# Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model

Ruoxi Shi[1] Hansheng Chen[2] Zhuoyang Zhang[3] Minghua Liu[1]
Chao Xu[4] Xinyue Wei[1] Linghao Chen[5] Chong Zeng[5] Hao Su[1]

[1]UC San Diego [2]Stanford University [3]Tsinghua University [4]UCLA [5]Zhejiang University

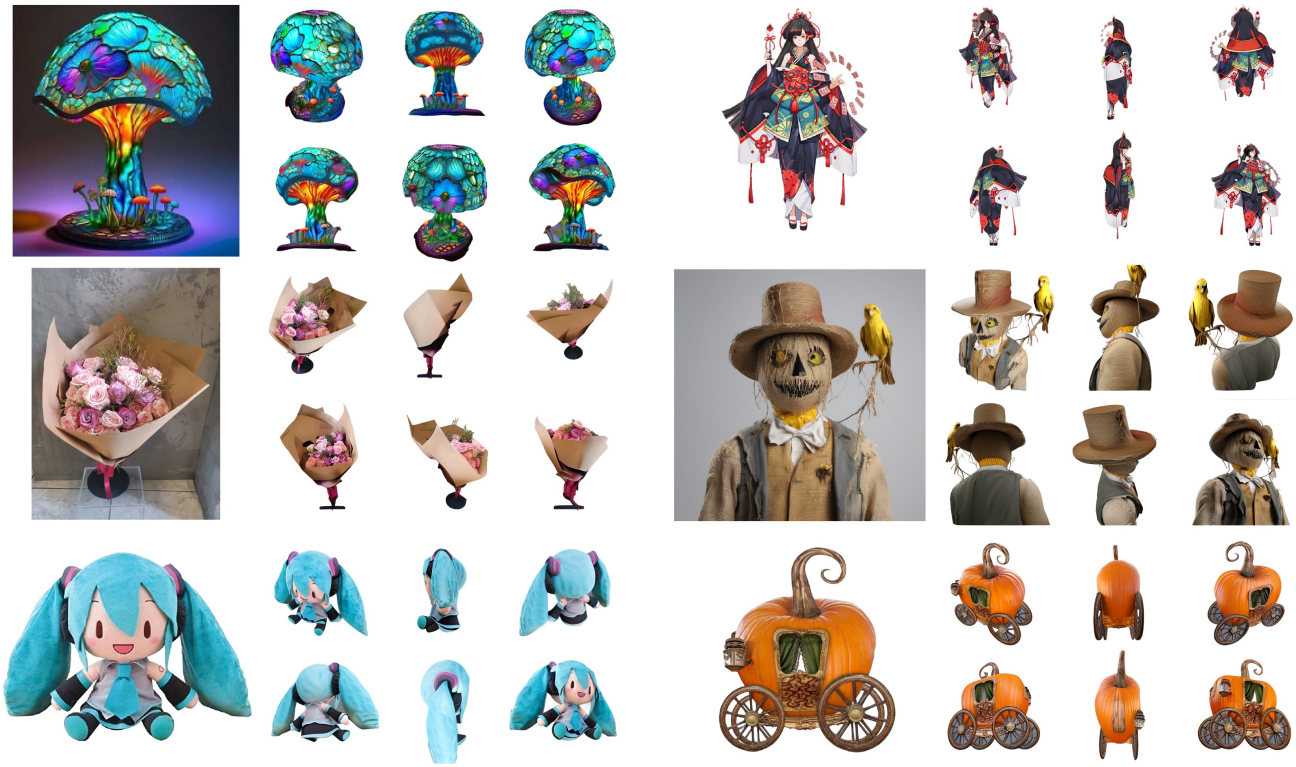

Figure 1. Zero123++ excels at generating high-quality, consistent multi-view 3D images, accommodating a wide range of open-world inputs, from real photos to generated images and 2D illustrations.

## Abstract

*We report Zero123++, an image-conditioned diffusion model for generating 3D-consistent multi-view images from a single input view. To take full advantage of pretrained 2D generative priors, we develop various conditioning and training schemes to minimize the effort of finetuning from off-the-shelf image diffusion models such as StableDiffusion. Zero123++ excels in producing high-quality, consistent multi-view images from a single image, overcoming common issues like texture degradation and geometric misalignment. Furthermore, we showcase the feasibility of training a ControlNet on Zero123++ for enhanced control over the generation process. The code is available at* https://github.com/SUDO-AI-3D/zero123plus.

## 1. Introduction

3D content generation has seen significant progress with the emerging novel view generative models, leveraging the powerful 2D diffusion generative priors learned from extensive datasets sourced from the Internet. Zero-1-to-

3 [12] (or Zero123) pioneers open-world single-image-to-3D conversion through zero-shot novel view synthesis. Despite promising performance, the geometric inconsistency in its generated images has yet to bridge the gap between multi-view images and 3D scenes. Recent works like One-2-3-45 [11], SyncDreamer [13] and Consistent123 [10] build extra layers upon Zero-1-to-3 to obtain more 3D-consistent results. Optimization-based methods like DreamFusion [16], ProlificDreamer [23] and DreamGaussian [22] distill a 3D representation from inconsistent models to obtain 3D results. While these techniques are effective, they could work even better with a base diffusion model that generates consistent multi-view images. In this light, we revisit Zero-1-to-3 and finetune a new multi-view consistent base diffusion model from Stable Diffusion [18].

Zero-1-to-3 generates each novel view independently. Due to the sampling nature of diffusion models, this approach leads to a breakdown in consistency between the generated views. To address this issue, we adopt a strategy of tiling six views surrounding the object into a single image. This tiling layout enables the correct modeling of the joint distribution of multi-view images of an object.

Another issue with Zero-1-to-3 is its underutilization of existing capabilities offered by Stable Diffusion. We attribute this to two design problems: a) During training with image conditions, Zero-1-to-3 does not effectively incorporate the global or local conditioning mechanisms provided by Stable Diffusion. In Zero123++, we have taken a careful approach, implementing various conditioning techniques to maximize the utilization of Stable Diffusion priors. b) Zero-1-to-3 uses a reduced resolution for training. It is widely recognized that reducing the output resolution below the training resolution can lead to a decline in image generation quality for Stable Diffusion models. However, the authors of Zero-1-to-3 encountered instability when training at the native resolution of 512 and opted for a lower resolution of 256. We have conducted an in-depth analysis of this behavior and proposed a series of strategies to address this issue.

## 2. Improving Consistency and Conditioning

In this section, we explore the techniques employed in Zero123++ to improve multi-view consistency and image conditioning, with a primary focus on reusing the priors from pretrained Stable Diffusion model.

### 2.1. Multi-view Generation

The essence of generating consistent multi-view images is the correct modeling of the joint distribution of multiple images. Zero-1-to-3 models the conditional marginal distribution of each image separately and independently, which ignores the correlations between multi-view images.

In Zero123++, we take the simplest form of tiling 6 images with a $3 \times 2$ layout into a single frame for multi-view generation (Fig. 2).

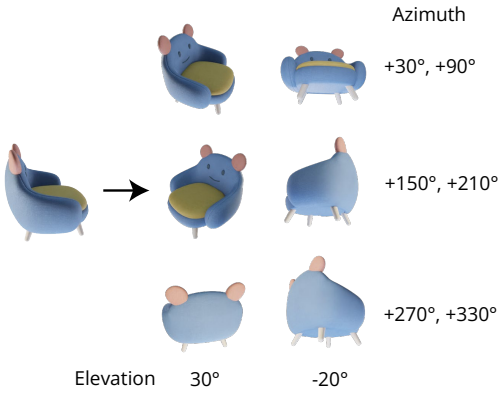


Figure 2. **Layout of Zero123++ prediction target.** We use a fixed set of relative azimuth and absolute elevation angles.

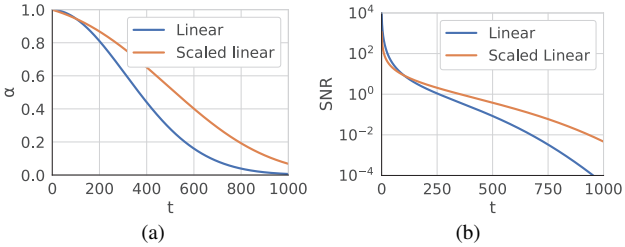


Figure 3. Comparison between the linear schedule and Stable Diffusion's scaled linear schedule.

In the context of object and camera poses, it's worth noting that the Objaverse dataset [4] does not consistently align objects in a canonical pose, despite them typically being oriented along the gravity axis. Consequently, there is a wide range of absolute orientations for objects. We have observed that training the model on absolute camera poses can lead to difficulties in disambiguating object orientations. Conversely, Zero-1-to-3 is trained on camera poses with relative azimuth and elevation angles to the input view. This formulation, however, requires knowing the elevation angle of the input view to determine the relative poses between novel views. As a result, various existing pipelines like One-2-3-45 [11] and DreamGaussian [22] have incorporated an additional elevation estimation module, which introduces extra error into the pipeline.

To address these issues, we use fixed absolute elevation angles and relative azimuth angles as the novel view poses, eliminating the orientation ambiguity without requiring additional elevation estimation. More specifically, the six poses consist of interleaving elevations of $30°$ downward and $20°$ upward, combined with azimuths that start at $30°$ and increase by $60°$ for each pose.


Figure 4. **Importance of the noise schedule.** Noise schedule strongly affects the model's capability to adapt to new global requirements (generating a pure white image from the prompt *a police car* in this case). Notably, both schedules produce highly similar images before fine-tuning; therefore, we present only the result of the $v$ model with linear schedule before fine-tuning.

Figure 5. **Swapping the noise schedule.** We swap the schedule of Stable Diffusion 2 $v$ (Left) and $\epsilon$-parameterized (Right) models from scaled-linear to linear at inference time without any finetuning. Prompt: *a blue clock with black numbers*. The $\epsilon$-parameterized model exhibits a significant decrease in quality, while the $v$ model produces a high-quality image with the linear noise schedule.

## 2.2. Consistency and Stability: Noise Schedule

The original noise schedule for Stable Diffusion, *i.e.*, the scaled-linear schedule, places emphasis on local details but has very few steps with lower Signal-to-Noise Ratio (SNR), as shown in Fig. 3. These low SNR steps occur in the early denoising stage, which is crucial for determining the global low-frequency structure of the content. A reduced number of steps in this stage, either during training or inference, can lead to greater structural variation. While this setup is suitable for single-image generation, we have observed that it limits the model's ability to ensure the global consistency between multiple views.

To empirically verify this, we perform a toy task by finetuning a LoRA [6] model on the Stable Diffusion 2 v-prediction model to overfit a blank white image given the prompt *a police car*. The results are presented in Fig. 4. Surprisingly, with the scaled-linear noise schedule, the LoRA model cannot overfit on this simple task; it only slightly whitened the image. In contrast, with the linear noise schedule, the LoRA model successfully generates a blank white image regardless of the prompt. While finetuning the full model may still be viable for the scaled-linear schedule, this example highlights the significant impact of the noise schedule on the model's ability to adapt to new global requirements.

As pointed out by Chen [2], high-resolution images appear less noisy compared to low-resolution images when subjected to the same absolute level of independent noise (see Fig. 2 in [2]). This phenomenon occurs because "higher resolution natural images tend to exhibit higher degree of redundancy in (nearby) pixels, therefore less information is destroyed with the same level of independent noise". Consequently, we can interpret the use of lower resolution in Zero-1-to-3 training as a modification of the noise schedule, placing greater emphasis on the global requirements of 3D-consistent multi-view generation. This also explains the instability issue of training Zero-1-to-3 with higher resolution [12].

In summary, we find it necessary to switch from the scaled-linear schedule to the linear schedule for noise in our model. However, this shift introduces another potential challenge: adapting the pretrained model to the new schedule. Fortunately, we have observed that the $v$-prediction model is quite robust when it comes to swapping the schedule, in contrast to the $x_0$- and $\epsilon$-parameterizations, as illustrated in Figure 5. It is also theoretically supported that the $v$-prediction is inherently more stable [19]. Therefore, we have opted to utilize the Stable Diffusion 2 $v$-prediction model as our base model for fine-tuning.

## 2.3. Local Condition: Scaled Reference Attention

In Zero-1-to-3, the conditioning image (single view input) is concatenated in the feature dimension with the noisy inputs to be denoised for local image conditioning. This imposes an incorrect pixel-wise spatial correspondence between the input and the target image.

We propose to use a scaled version of Reference Attention to provide proper local conditioning input.

As shown in Fig. 6, Reference Attention [24] refers to the operation of running the denoising UNet model on an extra reference image and appending the self-attention key and value matrices from the reference image to the corresponding attention layers when denoising the model input. The same level of Gaussian noise as the denoising input is added to the reference image to allow the UNet to attend to relevant features for denoising at the current noise level.

Without any finetuning, Reference Attention is already capable of guiding the diffusion model to generate images that share similar semantic content and texture with the reference image. When finetuned, we observed that the Reference Attention works better when we scale the latent (before adding noise). In Figure 7, we provide a comparison from experiments conducted on ShapeNet Cars [1] to demon-

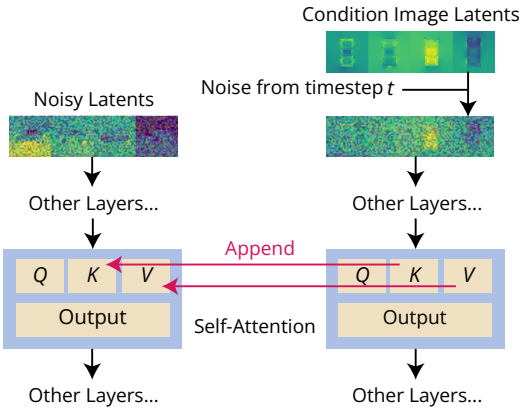


Figure 6. **Reference Attention.** It adds an additional conditioning branch and modifies key (K) and value (V) matrices of the self-attention layers to accept the extra condition image, which can fully reuse Stable Diffusion priors.


Figure 7. **Comparison on local conditioning.** We train Zero123++ with different levels of scaled reference attention on the ShapeNet Cars dataset. Output coherence with the input image is best on 5x scaled reference attention.

strate that the model achieves the highest consistency with the conditioning image when the reference latent is scaled by a factor of 5.

## 2.4. Global Condition: FlexDiffuse

In the original Stable Diffusion, global conditioning comes solely from text embeddings. Stable Diffusion employs CLIP [17] as the text encoder, and performs cross-attention between model latents and per-token CLIP text embeddings. As a result, we can make use of the alignment between CLIP image and text spaces to reuse the prior for global image conditioning.

We propose a trainable variant of the linear guidance mechanism introduced in FlexDiffuse [21] to incorporate global image conditioning into the model while minimizing the extent of fine-tuning. We start from the original prompt embeddings $T$ of shape $L \times D$ where $L$ is length of tokens and $D$ is the dimension of token embeddings, and add the CLIP global image embedding $I$ of shape $D$ multiplied by a trainable set of global weights $\{w_i\}_{i=1,\dots,L}$ (a shared set of weights for all tokens) to the original prompt embeddings, or formally,

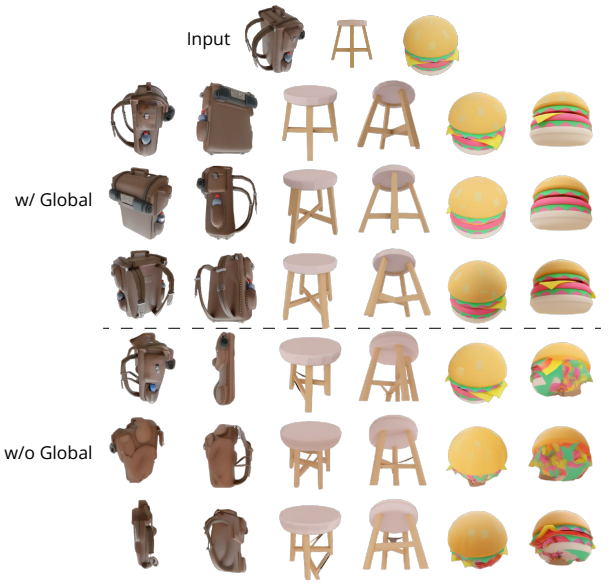


Figure 8. **Ablation on global conditioning.** In regions of the image that are not visible in the input, the results are significantly worse without global conditioning.

$$T_i' = T_i + w_i \cdot I, i = 1, 2, \dots, L. \quad (1)$$

We initialize the weights with FlexDiffuse's linear guidance:

$$w_i = \frac{i}{L}. \quad (2)$$

In the released Zero123++ models, we do not impose any text conditions, so $T$ is obtained by encoding an empty prompt.

We present the results of trained Zero123++ models with and without global conditioning in Figure 8. In the absence of the proposed global conditioning, the quality of generated content remains satisfactory for visible regions corresponding to the input image. However, the generation quality significantly deteriorates for unseen regions, as the model lacks the ability to infer the global semantics of the object.

## 2.5. Putting Everything Together

Starting from the Stable Diffusion 2 $v$-model, we train our Zero123++ model using all the techniques mentioned above. We train Zero123++ on Objaverse [4] data rendered with random HDRI environment lighting.

We adopt the phased training schedule from the Stable Diffusion Image Variations model [9] to further reduce the extent of finetuning and preserve as much prior in Stable Diffusion as possible. In the first phase, we only tune the self-attention layers and the KV matrices of cross-attention

layers of Stable Diffusion. We use the AdamW [7, 14] optimizer with cosine annealing learning rate schedule peaking at $7 \times 10^{-5}$ and 1000 warm-up steps. In the second phase, we employ a very conservative constant learning rate of $5 \times 10^{-6}$ and 2000 warm-up steps to tune the full UNet. We employ the Min-SNR weighting strategy [5] to make the training process more efficient.

## 3. Comparison to the State of the Art

### 3.1. Image to Multi-view

**Qualititative Comparison.** In Fig. 10 we show generation results of Zero-1-to-3 XL [3, 12], SyncDreamer [13] and our Zero123++ on four input images, including one image from the Objaverse dataset with large uncertainty on the back side of the object (an electric toy cat), one real photo (extinguisher), one image generated by SDXL [15] (a dog sitting on a rocket) and an anime illustration. We apply the elevation estimation method from One-2-3-45 [11] for the required elevation estimation steps in Zero-1-to-3 XL and SyncDreamer. We use SAM [8] for background removal. Zero123++ generates consistent and high-quality multi-view images, and can generalize to out-of-domain AI-generated and 2D illustration images.

**Quantitative Comparison.** We evaluate the LPIPS score [26] of different models on the validation split (subset of Objaverse) to quantitatively compare Zero-1-to-3 [12], Zero-1-to-3 XL [3, 12] and Zero123++. SyncDreamer [13] is excluded because it does not support changing the elevation. To evaluate the multi-view generation results, we tile 6 generated images and the ground truth reference images (rendered from Objaverse) respectively, and compute the LPIPS score between the tiled images. Note that Zero-1-to-3 models may have seen our validation split during training, and the XL variant is trained on much more data than Zero123++. Nevertheless, Zero123++ achieves the best LPIPS score on the validation split. This shows the effectiveness of our designs in Zero123++. The results are shown in Tab. 1.

Table 1. Quantitative results of models on our validation split.

| Model | LPIPS ↓ |
|---|---|
| Zero-1-to-3 | 0.210 ± 0.059 |
| Zero-1-to-3 XL | 0.188 ± 0.053 |
| Zero123++ (Ours) | **0.177** ± 0.066 |

### 3.2. Text to Multi-view

For text to multi-view, we first generate an image using SDXL with the text prompts, and then run Zero123++ upon the generated image. In Fig. 11, we compare our results to MVDream [20] and Zero-1-to-3 XL [3, 12]. We observe a texture style shift in MVDream to cartoonish and flat texture due to the bias in the Objaverse dataset, and the fact that Zero-1-to-3 can not guarantee multi-view consistency, while Zero123++ is able to generate realistic, consistent and highly detailed multi-view images using the text-to-image-to-multi-view pipeline.


Figure 9. **Example outputs from depth-controlled Zero123++.** In the first example, we generate the face for the fox from its back; in the second example, we start with only the geometry and generate both the input image to Zero123++ and the multi-view images conditioned on the geometry.

## 4. Depth ControlNet for Zero123++

In addition to the base Zero123++ model, we also release a depth-controlled version of Zero123++ built with ControlNet [25]. We render normalized linear depth images corresponding to the target RGB images and train a ControlNet to control Zero123++ on the geometry via depth. The trained model is able to achieve a superior LPIPS of 0.086 on our validation split.

Fig. 9 shows two example generations from depth controlled Zero123++. We may use a single view as the input image to Zero123++ (the first example) or generate the input image from depth with vanilla depth-controlled Stable Diffusion as well to eliminate any need for input colors (the second example).


Figure 10. Qualitative comparison of Zero123++ against various methods on single image to multi-view task.

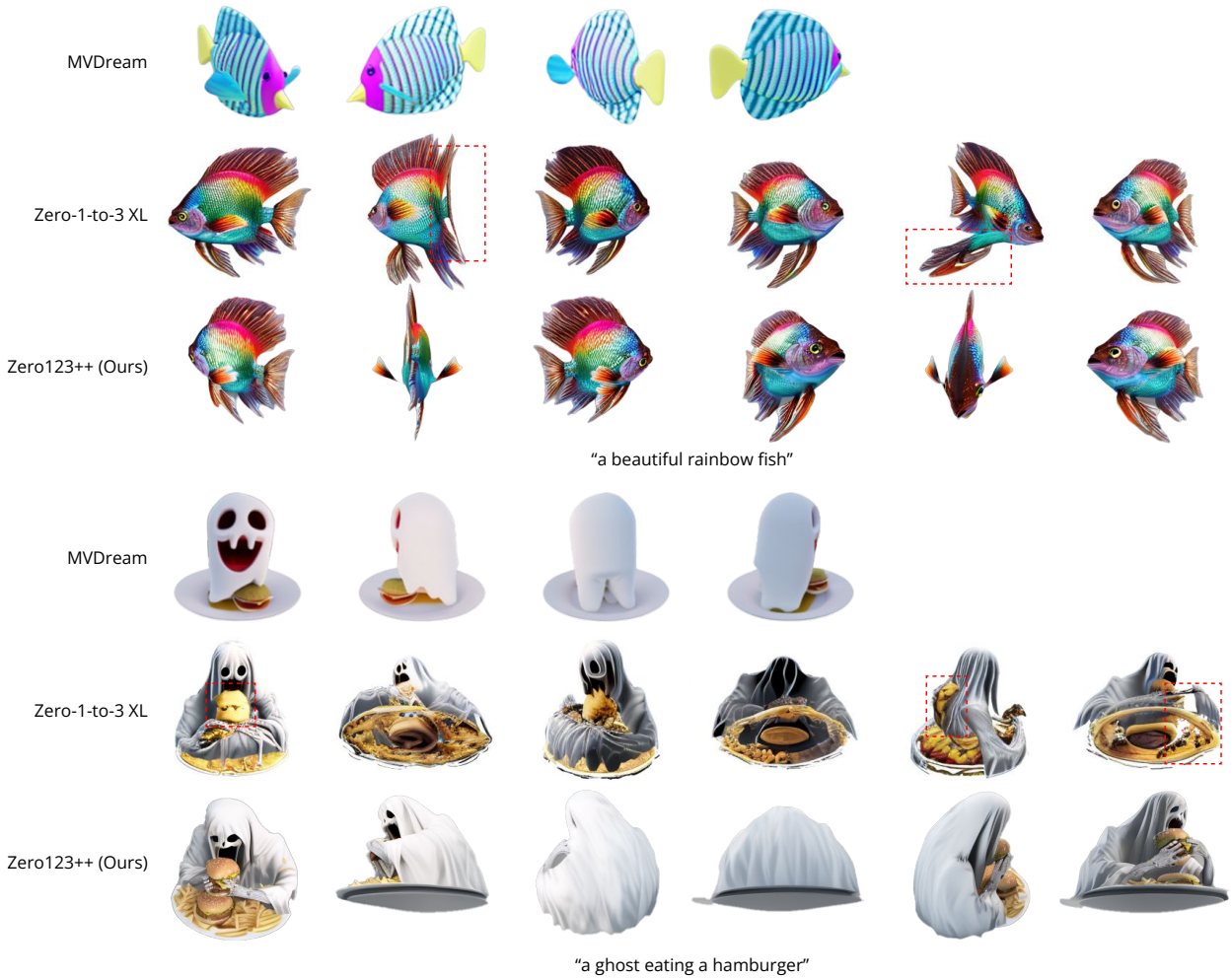


Figure 11. Qualitative comparison of Zero123++ against MVDream and Zero-1-to-3 XL on text to multi-view task.

## 5. Future Work

This report presents an array of analyses and enhancements for our new image-to-multi-view base diffusion model, Zero123++. While our model has already achieved significantly improved quality, consistency, and generalization compared to previous models, we'd like to highlight three areas of potential future work:

- Two-stage refiner model. Though $\epsilon$-parametrized models have trouble meeting the global requirements of consistency, they usually do better at generating local details. We may apply a two-stage generate-refine pipeline like SDXL [15], and employ the $\epsilon$-parametrized SDXL model as the base model for fine-tuning the refiner model, leveraging its stronger priors compared to the previous SD models.

- Further scaling-up. Currently Zero123++ is trained on the medium-scale Objaverse dataset, containing around 800k objects. To enhance our model's capabilities, we're looking into the possibility of scaling up our training to a larger dataset, such as Objaverse-XL [3].

- Utilizing Zero123++ for mesh reconstruction. There remains a gap between high quality multi-view images and high quality 3D meshes. We show some preliminary results utilizing Zero123++ for mesh generation in Fig. 12.

## References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[2] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 3


Figure 12. Preliminary mesh generation results with Zero123++.

[3] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 5, 7

[4] 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, 4

[5] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. *arXiv preprint arXiv:2303.09556*, 2023. 5

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[9] Lambda Labs. Stable diffusion image variations. https://huggingface.co/lambdalabs/sd-image-variations-diffusers, 2022. 4

[10] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 2

[11] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 5

[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*, pages 9298–9309, 2023. 2, 3, 5

[13] 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. 2, 5

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5, 7

[16] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[17] 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. 4

[18] 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, June 2022. 2

[19] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[20] 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. 5

[21] Tim Speed. Flexdiffuse: An adaptation of stable diffusion with image guidance. https://github.com/tim-speed/flexdiffuse, 2022. 4

[22] 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. 2

[23] 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. 2

[24] Lyumin Zhang. Reference-only control. https://github.com/Mikubill/sd-webui-controlnet/discussions/1236, 2023. 3

[25] 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. 5

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5