Supplementary Materials
MVD-Fusion: Single-view 3D via Depth-consistent Multi-view Generation

We provide a more detailed description of our architectural details and ablations. We include 360 visualizations on our project page: https://mvd-fusion.github.io/.

1. Architecture Details

Our network consists of modifications on top of Zero123 [2]. We describe each component of our network in detail.

VAE. We use the pretrained VAE from Stable Diffusion 1.4 [4]. We freeze the VAE.

UNet. We initialize our UNet with weights from Zero123 [2]. Zero123 has a novel view synthesis UNet that accepts one input image (4 channel latents) and one target noisy image (4 channel latents) along with camera pose and predicts a novel view image latent (4 channels). We modify the input and output blocks to accommodate prediction of an additional depth channel. Our UNet has 10 input channels and 5 output channels. For all experiments, we use only RGB images as input (4 channel latents) and pad the additional channel with zeros. The noisy target target image is always 5 channels.

CLIP and Camera Pose Embedding. We follow Zero123 [2] to use the frozen CLIP [3] image encoder along with camera information as one of the inputs to the cross attention layers in Stable Diffusion. However, instead of using azimuth and elevation angle representation, we directly use a flattened essential matrix as input. We use 3 fully connected layers to map CLIP image embedding and flattened essential matrix into cross attention input of dimension 768.

Depth-guided Multi-view Attention. After each of the existing cross attention layers in the UNet, we add additional cross attention layers that attend to view-aligned feature frustums sampled from our depth-guided multi-view attention module. Our depth-guided attention module is a 3 layers transformer that aggregates information across the noisy target latents from the current timestep and also input image latents. For each target view, we generated a feature frustum of shape (1, 256, 3, 32, 32), where the feature map is 32 by 32, with 3 depth samples, and feature dimension 256. The depth dimension represents the number of depth points sampled along each ray and can be reduced down to just 1. Our transformer uses a hidden dimension of 256 with 8 heads. We use an additional fully connected layer to project our features into 768 dimensions, making them compatible with existing cross attention layers. A key difference between our multi-view cross attention and text cross attention is that, in our multi-view attention, each latent patch independently attends to the corresponding patch in the feature frustum.

Table 1. Ablation study on Google Scanned Objects (GSO) dataset. We ablate the effect of the unconditional guidance scale during inference. We randomly chose 30 instances from the dataset for evaluation. 'Scale' denotes the unconditional guidance scale.

Scale	PSNR ↑	SSIM ↑	LPIPS ↓
1.0	19.34	0.775	0.199
2.0	**19.91**	**0.787**	**0.184**
3.0	19.38	0.778	0.189
5.0	18.66	0.771	0.194

2. Additional Results and Visualizations

Ablating Unconditional Guidance Scale. We further conduct experiments to ablate the effect of the unconditional guidance scale. Proposed in [1], classifier-free guidance jointly learns an unconditional model to enable higher-quality generation. In our method, we use the unconditional guidance scale ω to control the contribution of the unconditional model: $\hat{\epsilon}_{\phi'}(\mathbf{y}, \mathbf{x}_t^n, \pi^n, \mathbf{z}_t^n, t) = \omega\epsilon_{\phi'}(\mathbf{y}, \mathbf{x}_t^n, \pi^n, \mathbf{z}_t^n, t) + (1 - \omega)\epsilon_{\phi'}(\mathbf{x}_t^n, t)$, where $\epsilon_{\phi'}(\mathbf{y}, \mathbf{x}_t^n, \pi^n, \mathbf{z}_t^n, t)$ is our proposed multi-view diffusion model. In practice, we notice that a higher unconditional guidance scale leads to better multi-view consistency. As shown in Table 1, we find that adopting a scale of 2 yields the best performance. Therefore, we use this unconditional guidance scale for inference.

References

[1] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[2] 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 *ICCV*, 2023. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1