Distillation-free Text-to-3D via Guided Multi-view Diffusion

Yanbo Xu
Carnegie Mellon University

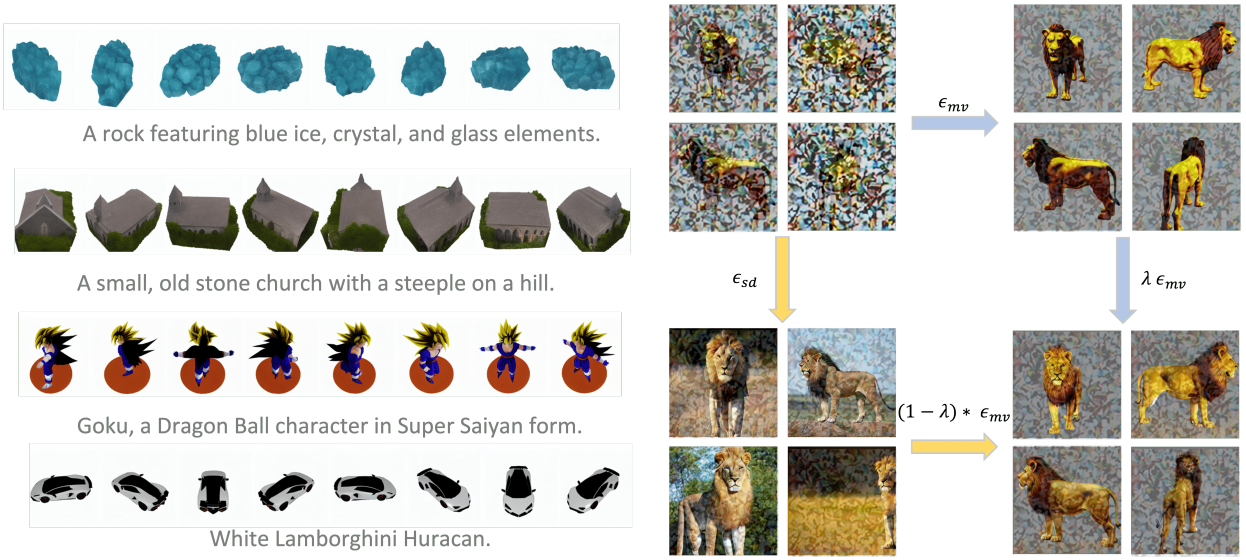

Figure 1

We propose a method that can generate multi-view images of objects from text prompts with high fidelity. There has been considerable progress in the field of 3D generation thanks to large synthetic 3D datasets such as Objaverse and the advancement of diffusion models. However, the generated multi-view images or 3D assets often suffer from simplistic textures, possibly because the textures of the objects in synthetic datasets are not as sophisticated as in-the-wild objects. This limits the practical usage of such models. On the other hand, text-to-2D methods are capable of generating images with high fidelity, partly due to the significantly larger and richer dataset they are trained on. We want to show that we can generate multi-view images with sophisticated textures by actively leveraging such a text-to-2D diffusion model in conjunction with a text-to-multi-view model while performing DDIM sampling.

1 Introduction

Large-scale 2D diffusion models [1], [2] have helped democratize image creation and editing. Building on this success, approaches aiming for text-conditioned 3D inference have 'distilled' such 2D generative models for inferring high-fidelity 3D representations [3]. While these methods have led to impressive 3D synthesis, their inference procedure is fundamentally different from the feed-forward probabilistic generation that the underlying 2D diffusion models are capable of. In particular, as these distillation-based methods do not capture the distribution over 3D structure, they cannot readily yield diverse 3D samples, and their reliance on optimization inherently limits inference efficiency. In this project, we want to pursue an alternate distillation-free approach and seek to learn a (text-conditioned) generative prior over 3D representations, thus allowing feed-forward diverse 3D generation.

While prior methods have pursued this task of learning 3D generative models [4], [5], they typically infer volumetric 3D or point cloud representations, relying solely on 3D data for learning. Unfortunately, the inherently limited diversity of this data limits their ability to produce complex 3D structures and these systems do not exhibit the strong generalization evident in the internet-driven 2D generative models and struggle. Our key insight is that instead of learning to generate such 3D representations, one can instead generate 3D-consistent multi-view images. In particular, we want to show that text-conditioned multi-view generation can be performed by adapting existing 2D generation models, thus allowing our approach to benefit from large-scale 2D pre-training. We will train a multi-view denoising diffusion model that leverages a 2D diffusion backbone while incorporating 3D bottleneck layers to ensure consistency across the generated views.

With successful training, our multi-view diffusion model should allow the direct generation of 3D-consistent views given an input text prompt. However, despite being initialized from a generic 2D generation model, the training on (simpler) multi-view data might hinder its ability to synthesize rich textures and complex structures. To overcome this, we propose an inference strategy to combine guidance from a frozen pre-trained generative model. More specifically, we bias the scores computed in the multi-view denoising diffusion sampling to also incorporate gradients from a generic 2D image generation model. We will show that this procedure is equivalent to sampling from a distribution that combines a per-image (pre-trained) prior with the learned multi-view generative distribution. Intuitively, our strategy allows us to generate multi-view images that are both 3D consistent and high-fidelity.

2 Related Works

2.1 Learning Generative 3D Models

There have been attempts to learn generative 3D models from 3D data [4], [5], such as voxel, point cloud, etc. However, the large-scale collection of 3D data is much harder than that of 2D, which limits the generative capability of models trained on these datasets.

2.2 3D Inference using 2D Diffusion

Dreamfusion [3] proposed a method called SDS to lift text-conditioned 2D diffusion models to 3D by optimizing a neural field such that its renderings have a high likelihood under the diffusion model. There are also similar formulations [6] that utilize naive SDS prompts to eliminate artifacts such as Janus effects, high saturation, etc. However, it has been shown [7] that SDS will reduce to a Maximize Likelihood Estimation problem, which means that its behavior is mode-seeking, limiting the diversity of optimized 3D assets.

Since then, there have been several works that focus on improving quality and reducing artifacts [8], [9]. [10] extended the distillation to the image domain to enable sparse image reconstruction. While these lead to impressive results, they require time-consuming per-instance optimization. In addition, these approaches often fail to generate diverse results given the probabilistic 2D model.

2.3 Distillation-free 3D Inference

Instead of learning from 3D data such as voxel or point cloud, supervision from view-consistent images enables large 2D pre-training. Zero123 [11] directly finetune a 2D model using the multiview dataset, conditioning the training process on camera pose. However, as the generation processes of different views are independent, generated images could correspond to different objects. MVDream [12] treats view as an additional dimension added to 2D generation and utilizes attention layers to attend all views during the generation process, which enables higher correlation across views. Syncdreamer [13] uses a voxel grid as its 3D bottleneck for image-conditioned generation. The explicit 3D bottleneck enhances the view consistency. We will show that the trained model on a synthetic dataset could not generalize well to the diverse generated 2D images from 2D data.

3 Proposed Method

In this section, we will introduce our text-to-multiview generator (Section 3.1, Section 3.2) and the proposed joint inference (Section 3.3, Section 3.4) using the trained text-to-multiview model and text-2d model.

3.1 Text to Multiview Diffusion Model

Figure 2 illustrates our proposed method that generated multiple images from different views using text conditions. Compared with the original Stbale Diffusion [4] that consists of blocks of self-attention layer and text-cross-attention networks, we additionally utilize Epipolar Consistency Blocks conditioned on cameras to correlate features in a 3D-aware manner.

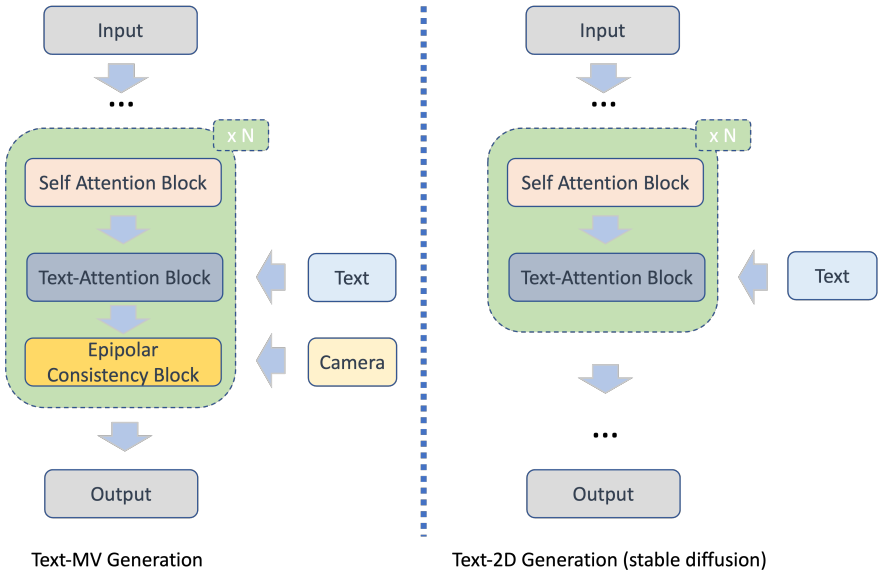

Figure 2: Our Text to Multiview Generator.

Compared with the reverse process of 2D diffusion models that independently sample an image:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

, the formulation of a multiview diffusion model is:

$$\begin{aligned} p_\theta(x_{0:T}^{(1:N)}) &= p(x_T^{(1:N)}) \prod_{t=1}^{T} p_\theta(x_{t-1}^{(1:N)}|x_t^{(1:N)}) \\ &= p(x_T^{(1:N)}) \prod_{t=1}^{T} \prod_{n=1}^{N} p_\theta(x_{t-1}^{(n)}|x_t^{(1:N)}) \end{aligned}$$

, which samples N images with correlations jointly.

In our text-to-multiview diffusion model, the formulation is now:

$$p_\theta(x_{0:T}^{(1:N)}|y, c^{(1:N)}) = p(x_T^{(1:N)}) \prod_{t=1}^{T} \prod_{n=1}^{N} p_\theta(x_{t-1}^{(n)}|x_t^{(1:N)}, y, c^{(1:N)}) \tag{1}$$

, where $c^{1:N}$ represents the set of input cameras and y is the encoded text condition. And the training loss of our model is:

$$l = E_{t, x_o^{(1:N)}, \epsilon^{(1:N)}, y, c^{(1:N)}, n}[||\epsilon^n - \epsilon_\theta^n(x_t^{(1:N)}, t, y, c^n)||^2]$$

.

3.2 Epipolar Consistency Block

Given N features from layer i, denoted as $F_i^{1:N}$, the Epipolar Consistency Block at layer i will correlate them using transformer layers [14] conditioned on input cameras, as illustrated in Fig. 3. Note that we append positional encoding to the transformer inputs.

Epipolar Attention. For each feature at pixel position (x,y) corresponds to the nth camera position, we will sample a ray $R^n(x,y)$ from this location, and M points $p^n_{1:M}(R^n(x,y)) = X^n_{1:M}$ along the ray to form M features, denoted as $V(X^n_{1:M})$. Then we have

$$\bar{V}(X^n_m) = CrossAttention(V(X^n_m), \{V(\pi_c(X^n_m)\})^{1:N})$$

. This process correlates each point along the ray to other features from other views using explicit camera projections by projecting the point X^n_m to other input features and sampling the feature at the project points.

Ray Aggregation. The Epipolar Attention process will produce m features for each ray. To aggregate those features back to the pixel at position (x,y), we will use:

$$\bar{F}^n_i(x,y) = CrossAttention(F^n_i(x,y), \{\bar{V}(X^n_m))\}^{1:M})$$

.

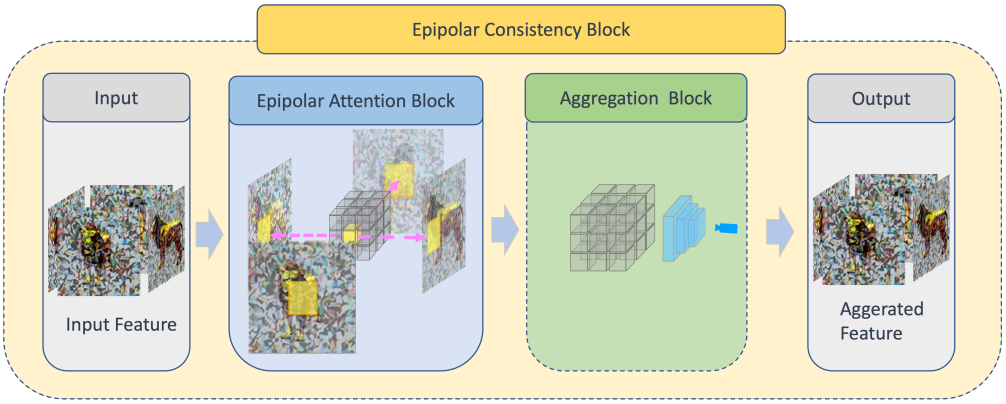

Figure 3: Epipolar Attention Block

3.3 Joint Inference

Having the text-to-multiview generator allows us to sample images that are multiview consistent. However, because the training distribution of multiview images often consists of synthetic images, the sampled results could also look unrealistic. To improve the image quality, we propose to jointly sample images using our trained text-to-multiview with a trained 2D text-to-image diffusion model, as shown in the right part of Fig. 1.

Specifically, we can sample multiview images $\{I_{mv}\}^n$ from the multiview distribution $p_{3d}(x|y)$ using the score from trained multivew diffusion model: $\epsilon_{mv}(x_t|\{x_t\}^N, y, c)$. One can also sample images I_{2d} from 2d image distributions $p_{2d}(x|y)$ using the score from pre-traiend text-image diffusion $\epsilon_{2d}(x_t|y)$.

Here is our mixed score:

$$\epsilon_{mix}(x_t|\{x_t\}^N, y, c) = \lambda \cdot \epsilon_{mv}(x_t|\{x_t\}^N, y, c) + (1-\lambda) \cdot \epsilon_{2d}(x_t|y) \tag{2}$$

, where λ being the weight of multiview score. Using it, we will be effectively sampling from the distribution:

$$\{I_{mix}\}^N \sim p_{mv}(x|y)^{\lambda} \cdot p_{2d}(x|y)^{1-\lambda} \tag{3}$$

$$\epsilon_{consis}(x_t|\{x_t\}^N, y, c) = \epsilon_{mv}(x_t|\{x_t\}^N, y, c) - \epsilon_{mv}(x_t|x_t, y, c)$$

Method	FID (↓)	MVC (↓)
SD + Z123	248.67	0.160**
SD + SyncD	270.57	0.011
MVDream	242.34	0.187
Ours(MVDream)	216.26	0.192*

Table 1: Quantitative comparison of text to multi-view methods. For each method, we evaluate the visual quality of the generated images and well as their multi-view consistency across N prompts.

3.4 Lambda Scheduling

The amount of λ controls the weight of sampling from trained MV distribution. Intuitively, having a large λ value will result in more consistent but less realistic images, and vice-versa for a smaller λ value. We also notice that varying λ during the inference could help to get realistic images while preserving consistency. We empirically find the optimal by setting the λ to a relatively smaller value and gradually increasing it as T decreases.

4 Experiments

4.1 Sampling Multiview Images with Text Condition

Here we show the result (Fig. 4) of sampled images using our method. Note that as our proposed text-to-multiview model Section 3.1 is still training, the shown results are based on the model from [15].

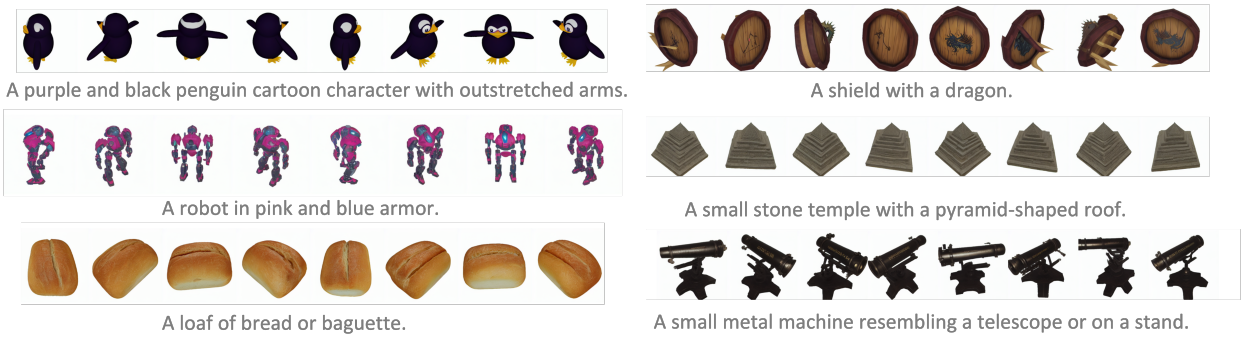

Figure 4: More results of sampled images.

4.2 Effectiveness of Joint Inference

Fig. 5 compares the effectiveness of joint inference. It can be observed that joint inference could improve the image quality.

We also have quantitative results as in Table 1 evaluated using [12], where the FID [16] can be used to evaluate the image quality, and MVC quantifies the multiview consistency by constructing a 3D assets and calculate the reprojection error. We can observe the improvement of the image quality by the 2D model while preserving similar consistency to the multiview model.

5 Conclusion

In conclusion, we have introduced a novel method for generating high-fidelity multi-view images of objects from textual prompts. By leveraging large-scale 2D diffusion models and incorporating 3D bottleneck layers, our approach

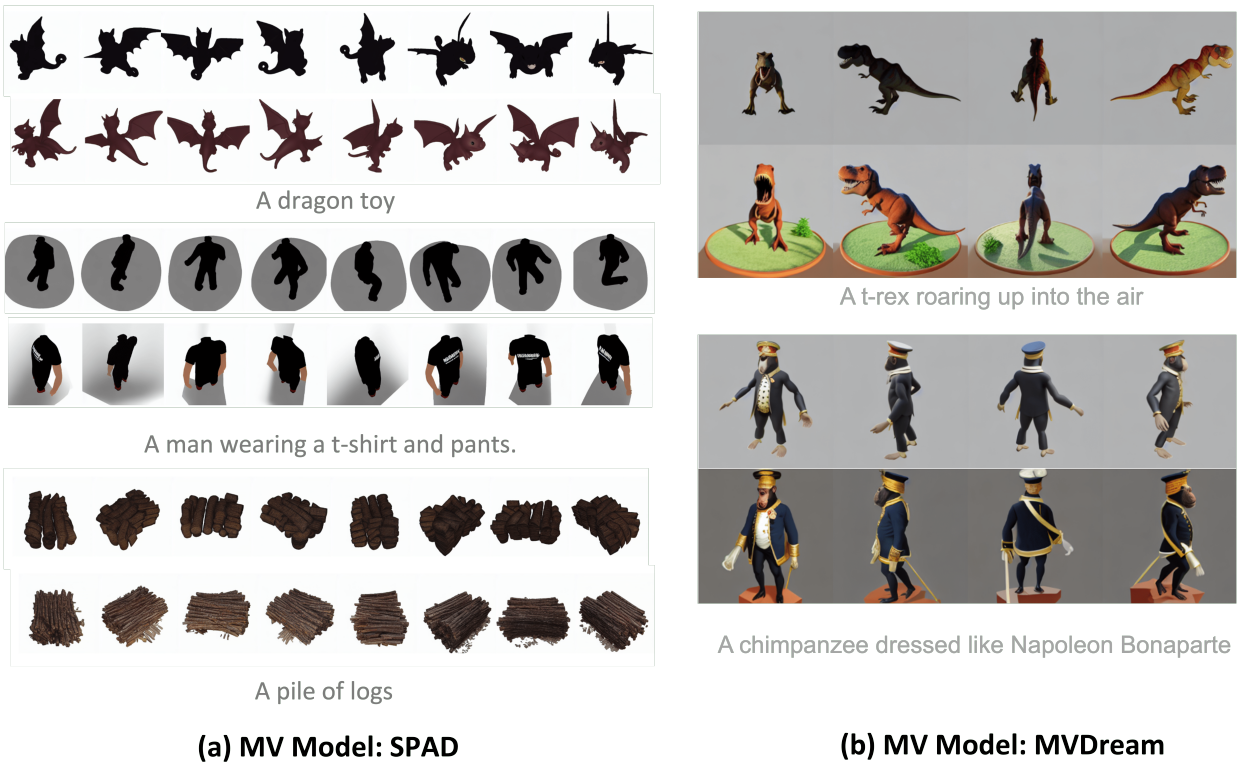

Figure 5: Effectiveness of Joint Inference, the upper row is one with the mv model only, and the lower row is the one with both the 2d and mv model. The mv model in (a) is Spad, and the mv model in (b) is MV dream, YX: cite mv dream

enables the direct generation of 3D-consistent views while benefiting from rich 2D pre-training. Our approach aims to bridge the gap between 2D and 3D generative models, offering a promising avenue for creating multi-view images that are both 3D consistent and high-fidelity.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[3] Ben Poole, Ajay Jain, Jonathan T. Barron, et al. "DreamFusion: Text-to-3D using 2D Diffusion". In: *arXiv* (2022).

[4] Shitong Luo and Wei Hu. *Diffusion Probabilistic Models for 3D Point Cloud Generation*. 2021. arXiv: 2103.01458 [cs.CV].

[5] Linqi Zhou, Yilun Du, and Jiajun Wu. *3D Shape Generation and Completion through Point-Voxel Diffusion*. 2021. arXiv: 2104.03670 [cs.CV].

[6] Haochen Wang, Xiaodan Du, Jiahao Li, et al. "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation". In: *arXiv preprint arXiv:2212.00774* (2022).

[7] Peihao Wang, Dejia Xu, Zhiwen Fan, et al. *Taming Mode Collapse in Score Distillation for Text-to-3D Generation*. 2024. arXiv: 2401.00909 [cs.CV].

[8] Zhengyi Wang, Cheng Lu, Yikai Wang, et al. *ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation*. 2023. arXiv: 2305.16213 [cs.LG].

[9] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, et al. *Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors*. 2023. arXiv: 2306.17843 [cs.CV].

[10] Zhizhuo Zhou and Shubham Tulsiani. "SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction". In: *CVPR*. 2023.

[11] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, et al. *Zero-1-to-3: Zero-shot One Image to 3D Object*. 2023. arXiv: 2303.11328 [cs.CV].

[12] Yichun Shi, Peng Wang, Jianglong Ye, et al. *MVDream: Multi-view Diffusion for 3D Generation*. 2023. arXiv: 2308.16512 [cs.CV].

[13] Yuan Liu, Cheng Lin, Zijiao Zeng, et al. "SyncDreamer: Generating Multiview-consistent Images from a Single-view Image". In: *arXiv preprint arXiv:2309.03453* (2023).

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[15] Yash Kant, Ziyi Wu, Michael Vasilkovsky, et al. *SPAD : Spatially Aware Multiview Diffusers*. 2024. arXiv: 2402.05235 [cs.CV].

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: 1706.08500 [cs.LG].