MVControl: Adding Conditional Control to Multi-view Diffusion for Controllable Text-to-3D Generation

Zhiqi Li[1,2] Yiming Chen[2,3] Lingzhe Zhao[2] Peidong Liu[2,†]
[1]Zhejiang University [2]Westlake University [3]Tongji University
{lizhiqi49, chenyiming, zhaolingzhe, liupeidong}@westlake.edu.cn

Abstract

We introduce MVControl, a novel neural network architecture that enhances existing pre-trained multi-view 2D diffusion models by incorporating additional input conditions, e.g. edge maps. Our approach enables the generation of controllable multi-view images and view-consistent 3D content. To achieve controllable multi-view image generation, we leverage MVDream as our base model, and train a new neural network module as additional plugin for end-to-end task-specific condition learning. To precisely control the shapes and views of generated images, we innovatively propose a new conditioning mechanism that predicts an embedding encapsulating the input spatial and view conditions, which is then injected to the network globally. Once MVControl is trained, score-distillation (SDS) loss based optimization can be performed to generate 3D content, in which process we propose to use a hybrid diffusion prior. The hybrid prior relies on a pre-trained Stable-Diffusion network and our trained MVControl for additional guidance. Extensive experiments demonstrate that our method achieves robust generalization and enables the controllable generation of high-quality 3D content. Our project page is ***https://lizhiqi49.github.io/MVControl/****.*

1. Introduction

Remarkable progress has been made in the field of 2D image generation recently. High fidelity images can be easily generated via input text prompts [52]. Due to the scarcity of 3D training data, the success in text-to-image generation is hardly copied to the text-to-3D domain. Instead of training a large text-to-3D generative model from scratch with large amounts of 3D data, due to the nice properties of diffusion models [21, 61] and differentiable 3D representations [27, 40, 57, 71], recent score distillation optimization (SDS) [48] based methods [12, 31, 39, 48, 65, 66, 73], attempt to distill 3D knowledge from a pre-trained large text-to-image generative model [52] and have achieved impressive results. The representative work is DreamFusion [48], which starts a new paradigm for 3D asset generation.

Following the 2D-to-3D distillation methodology, the techniques are rapidly evolving over the past year. Many works have been proposed to further improve the generation quality by applying multiple optimization stages [12, 31], optimizing the diffusion prior with the 3D representation simultaneously [62, 73], deriving more precise formulation of score distillation algorithm [26, 83], or enhancing the details of whole pipeline [4, 23, 87]. Although the above mentioned efforts can get access to delicate texture, the view consistency of generated 3D content is hard to be achieved, since the 2D diffusion prior is not view-dependent. Hence, there is a series of works hammering at introducing multi-view knowledge to the pre-trained diffusion models [30, 33–35, 58, 59]. Although they can deliver impressive text controlled multiview images and 3D assets, they still cannot achieve fine-grained control over the generated content via an edge map for example, as its counterpart in text-to-image generation, i.e. ControlNet [85]. In this work, we therefore propose MVControl, a multi-view version of ControlNet, to enable controllable text-to-multi-view image generation. Once MVControl is trained, we can exploit it to the score distillation optimization pipeline, so as to achieve controllable text-to-3D content generation via an input condition image, e.g. edge map.

Inspired by 2D ControlNet [85], which works as a plug-in module of Stable-Diffusion [52], we choose a recently released multi-view diffusion network, MVDream [59], as our base model. A control network is then designed to interact with the base model to achieve controllable text-to-multi-view image generation. Similarly to [85], the weights of MVDream is all frozen and we only train the control network. While MVDream is trained with camera poses defined in the absolute world coordinate system, we experimentally find that the relative pose condition with respect to the condition image is more proper for controllable text-to-multi-view generation. However, it conflicts with the definition of the pretrained MVDream network. Furthermore,

† Corresponding author.

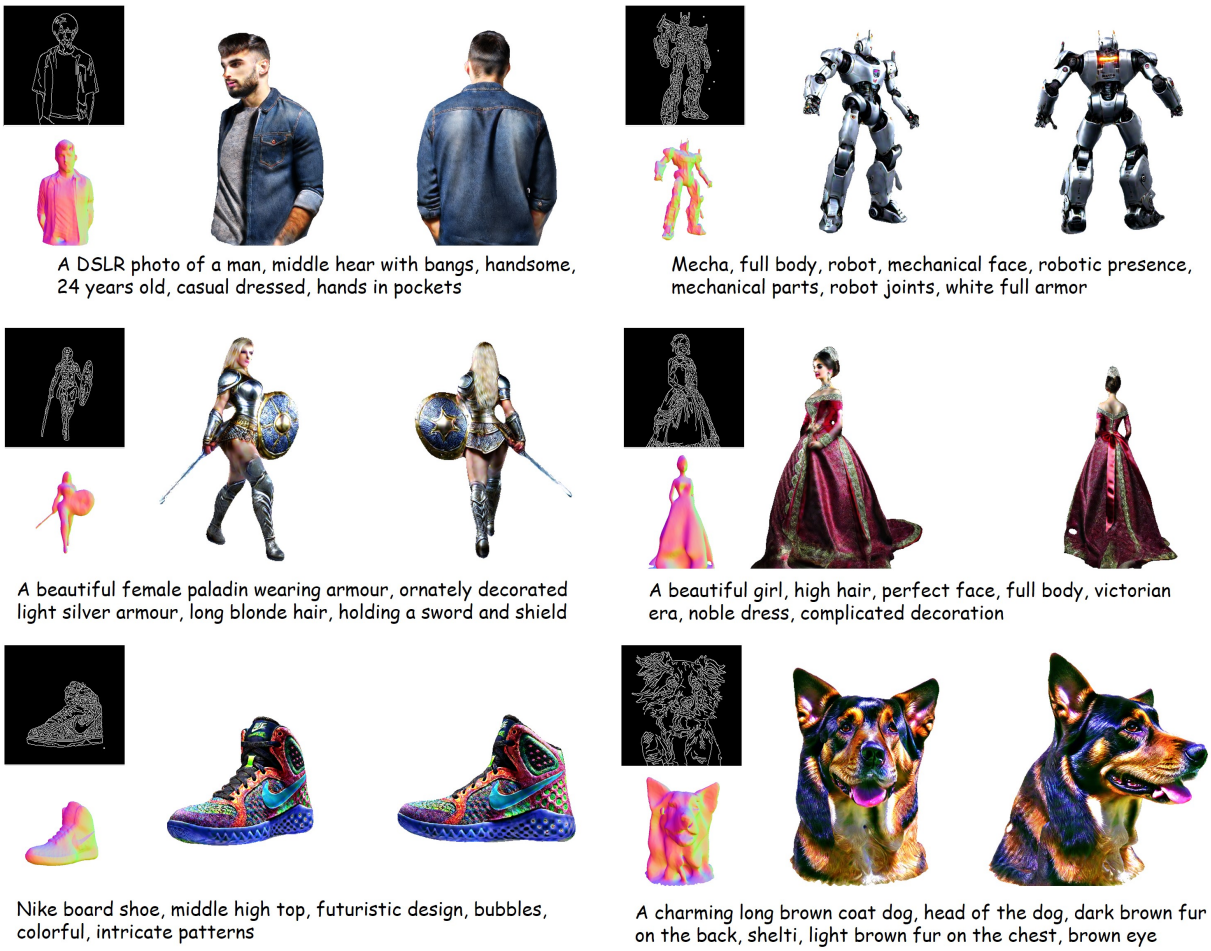


Figure 1. **MVControl:** Given an input text prompt and an edge map, our method is able to generate high-fidelity controllable multi-view images and view-consistent 3D content.

since the conditioning mechanism of 2D ControlNet is designed for single image generation and does not consider the multi-view scenario, view-consistency cannot be easily achieved by directly adopting its control network to interact with the base model. To overcome these issues, we design a simple but effective novel conditioning strategy based on the original ControlNet architecture to achieve controllable text-to-multi-view generation. MVControl is jointly trained on a subset of the large-scale 2D dataset LAION [55] and 3D dataset Objaverse [16] as what [59] does. We only explore to use the edge map as conditional input in this work. However, our network has no restriction to use other type of input conditions, e.g. depth map, sketch image etc. Once MVControl is trained, we can exploit it to provide 3D priors for controllable text-to-3D asset generation. In particular, we employ a hybrid diffusion prior relying on pretrained Stable-Diffusion model and MVControl network. The generation is conducted in a coarse-to-fine manner. After we get a good geometry after the coarse stage, we fix it and only optimize the texture during the fine stage. Our extensive experiments demonstrate that our proposed method can generate fine-grain controlled high-fidelity multi-view images and 3D content via an input condition image as well as textual description.

In summary, our main contributions are as follows.

- We propose a novel network architecture to achieve fine-grain controlled text-to-multi-view image generation;
- Once our network is trained, it can be exploited to serve as a part of hybrid diffusion prior for controllable text-to-3D content generation vis SDS optimization;
- Extensive experimental results demonstrate that our method is able to deliver high-fidelity multi-view image and 3D asset, which can be fine-grain controlled by an input condition image and text prompt;
- Besides being used to generate 3D asset via SDS optimization, we believe our MVControl network could benefit the general 3D vision/graphic community for broad application scenarios.

2. Related Work

We review the related works in 3D generation and classify them into three categories: diffusion-based novel view synthesis, 3D generative models and text-to-3D generation, which are the most related to our method.

Diffusion-based novel view synthesis. The success of text-to-image generation via large diffusion models inspires the development of pose-guided novel view image synthesis. Commonly adopted approach is to condition on a diffusion model by an additional input image and target pose [33, 68, 74, 78]. Different from those methods, Chan et al. recently proposes to learn 3D scene representation from a single or multiple input images and then exploit a diffusion model for target novel view image synthesis [10]. Instead of generate a single target view image, MVDiffusion [67] proposes to generate multi-view consistent images in one feed-forward pass. They build upon a pre-trained diffusion model to have better generalization capability. MVDream [59] also proposes to generate consistent multi-view images from a text prompt recently, by fine-tuning a pre-trained diffusion model with a 3D dataset. They then exploit the trained model to serve as a 3D prior to optimize the 3D representation via Score Distillation Sampling. While prior work can generate impressive novel/multi-view consistent images, fine-grained control over the generated text-to-multi-view images is still difficult to achieve, as what ControlNet [85] has achieved for text-to-image generation. Therefore, we propose a multi-view ControlNet (i.e. MVControl) in this work to further advance diffusion-based multi-view image generation.

3D generative models. Current 3D generative models usually exploit existing 3D datasets to train generative models with different 3D representations. Commonly used 3D representations are volumetric representation [8, 17, 76, 79], triangular mesh [6, 15, 18, 64, 82], point cloud [1, 2, 45, 60, 81] as well as the recent implicit neural representation [9, 13, 38, 46, 56, 72]. Various generative modeling techniques have also been explored to 3D data as their success in 2D image synthesis, which range from variational auto-encoder [5, 18, 64, 77], generative adversarial network [1, 9, 15, 44, 76, 79], flow-based method [2, 3, 29, 80], and the recent popular diffusion based method [14, 24, 36, 42, 84, 86]. Different from image generative modeling, which has large amount of training images, those 3D generative methods usually lack sufficient 3D assets for training. They are usually limited to category-specific dataset, e.g. shapeNet [11]. Although Objaverse [16] released a million-scale 3D asset dataset recently, its size is still infancy compared to the 2D training data [55] used by modern generative models for image synthesis. Due to the lack of large amount of training data, they usually cannot generate arbitrary type of objects to satisfy the requirements of end consumers. Instead of relying on large amount of 3D data as those methods, we propose to exploit a pre-trained large image generative model to distill 3D knowledge for controlled text-to-3D generation.

Text-to-3D generation. Due to the scarcity of 3D data, researchers attempt to distill knowledge for 3D generation from pre-trained large image models. The initial attempt was to exploit a pre-trained CLIP [49] model to align the input text prompt and rendered images for the supervision of 3D object generation [25, 41, 54]. However, the generated 3D objects usually tend to be less realistic due to that CLIP [49] can only offer high-level semantic guidance. With the advancement of large text-to-image diffusion models [53], DreamFusion [48] demonstrates the potential to generate more realistic 3D objects via knowledge distillation. Follow-up works continue to push the performance to generate photo-realistic 3D objects that closely match the provided text prompts [12, 23, 30, 31, 50, 65, 70, 73, 87]. The main insights of those methods are usually to develop more advanced score distillation loss or better optimization strategy etc. to further improve the generation quality. Although those methods generate high-fidelity 3D shapes via text description, fine-grained control on the text-to-3D shape generation is still lacking. We therefore propose to exploit our pre-trained MVControl network to provide a 3D prior for controllable text-to-3D generation.

3. Method

We first review relevant methods, including diffusion model [21, 61], MVDream [59], ControlNet [85] and score distillation sampling [48] in Section 3.1. Then, we analyze the strategy of introducing additional spatial conditioning to MVDream by training a multi-view ControlNet in Section 3.2. Finally in Section 3.3, based on the trained multi-view ControlNet, we propose the realization of controllable 3D content generation using SDS loss with hybrid diffusion priors as guidance.

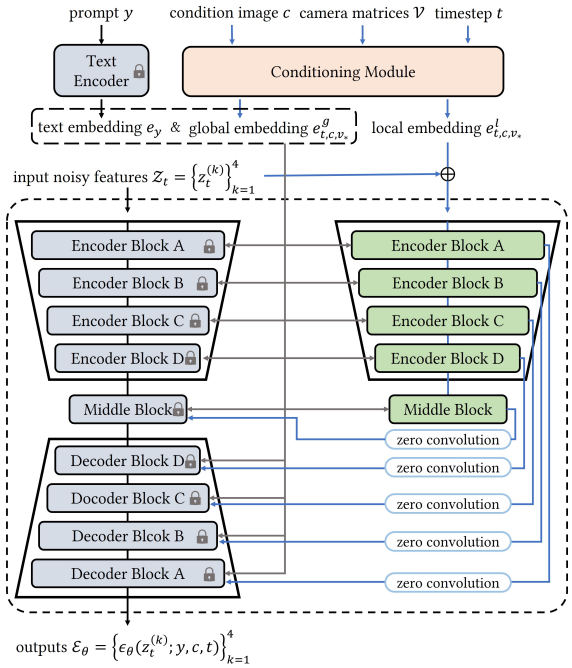
3.1. Preliminary

Diffusion model. Diffusion model predicts the score function $\nabla_{\mathbf{x_t}} \log p_{data}(\mathbf{x_t})$ in the data space under different noise level $t \sim \mathcal{U}(0, T)$, so as to guide the sampling process to progressively denoise a pure noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a clean data x_0. To learn the denoising score, noises at different scales are added to x_0 with pre-defined noise schedule according to:

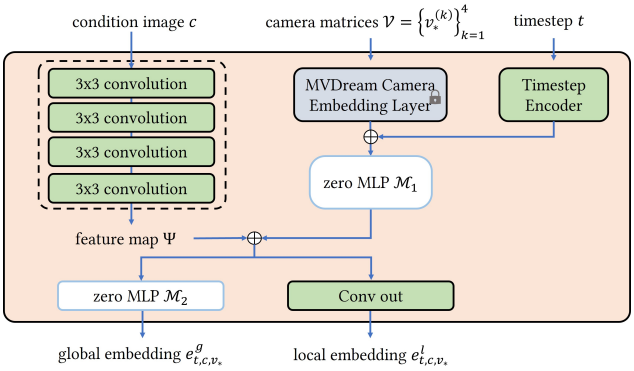
$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\alpha_t \in (0, 1)$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The diffusion model parameterized by ϕ can then be trained by minimizing the noise reconstruction loss:

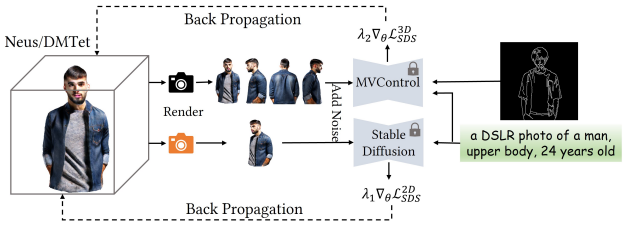
$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,\epsilon}[\|\epsilon_\phi(x_t, t) - \epsilon\|_2^2]. \tag{2}$$



(a) Framework of MVControl



(b) Architecture of conditioning module



(c) Controllable 3D generation with hybrid diffusion prior

Figure 2. **Overview of proposed method.** (a) MVControl consists of a frozen multi-view diffusion model and a trainable MVControl. (b) Our model takes care of all input conditions to control the generation process both locally and globally through a conditioning module. (c) Once MVControl is trained, we can exploit it to serve a hybrid diffusion prior for controllable text-to-3D content generation via SDS optimization procedure.

Once the model is trained, it can be iteratively applied to denoise a sampled noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to obtain a clean data. The readers can refer to [21, 61] for more detailed illustration about diffusion models.

MVDream. MVDream [59] is a recently proposed text-to-multi-view diffusion model. It is fine-tuned based on a large-scale pre-trained image diffusion model, i.e. Stable-Diffusion (SD) [52], via text labled multi-view image data. While SD generates one image for a sampling process, MVDream generates four images simultaneously. The four images corresponding to four consistent views of an object, and are 90 degrees apart from each other in the longitude direction with the same elevation. To ensure the interaction of features among different views for view-consistent generation, MVDream replaces self attention layers of SD with cross-view attentions, which concatenate the patches of all views before attention computation. To enable view conditioning, they further exploit a MLP (Multi-layer Perceptron) network to transform camera extrinsic parameters into an embedding vector, which are then concatenated with the timestep embedding to be injected to the whole model.

ControlNet. ControlNet [85] enables pretrained large diffusion models to support additional input conditions (e.g. canny edges, sketches, depth maps, etc.) beside the text prompts. It is constructed by directly copying the structure and weights of SD's encoder blocks and mid block, and adding zero convolution layers to connect it with the pre-trained SD. With those connections, the feature map computed by each inner layer of ControlNet can then be injected to its corresponding symmetric layer in SD's UNet decoder, so as to control the sampling process of SD once it is trained.

Score distillation sampling. Score distillation sampling (SDS) [31, 48] leverages pretrained text-to-image diffusion model as prior to guide text-conditioned 3D asset generation. In particular, given a pre-trained diffusion model ϵ_ϕ, SDS optimizes the parameters θ of a differentiable 3D representation, e.g. neural radiance field, using the gradient of the loss $\mathcal{L}_{SDS}$ with respect to θ:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x}) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi - \epsilon)\frac{\partial z_t}{\partial \theta}], \tag{3}$$

where $\mathbf{x} = g(\theta, c)$ is an image rendered by g under a camera pose c, $w(t)$ is a weighting function depending on the timestep t and z_t is the noisy image input to diffusion model by adding Gaussian noise ϵ to $\mathbf{x}$ corresponding to the t-th timestep according to Eq. (1). The main insight is to enforce the rendered image of the learnable 3D representation to satisfy the distribution of the pretrained diffusion model. In practice, the values of timestep t and Gaussian noise ϵ

are randomly sampled at every optimization step.

3.2. Multi-view ControlNet

Inspired by ControlNet in controlled text-to-image generation and recently released text-to-multi-view image diffusion model (e.g. MVDream), we aim to design a multi-view version of ControlNet (i.e. MVControl) to achieve controlled text-to-multi-view generation. As shown in Fig. 2a, we follow similar architecture style as ControlNet, i.e. a locked pre-trained MVDream and a trainable control network. The main insight is to preserve the learned prior knowledge of MVDream, while train the control network to learn the inductive bias with small amount of data. The control network consists of a conditioning module and a copy of the encoder network of MVDream. Our main contribution lies at the conditioning module and we will detail it as follows.

The conditioning module (Fig. 2b) receives the condition image c, four camera matrices $\mathcal{V}_* \in \mathbb{R}^{4\times4\times4}$ and timestep t as input, and outputs four local control embeddings e^l_{t,c,v_*} and global control embeddings e^g_{t,c,v_*}. The local embedding is then added with the input noisy latent features $\mathcal{Z}_t \in \mathbb{R}^{4\times C\times H\times W}$ as the input to the control network, and the global embedding e^g_{t,c,v_*} is injected to each layer of MVDream and MVControl to globally control generation.

The condition image c (i.e. edge map, depth map etc.) is processed by four convolution layers to obtain a feature map Ψ. Instead of using the absolute camera pose matrices embedding of MVDream, we move the embedding into the conditioning module. To make the network better understand the spatial relationship among different views, the relative camera poses with respect to the condition image are used for the camera matrices $\mathcal{V}_*$. The experimental results also validate the effectiveness of the design. The camera matrices embedding is combined with the timestep embedding, and is then mapped to have the same dimension as the feature map Ψ by a zero-initialized module $\mathcal{M}_1$. The sum of these two parts is projected to the local embedding e^l_{t,c,v_*} through a convolution layer.

While MVDream is pretrained with absolute camera poses, the conditioning module exploit relative poses as input. We experimentally find that the network hardly converges due to the mismatch of both coordinate frames. We therefore exploit an additional network $\mathcal{M}_2$ to learn the transformation and output a global embedding e^g_{t,c,v_*} to replace the original camera matrix embedding of MVDream and add on timestep embeddings of both MVDream and MVControl part so that inject semantical and view-dependent features globally.

3.3. Controllable 3D Content Generation

Once MVControl is trained, it can be utilized for controllable text-to-3D content generation via SDS optimization [48]. We adopt a hybrid diffusion prior from both Stable-Diffusion and MVControl, to better guide the 3D generation. MVControl provides a strong consistent geometry guidance over four canonical views of the optimizing 3D object, while Stable-Diffusion provides fine geometry and texture sculpting at the other randomly sampled views. As is shown in Fig. 2c, the hybrid SDS gradient can be calculated as:

$$\nabla_\theta \mathcal{L}^{hybrid}_{SDS} = \lambda_1 \nabla_\theta \mathcal{L}^{2D}_{SDS} + \lambda_2 \nabla_\theta \mathcal{L}^{3D}_{SDS}, \tag{4}$$

where λ_1 and λ_2 are the strength of 2D and 3D prior respectively. The optimization procedure consists of two stages: a coarse stage for initial model generation and a fine stage for texture refinement.

During the coarse stage, we exploit a coarse neural surface, i.e. NeuS [71], to represent the 3D asset. SDS loss is then computed to optimize the neural 3D representation. To encourage the smoothness of the surface, we also exploit eikonal loss [71] to regularize the training process. To obtain a high-fidelity 3D asset, we extract the coarse neural surface to a hybrid mesh representation via DMTet [57]. The texture is further refined by SDS loss with fixed geometry. We use the conventional SDS loss [48] for coarse stage generation, and we use the recently proposed noise-free score distillation [26] for our fine stage, which delivers similar performance with conventional SDS but can use normal scale for classifier-free guidance (CFG) [22].

4. Experiments

4.1. Implementation Details

Data Preparation. We use the multi-view renderings of the public large 3D dataset, Objaverse [16] to train our MVControl. Firstly we clean the dataset by excluding all samples with CLIP-score lower than 22 based on the labeling of [63] and finally we have about 400k samples left. Instead of using the name and tags of the 3D assets, we refer to the captions of [37] as text descriptions of our kept objects. For each object, we first normalize its scene bounding box to unit cube at the world center, and then randomly sample camera distance between [1.4, 1.6], fov between [40, 60] and elevation between [0, 30]. Finally, we randomly sample from 32 uniformly distributed azimuths as starting point to sample 4 orthogonal views each time.

Training details of MVControl network. We exploit the weights of pretrained MVDream and ControlNet to initialize our network. All the connections between the locked and trainable networks are initialized with zero. Our network is then trained with both 2D and 3D datasets. In particular, We sample images from the AES v2 subset of LAION [55] with a 30% probability for training, such that the network will not lose its learned 2D image priors. We

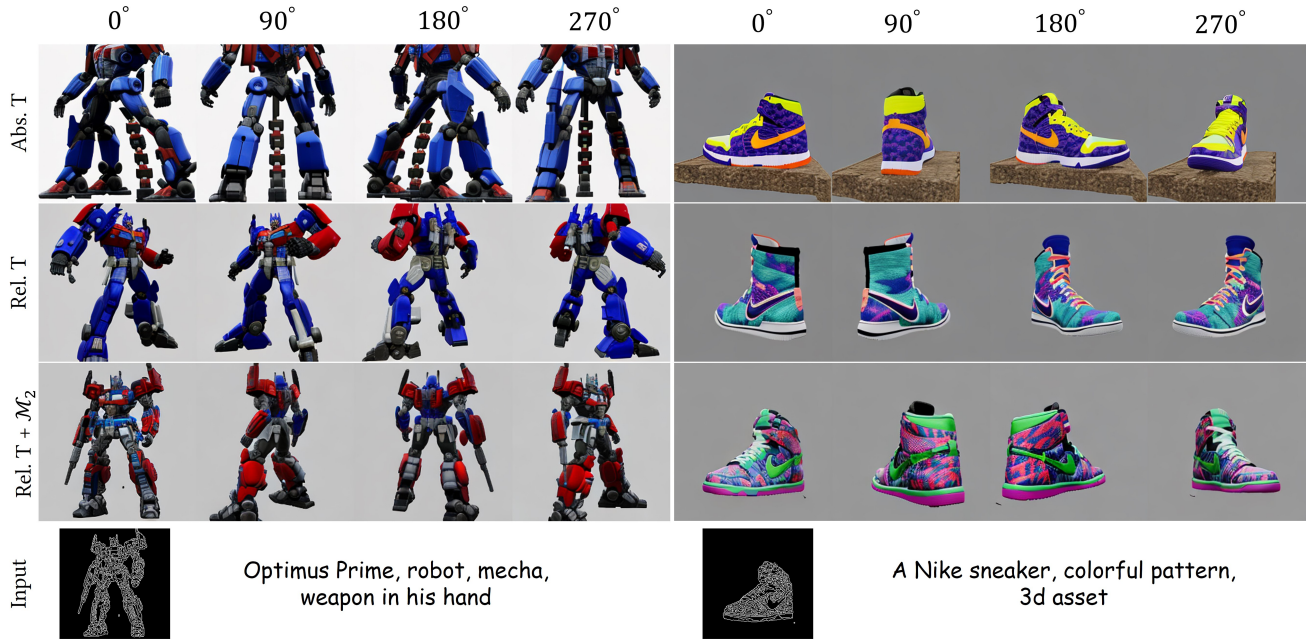


Figure 3. **The necessity on the design of camera pose condition.** It demonstrates that only the complete conditioning module can properly control the generation of the posed images, which is defined relative to the input condition edge image. **Abs. T** denotes the conditioning module does not accept any pose condition as input, the whole network relies on the absolute pose condition of MVDream for pose control; **Rel. T** denotes the MLP network $\mathcal{M}_2$ is removed from the conditioning module, which is used to bridge the relative pose condition and the base model, which is pretrained with absolute pose condition; **Rel. T +** $\mathcal{M}_2$ denotes the complete module.

then sample from our prepared 3D/mult-view image dataset with a 70% probability to learn the 3D knowledge. We exploit the Canny edge map of a sampled image as the conditioning image for training. Other options, e.g. depth image, sketch etc., can also be exploited without any modification of the method.

The training images have a resolution at 256x256 pixels, and batch size is chosen as 2560 images. The model is fine-tuned for 50,000 steps under a conservative learning rate, $4e \times 10^{-5}$, on 8 Nvidia Tesla A100 GPUs with AdamW optimizer [28]. Following [85], we also drop the text prompt of one sample as empty with 50% chance for classifier-free training, such that the model can be trained to better understand the semantics of input condition images.

3D Content Generation. We choose Stable-Diffusion-v2.1-base [52] as the 2D part of our hybrid diffusion prior. In the coarse stage, we use the NeuS 3D representation and Instant-NGP [43] as its implementation for training efficiency. The neural surface is optimized for 8,000 steps with AdamW optimizer. Its rendering resolution is increased from 64×64 to 256×256 after 5,000 steps. We also employ timestep annealing strategy. The timestep sampling range is gradually decreased from $(0.98, 0.98)$ to $(0.5, 0.02)$ over the optimization process. The CFG scale for 2D and 3D part of our hybrid diffusion prior is empirically chosen as 10 and 50 respectively. For the computation of the SDS loss of MVControl, we use the x_0-reconstruction formulation proposed in [59] to use CFG rescale trick [32], and the rescale factor is set as 0.5. In the fine stage, we extract the neural surface to 128 grid DMTet and the rendered image resolution is set to be 512×512 pixels. The CFG scale for 2D and 3D part is 7.5 and 10 respectively. For both stages, we set the strength of 2D diffusion guidance as 1, and that of MVControl guidance as 0.1 or 0.2 empirically.

4.2. Ablation Studies

The necessity on the design of camera pose condition. We train our network under three different settings: 1) we exploit the absolute pose condition (i.e. Abs. T) of MVDream [59], and only have the local embedding of the input condition image as the output from the conditioning module; 2) we remove the zero MLP $\mathcal{M}_2$ (i.e. Rel. T), which is used to bridge the relative pose embedding of the conditioning module and that of MVDream base model; 3) complete conditioning module (i.e. Rel. T+$\mathcal{M}_2$). The experimental results are shown in Fig. 3, it demonstrates that only the complete conditioning module can have good control over the pose of the generated images. The pose is defined as relative pose with respect to the input condition image.

Coarse-to-fine optimization strategy for 3D generation. We study the benefit to exploit the coarse-to-fine strategy for text-to-3D generation via SDS optimization. The ex-

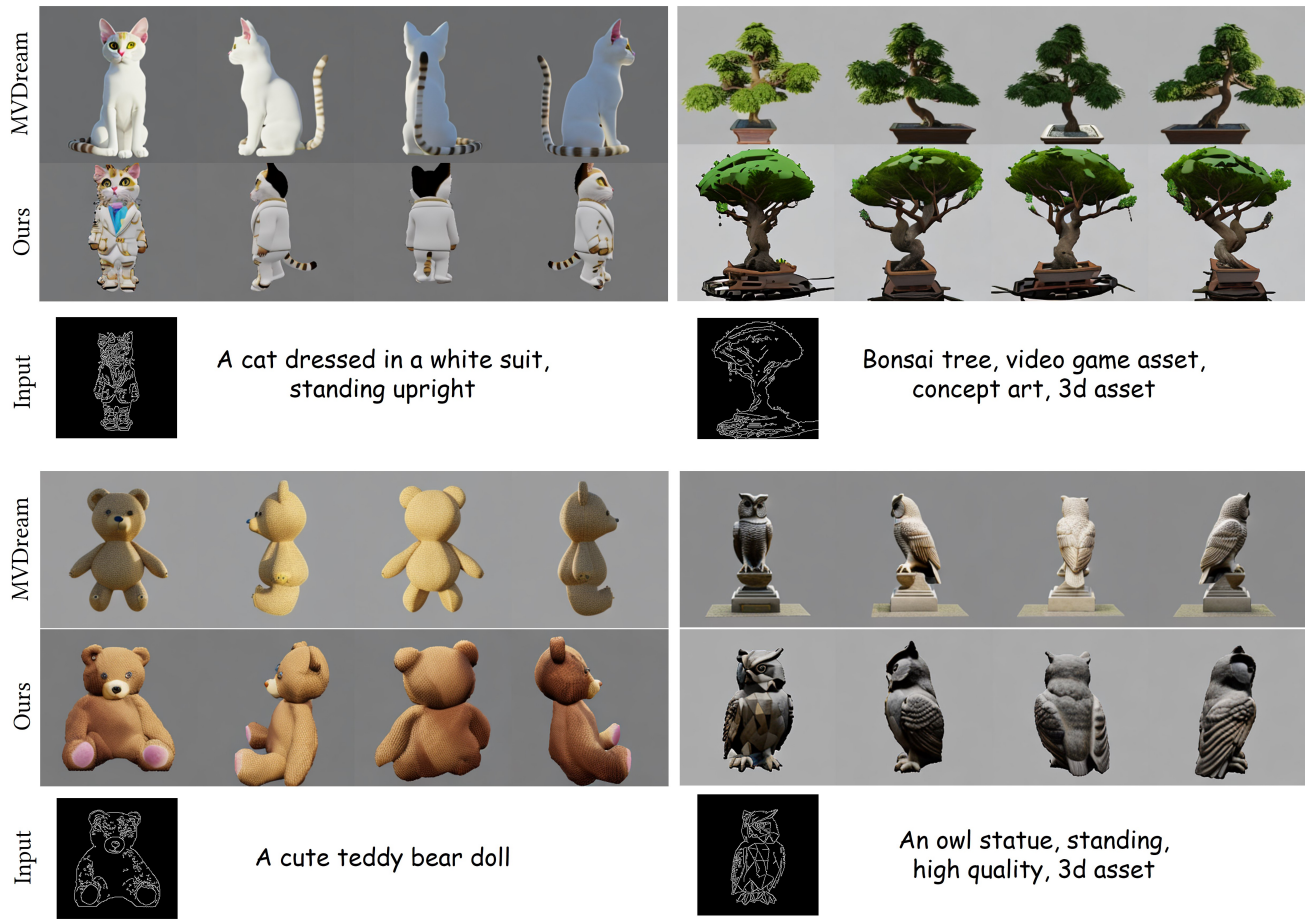


Figure 4. **Controllable multi-view image generation.** It demonstrates that our method is able to generate controllable high-fidelity multi-view images, satisfying both the input condition image and text prompt.

perimental results are shown in Fig. 5, i.e. *Ours (Coarse Stage)* and *Ours (Full Stage)*. It demonstrates that we can obtain a good geometry of the generated 3D asset during the coarse stage. However, the texture lacks details and is over-smoothed. After we converted the generated 3D asset to deformable mesh [57] and optimize the texture only at a higher resolution, the 3D asset looks more photo-realistic and contains richer texture details. It demonstrates the necessity of the fine texture optimization stage.

4.3. Controllable Multi-view Image Generation

We compare the performance of our network against its base model, i.e. MVDream. The experimental results present in Fig. 4 demonstrate both networks are able to generate multi-view consistent images, which satisfy the input text prompt description. However, since MVDream cannot accept additional condition image, they are unable to control the shapes of the generated images. In contrary, our method is able to generate controllable multi-view consistent images with Canny edge image as additional input.

4.4. Controllable 3D Content Generation

We compare against prior state-of-the-art text-to-3D generation methods, i.e. DreamFusion [48], Fantasia3D [12], ProlificDreamer [73] and MVDream [59]. The results shown in Fig. 5 demonstrate that prior works can generate reasonable 3D assets, but suffer from the Janus problem and lack proper control via a condition image. We also compare against an image-based method, i.e. Zero123 [33]. For proper comparison, we render the reference image of our generated asset after the fine stage for Zero123 [33]. The results demonstrate that it can generate proper 3D asset satisfying the edge map. However, it lacks details at the back of the 3D assets. We use their default setup implemented in the threestudio repository. It demonstrates that our method is able to generate controllable high-fidelity 3D assets over prior methods.

5. Conclusion

We present a novel network architecture for controllable text-to-multiview image generation. Our network exploits

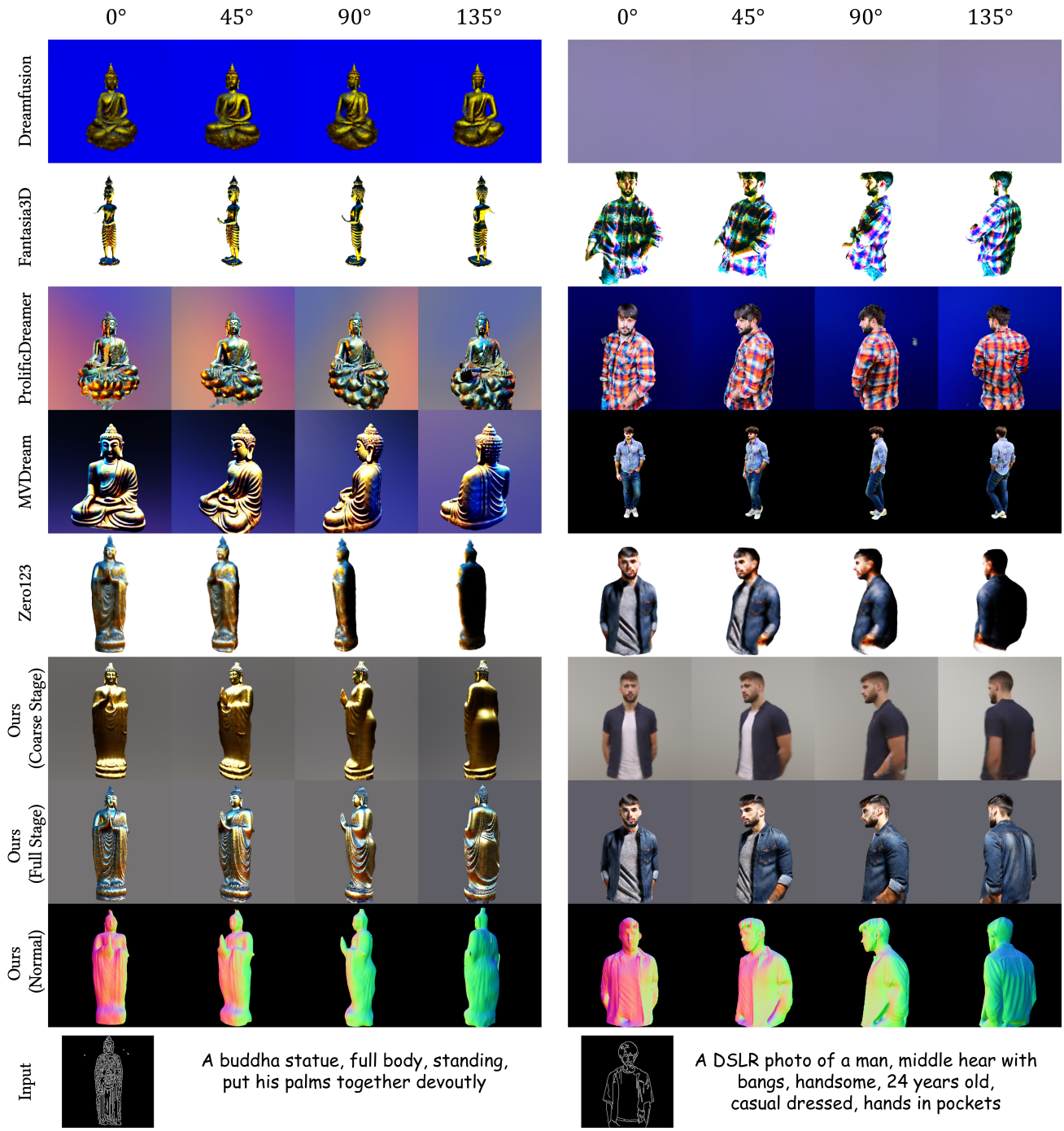


Figure 5. **Controllable text-to-3D content generation.** It demonstrates that our method is able to generate controllable high-fidelity 3D assets over prior methods. We use the rendered reference image of our final model for Zero123 to have proper comparison.

a pretrained image diffusion network as base model. A novel trainable control network is proposed to interact with the base model for controllable multiview image generation. Once it is trained, our network can provide 3D prior for controllable text-to-3D generation via SDS optimization. The experimental results demonstrate that our method can generate controllable high-fidelity text-to-multiview images and text-to-3D assets. Besides being used for controllable 3D generation via SDS optimization, we believe our network would be applicable for more broad 3D vision/graphic application scenarios in future.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 3

[2] Francesc Moreno-Noguer Albert Pumarola, Stefan Popov and Vittorio Ferrari. C-Flow: Conditional generative flow models for images and 3D point clouds. In *CVPR*, 2020. 3

[3] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J. Cashman. FLAG: Flow-based 3D Avatar generation from sparse observations. In *CVPR*, 2022. 3

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

[5] Elena Balashova, Vivek Singh, Jiangping Wang, Brian Teixeira, Terrence Chen, and Thomas Funkhouser. Structure-aware shape synthesis. In *2018 International Conference on 3D Vision (3DV)*, pages 140–149. IEEE, 2018. 3

[6] Heli Ben-Hamu, Haggai Maron, Itay Kezurer, Gal Avineri, and Yaron Lipman. Multi-chart generative surface modeling. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 3

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam. 13

[8] Andrew Brock, Theodore Lim, J.M. Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. In *NeurIPS*, 2016. 3

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[10] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3D aware diffusion models. In *ICCV*, 2023. 3

[11] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[12] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 1, 3, 7

[13] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 3

[14] Gene Chou, Yuval Bahat, and Felix Heide. DiffusionSDF: Conditional generative modeling of signed distance functions. In *ICCV*, 2023. 3

[15] Thomas Hofmann Dario Pavllo, Jonas Kohler and Aurelien Lucchi. Learning generative models of textured 3D meshes from real-world images. In *ICCV*, 2021. 3

[16] 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, 3, 5

[17] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 3

[18] Lin Gao, Jie Yang, Tong Wu, Yujie Yuan, Hongbo Fu, Yukun Lai, and Hao Zhang. SDM-Net: Deep generative network for structured deformable mesh. In *ACM TOG*, 2019. 3

[19] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022. 13

[20] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 14

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 4

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[23] 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. *arXiv preprint arXiv:2306.12422*, 2023. 1, 3

[24] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[25] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3

[26] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 1, 5, 13

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 1

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete point flow networks for efficient point cloud generation. In *ECCV*, 2020. 3

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

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

[32] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 6, 13

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

[34] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1

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

[36] Shitong Luo and Wei Hu. Diffusion Probabilistic Models for 3D Point Cloud Generation. In *CVPR*, 2021. 3

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

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastuan Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 3

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

[40] 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. 1

[41] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 3

[42] Norman Muller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Niebner. DiffRF: Rendering-Guided 3D Radiance Field Diffusion. In *CVPR*, 2023. 3

[43] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 6

[44] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems*, 33:6767–6778, 2020. 3

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

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 13

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 3, 4, 5, 7

[49] 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. 3

[50] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. In *ICCV*, 2023. 3

[51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 13

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 4, 6

[53] 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. 3

[54] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chinyi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[55] 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. 2, 3, 5

[56] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In *NeurIPS*, 2022. 3

[57] 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. 1, 5, 7, 13

[58] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1

[59] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 3, 4, 6, 7

[60] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 3

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 3, 4

[62] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 1

[63] Qinghong Sun, Yangguang Li, ZeXiang Liu, Xiaoshui Huang, Fenggang Liu, Xihui Liu, Wanli Ouyang, and Jing Shao. Unig3d: A unified 3d object generation dataset. *arXiv preprint arXiv:2306.10730*, 2023. 5

[64] Qingyang Tan, Lin Gao, Yukun Lai, and Shihong Xia. Variational autoencoders for deforming 3D mesh models. In *CVPR*, 2018. 3

[65] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 3

[66] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 1

[67] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. MVDiffusion: enabling holistic multi-view image generation with correspondence aware diffusion. In *NeurIPS*, 2023. 3

[68] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-bin Huang, and Johannes Kopf. Consistent view synthesis with pose guided diffusion models. In *CVPR*, 2023. 3

[69] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 13

[70] 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*, 2023. 3

[71] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 5, 13

[72] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 3

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

[74] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel View Synthesis with Diffusion Models. In *ICLR*, 2023. 3

[75] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 13

[76] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 3

[77] Zhijie Wu, Xiang Wang, Di Lin, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. SAGNet: Structure aware generative network for 3D shape modeling. In *ACM TOG*, 2019. 3

[78] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3D-aware image generation using 2D diffusion models. In *ICCV*, 2023. 3

[79] Pieter Peers Xiao Li, Yue Dong and Xin Tong. Synthesizing 3D shapes from silhouette image collections using multi-projection generative adversarial networks. In *CVPR*, 2019. 3

[80] Guandao Yang, Xun Huang, Zekun Hao, Mingyu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows. In *ICCV*, 2019. 3

[81] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. 3

[82] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. CLIP-Actor: text driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. 3

[83] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023. 1

[84] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. LION: latent point diffusion models for 3D shape generation. In *NeurIPS*, 2022. 3

[85] 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. 1, 3, 4, 6

[86] Linqi Zhou, Yilun Du, and Jiajun Wu. 3D Shape Generation and Completion through Point-Voxel Diffusion. In *ICCV*, 2021. 3

[87] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 1, 3

Appendix

A. Introduction

In this supplementary material, we provide more details on the method of controlled text-to-3D generation, and implementations of our work and the compared baselines. Furthermore, we also train a new MVControl network which is conditioned on the dense depth map, for depth controlled text-to-multiview image generation. Additional qualitative results on both the depth map controlled text-to-multiview image generation, and edge map controlled text-to-3D generation are presented.

B. Controlled Text-to-3D Generation

Controllable 3D content generation is realized through a coarse-to-fine optimization process. In particular, we first optimize a coarse neural surface [71], and then conduct texture refinement in the fine optimization stage. The coarse geometry is transformed to a deformable mesh [57] for the texture refinement at the fine stage. As for the hybrid diffusion prior, the pretrained MVControl network works as a strong consistent geometry guidance at four canonical views of the 3D object, and Stable-Diffusion network provides fine geometry and texture sculpting at the other randomly sampled views. In the following part, we denote the four canonical views with $\mathcal{V}_*$ and the images rendered under those views as $\mathcal{X}_* \in \mathbb{R}^{4\times H\times W\times C}$.

B.1. Coarse Geometry Stage

At this stage, we aim to generate a 3D model whose geometry is consistent with the input condition image. While our MVControl can already provide consistent geometry constrains from four canonical views, it's still not sufficient to recover a plausible geometry from the four sparse views only. Hence, we propose to incorporate the 2D diffusion model SD to provide a semantic guidance under those views other than the four canonical views, and so as to sculpt the geometry to satisfy the distribution described by the condition image.

Specifically, we denote a differentiable renderer with $g(\cdot)$ and the parameters of 3D representation as θ. We render the images $\mathcal{X}_* = g(\theta, \mathcal{V}_*)$ under four canonical views and image $x_r = g(\theta, v_r)$ under a randomly sampled view v_r. Then the gradient of hybrid SDS loss can be computed as:

$$\nabla_\theta \mathcal{L}^{hybrid}_{SDS} = \lambda_1 \nabla_\theta \mathcal{L}^{2D}_{SDS} + \lambda_2 \nabla_\theta \mathcal{L}^{3D}_{SDS}, \tag{5}$$

where $\nabla_\theta \mathcal{L}^{2D}_{SDS}$ is the SDS gradient distillated from SD taking the rendering x_r as input, and $\nabla_\theta \mathcal{L}^{3D}_{SDS}$ is that distillated from MVControl network with $\mathcal{X}_*$ and $\mathcal{V}_*$ as input. λ_1 and λ_2 are two hyperparameters and chosen empirically.

While classifier-free guidance (CFG) has become a necessary technique when doing diffusion sampling, we should consider the CFG scale for each of our diffusion priors. In order to enforce the optimization process to align with the distribution defined by MVControl, we apply a large CFG scale for $\nabla_\theta \mathcal{L}^{3D}_{SDS}$. To avoid the discrepancy between the guidance from both SD (2D) and MVControl (3D), we choose to use a relatively small CFG scale for $\nabla_\theta \mathcal{L}^{2D}_{SDS}$. Here, following MVDream, we compute $\nabla_\theta \mathcal{L}^{MV}_{SDS}$ through x_0-reconstruction formulation to alleviate the color saturation from large CFG scale by applying CFG rescale trick [32]:

$$\nabla_\theta \mathcal{L}^{3D}_{SDS}(\psi, \mathcal{X}_*, \mathcal{V}_*) = \mathbb{E}_{t,\epsilon}[\|\mathcal{X}_* - \hat{\mathcal{X}}_0\|_2^2], \tag{6}$$

where $\hat{\mathcal{X}}_0$ is the estimated clean images of the four noisy input from $\epsilon_\psi(z_t(\mathcal{X}_*); y, t, \mathcal{V}_*)$ and its gradient is detached from the optimization step. Regarding the computation of $\nabla_\theta \mathcal{L}^{2D}_{SDS}$, we refer to the normal SDS calculation since it uses a small CFG scale. We also exploit the Eikonal loss proposed by [71] to regularize the SDF values to be more plausible.

B.2. Fine Texture Stage

In this stage, the coarse geometry is converted to a deformable mesh [57] for further texture refinement under high rendering resolution with geometry fixed. For the computation of score distillation gradients, we refer to the recently released Noise-free Score Distillation (NFSD) technique [26]. The only difference between our implementation and theirs is that we replace the null prompt $\oslash$ with negative prompt p_{neg} in the δ_C part, with which we observe a quality improvement. For more details, please refer to [26]. Empirically, our method achieves similar results with simple SDS gradient which usually requires large CFG scale, and we choose this strategy due to its normal CFG scales.

C. Additional Implementation Details

For training data creation, we use Blender [7] to render images from Objaverse objects. The rendering scripts are based on a public repository. The implementations of our model and training code are based on Pytorch [47] and heavily rely on the public projects [19, 69, 75] by Hugging Face Organization. Our MVControl networks conditioned on edge map and depth map respectively are fine-tuned from public ControlNet checkpoints. And for depth prediction on the training images, we use off-the-shelf depth estimation network [51]. The implementations of our 3D generation part and all compared 3D generation baselines are

https://github.com/allenai/objaverse-rendering

https://huggingface.co/thibaud/controlnet-sd21-canny-diffusers

https://huggingface.co/thibaud/controlnet-sd21-depth-diffusers

Figure 6. Additional 2D results of depth-conditioned MVControl.

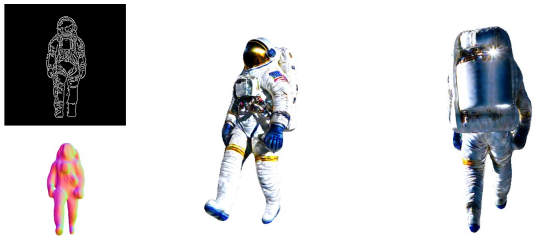
based on the ThreeStudio project [20] except MVDream, which is from their official implementation. All the experiments of baselines are conducted under their default setup.

D. Additional Qualitative Results

While we mainly focus on using canny edge maps as the additional condition for MVControl, we also trained another depth-conditioned version of MVControl under the same training settings. Its multi-view image generation results are shown in Fig. 6. The figure shows that our method is also able to generate high-fidelity multi-view images with depth map as the additional conditioning input together with the text prompt, which demonstrates that our MVControl has the potential to be generalized to different types of conditions.

We also provide more qualitative results of controlled text-to-3D generation via MVControl in Fig. 7 and Fig. 8. The results demonstrate that our method has the capacity to generate high-fidelity view-consistent 3D assets with high-quality texture, which can be controlled by both the text prompt and additional control input (e.g. edge map). The reader can refer to our project page for more results.

https://github.com/bytedance/MVDream-threestudio



An astronaut, full body, extremely detailed spacesuit, dynamic shot, walking pose, best quality

Photorealistic photo of the male witcher, upper body, middle aged, long white hair, Medieval black light armor, a sword on the back, perfect face, scar, necklace in the shape of a wolf, detailed face and body

Skull, masterpiece, a human skull made of broccoli, best quality, highest details, 8k

Full body, fat cat, in a T-shirt and jeans, caricatural, hyper-detailed, hyper-realism, sharp frame, cinematics

A vanilla coffee ice cream cone with whipped cream, colorful, delicious, digital art, best quality, highest details

BJ_Sacred_beast, Drawing, film grain, Belted Kingfisher in temperate ocean, standing on a branch, masterpiece, best quality, highest details, 4k, 8k

Steampunk style Cute, Black dog, seated dog, Black,Antique, mechanical, brass and copper tones, gears, intricate, detailed

1girl, upper body, perfect face, looking straight ahead, photorealistic highly detailed 8k photography, young woman in rwtcdress, extremely detailed

Figure 7. Additional qualitative results of MVControl 3D generation.

1girl, full body of virgin mary, holy, open-armed, in digital art illustration style, hyper detailed, smooth, sharp focus, masterpiece, best quality

A DSLR photo of a man, walking, strong, cosplaying a sheet ghost, sheet ghost face, long white cloth on upper body, jeans under knees, extremely detailed

SDPA Pixel Art of a blue vase of sunflowers

Realistic, full body, standing, TMNTDonatello, a teenage mutant ninja turtle, extremely detailed

Masterpiece, full body, a beautiful female demon with deep black eyes, very pale skin, long billowing hair, very long floating delicate black dress, long intricate horns, best quality

A cute penguin wearing smoking is riding skateboard, Adorable Character

Oldschool vintage medieval knight, full body armor, Bright silver armor, helmet, portrait, photorealistic, 8k

Military backpack, video game asset, concept art, masterpiece

Figure 8. Additional qualitative results of MVControl 3D generation.