A. More Implementation Details

DDIM Initialization. We use a black Gaussian-blob on a white background to initialize the first 20 steps (out of 200) of our DDIM sampler, which ensures that our model correctly generates a single object placed at the center on the white background (similar to images during training). This trick is similar to the one used in iNVS [39] which starts inpainting with the partial image warped with depth. Our hyper-parameter choices and additional training details are shown in Tab. 4.

Hyper-parameters	Value
Base Learning rate	1e-4
Learning rate decay	None
Loss Type	L2
Classifier-free guidance	7.5 (text-only)
Effective batch size	1152
DDIM Steps	200
Gaussian Blob Initialization Steps	20
CLIP Frozen	True
Renders background color	White
Image Resolution	256
Learning rate linear warmup	100 steps

Table 4. Hyperparameter choices for SPAD.

B. Additional Experiments and Results

B.1. Qualitative Results

Close viewpoints generations from SPAD. In Fig. B.10, we put text-conditioned multi-view generations from SPAD where we increment the azimuth angle by 10 degrees per view. We find that SPAD can synthesize continuous moving views well, without content copying issues.

B.2. Baseline Comparisons

Text-conditioned multi-view generations and comparison with MVDream [73]. Fig. B.8 presents the results. We find SPAD synthesizes images with higher quality details and better alignment with the text prompt.

NVS Comparison with Zero123 [45]. Fig. B.9 presents the results. We find SPAD preserves the structural and perceptual details of objects and exhibits better 3D consistency.

NVS setup with SyncDreamer [46]. The official inference code of SyncDreamer always generates 16 views at fixed azimuth angles uniformly distributed in $[0°, 360°]$, which is incompatible with our random view generation setup. We modified their code to consider the exact target camera pose as model input, but found it performed worse than choosing the prediction at the azimuth that is closest to the target azimuth. Therefore, we report SyncDreamer results using the closest view, where the error is usually smaller than $10°$.

B.3. User Study comparing SPAD with MVDream

We conducted a user study on the visual quality, 3D consistency, and text alignment of multi-view generations. We distributed our questions via Amazon Mechanical Turk, where participants were given 4-view generations of SPAD and MVDream [73], and asked to choose the better one satisfying the above properties. We found that SPAD is preferred over MVDream with 59% vs 41%.

Exact Instructions: You are shown a text promt and two sets of images corresponding to 4 different views of the same object. The views is front, left, right and back. Your task is to choose which of the sets of views is better, based on (1) consistency between different views (e.g it should represent the same object, have the same structure and colors) (2) looks better visually, (3) describes what is written in the text accurately, either Option A or Option B.

B.4. Training with Stable Diffusion v2.1 Weight

The SPAD model we evaluated in the main paper and Appendix B.2 is initialized from the weight of Stable Diffusion (SD) v1.5. Here, we train another model initializing from the weight of the stronger SD v2.1 release. Fig. B.12 presents the multi-view generation results of this model. Indeed, we observe better alignment with the text input, especially with longer and more complicated prompts.

This is also verified by the quantitative result. SPAD with SD v1.5 achieves a CLIP-score of 29.87$\pm$3.33. SPAD with SD v2.1 achieves a better CLIP-score of 30.39$\pm$3.30, which is also higher than MVDream [73] initialized from the same SD v2.1 weight (30.22$\pm$3.83).

B.5. Classifier-free Guidance

Classifier-free diffusion guidance [33] is a technique used to balance the quality and diversity of images produced by diffusion models. This method is particularly effective in class-conditional and text-conditional image generation, enhancing both the visual quality of images and their alignment with given conditions. Inspired by [5] we explore the integration of classifier-free guidance with Epipolar Attention and Plücker Embedding. Implementing classifier-free guidance involves simultaneous training of the diffusion model for both conditional and unconditional denoising tasks. During inference, these models' score estimates are merged. We have four different types of conditioning injected into our system:

- Text (c_T): Injected from CLIP text-encoder similar to Vanilla Stable Diffusion.
- Camera (c_C): Injected with timestep via Residual blocks.
- Epipolar Attention (c_E): Injected by applying mask during self-attention.

Figure B.6. **Text-to-3D generation using multi-view Triplane generator with SPAD.** Following [35, 43] we trained a multi-view conditioned triplane generator that outputs a NeRF using four outputs of SPAD in a single feed-forward pass. We show the rendered NeRF on the top row (zoomed) and corresponding multi-view outputs from SPAD in the bottom row. For entire 360-degree videos see our website.

- Plücker Embedding (c_P): Injected by concatenation during self-attention.

During training, we extend classifier-free guidance over all these conditions. Therefore, our modified score estimate during inference is as follows:

$$\begin{aligned}\tilde{e_\theta}(z_t, c_T, c_C, c_E, c_P) &= e_\theta(z_t, \varnothing, \varnothing, \varnothing, \varnothing) \\ +s_T \cdot (e_\theta(z_t, c_T, \varnothing, \varnothing, \varnothing) &- e_\theta(z_t, \varnothing, \varnothing, \varnothing, \varnothing)) \\ +s_C \cdot (e_\theta(z_t, c_T, c_C, \varnothing, \varnothing) &- e_\theta(z_t, c_T, \varnothing, \varnothing, \varnothing)) \\ +s_E \cdot (e_\theta(z_t, c_T, c_C, c_E, \varnothing) &- e_\theta(z_t, c_T, c_C, \varnothing, \varnothing)) \\ +s_P \cdot (e_\theta(z_t, c_T, c_C, c_E, c_P) &- e_\theta(z_t, c_T, c_C, c_E, \varnothing))\end{aligned}$$

Outcome: As shown in Fig. B.13, we find that classifier-free guidance beyond text conditioning does not provide additional benefits, and rather leads to over-saturated generations. This also aligns with our observations on MVDream.

B.6. Joint Multi-View Inference

Concurrent multi-view diffusion models [46, 73] are limited to generating the same number of views they were trained on during testing. However, generating a high-quality 3D asset by e.g. training a NeRF model usually requires more than ten views of the asset. A naive solution is to use more views during training, which leads to quadratically increasing training costs due to the use of 3D self-attention. Instead, we propose a joint multi-view inference technique, which enables generating an infinite number of views using a model trained with fewer views.

Assume that we want to generate M views with a two-view model. We first initialize M noise maps $\{\boldsymbol{x}_T^i\}_{i=1}^M$, and then iteratively denoise all possible pairs of views, i.e.

$$(\boldsymbol{x}_{t-1}^i, \boldsymbol{x}_{t-1}^j) = \text{Denoise}(\boldsymbol{x}_t^i, \boldsymbol{x}_t^j, \epsilon_\theta) \tag{6}$$

for all $i, j \in [1, M]$ with $i \neq j$. Since the model is only trained on both views with the same noise level (i.e., timestep t), we sample (i, j) pairs without replacement and make sure to go over all possible combinations uniformly via simple heuristics.

Outcome: We find that this experiment trades off 3D consistency, as it only allows cross-view communication between two views at any given timestep of generation.

B.7. Fréchet Inception Distance (FID) Results

Compared to Vanilla MV-DM with an FID score of 55.25, our full model SPAD achieves a better FID score of 52.77 which shows further evidence of improvement in 2D generation quality.

FID Comparison with MVDream. Since our model generates images at random views, it has a much larger pose distribution mismatch compared to MVDream which uses

Figure B.7. **Text-to-3D generation using multi-view SDS with SPAD.** We adopt the multi-view SDS proposed in MVDream [73] to train a NeRF model. Thanks to the 3D consistency of our model, we do not suffer from the multi-face Janus issue.

orthogonal (90-degree varying) views in both ground-truth and generated images. Due to this reason, our FID cannot be compared directly with MVDream (trained with v2.1) which is reported to be 32.06 in the original work.

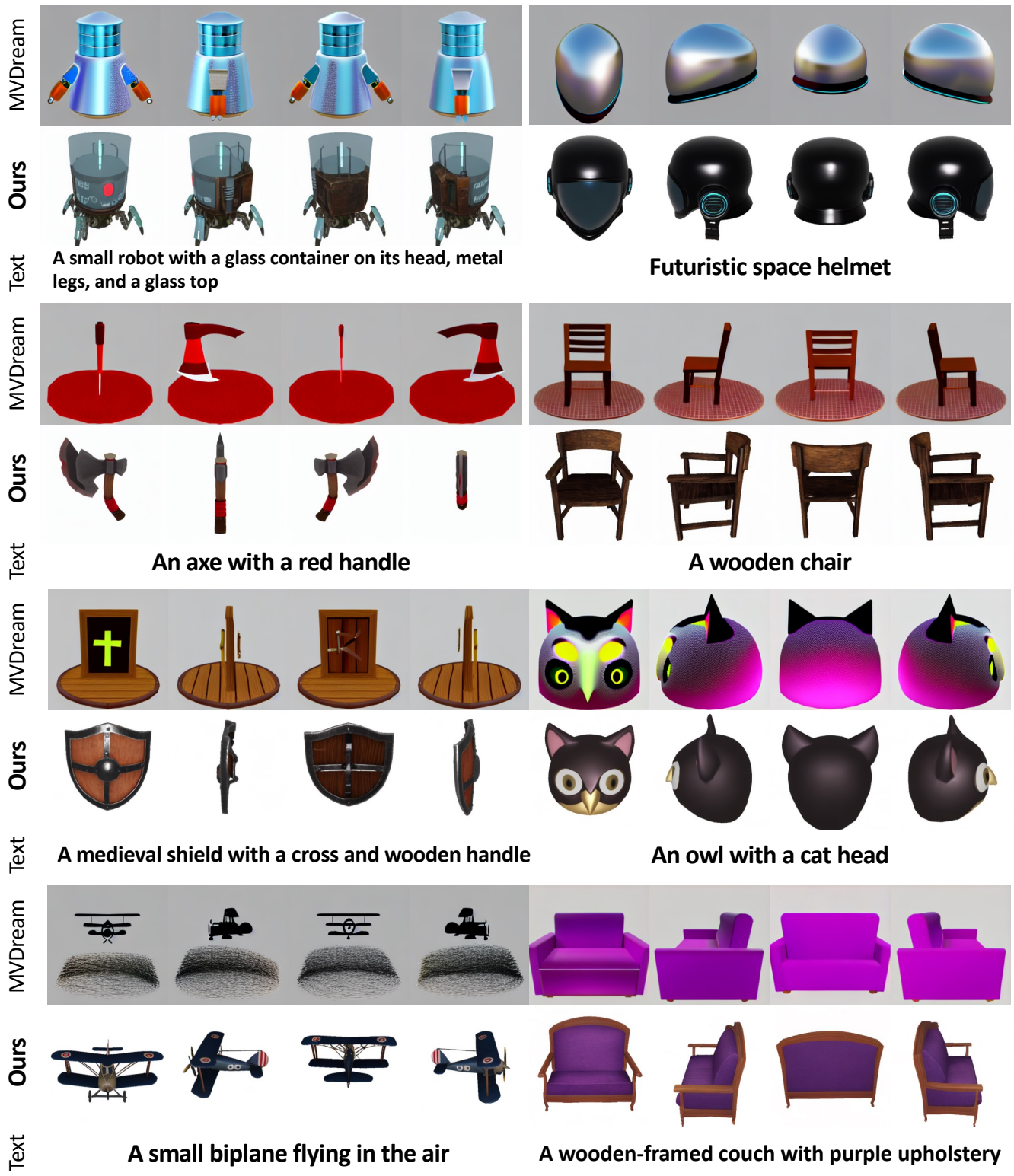

Figure B.8. **Comparison of text-conditioned multi-view generation with MVDream [73].**

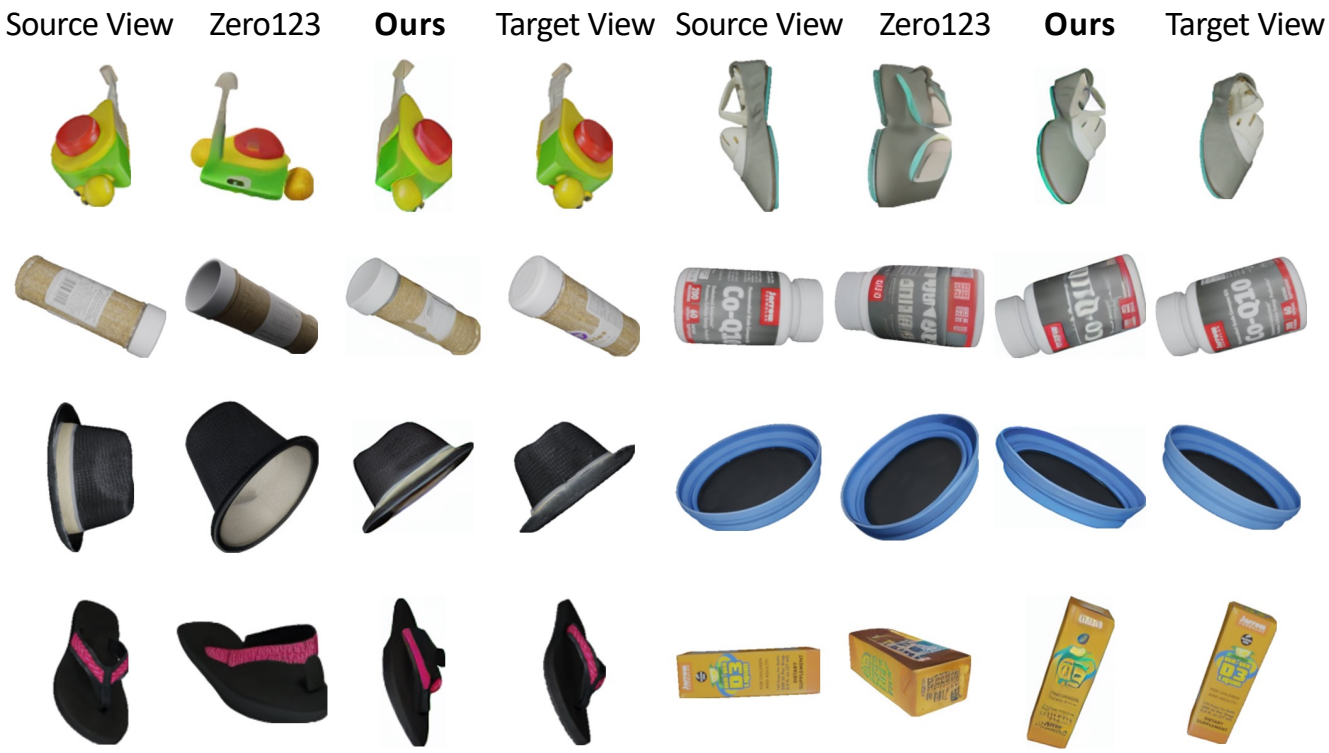

Figure B.9. **Comparison of image-conditioned novel view synthesis with Zero123 [45].**

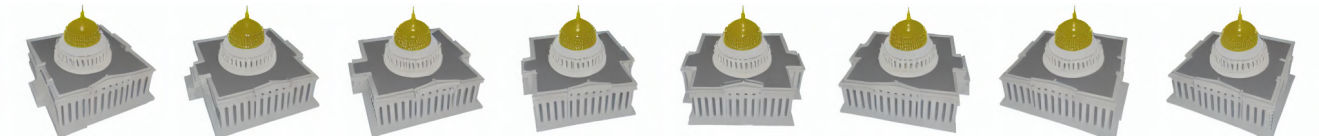

Figure B.10. **Close-view generation results from SPAD**. We generate images at continuous viewpoints with an offset of 10 degrees.

Figure B.11. **More multi-view generation results with SPAD.** The tested model is initialized with the weight of Stable Diffusion v1.5, and fine-tuned on Objaverse rendered images (same as Fig. 1 in the main paper).

Figure B.12. **More multi-view generation results with SPAD.** The tested model is initialized with the weight of Stable Diffusion v2.1, and fine-tuned on Objaverse rendered images. Compared to results in Fig. B.11 which adopts the weight of Stable Diffusion v1.5, this model is able to follow more complicated text prompts.

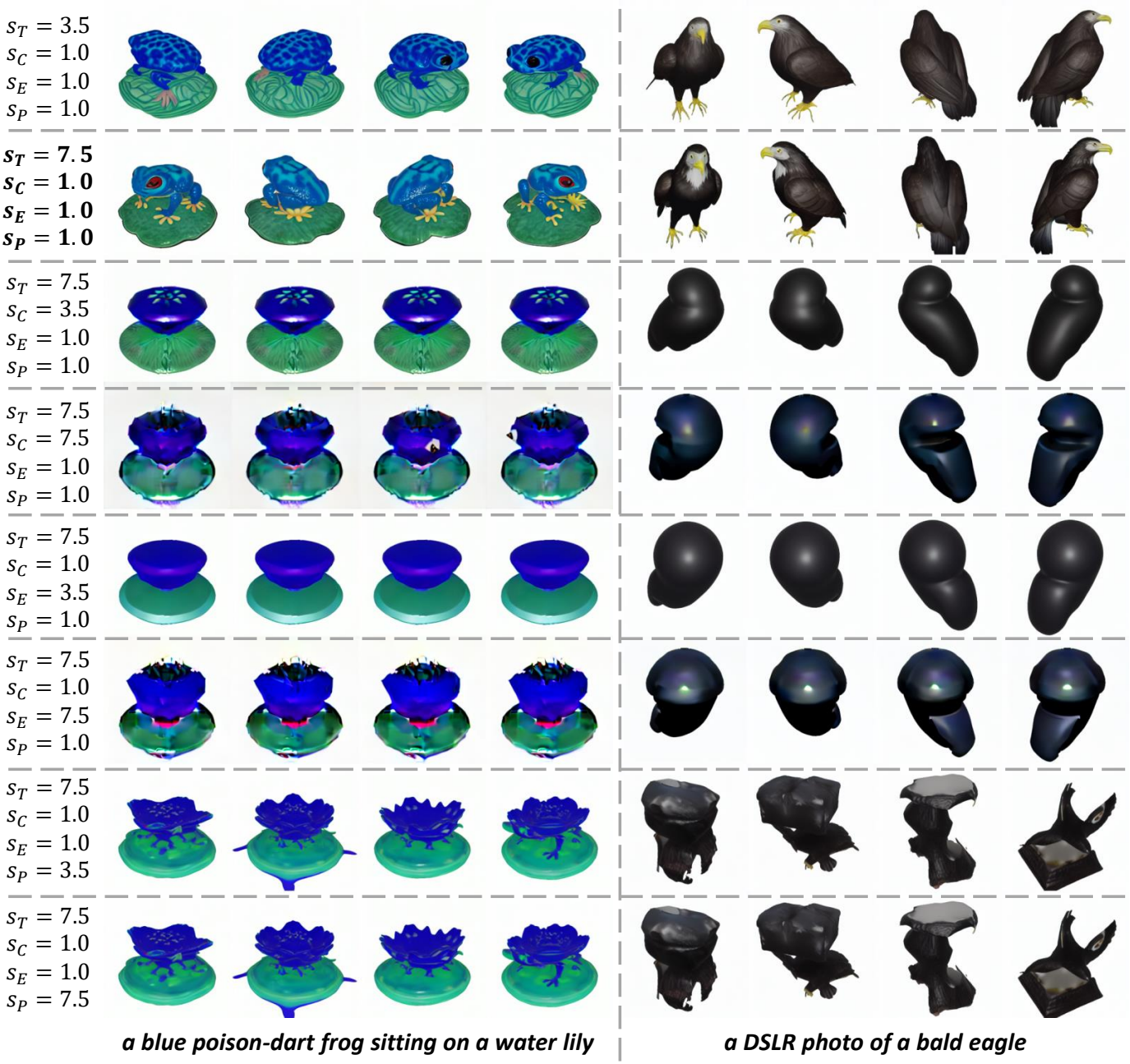

Figure B.13. **Ablation study regarding the classifier-free guidance scales.** Using a large scale of $s_T = 7.5$ for text conditioning works the best (row 2), while increasing scales for camera embedding, Epipolar Attention, and Plücker Embedding all leads to over-saturated images.