DGBD: Depth Guided Branched Diffusion for Comprehensive Controllability in Multi-View Generation

Hovhannes Margaryan[1] Daniil Hayrapetyan[1] Wenyan Cong[2] Zhangyang Wang[1,2] Humphrey Shi[1,3]
[1]Picsart AI Research (PAIR) [2]UT Austin [3]SHI Labs @ Georgia Tech, Oregon & UIUC

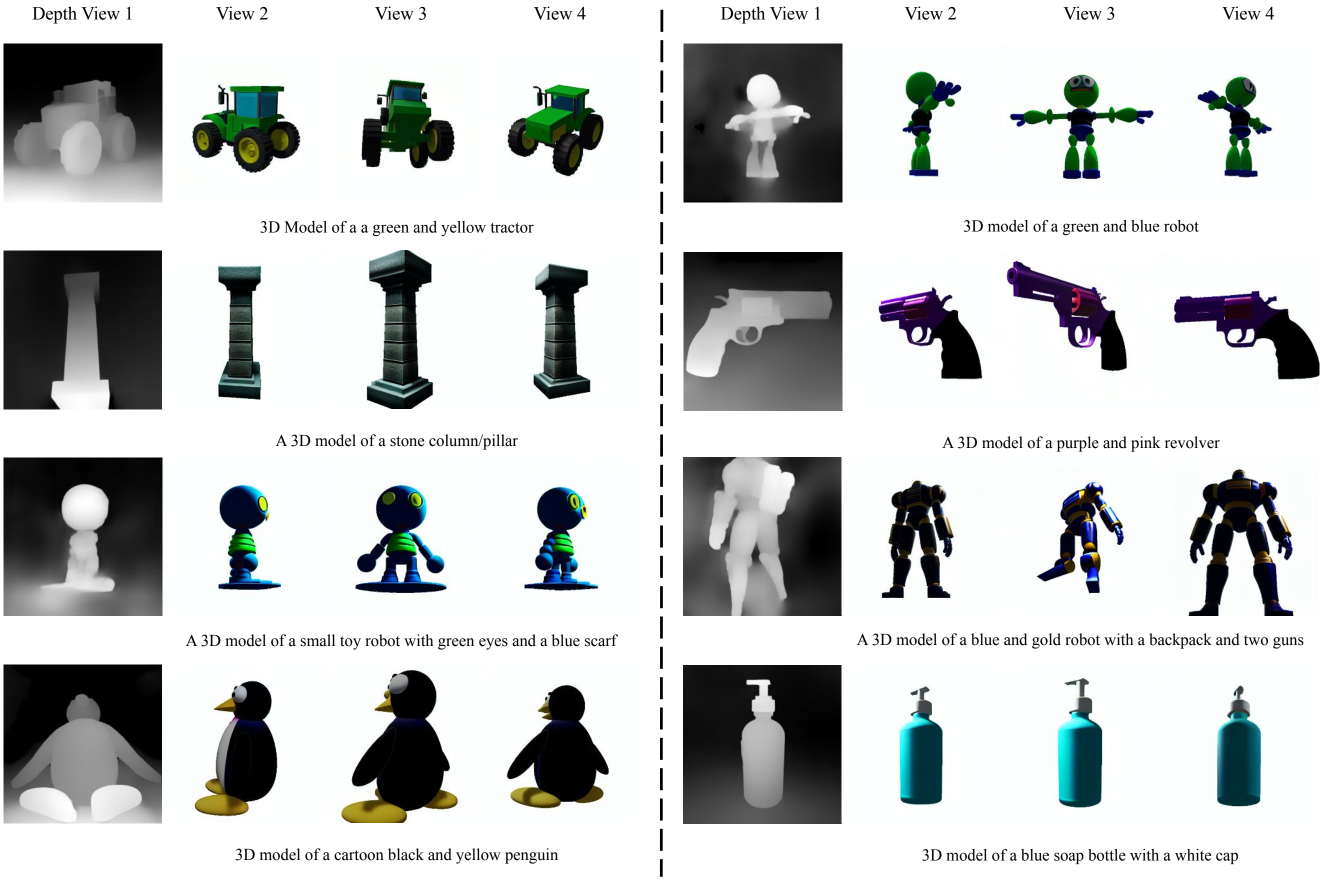

Depth View 1
View 2
View 3
View 4
Depth View 1
View 2
View 3
View 4
3D Model of a a green and yellow tractor
3D model of a green and blue robot
A 3D model of a stone column/pillar
A 3D model of a purple and pink revolver
A 3D model of a small toy robot with green eyes and a blue scarf
A 3D model of a blue and gold robot with a backpack and two guns
3D model of a cartoon black and yellow penguin
3D model of a blue soap bottle with a white cap

Figure 1. The proposed Depth Guided Branched Diffusion (DGBD) enables simultaneous control over non-perspective and perspective information. Given a prompt, depth map from one view and camera location of other views the framework generates views by propagating shape and size attributes from the depth map and aligning with the prompt and perspective. Moreover, the generated results are multi-view consistent.

Abstract

*This paper presents an innovative approach to multi-view generation that can be comprehensively controlled over both perspectives (viewpoints) and non-perspective attributes (such as depth maps). Our controllable dual-branch pipeline, named Depth Guided Branched Diffusion (**DGBD**), leverages depth maps and perspective information to generate images from alternative viewpoints while preserving shape and size fidelity. In the first DGBD branch, we fine-tune a pre-trained diffusion model on multi-view data, introducing a regularized batch-aware self-attention mechanism for multi-view consistency and generalization. Direct control over perspective is then* *achieved through cross-attention conditioned on camera position. Meanwhile, the second DGBD branch introduces non-perspective control using depth maps. Qualitative and quantitative experiments validate the effectiveness of our approach, surpassing or matching the performance of state-of-the-art novel view and multi-view synthesis methods.*

1. Introduction

Generative models have achieved significant success in high-fidelity image generation. This progress has primarily been facilitated by the arrival of diffusion models [2, 7, 28–30]. Recent endeavors within the domain of diffusion models have sought to adapt the pre-trained diffusion models, specifically, text-to-image diffusion models [4, 20, 23, 24, 35], originally conceived for 2D image generation, to address the task of multi-view and 3D synthesis [8–12, 16, 17, 26, 31, 33, 38]. Text-to-multi-view generation involves the generation of a set of views based on a prompt and a specified viewpoint. In other words, given a text and a designated camera position, the objective of text-to-multi-view is to produce an image that aligns with both the provided caption and geometric parameters, while maintaining coherence across multiple viewpoints. The viewpoint, in this context, refers to the camera location of the target view. In addition, controllable text-to-multi-view generation offers simultaneous manipulation of the shape and size of the object in the generated view through depth maps and adjustment of the viewpoint based on perspective information. One application of multi-view generation lies in its integration into the 3D artist's asset generation pipeline.

Diffusion-based multi-view synthesis methods provide direct control over perspective, enabling the generation of views from multiple camera locations. Additionally, beyond offering perspective control, they leverage either an image or text as guiding inputs. While these methods exhibit promising results, they still encounter challenges related to controllability and maintenance of multi-view consistency. To address these issues, this paper presents a holistic dual-branch pipeline for multi-view generation, called **Depth Guided Branched Diffusion (DGBD)**. Our method receives a textual caption, depth map from the first view, and camera location of the second view, and its objective is to synthesize an output view that represents the prompt by transferring the shape and size information from the given depth and incorporating the perspective information from the given viewpoint: see Fig. 1 for an overview.

The proposed DGBD pipeline comprises of two branches: the perspective branch and the non-perspective branch. In the perspective branch, a pre-trained diffusion model, Stable Diffusion [23], is modified and fine-tuned on the 3D dataset, Objaverse [1]. The following two alterations are applied to the U-Net of Stable Diffusion. First, the self-attention modules are replaced by a novel regularized batch-aware self-attention to introduce multi-view consistency and generalizability. Second, the camera position is injected into the U-Net of Stable Diffusion through a cross-attention mechanism to achieve a geometry-aware model. The non-perspective branch makes a trainable copy of the encoder of the pre-trained perspective branch's U-Net without the perspective projection layer and fine-tunes it using depth information motivated by Controlnet's [36] set-up.

The contributions of this paper are multi-fold:

- A novel formulation of multi-view generation with concurrent control over text, shape, size, and perspective.
- A controllable dual-branch pipeline allowing control over shape and size through non-perspective features, and viewpoint control by means of camera location.
- Introducing regularized batch-aware self-attention (RBA) mechanism for view consistency and generalization.
- Extensive experiments and an ablation study are conducted on the Objaverse and Google Scanned Objects [3] datasets to demonstrate the proficiency of the proposed method, exceeding or equating the performance of state-of-the-art diffusion-based novel view and multi-view synthesis methods.

2. Related Work

2D diffusion models. Diffusion models [7, 28, 29], particularly text-to-image diffusion models [19, 21, 23], have demonstrated their capability for enabling high-quality and diverse 2D image generation by training on extensive image datasets. GLIDE [15] enhances image fidelity by integrating classifier-free guidance [6] in diffusion models. Conversely, DALLE-2 [20] leverages CLIP [18] features to improve text-to-image alignment. Additionally, Latent Diffusion Models (LDMs) [23] propose training within an autoencoder's latent space for faster training. These models are inherently geometry-unaware. Thus, their direct applicability to 3D and multi-view synthesis remains limited. To mitigate this restriction, substantial modifications to the existing frameworks are necessary. Moreover, additional mechanisms are required to address multi-view consistency.

Diffusion-based 3D generation. State-of-the-art diffusion-based 3D generation techniques [9, 11, 17, 26, 33, 38] utilize differentiable image generators such as Neural Radiance Fields (NeRF) [14] and the knowledge of a pre-trained text-to-image diffusion model as a prior for 3D generation. This category of methods mandates a time-consuming optimization process to derive an object-specific 3D representation. Particularly, Dreamfusion [17] and Score Jacobian Chaining (SJC) initialize [33] a NeRF-based differentiable generator and optimize it using the gradients of a pre-trained diffusion model. Primary issues associated with these score distillation methods are color saturation, diversity across samples, and consistency of objects from

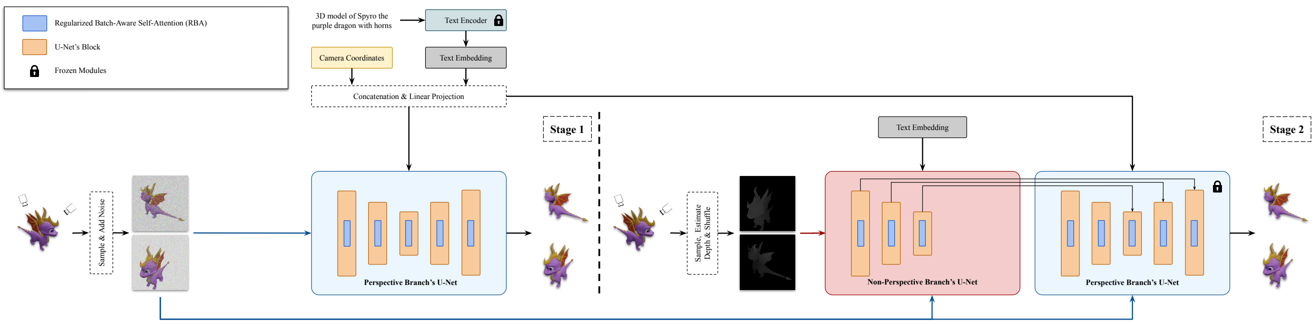

Regularized Batch-Aware Self-Attention (RBA)
U-Net's Block
Frozen Modules
3D model of Spyro the purple dragon with horns
Text Encoder
Camera Coordinates
Text Embedding
Concatenation & Linear Projection
Stage 1
Sample & Add Noise
Perspective Branch's U-Net
Text Embedding
Stage 2
Sample, Estimate Depth & Shuffle
Non-Perspective Branch's U-Net
Perspective Branch's U-Net

Figure 2. Two-stage training procedure for the DGBD pipeline: (1) The initial phase involves selecting views, prompts, and camera coordinates as inputs to the first branch. (2) In the second stage, the perspective branch's encoder is duplicated, and permuted depth maps, selected views, and prompts enter the second branch. The dual-branch pipeline generates views aligned with camera location and attributes like shape and size propagated from depth. It uses a regularized batch-aware self-attention in the U-Net. Though visualized in pixel space, in practice we work within latent space.

view to view. The first two challenges are addressed by DreamTime [9] through the substitution of time-based sampling with non-uniform sampling, aligning the sampling process with NeRF optimization. However, it still faces challenges with object-consistency. Collaborative Score Distillation [11] proposes a generalization of DreamFusion's framework by modifying the objective function to include multiple samples of the same scene, yielding 3D synthesis with object consistency maintained across multiple views. Meanwhile, 3DFuse [26] obtains and injects coarse view-specific depth information in a pre-trained LDM from the generated image and optimized prompt embedding to tackle 3D incoherence. On the other hand, SparseFusion [38] presents a two-stage approach to address the issue of 3D inconsistency. The process initially employs a view-conditioned LDM to obtain a distribution over possible images given reference views and a random query viewpoint. Next, distillation is applied aimed at mode seeking, which yields a 3D representation specific to an object.

Diffusion-based novel view and multi-view generation. In this stream of approaches [12, 27, 31] geometry information is directly incorporated into a pre-trained LDM conditioned on either a reference image or text and fine-tuned to facilitate novel view or multi-view synthesis. On the one hand, Zero-1-to-3 [12] tackles novel view synthesis given a condition view and a relative viewpoint by injecting camera information into a pre-trained LDM via cross-attention modules and concatenating the given input view with the image being denoised to obtain object-consistency across viewpoints. On the other hand, a correspondence-aware attention mechanism is introduced by MVDiffusion [31] between the modules of the U-Net, aiming to yield view-consistent images. Additionally, in MVDREAM [27], the authors condition the diffusion process on the camera extrinsic matrix and use 3D attentions with simultaneous training on the LAION dataset [25].

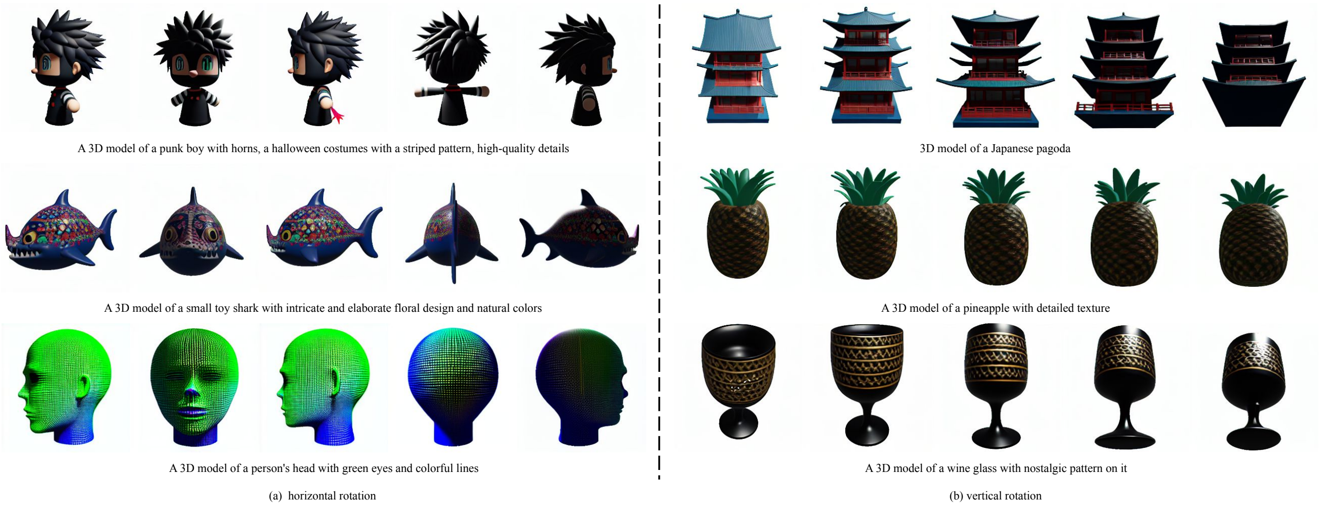
3. Method

This section begins with a problem definition of controllable multi-view synthesis that allows simultaneous control over text, depth, and perspective. Then, each branch of the proposed DGBD multi-view generation method is described: the perspective branch and the non-perspective branch.

3.1. Controllable Multi-View Generation

Controllable multi-view generation aims at synthesizing an image aligning with the given prompt and propagating the given depth information from the first view and the perceptive information from the second view. Formally, given a depth map $D_1 \in R^{1\times H\times W}$ from the first view and camera location $V_2 \in R^3$ from the second view, the proposed pipeline generates an output image $X_2 \in R^{3\times H\times W}$ from the second view coherent with the shape and size information of the given depth and viewpoint from camera location where H and W are the height and width of the depth map.

The naive approach of training both branches from the beginning does not produce satisfactory results: neither the perceptive nor the non-perspective information is propagated to the generated output. Hence, we introduce a two-stage dual-branch training strategy. An overview of the proposed pipeline's training is shown in Fig. 2. In the first stage, the perspective branch is fine-tuned from Stable-Diffusion-v1-5 (SD-v1-5) [23] by conditioning the model on text prompt, perspective information and using regularized batch-aware self-attention for multi-view consistency (see Sec. 3.3). In the second stage, inspired by ControlNet [36], a trainable copy of the latter without perspective injection layer is made and fine-tuned using depth guidance. Specifically, in the second stage of the training, a batch of



A 3D model of a punk boy with horns, a halloween costumes with a striped pattern, high-quality details

3D model of a Japanese pagoda

A 3D model of a small toy shark with intricate and elaborate floral design and natural colors

A 3D model of a pineapple with detailed texture

A 3D model of a person's head with green eyes and colorful lines

A 3D model of a wine glass with nostalgic pattern on it

(a) horizontal rotation

(b) vertical rotation

Figure 3. Visual results of the perspective branch of the proposed DGBD framework on different prompts. Given a textual prompt and camera locations (five in this case), this branch generates images of the same object consistently from the given viewpoints.

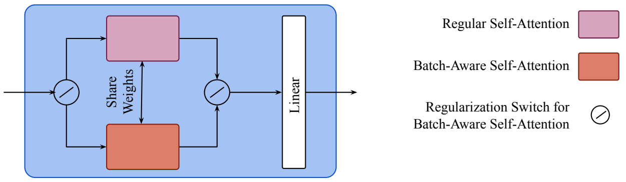

Share Weights
Linear
Regular Self-Attention
Batch-Aware Self-Attention
Regularization Switch for Batch-Aware Self-Attention

Figure 4. The proposed RBA module. The RBA module comprises regular self-attention and batch-aware self-attention with a stochastic switch.

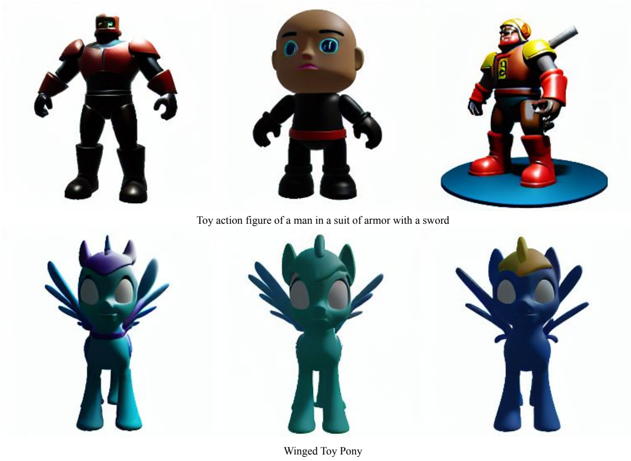
images from the same object is given to the first and second branches, corresponding camera locations are given to the first branch, and shuffled depth maps are given to the second branch. At inference, the influence of the depth map can be controlled via depth control scale similar to how conditioning maps are controlled in ControlNet. An ablation study is provided in Sec. 4.5 to demonstrate the effect of the depth control scale with additional results in the appendix.

3.2. Perspective Branch

The goal of the perspective branch is to generate a view conditioned on the given text and camera location. Formally, given a prompt y and a viewpoint $V \in R^{N\times 3}$, this branch aims to train a model, f, to generate views of the same object from N camera positions consistent with the text:

$$\hat{X} = f(y, V), \tag{1}$$

where $\hat{X} \in R^{N\times 3\times H\times W}$. Additionally, the object should be coherent across the viewpoints. In this branch, pretrained SD-v1-5 is fine-tuned after modifying its U-Net to handle view consistency and geometry. It is hypothesized that the knowledge of SD-v1-5 trained on millions of 2D images can be transferred to tackle geometry-aware multi-view generation. The following two modifications are applied to the U-Net of SD-v1-5. First, regularized batch-aware self-attention is introduced and all self-attention modules in the U-Net are replaced to tackle multi-view consistency and generalization. Second, geometry control is introduced by conditioning the model on the camera position using cross-attention. Fig. 3 shows some results of the perspective branch (see the appendix for additional results).



Toy action figure of a man in a suit of armor with a sword

Winged Toy Pony

Figure 5. Diversity of generated samples of the perspective branch. The camera location and caption are fixed.

3.3. Choice of self-attention: regularized batch-aware self-attention

The proposed regularized batch-aware self-attention (RBA) is shown in Fig. 4. The RBA module consists of reg-

	Zero-1-to-3	MVDREAM	Ours
Image	✓	✗	✗
Text	✗	✓	✓
Perspective	✓	✓	✓
Non-Perspective	✗	✗	✓

Table 1. Controllability analysis among state-of-the-art novel view and multi-view generation methods and ours.

ular self-attention and batch-aware self-attention. Regular self-attention module [32] in the U-Net architecture of SD-v1-5 is defined as follows. Given a feature map $x \in R^{B\times(H\times W)\times C}$, where B is the batch-size, H and W are the height and width of the feature map, and C is the number of channels, it is first linearly projected to query, key and value $Q, K, V \in R^{B\times(H\times W)\times C}$, and then the output of the self-attention module is computed as follows:

$$\text{Self-Attn}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{C}})V. \qquad (2)$$

In the case of batch-aware attention, the feature map is $x \in R^{1\times(B\times H\times W)\times C}$ and self-attention is computed similar to Eq. (2). Hence, in the case of batch-aware attention, each feature map in a batch attends itself and all other feature maps in a batch. RBA applies batch-aware attention with probability p and regular attention with probability $1-p$. Thus, the RBA module is formalized as:

$$\text{RBA}(B, R) = ZB + (1-Z)R, \qquad (3)$$

where B and R are the outputs of batch-aware self-attention and regular self-attention, and $Z \sim \text{Bernoulli}(p)$. In our experiments, we empirically set $p = 0.1$ during training and find that it helps with view consistency and generalization. During the inference stage p is set to 1.

It is empirically discovered that training only with batch-aware self-attention results in generation of box-like objects and forgetting of concepts. Additionally, the generated results are not consistent with the given prompt. On the contrary, training with the proposed RBA module avoids overfitting, introduces multi-view consistency, and improves convergence. An ablation study on RBA is presented in Sec. 4.5. On the implementation level, all components of the RBA module are initialized and fine-tuned from the regular self-attention modules of SD-v1-5.

Additionally, replacing the regular self-attention modules in SD-v1-5's U-Net with the proposed RBA module keeps the characteristic of generating diverse results across samples of the original model. Fig. 5 shows two examples with fixed camera position, prompt, and different latents. The generated samples differ in detail and are diverse.

	Zero-1-to-3	MVDREAM	Ours
PSNR ↑	10.30	9.69	**10.90**
SSIM ↑	**0.67**	0.62	**0.67**
LPIPS ↓	**0.26**	0.33	0.27
A-LPIPS ↓	0.15	0.24	**0.10**
CLIPScore ↑	NA	20.98	**21.27**

Table 2. Quantitative comparison among state-of-the-art diffusion-based novel view and multi-view generation methods and ours with zero depth control scale (i.e. no depth guidance) on a random sample of 400 objects from Google Scanned Objects dataset. NA stands for not applicable.

Geometry control: The U-Net of SD-v1-5 does not possess information about the geometry of the generated samples. To introduce perspective control over the generated examples the absolute camera location, V, is concatenated with the textual embedding and linearly projected to the dimensionality of the embedding space of the text and then injected into the model via cross-attention similar to how Zero-1-to-3 [12] injects relative camera position with the given condition view. The camera position is expressed in a spherical coordinate system. We assume the object's center is the origin and the camera faces it.

3.4. Non-perspective Branch

In the second branch of DGBD, non-perceptive control is introduced to SD-v1-5. After training the perspective branch, a trainable copy of the U-Net's encoder without the geometry control is made and it is fine-tuned by adding depth guidance. In this stage of the training, the objective is to propagate non-perceptive information from the given depth D_1 and generate a view aligning with the given viewpoint V_2. During training, views are sampled from the given object for a batch and they are input to the first branch along with prompts and camera locations. Additionally, the depth maps are shuffled and they are input to the second branch with a batch of views and prompts. Specifically, the depth maps are fed to the learnable encoder mirroring the method in Controlnet with zero convolution layers. The output of the dual-branch pipeline is a batch of views where each element of the batch possesses non-perspective information such as shape and size from the given depth maps and aligns with the given camera location. Fig. 1 provides results of the DGBD method (see the appendix for additional results).

4. Experiments

4.1. Dataset

The proposed DGBD pipeline is fine-tuned on a subset of Objaverse [1], a large-scale dataset of 3D objects, with prompts provided by Cap3D [13]. Almost 413K objects are

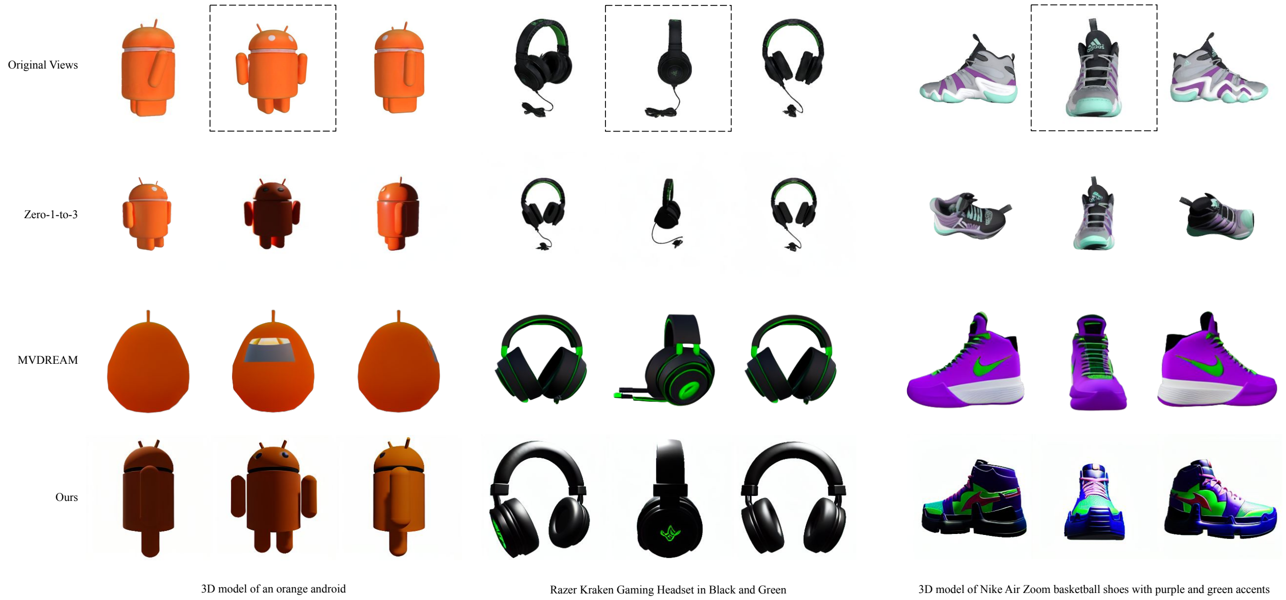

Original Views
Zero-1-to-3
MVDREAM
Ours
3D model of an orange android
Razer Kraken Gaming Headset in Black and Green
3D model of Nike Air Zoom basketball shoes with purple and green accents

Figure 6. Visual comparison of baseline diffusion-based novel view and multi-view synthesis methods and ours with zero depth control scale (i.e. no depth guidance). Our method achieves better visual results with high-fidelity than Zero-1-to-3 and similar or more detailed and illuminated objects (e.g. headset) than MVDREAM. Input to Zero-1-to-3 is indicated using a dashed outline.

used for training. For each object, 32 views are rendered. The polar angle, azimuth angle, and the distance from the center are uniformly sampled from $[60, 120]$, $[0, 360]$ and $[0.8, 1.5]$. The horizontal field of view is fixed to $49.1°$. All images are rendered at 512×512 resolution with random area lighting using the BLENDER EEVEE engine. Our dataset comprises roughly 13M images. Depth maps are approximated using MiDaS [22].

4.2. Baselines and Evaluation Metrics

In this section an analysis is presented, encompassing aspects of controllability, quantitative and qualitative evaluation with a focus on comparing the proposed DGBD approach to state-of-the-art diffusion-based novel view and multi-view generation methods: Zero-1-to-3 [12] and MVDREAM [27].

A random sample of 400 objects are used from Google Scanned Objects (GSO), a dataset of scanned photo-realistic items [3]. For each object, five images are rendered from different angles covering frontal and lateral views. Cap3D is applied to obtain text prompts. For Zero-1-to-3 an original view is input to the method. v1.5 model released by MVDREAM is used for a fair comparison. The following five azimuth angles are used in MVDREAM and our method: $330°, 270°, 0°, 90°, 180°$. The elevation angle is fixed to $90°$ for our method which corresponds to $0°$ in Zero-1-to-3 and MVDREAM. The corresponding azimuth angles in Zero-1-to-3 are $-30°, -90°, 0°$ (given as an input), $90°, 180°$. In this section, the depth control scale is set to zero (only the perspective branch works at inference), unless mentioned otherwise. Thus, the results obtained by our model are fairly comparable with MVDREAM. We also include a comparison with Zero-1-to-3 as, to the best of our knowledge, it is the first known work fine-tuning a pre-trained diffusion model for novel view synthesis. The following metrics are computed for quantitative comparison: PSNR, SSIM [34] and LPIPS [37]. A-LPIPS [8] is reported as an assessment of multi-view consistency among the generated views. Moreover, CLIPScore [5] is used to compare text-to-image alignment of our method and MVDREAM.

4.3. Comparison with Baselines

Tab. 1 presents a controllability analysis between the baselines and the proposed method. Zero-1-to-3 conducts novel view synthesis given only an input view and hence offers control over the input image and perspective. MVDREAM provides control over a prompt and camera position. Alternatively, DGBD affords guidance over perspective, non-perspective (shape and size), and textual information.

Tab. 2 demonstrates a quantitative comparison between the baselines and ours on the GSO dataset. PSNR, SSIM, and LPIPS are computed by comparing the generated output against a ground truth render from the same viewpoint. To get the A-LPIPS score, LPIPS is calculated and averaged across five adjacent views of the same object. The proposed DGBD pipeline achieves better PSNR and A-LPIPS values and similar SSIM and LPIPS scores in comparison to baselines. Also, DGBD suppresses MVDREAM by CLIPScore.

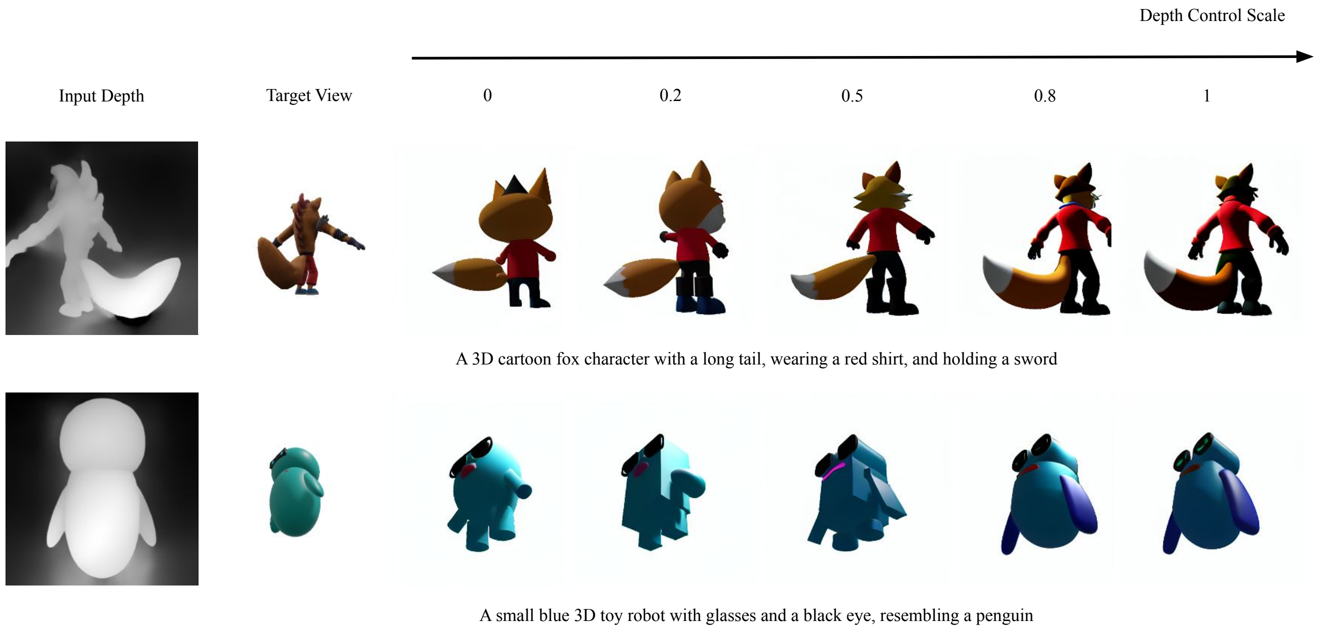

Depth Control Scale
Input Depth
Target View
0
0.2
0.5
0.8
1
A 3D cartoon fox character with a long tail, wearing a red shirt, and holding a sword
A small blue 3D toy robot with glasses and a black eye, resembling a penguin

Figure 7. Effect of depth control scale of the proposed DGBD framework. Given a textual caption, input depth, and the camera viewpoint of the target view, the method generates an output view corresponding with the given prompt, perceptive and transferring shape and size information from the depth map, and as the depth control scale increases the propagation of this information in the generated view increases.

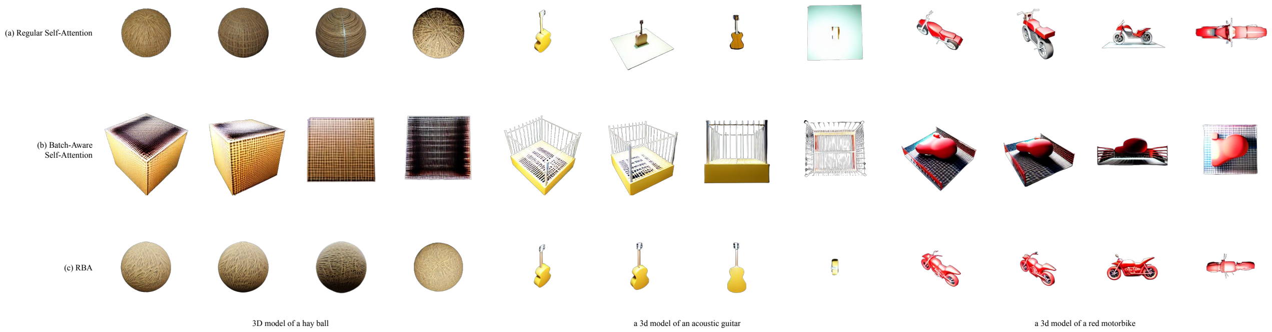

(a) Regular Self-Attention
(b) Batch-Aware Self-Attention
(c) RBA
3D model of a hay ball
a 3d model of an acoustic guitar
a 3d model of a red motorbike

Figure 8. Ablation study on the proposed RBA module. (a) Regular self-attention lacks multi-view consistency. (b) Training with batch-aware self-attention results in catastrophic forgetting. (c) Training with RBA improves multi-view consistency and generalizability. In this example, the depth control scale is set to zero (i.e. no depth guidance).

Fig. 6 demonstrates a qualitative comparison among the baseline methods, ours and the original renders. The input view to Zero-1-to-3 is indicated using a dashed outline. The results generated with the proposed DGBD method with zero depth control scale have higher visual quality with more details, shadows, and light information and better multi-view consistency than Zero-1-to-3. In certain instances, the proposed DGBD method yields output views with highly detailed objects and more light information such as in the example of shoe and headset than MVDREAM. Conversely, MVDREAM pays more attention to details available in the text (e.g. Nike logo). In some cases, it is challenging for MVDREAM to generate a view from an oblique angle as in the ”orange android” example. Additionally, in our experiments, we notice that both MVDREAM and the proposed DGBD pipeline can hallucinate the given camera location and generate the same view given two different camera locations in a batch.

4.4. Influence of depth control scale

Fig. 7 (see additional results in the appendix) provides a visual comparison of the generated results by the proposed method with different depth control scales (from 0 to 1) on a validation set from Objaverse. For smaller values of the depth control scale, the influence of the given depth map reduces and the method has more freedom in generating an

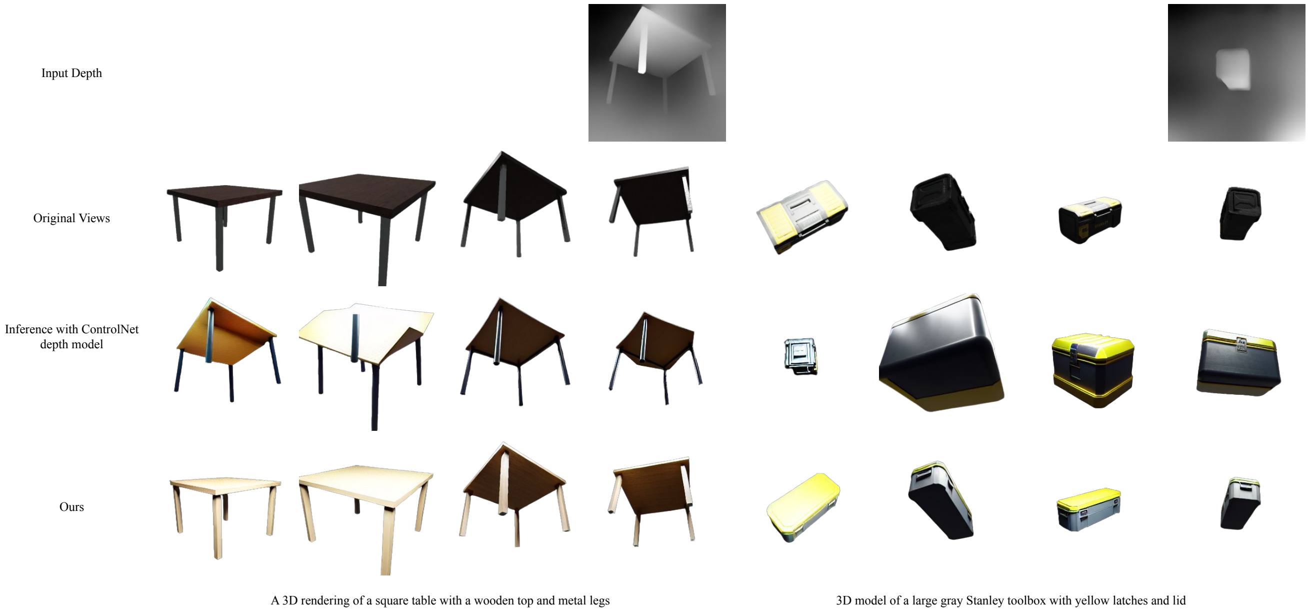

Input Depth
Original Views
Inference with ControlNet depth model
Ours
A 3D rendering of a square table with a wooden top and metal legs
3D model of a large gray Stanley toolbox with yellow latches and lid

Figure 9. Ablation study on the proposed DGBD pipeline. Combining its perspective branch unwittingly with the ControlNet depth model at inference generates inconsistent views relying on depth and disregarding the provided camera location. The proposed pipeline introduces view consistency, transfers shape, and size attributes from the depth map, and aligns with the given viewpoint. In this experiment, the depth control scale is set to one.

image aligning with the prompt and camera location. Particularly, when the depth control scale is set to zero, the method performs multi-view generation without considering the structural attributes form the depth map. Higher depth control scale values intensify depth map impact, propagating shape and size to the output view.

4.5. Ablation Study

An ablation study is conducted in this section on two primary components of the proposed method on a validation set from Objaverse. First, the proposed RBA module is juxtaposed with regular and batch-aware self-attention (Fig. 8). The regular self-attention module inherently lacks multi-view consistency (Fig. 8 (a)). This is vividly exemplified in the ”ball” and ”guitar” cases. Training the model with batch-aware self-attention results in concept forgetting and generation of box-like objects without considering the given prompt (Fig. 8 (b)). On the other hand, training the model using the proposed regularized batch-aware self-attention with $p = 0.1$ improves generalizability and introduces multi-view consistency (Fig. 8 (c)).

Second, the proposed DGBD pipeline is compared against combining the perspective branch with ControlNet depth model [36] during inference (Fig. 9). Naively combining the perspective branch with the ControlNet depth model does not handle the proposed problem of controllable multi-view synthesis. First, the generated output follows the given depth and ignores the perspective information. Second, the generated views are not consistent. In contrast, the generated output by the proposed dual-branch method successfully propagates shape and size information from the input depth and generates views aligning with the given perspective. Moreover, the proposed pipeline ensures multi-view consistency.

5. Conclusion

This paper introduces DGBD, a novel controllable multi-view generation method. DGBD is a dual-branch pipeline that adjusts perspective and non-perspective attributes concurrently. Given a prompt, depth map from one view, and camera position of another view, DGBD generates an output view with shape and size propagated from the depth map and aligns with caption and perspective. The perspective branch of DGBD is fine-tuned from Stable Diffusion, a text-to-image model, with two key modifications: replacing self-attention modules with novel regularized batch-aware self-attention for multi-view consistency and introducing geometry control via cross-attention. In the non-perspective branch, an encoder copy is created, omitting the geometry projection layer, and fine-tuned using shuffled depth maps for alternative view generation. Our approach achieves superior or comparable results, validated through qualitative and quantitative assessments. Our future work will extend DGBD to controlling multi-view generation in the wild.

References

[1] 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. 2, 5

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[3] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Michael Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022. 2, 6

[4] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation withnbsp;human priors. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, page 89–106, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 6

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[8] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation, 2023. 2, 6

[9] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation, 2023. 2, 3

[10] Zutao Jiang, Guansong Lu, Xiaodan Liang, Jihua Zhu, Wei Zhang, Xiaojun Chang, and Hang Xu. 3d-togo: Towards text-guided cross-category 3d object generation, 2023.

[11] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual synthesis, 2023. 2, 3

[12] 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 (ICCV)*, pages 9298–9309, 2023. 2, 3, 5, 6

[13] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 5

[14] 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, 2020. cite arxiv:2003.08934Comment: ECCV 2020 (oral). Project page with videos and code: http://tancik.com/nerf. 2

[15] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 2

[16] Yichen Ouyang, Wenhao Chai, Jiayi Ye, Dapeng Tao, Yibing Zhan, and Gaoang Wang. Chasing consistency in text-to-3d generation from a single image, 2023. 2

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

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 2

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[22] 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*, 44:1623–1637, 2019. 6

[23] 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 (CVPR)*, pages 10684–10695, 2022. 2, 3

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 2

[25] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294. Curran Associates, Inc., 2022. 3

[26] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee,

and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2, 3

[27] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 3, 6

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 2

[31] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023. 2, 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[33] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12619–12629, 2023. 2

[34] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[35] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7754–7765, 2023. 2

[36] 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 (ICCV)*, pages 3836–3847, 2023. 2, 3, 8

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[38] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12588–12597, 2023. 2, 3