A. Appendix

A.1. Additional Implementation Details

Camera pose fine-tuning. At test time, the camera poses are fine-tuned jointly with the diffusion model, by simply back-propagating gradients to camera parameters, represented as azimuth, elevation, and radius. We update them using a much higher learning rate of 100, which is 1000× the learning rate of the diffusion model. Additionally, we constrain the camera elevation and scale to make optimization more stable: Elevation is projected to $[0, \pi]$ every iteration, and the radius is mapped to a range of $[1.5, 2.2]$ via a SoftMax.

3D reconstruction. We regularize our 3D reconstruction using a number of loss terms. To get smooth surfaces, we regularize surface normals $\hat{n}$ to be smooth. $\mathcal{L}_{\hat{n},1}$ regularizes normals at sampled 3D points X to be smooth to small perturbations δ, while $\mathcal{L}_{\hat{n},2}$ regularizes rendered normals $\mathcal{R}^{\hat{n}}_{\psi}(\pi)$ from random camera viewpoints π to be smooth.

$$\mathcal{L}_{\hat{n},1} = \mathbb{E}_{X,\delta \in \mathcal{N}(0,1)} \left\|\hat{n}(X) - \hat{n}(X+\delta)\right\|^2$$

$$\mathcal{L}_{\hat{n},2} = \mathbb{E}_{\pi} \left\|\Delta \mathcal{R}^{\hat{n}}_{\psi}(\pi)\right\|^2$$

Additionally, we regularize the density field to form opaque surfaces without floating artifacts. $\mathcal{L}_{\text{Sparse}}$ nudges rendered masks $\mathcal{R}^{\text{Mask}}_{\psi}(\pi)$ to be sparse with an L1-regularization loss to prevent floaters, while $\mathcal{L}_{\text{Opaque}}$ minimizes their entropy to make the closer to 0/1, from random cameras π.

$$\mathcal{L}_{\text{Sparse}} = \mathbb{E}_{\pi} \left\|\mathcal{R}^{\text{Mask}}_{\psi}(\pi)\right\|_1$$

$$\mathcal{L}_{\text{Opaque}} = \mathbb{E}_{\pi}[H(\mathcal{R}^{\text{Mask}}_{\psi}(\pi))]$$

Total regularization is a weighted sum of the normal and mask loss terms with $\lambda_{\hat{n},1} = 0.1$, $\lambda_{\hat{n},2} = 0.1$, $\lambda_{\text{Sparse}} = 1$, and $\lambda_{\text{Opaque}} = 1$:

$$\mathcal{L}_{\text{Reg}} = \lambda_{\hat{n},1}\mathcal{L}_{\hat{n},1} + \lambda_{\hat{n},2}\mathcal{L}_{\hat{n},2} + \lambda_{\text{Sparse}}\mathcal{L}_{\text{Sparse}} + \lambda_{\text{Opaque}}\mathcal{L}_{\text{Opaque}}$$

A.2. Additional Novel View Synthesis Ablations

We provide an additional ablation study of the components of our system, evaluated on novel view synthesis. Table 6 presents all of the ablated components.

3D preservation loss. We compare different versions of regularization losses, applied during the fine-tuning stage of the view-conditioned diffusion model. Specifically, we evaluate three types: no regularization (`b2`), regularization by sampling random pairs from the pre-training set and incorporating them during the fine-tuning (`b2`), and regularization by incorporating nearest-neighbors according to CLIP similarity score (3D reservation loss, as described in Section 3.2).

As presented in Table 6, incorporating random images from the training set during the fine-tuning results in comparable performance to not applying regularization. Differently from these two options, using CLIP as a metric for retrieving nearest neighbors from the training data results in an improvement in all metrics.

Camera initialization and refinement. Similarly to the ablations presented in Section 4.3, we compare the downstream performance with different camera pose initialization - RelPose++ (`d2`) and RelPose++* (SAP3D). Differently from Section 4.3, we apply the fine-tuning stage in both cases. As shown in Table 6, scaling the training data of RelPose++ results in significant improvements in novel view synthesis - 4.3dB increase in PSNR. Additionally, the results of initialization from RelPose++* are comparable to initialization with ground-truth camera poses (`d3`).

We evaluate the effect of *not* fine-tuning the camera poses together with the view-conditioned diffusion model (`d1`). The camera-pose fine-tuning that is done in SAP3D results in 0.4dB improvement in PSNR. Therefore, fine-tuning the camera pose is beneficial for downstream novel view synthesis.

Sampling conditioning. During the novel view generation process, we use stochastic conditioning - each sampling step from the diffusion model is conditioned on a randomly sampled image from the input images. We compare this conditioning strategy to two different conditioning strategies - using one random image for all the diffusion sampling steps (`c2`), and conditioning the diffusion process on the closest input image, computed according to the camera pose (`c1`). As shown in Table 6, conditioning on the nearest image results in better performance than conditioning on one random image, and stochastic sampling (SAP3D) is better than both.

A.3. Additional Qualitative Results

In Figure 9, We show two sets of qualitative reconstruction comparisons: one on the ABO dataset [2] and another on the Tanks and Temples dataset [17]. These comparisons evaluate our proposed method SAP3D against Zero123 [23] and One2345 [21].

We provide additional qualitative results for novel view synthesis and 3D reconstruction, as well as additional qualitative comparisons for our ablation study in https://sap3d.github.io/supp.html.

Acknowledgments. The authors would like to thank Yutong Bai for the helpful discussions. YG is funded by the Google Fellowship.

	Model	Regularization	Sampling	Cameras		Novel View Quality		
	Finetuned		Conditioning	Initialization	Refine	PSNR↑	SSIM↑	LPIPS↓
SAP3D	✓	CLIP-retrieved	Stochastic	RelPose++*	✓	17.7	0.83	0.13
a			Stochastic	RelPose++*		16.3	0.78	0.18
b1	✓	Random	Stochastic	RelPose++*	✓	16.8	0.79	0.16
b2	✓	None	Stochastic	RelPose++*	✓	16.4	0.79	0.17
c1	✓	CLIP-retrieved	Nearest	RelPose++*	✓	16.8	0.72	0.16
c2	✓	CLIP-retrieved	Random single	RelPose++*	✓	15.2	0.79	0.23
d1	✓	CLIP-retrieved	Stochastic	RelPose++*		17.3	0.80	0.15
d2	✓	CLIP-retrieved	Stochastic	RelPose++	✓	13.4	0.74	0.29
d3	✓	CLIP-retrieved	Stochastic	Ground Truth	✓	17.8	0.84	0.13

Table 6. **Ablation study on novel view synthesis.** We evaluate the effect of various design choices on novel view synthesis. RelPose++* denotes our RelPose++ model trained on Objaverse. In the paper, we refer to `a` as "SAP3D w/o adaptation". Please see the text for details.

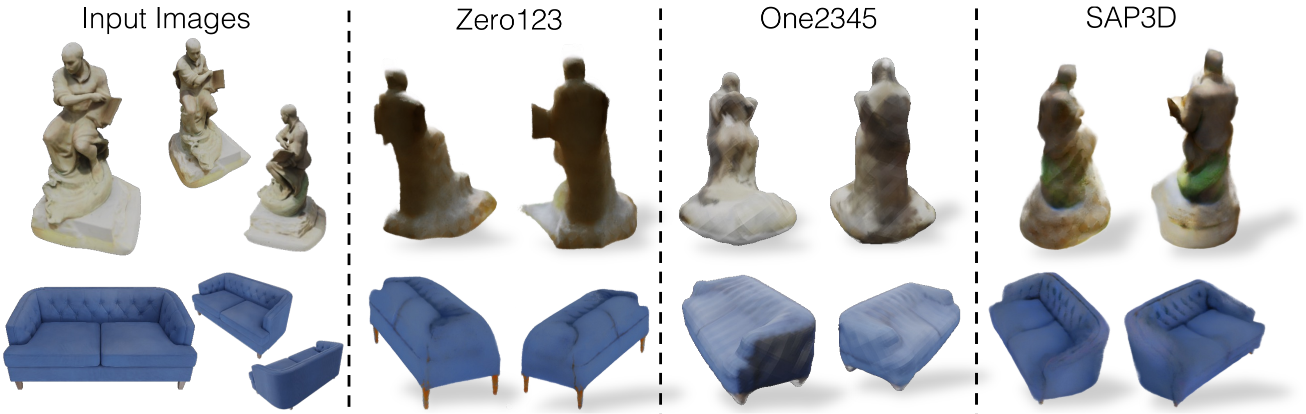

Figure 9. 3D reconstruction comparisons on Tanks and Temples and ABO respectively between SAP3D, Zero123 [23] and One2345 [21].