A Details: Diffusion model training

A.1 Model

We train diffusion models for various camera conditioning parameterizations: $\mathbf{M}_{\text{Zero-1-to-3}}$, $\mathbf{M}_{\text{6DoF+1}}$, $\mathbf{M}_{\text{6DoF+1, norm.}}$, $\mathbf{M}_{\text{6DoF+1, agg.}}$, and $\mathbf{M}_{\text{6DoF+1, viewer}}$. Our runtime is identical to Zero-1-to-3 (Liu et al., 2023) as the camera conditioning novelties we introduce add negligible overhead and can be done mainly in the dataloader. We train our main model for $60,000$ steps with batch size 1536. We find that performance tends to saturate after about $20,000$ steps for all models, though it does not decrease. For inference of the 2D diffusion model, we use 500 DDIM steps and guidance scale 3.0.

Details for $\mathbf{M}_{\text{6DoF+1}}$: To embed the field of view f in radians, we use a 3-dimensional vector consisting of $[f, \sin(f), \cos(f)]$. When concatenated with the $4 \times 4 = 16$-dimensional relative pose matrix, this gives a 19-dimensional conditioning vector.

Details for $\mathbf{M}_{\text{6DoF+1, viewer}}$: We use the DPT-SwinV2-256 depth model (Ranftl et al., 2021) to infill depth maps from ORB-SLAM and COLMAP on the ACID, RealEstate10K, and CO3D datasets. We infill the invalid depth map regions only after aligning the disparity from the monodepth estimator to the ground-truth sparse depth map via the optimal scale and shift following Ranftl et al. (2022). We downsample the depth map $4\times$ so that the quantile function is evaluated quickly.

At inference time, the value of $\mathbf{Q}_{20}(\bar{D})$ may not be known since input depth map D is unknown. Therefore there is a question of how to compute the conditioning embedding at inference time. Values of $\mathbf{Q}_{20}(\bar{D})$ between $.7 - 1.$ work for most images and it can be chosen heuristically. For instance, for DTU we uniformly assume a value of .7, which seems to work well. Note that any value of $\mathbf{Q}_{20}(\bar{D})$ is presumably possible; it is only when this value is incompatible with the desired SDS camera radius that distillation may fail, since the cameras may intersect the visible content.

A.2 Dataloader

One significant engineering component of our work is our design of a streaming dataloader for multiview data, built on top of WebDataset (Breuel, 2020). Each dataset is sharded and each shard consists of a sequential tar archive of scenes. The shards can be streamed in parallel via multiprocessing. As a shard is streamed, we yield random pairs of views from scenes according to a "rate" parameter that determines how densely to sample each scene. This parameter allows a trade-off between fully random sampling (lower rate) and biased sampling (higher rate) which can be tuned according to the available network bandwidth. Individual streams from each dataset are then combined and sampled randomly to yield the mixture dataset. We will release the code together with our main code release.

B Details: NeRF prediction and distillation

B.1 SDS Anchoring

We propose SDS anchoring in order to increase the diversity of synthesized scenes. We sample 2 anchors at 120 and 240 degrees of azimuth relative to the input camera.

One potential issue with SDS anchoring is that if the samples are 3D-inconsistent, the resulting generations may look unusual. Furthermore, traditional SDS already performs quite well except if the criterion is diverse backgrounds. Therefore, to implement anchoring, we randomly choose with probability $.5$ either the input camera and view or the nearest sampled anchor camera and view as guidance. If the guidance is an anchor, we "gate" the gradients flowing back from SDS according to the depth of the NeRF render, so that only depths above a certain threshold (1.0 in our experiments) receive guidance from the anchors. This seems to mostly mitigate artifacts from 3D-inconsistency of foreground content, while still allowing for rich backgrounds. We show video results for SDS anchoring on the webpage.

B.2 HYPERPARAMETERS

NeRF distillation via involves numerous hyperparameters such as for controlling lighting, shading, camera sampling, number of training steps, training at progressively increasing resolutions, loss weights, density blob initializations, optimizers, guidance weight, and more. We will share a few insights about choosing hyperparameters for scenes here, and release the full configs as part of our code release.

Noise scheduling: We found that ending training with very low maximum noise levels such as $.025$ seemed to benefit results, particularly perceptual metrics like LPIPS. We additionaly found a significant benefit on 360-degree scenes such as in the Mip-NeRF 360 dataset to scheduling the noise "anisotropically;" that is, reducing the noise level more slowly on the opposite end from the input view. This seems to give the optimization more time to solve the challenging 180-degree views at higher noise levels before refining the predictions at low noise levels.

Miscellaneous: Progressive azimuth and elevation sampling following (Qian et al., 2023) was also found to be very important for training stability. Training resolution progresses stagewise, first with batch size 6 at 128x128 and then with batch size 1 at 256×256.

C EXPERIMENTAL SETUPS

For our main results on DTU and Mip-NeRF 360, we train our model and Zero-1-to-3 for $60,000$ steps. Performance for our method seems to saturate earlier than for Zero-1-to-3, which trained for about $100,000$ steps; this may be due to the larger dataset size. Objaverse, with $800,000$ scenes, is much larger than the combination of RealEstate10K, ACID, and CO3D, which are only about $95,000$ scenes in total.

For the retrained PixelNeRF baseline, we retrained it on our mixture dataset of CO3D, ACID, and RealEstate10K for about $560,000$ steps.

C.1 MAIN RESULTS

For all single-image NeRF distillation results, we assume the camera elevation, field of view, and content scale are given. These parameters are identical for all DTU scenes but vary across the Mip-NeRF 360 dataset. For DTU, we use the standard input views and test split from from prior work. We select Mip-NeRF 360 input view indices manually based on two criteria. First, the views are well-approximated by a 3DoF pose representation in the sense of geodesic distance between rotations. This is to ensure fair comparison with Zero-1-to-3, and for compatibility with Threestudio's SDS sampling scheme, which also uses 3 degrees of freedom. Second, as much of the scene content as possible must be visible in the view. The exact values of the input view indices are given in Table 1.

The field of view is obtained via COLMAP. The camera elevation is set automatically via computing the angle between the forward axis of the camera and the world's XY-plane, after the cameras have been standardized via PCA following Barron et al. (2022).

One challenge is that for both the Mip-NeRF 360 and DTU datasets, the scene scales are not known by the zero-shot methods, namely Zero-1-to-3, our method, and our retrained PixelNeRF. Therefore, for the zero-shot methods, we manually grid search for the optimal world scale in intervals of $.1$ to find the appropriate world scale for each scene in order to align the predictions to the generated scenes. Between five to nine samples within $[.3, .4, .5, .6, .7, .8, .9, 1., 1.1, 1.2, 1.3, 1.4, 1.5]$ generally suffices to find the appropriate scale. Even correcting for the scale misalignment issue in this way, the zero-shot methods generally do worse on pixel-aligned metrics like SSIM and PSNR compared with methods that have been fine-tuned on DTU.

C.2 USER STUDY

We conduct a user study on the seven Mip-NeRF 360 scenes, comparing our method with and without SDS anchoring. We received 21 respondents. For each scene, respondents were shown 360-

Scene name	Input view index	Content scale
bicycle	98	.9
bonsai	204	.9
counter	95	.9
garden	63	.9
kitchen	65	.9
room	151	2.
stump	34	.9

Table 1: Setup for the Mip-NeRF 360 dataset

.

degree novel view videos of the scene inferred both with and without SDS anchoring. The videos were shown in a random order and respondents were unaware which video corresponded to the use of SDS anchoring. Respondents were asked:

1. Which scene seems more realistic?
2. Which scene seems more creative?
3. Which scene do you prefer?

Respondents generally preferred the scenes produced by SDS anchoring, especially with respect to "Which scene seems more creative?"

C.3 Ablation studies

We perform ablation studies on dataset selection and camera representations. For 2D novel view synthesis metrics, we compute metrics on a held-out subset of scenes from the respective datasets, randomly sampling pairs of input and target novel views from each scene. For 3D SDS distillation and novel view synthesis, our settings are identical to the NeRF distillation settings for our main results except that we use shorter-trained diffusion models. We train them for 25,000 steps as opposed to 60,000 steps for computational constraint reasons.

References

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

Thomas Breuel. Webdataset library, 2020.

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 *CVPR*, 2023.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3), 2022.