Eq. 1 can be written as the following equation:

$$\begin{aligned}
\hat{\varepsilon}_\phi(z_t;y_{\text{geo}}) &= s\cdot(\varepsilon_\phi(z_t;y_{\text{geo}})-\varepsilon_\phi(z_t))+\varepsilon_\phi(z_t) && (3)\\
&\propto \nabla_{z_t}(s\cdot(\log p(z_t|y_{\text{geo}})-\log p(z_t))+\log p(z_t)) && (4)\\
&= \nabla_{z_t}(s\cdot\log p(y_{\text{geo}}|z_t)+\log p(z_t)), && (5)
\end{aligned}$$

due to Bayes' rule. Here, s represents the guidance scale and c and t are omitted for brevity. The equation indicates strong guidance encourages higher text alignment $p(y_{\text{geo}}|z_t)$.

The effect of integrating our MV-ControlNet into score distillation can be demonstrated similarly. The guided score in Eq. 2 can be expressed as follows:

$$\begin{aligned}
\hat{\varepsilon}_\phi(z_t;\mathcal{D}(x),y_{\text{style}}) &\propto \nabla_{z_t}(s\cdot(\log p(z_t|\mathcal{D}(x),y_{\text{style}})-\log p(z_t|\mathcal{D}(x)))+\log p(z_t|\mathcal{D}(x))) && (6)\\
&= \nabla_{z_t}(s\cdot\log \underbrace{p(y_{\text{style}}|z_t,\mathcal{D}(x))}_{\text{style blending}}+\log \underbrace{p(z_t|\mathcal{D}(x))}_{\text{geometry}}), && (7)
\end{aligned}$$

also due to Bayes' rule. Here, maximizing $p(z_t|\mathcal{D}(x))$ ensure that the optimized 3D representation maintains the geometry, while maximizing $p(y_{\text{style}}|z_t,\mathcal{D}(x))$ guides the blending of a style that suits the geometry.

3.2 Multi-View ControlNet

As elaborated in the previous section, our pipeline generates style in the second stage by refining the DMTet, converted from NeRF. We note the limitations of MVDream in refining DMTet, stemming from a lack of understanding of the geometry. To address this, we introduce the Multi-View (MV) ControlNet, crucial for generating fine-grained style by understanding coarse-grained geometry.

3.2.1 Generating Dataset

To train a depth-aware multi-view model, we construct a dataset using the curated 100K high-quality texts from OpenShape [20], filtered with GPT-4 [27] for quality assurance. Initially, these texts are refined using BLIP [18] and Azure Cognition Services for 2D image captions and subsequently enriched by GPT-4 to filter out uninformative content. Using MVDream, we generate multi-view images from these texts, matching MVDream's camera parameters to maintain its prior knowledge. This involves generating four orthogonal views for each text, each with uniformly distributed azimuth angles and elevation. Specifically, we randomly sample the camera parameters in a range of [0.9, 1.1] for distance, [15, 60] for fov, and [0, 30] for elevation. Then, we generate corresponding depth maps from the generated images using MiDAS [31]. This process, completed in 1.5 days on four A40 GPUs, involves texts that either explicitly mention '3D' or have '3D asset' appended as a postfix.

3.2.2 Training MV-ControlNet

Our goal is to train a depth encoder that incorporates multi-view depth conditions into MVDream, leveraging the model's 3D consistency and generation quality. Following the approach of ControlNet [42], which involves training an additional encoder atop a frozen diffusion model, we introduce multi-view depths as additional inputs to the pre-trained MVDream. Specifically, a trainable copy of MVDream's U-Net [34] encoder, including 3D attentions, is tasked with encoding depths and integrated into the U-Net decoder using zero-initialized convolutions. We train this encoder with the standard diffusion objective [13] formalized as follows:

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),t,\varepsilon\sim\mathcal{N}(0,1)}\left[\left\|\varepsilon - \varepsilon_\phi(z_t;\mathcal{D}(x),y,c,t)\right\|^2\right], \tag{8}$$

where $\mathcal{E}$ denotes the frozen VAE encoder [33], $\mathcal{D}$ is the trainable depth encoder, ε_ϕ is the frozen MVDream's U-Net, y is the text prompt, and c is the camera parameter.

4 Experiments

For both qualitative and quantitative assessments, we have created a benchmark dataset comprising a total of 30 pairs of geometry and style prompts. These pairs span across five categories: *Animals, Characters, Foods, General*, and *Objects*. Qualitative samples of our pipeline are presented in Fig. 2.

4.1 Qualitative Comparison of Multi-View Image Generation

Before evaluating 3D generation, we assess MV-ControlNet's multi-view image generation capabilities by comparing with several baselines as shown in Fig. 4. The baselines include

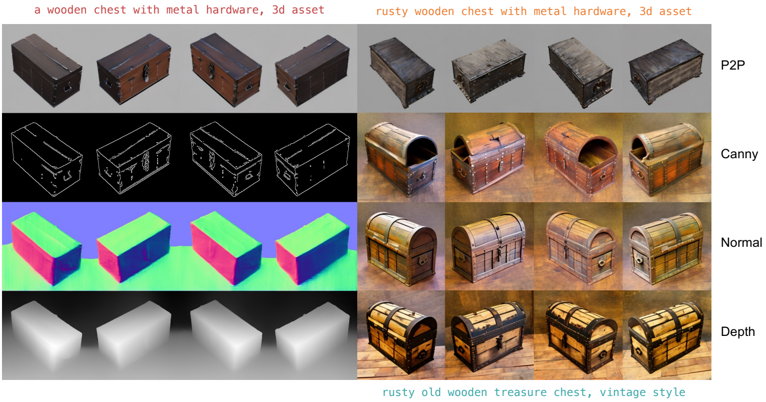

Figure 4: We compare the depth-aware MV-ControlNet with the P2P [] approach on MV-Dream, and also against MV-ControlNet variants trained under edge and normal conditions. On the left, source images are displayed alongside their respective conditions. Among these, the depth-conditioned multi-view images display the most visually appealing results.

MVDream using P2P [], an image editing method, and MV-ControlNet variants trained under the canny edge and normal maps. Regarding MV-ControlNet variants, we observe that canny edge maps result in unrealistic and 3D-inconsistent images, whereas normal maps provide 3D consistency but show degraded text alignment. Conversely, using depth conditioning proves most effective, leading to visually appealing and well-aligned results. This success motivated us to choose depth-conditioned MV-ControlNet as our model.

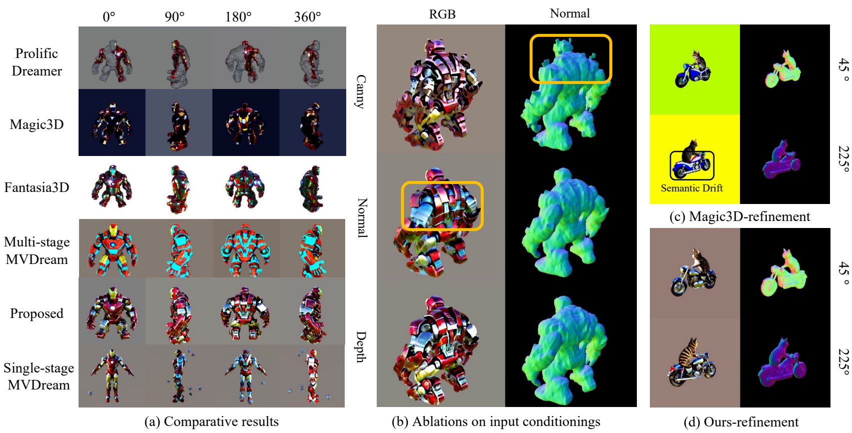

Figure 5: In (a), we present comparisons with previous pipelines. Hulk's geometry from Fig. 1, styled as Ironman, reveals that Magic3D and ProlificDreamer often produce texture artifacts, while Fantasia3D and MVDream are prone to color oversaturation. (b) illustrates the results under various input conditions. (c) shows the refinement process using Magic3D, while (d) highlights the superior results achieved with our ControlDreamer.

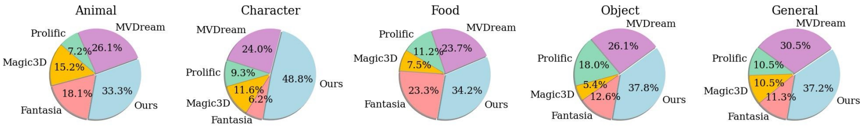

Figure 6: Human preference study on editability and text alignment in five prompt domains. ControlDreamer consistently ranks highest in user preferences across all domains.

Table 2: Comparison of directional CLIP similarity on five different categories.

Domain	Animals	Characters	Foods	General	Objects
Magic3D	0.218	0.280	0.236	0.247	0.217
Fantasia3D	0.191	0.250	0.256	0.240	0.208
ProlificDreamer	0.224	0.265	0.239	0.231	0.163
MVDream 2-stage	0.160	0.235	0.150	0.211	0.122
ControlDreamer	**0.241**	**0.353**	**0.270**	**0.341**	**0.255**

4.2 Qualitative Comparison of 3D Generation

We compare against several baselines, namely Magic3D [19], Fantasia3D [6], and ProlificDreamer [40]. All these methods, including our ControlDreamer, employ the DMTet [37] algorithm to refine textured meshes. The results of our stylization experiments using these baseline models are illustrated in Fig. 5. With Magic3D and ProlificDreamer, we observe some instances of failure, characterized by only partial texture changes in the geometry. This issue arises from a misalignment between the prior knowledge embedded in Stable Diffusion for the style prompt (e.g., the Ironman is slim) and the geometry from the previous stage. Our MV-ControlNet addresses this misalignment by understanding the depth information of the geometry, thereby enhancing alignment and coherence. While Fantasia3D and MVDream successfully stylize the overall geometry, they tend to produce results with issues like oversaturation or residual traits from the original geometry prompt. We invite the readers to explore additional results on the project page.

4.3 Quantitative Comparison of 3D Generation

4.3.1 Directional CLIP Similarity

To quantitatively compare the alignment between prompt modifications and image changes in ControlDreamer and MVDream, we evaluate the directional CLIP similarity [9] between first and second stage 3D models on our benchmark dataset. This score measures how well the change of rendered images align with the text directions within the CLIP space. We use texts with three view-oriented postfixes: ‘a front view’, ‘a side view’, and ‘a back view’, utilizing four views from each azimuth angles (0°, 90°, 180°, 270°) for the 3D samples.

In Table 2, averaged similarities and standard deviations are presented for each domain, across all six prompt candidates and directions. We employ the OpenCLIP ViT-L/14 model [7], trained on the LAION-2B dataset, which is different from the text encoder used in ControlDreamer. Our method demonstrates superior performance, excelling in generating text-aligned 3D models across all evaluated domains in our benchmark. Detailed explanations of the used prompts are presented in the Appendix.

4.3.2 Human Evaluations

We assess preference scores using Amazon Mechanical Turk (AMT) across five domains in Fig. 6. As a result, the proposed method consistently achieves the highest preference scores in all domains, demonstrating its effectiveness. This assessment allows for a comparative ranking of MVDream, Fantasia3D, Magic3D, and ProlificDreamer. We assess each methodology using 30 3D models from each domain, with 200 human evaluators participating. Each evaluator encounters three random samples and repeats the evaluation task three times, resulting in about 20 evaluations per domain. For additional details, the templates for the human evaluation are included in the Appendix.

4.4 Ablation Study

4.4.1 Variants of MV-ControlNet

In Fig. 5 (b), we compare the 3D stylization results using MV-ControlNet variants trained under canny edge, normal, and depth conditions. With canny edge inputs, significant artifacts appear in both RGB and normals. When using normal maps, the RGB colors in the 3D models are less photorealistic. However, depth maps consistently produce 3D models without artifacts in both color and normal maps, showcasing their superiority.

4.4.2 MV-ControlNet as a plug-in for prior methods

Additionally, we demonstrate a scenario where MV-ControlNet is used in the refinement stage of existing methodologies, with identical geometry and style prompts. DreamFusion [29] is employed to create the geometry model. We then compare the performance of MV-ControlNet against Magic3D [19]. Fig. 5 (d) highlights MV-ControlNet's adaptability as a plug-in component across various methods.

5 Discussion

We introduce ControlDreamer, a two-stage pipeline to improve text alignment in text-to-3D by decomposing the text into geometry and style prompts. To blend the geometry and style, we train MV-ControlNet, a depth-aware multi-view diffusion model, to enhance the geometry understanding during score distillation in the second stage. This integration effectively generates stylized 3D models by aligning diverse geometries and styles given by textual descriptions. Comparative results and theoretical analysis confirm that ControlDreamer outperforms existing text-to-3D methods in stylizing 3D models. Additionally, through comparisons with normal map and canny edge variants of MV-ControlNet, we demonstrated that depth conditioning is the most suitable for blending geometry and style.

While we improved text alignment in text-to-3D generation, some limitations remain. First, our training on single-object datasets limited our ability to experiment with complex scenes typical in 3D reconstruction. If multi-view diffusion models like ReconFusion [41], trained on complex scenes, become publicly available, we could better assess ControlDreamer's effectiveness. Additionally, though not our primary focus, future research could explore on various NeRF resolutions as these multi-view diffusion models advance to multiple resolutions and aspect ratios, similar to recent 2D image generative models [28].

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024.

[3] Cheng Chen, Xiaofeng Yang, Fan Yang, Chengzeng Feng, Zhoujie Fu, Chuan-Sheng Foo, Guosheng Lin, and Fayao Liu. Sculpt3d: Multi-view consistent text-to-3d generation with sparse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10228–10237, 2024.

[4] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18558–18568, October 2023.

[5] 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, October 2023.

[6] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1237–1244, 2024.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

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

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[10] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19740–19750, October 2023.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023.

[15] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[19] 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.

[20] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems*, 36, 2023.

[21] 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.

[22] 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.

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15, July 2022. doi: 10.1145/3528223.3530127. URL https://doi.org/10.1145/3528223.3530127.

[24] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, June 2022.

[25] Yeongtak Oh, Jonghyun Lee, Jooyoung Choi, Dahuin Jung, Uiwon Hwang, and Sungroh Yoon. Efficient diffusion-driven corruption editor for test-time adaptation. *arXiv preprint arXiv:2403.10911*, 2024.

[26] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR, 2018.

[27] OpenAI. Gpt-4 technical report. 2023.

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023.

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

[31] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[32] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

[33] 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.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[35] 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.

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[37] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

[38] 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.

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

[40] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2023.

[41] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024.

[42] 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.

Acknowledgement

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) through a grant from the Korea government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea (NRF) grant from the Korea government (MSIT) (No. 2022R1A3B1077720 and No. 2022R1A5A7083908), and the BK21 FOUR program for the Education and Research Program for Future ICT Pioneers at Seoul National University in 2024.