Coin3D: Controllable and Interactive 3D Assets Generation with Proxy-Guided Conditioning

Wenqi Dong[1,2*†] Bangbang Yang[2*] Lin Ma[2] Xiao Liu[2] Liyuan Cui[1]
Hujun Bao[1] Yuewen Ma[2] Zhaopeng Cui[1‡]
[1]Zhejiang University [2]ByteDance

Figure 1: Coin3D allows users to add 3D-aware control to the object generation using coarse proxies assembled from basic shapes, enabling an interactive generation workflow with fine-grained part editing and responsive 3D previewing.

ABSTRACT

As humans, we aspire to create media content that is both freely willed and readily controlled. Thanks to the prominent development of generative techniques, we now can easily utilize 2D diffusion methods to synthesize images controlled by raw sketch or designated human poses, and even progressively edit/regenerate local regions with masked inpainting. However, similar workflows in 3D modeling tasks are still unavailable due to the lack of controllability and efficiency in 3D generation. In this paper, we present a novel controllable and interactive 3D assets modeling framework, named Coin3D. Coin3D allows users to control the 3D generation using a coarse geometry proxy assembled from basic shapes, and introduces an interactive generation workflow to support seamless local part editing while delivering responsive 3D object previewing within a few seconds. To this end, we develop several techniques, including the 3D adapter that applies volumetric coarse shape control to the diffusion model, proxy-bounded editing strategy for precise part editing, progressive volume cache to support responsive preview, and volume-SDS to ensure consistent mesh reconstruction. Extensive experiments of interactive generation and editing on diverse shape proxies demonstrate that our method achieves superior controllability and flexibility in the 3D assets generation task. Code and data are available on the project webpage: https://zju3dv.github.io/coin3d/.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Artificial intelligence**.

KEYWORDS

AI-based 3D modeling, generative model

1 INTRODUCTION

As a child, we are born with the instinct to create things with our imagination, i.e., building houses or vehicles using different Lego bricks, or doodling pictures with pencils [Nath and Szücs 2014]. Yet only a few people learn drawing or modeling skills, which eventually develop the ability to create qualified artworks. Fortunately, the rapid development of generative techniques grants everyone a chance to create fantasy content, i.e., using LLM for automatic manuscripting [Radford et al. 2018; Wei et al. 2022] or 2D diffusion methods for text-to-image/video generation [Guo et al. 2023; Lugmayr et al. 2022; Rombach et al. 2022]. To enable the controllability of the generative model, recent advances in 2D diffusion (such as ControlNet [Zhang et al. 2023], T2I-Adapter [Mou et al. 2023], SDEdit [Meng et al. 2021], and etc.) allow users to take

*Wenqi Dong and Bangbang Yang contributed equally to this work.
†Wenqi Dong conducted this work during his internship at PICO, ByteDance.
‡Corresponding authors.

depth, sketches or human poses to control the generation process, enables iteratively editing designated region with inpainting and re-generating mechanism. However, in the field of 3D asset generation [Poole et al. 2022], existing 3D generative methods still lack controls for artistic creation. First, they are usually conditioned with text prompts [Poole et al. 2022] or perspective images [Liu et al. 2023d,c,a; Long et al. 2023; Qian et al. 2023], which is not sufficient to express 3D objects accurately. Second, when performing high-level tasks such as generative editing [Cheng et al. 2023b; Haque et al. 2023; Kamata et al. 2023; Li et al. 2023a] or inpainting [Zhou et al. 2023], existing approaches usually require a significant amount of time for reconstruction before previewing the editing operation.

Given this observation, we believe that a controllable and user-friendly 3D assets generation framework should have these properties. **(1) 3D-aware controllable**: similar to a child stacking Lego bricks and picturing the vivid appearance in its mind, a controllable 3D generation can be started by assembling basic shapes (e.g., cuboids, spheres, or cylinders), which serves as coarse shape guidance for the detailed generation. Therefore, it reduces the difficulty of 3D modeling for common users and also provides sufficient control over the generation. **(2) flexible**: the framework should allow users to interactively composite or adjust local regions in a 3D-aware manner, ideally as easy as image inpainting tools [Meng et al. 2021]. **(3) responsive**: once the user's editing is temporarily finished, the framework should instantly deliver preview images of the generated object from the desired viewpoints, rather than waiting for a long reconstruction period. In this paper, we propose a novel **CO**ntrollable and **IN**teractive 3D assets generation framework, named Coin3D. Instead of using text prompts or images as conditions, Coin3D allows users to add 3D-aware conditions into a typical multiview diffusion process in the 3D generation task, i.e., using a coarse 3D proxy assembled from basic shapes to guide the object generation, as illustrated in Fig. 1. Based on proxy-guided conditioning, Coin3D introduces a novel generative and interactive 3D modeling workflow. Specifically, users can depict the desired object by typing in text prompts and assembling basic shapes with their familiar modeling software (such as Tinkercad, Blender, and SketchUp). Then, Coin3D would construct the on-the-fly feature volume in a few seconds, which enables the preview of the result from arbitrary viewpoints or even progressively adjust/regenerate the designated local part of the object. For example, we can generate a bronze car by assembling basic shapes and incrementally adding tubes or changing tires as shown in Fig. 1. However, even though adding 3D-aware conditions is technically plausible, there are still some challenges to an interactive 3D modeling workflow, which will be addressed in this work:

Coarse Shape Guidance. Since we only use simple basic shapes (e.g., stacked spheres or cuboids) instead of intricate CAD models for 3D guidance, the proxy-guided conditioning should allow some freedom during the generation rather than being strict to the given basic shapes, e.g., growing animal ears from the sphere head as shown in Fig. 1. To achieve this goal, we design a novel 3D adapter to process the 3D control, where the 3D proxies (basic shapes) are first voxelized and extracted to 3D features, and then integrated into the spatial features of a multiview generation pipeline [Liu et al. 2023a]. In this way, users can manipulate the control strength by changing the plug-in weights, enabling controlling the generated object more or less close to the given proxy.

Interactive Modeling. An interactive and productive 3D generation workflow should support progressive modeling operations and responsive preview, i.e. seamlessly adding/adjusting shape primitives or precisely regenerating local parts without touching others, while all the operated results should be previewed as quickly as possible without time-consuming reconstruction. To fulfill the demands, we first develop a novel proxy-bounded editing strategy, which ensures precise bounded control and natural style blending when modifying part of the object, and then utilize a progressive volume caching mechanism by memorizing stepwise 3D features to enable responsive preview.

Consistent Reconstruction. To facilitate the standard CG workflow, one might need to export the generated assets into textured mesh with reconstruction. However, even with 3D-aware conditioning, there might still be poor reconstructions when naïvely reconstructing objects using synthesized multiview images due to the limited viewpoints (see Sec. 4.4). To tackle this issue, we propose to leverage the proxy-guided feature volume during reconstruction with a novel volume-SDS loss. This strategy effectively exploits the controlled 3D context during the score distillation sampling [Poole et al. 2022] and faithfully improves the reconstruction quality.

Our contributions can be summarized as follows. **1)** We propose a novel controllable and interactive 3D assets generation framework, named Coin3D. Our method designs a 3D-aware adapter to take simple 3D shape proxies as guidance to control the object generation, which supports interactive generation operations such as altering prompts, adjusting shapes, or fine-grained local part regeneration. **2)** To ensure an interactive and consistent experience of generative 3D modeling, we develop several techniques, including proxy-bounded editing for precise and seamless part editing, progressive volume cache to support responsive preview from arbitrary views, and a conditioned volume-SDS to improve the mesh reconstruction quality. **3)** Extensive experiments of interactive generation with various shape proxies and the interactive workflow deployed on the 3D modeling software (e.g., Blender) demonstrate the controllability and productivity of our method on generative 3D modeling.

2 RELATED WORKS

2.1 3D Object Generation

3D object generation is a popular task in computer vision and graphics. Early works [Achlioptas et al. 2018; Dubrovina et al. 2019; Kluger et al. 2021] mainly focus on natively generating 3D representations from models, such as polygon meshes [Gao et al. 2022; Groueix et al. 2018; Kanazawa et al. 2018; Nash et al. 2020; Wang et al. 2018], pointclouds [Achlioptas et al. 2018; Fan et al. 2017; Nichol et al. 2022; Yu et al. 2023], parametric models [Hong et al. 2022; Jiang et al. 2022], voxels [Choy et al. 2016; Sanghi et al. 2022; Wu et al. 2017; Xie et al. 2019], or implicit fields [Chan et al. 2022, 2021; Cheng et al. 2023a; Gu et al. 2021; Jun and Nichol 2023; Li et al. 2023b; Mescheder et al. 2019; Park et al. 2019; Skorokhodov et al. 2022], which learn from specific CAD database [Chang et al. 2015] and are often bounded by specific categories (e.g., chairs, cars,

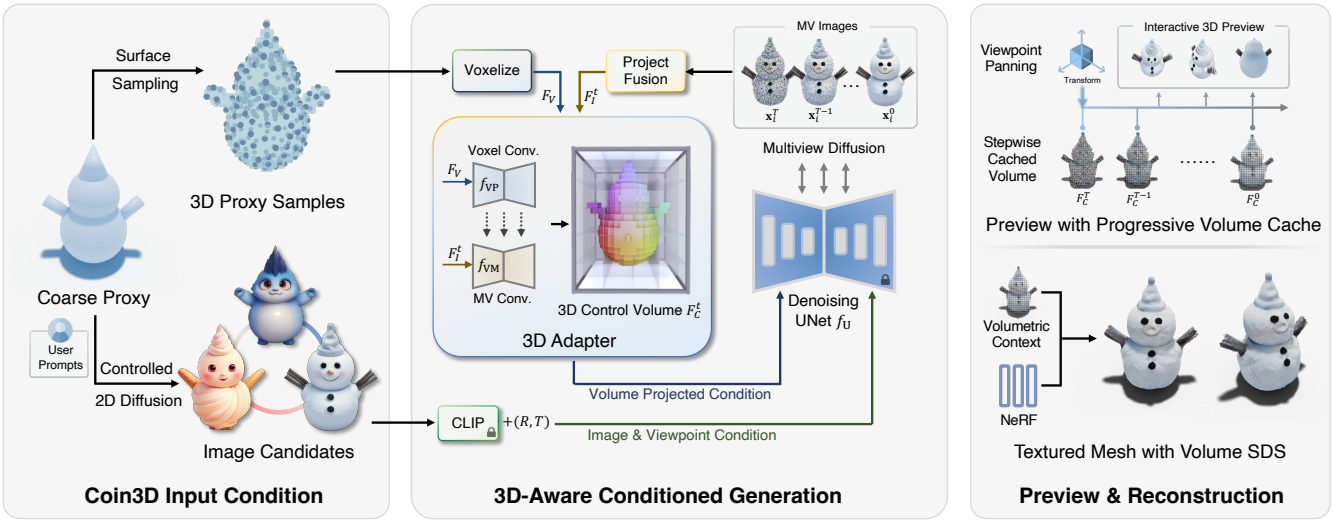

Figure 2: Overview. Given a coarse shape proxy and user prompts that describe the identity, our method first constructs 2D image candidates from the proxy's silhouette and 3D proxy samples as input conditions. Then, we employ a 3D adapter to integrate 3D-aware control to the diffusion's denoising process with a 3D control volume F_C, yielding multiview images of the object. By fully leveraging F_C, we realize accelerated 3D previewing with volume cache and also improve mesh reconstruction quality.

and etc.) due to the limited network capacity and data diversity. Recently, with the rapid evolution in large-scale generative models, especially the great success in 2D diffusion models [Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022], methods like DreamFusion [Poole et al. 2022], SJC [Wang et al. 2023a] and their follow-up works [Chen et al. 2023a; Lin et al. 2023; Melas-Kyriazi et al. 2023; Raj et al. 2023; Seo et al. 2023; Tang et al. 2023a,b; Xu et al. 2023c] attempt to distill 2D gradient priors from the denoising process using score distillation sampling loss (SDS loss) or its variants, which guide the per-shape neural reconstruction following users' text prompts. While being generic to unlimited categories and diverse composited results with prompt engineering, these lines of work often suffer from unstable convergence due to the noisy and inconsistent gradient signal, which often leads to incomplete results or "multi-face Janus problem" [Chen et al. 2023a]. Subsequently, Zero123 [Liu et al. 2023c] analyzes the viewpoint bias problem of the generic 2D latent diffusion model (LDM), and proposes to train an object-specific LDM with relative viewpoint as a condition using Objaverse dataset [Deitke et al. 2023], which shows promising results in the image-to-3D tasks, and has been widely adopted in the follow-up 3D generation works [Liu et al. 2023d; Qian et al. 2023]. While being fine-tuned on multiview images, Zero123 still suffers from the cross-view inconsistency issue as its resulting images cannot satisfy the requirements for reconstruction. Hence, later works such as MVDream [Shi et al. 2023b], SyncDreamer [Liu et al. 2023a], Zero123++ [Shi et al. 2023a] and Wonder3D [Long et al. 2023] propose to enhance multiview image generation, which either trains with stacked views [Long et al. 2023; Shi et al. 2023a,b] or builds synchronized volumes online to condition the diffusion process [Liu et al. 2023a], and usually enables to produce highly consistent images or yields 3D reconstructions in few seconds. Very recently, LRM [Hong et al. 2023] and its variant methods [Wang et al. 2023b; Xu et al. 2023b] propose to train an end-to-end transformer-based model, which directly produces neural reconstruction given one or few perspective images. Nevertheless, existing 3D object generation methods primarily focus on using text prompts (text-to-3D) or images (image-to-3D) as the input, which cannot accurately convey exact 3D shapes or precisely control the generation in a 3D manner. By contrast, our method first adds 3D-aware control to the multiview diffusion process without compromising generation speed, which realizes interactive generation workflow with 3D proxy as conditions.

2.2 Controllable and Interactive Generation

Adding precise control to the generative methods is crucial for productive content creation [Bao et al. 2023; Epstein et al. 2022; Yang et al. 2022a, 2024, 2022b, 2021]. Previous generative works [Bao et al. 2024; Chen et al. 2022; Deng et al. 2023; Hao et al. 2021; Melnik et al. 2024] mainly learn a latent mapping of the attributes to add control to the generation, but are limited to specific categories (e.g., human faces or nature landscape). Recent progress in 2D diffusion models, such as ControlNet [Zhang et al. 2023] and T2I-Adapter [Mou et al. 2023], enables various 2D image hints (e.g., depth, normal, soft-edge, human poses, color grids, and etc.), to interactively control the denoising process of the image generation. However, similar controllable capabilities [Bhat et al. 2023; Cohen-Bar et al. 2023; Pandey et al. 2023] in 3D generation are far from applicable. For generative 3D editing, recent works [Cheng et al. 2023b; Li et al. 2023a] propose to constrain the text-driven 3D generation at the desired region, but cannot support controlling the exact geometry shape. For Controllable 3D generation, the most related works

to our methods are Latent-NeRF [Metzer et al. 2023] and Fantasia3D [Chen et al. 2023a]. However, these two works cannot ensure steady convergence and the generated results are usually far from the given 3D shape (see Sec. 4.2), as they naïvely add control to the 3D representation regardless of altering the supervision of 2D priors (i.e., SDS loss). Other works such as Control3D [Chen et al. 2023b] only add control from 2D sketches/silhouettes instead of 3D space. Moreover, all these methods require a long time of reconstruction (e.g., from dozens of minutes to hours) to inspect the effect of editing or controlling, which cannot fulfill the demand for interactive modeling. On the contrary, our method directly integrates the 3D-aware control into the diffusion process, which not only ensures faithful and adjustable control over the 3D generation but also allows to interactively preview the generated/edited 3D object in a few seconds.

3 METHOD

We introduce Coin3D, a novel Controllable 3D assets generation framework, which adds 3D-aware control to the multiview diffusion process in object generation tasks, enabling an interactive modeling workflow for fine-grained customizable object generation. An overview of Coin3D is shown in Fig. 2. Instead of using conventional text prompts or a perspective image as a condition, our framework employs a coarse geometry proxy made from basic shapes (e.g., a snowman composed of two spheres, two sticks, and one cone), complemented by user prompts that describe the object's identity. Then, the diffusion-based generation will be conditioned on both a voxelized 3D proxy and 2D image candidates generated by controlled 2D diffusion with the proxy's silhouette (e.g., images with different appearances in the left bottom of Fig. 2). During the 3D-aware conditioned generation, we use a novel 3D adapter module that seamlessly integrates proxy-guided controls with adjustable strength into the diffusion pipeline (Sec 3.1). To deliver an interactive generation workflow with fine-grained 3D-aware part editing and the responsive previewing ability, we also introduce proxy-bounded editing to precisely control the volume update, and employ an efficient volume cache mechanism to accelerate the image previewing at arbitrary viewpoints (Sec 3.2). Furthermore, we propose a volume-conditioned reconstruction strategy, which effectively leverages the 3D context from feature volume to improve the reconstruction quality (Sec 3.3).

3.1 Proxy-Guided 3D Conditioning for Diffusion

3D proxy as initial condition. As illustrated in Fig. 2, our method uses a 3D coarse shape proxy assembled from basic elements (e.g., cubes, cylinders, cones, spheres, etc.) and user prompts to condition the multiview diffusion process. More specifically, given the coarse shape P and prompts y, we want to predict N_v consistent images $\{\mathbf{x}_i | i = 1, 2, \ldots, N_v\}$ under the camera poses $\{\mathbf{c}_i | i = 1, 2, \ldots, N_v\}$ using a multiview diffusion-based generator f as the following:

$$\mathbf{x}_{(i:N_v)} = f(P, y, \mathbf{c}_{(i:N_v)}). \tag{1}$$

Note that, unlike regular 2D diffusion, multiview diffusion synchronously performs denoising iterations on all the preset views, which integrates cross-view correlations with view-dependent self-attention [Long et al. 2023; Shi et al. 2023a] or spatial volume [Liu et al. 2023d,a]. To simplify the preparation of proxy shapes, our method allows the user to realize the input by simply scaling and assembling basic shapes in 3D modeling software (e.g., Tinkercad, SketchUp, or Blender) without relying on complex modeling skills. Hence, to adapt the coarse proxy inputs (i.e., a coarse polygon mesh) for 3D generation tasks, we develop a two-pathway condition preprocess for the proxy. First, we sample N_S surface points $\mathcal{P} = \{\mathbf{p}_i | i = 1, 2, \ldots, N_S\}$ on the proxy mesh, which will be used for the 3D-aware control for the generation pipeline. Second, we use the proxy's rendered silhouette and users' prompts as a condition, and generate multiple 2D image candidates for interactive appearance selection [Rombach et al. 2022; Zhang et al. 2023].

3D-aware control with 3D adapter. We introduce the 3D adapter to add 3D-aware control from coarse proxy samples to the multiview diffusion pipeline, which yields multiview images of the object following the given proxy shape. To achieve the lossless 3D control of the diffusion model, inspired by volumetric multiview diffusion works [Liu et al. 2023b,a], we construct a 3D control volume to add the 3D-aware context into the diffusion pipeline, where the volume is a voxel feature grid containing v^3 vertices. As shown in the middle part of Fig. 1, the 3D adapter receives two inputs from proxy feature volume F_V and the multiview image fused volume F_I^t. Specifically, F_V is constructed by first voxelizing the proxy samples $\mathcal{P}$ to fill in the zero-initialized occupancy grid, where each grid will be assigned to 1 if containing any point. F_I^t is the multiview feature volume constructed by unprojecting and fusing multiview images $\mathbf{x}_{(i:N)}^t$ produced by the denoising UNet f_{U} at timestamp t. F_I^t is the multiview feature volume, which is constructed by first projecting vertices of V onto the multi-view images $\mathbf{x}_{(i:N)}^t$ to obtain interpolated image-plane features, and then fusing them with a 3D CNN module. $\mathbf{x}_{(i:N)}^t$ are produced from the denoising process of UNet f_{U} at timestamp t. Then, for each denoising step t in the 3D adapter, we first perform 3D convolution (with 3D UNet f_{VP}) on the volume F_V, and hierarchically add the intermediate layer outputs to the 3D convolution (with 3D UNet f_{VM}) of multiview feature volume F_I^t, which yields the final 3D control volume F_C^t. Then, during the multiview denoising of the 2D diffusion model, we first project the F_C^t to align the corresponding view to obtain 2D feature map F_p^t [Yao et al. 2018], and then integrate the F_p^t (with depthwise attention) along with CLIP-embedded candidate image feature and viewpoint embedding [Liu et al. 2023c] to the Zero123's diffusion UNet [Liu et al. 2023c].

Training 3D adapter with proxy samples. To train the 3D adapter with coarse proxy as conditions, we preprocess each training object item into several multiview images and uniformly sampled points on the surface. For each training step, we randomly select B conditioning and target images with the corresponding point samples, and also B timestamp with the Gaussian noise $\epsilon^{(1:B)} \sim \mathcal{N} \in (0, 1)$. We enforce the network to predict the added noises following [Rombach et al. 2022; Song et al. 2020], which is defined as:

$$\min_{\theta} \mathbb{E}_{t, \mathbf{x}_{(1:N_v)}, \epsilon_{(1:N_v)}} \|\epsilon_i - \epsilon_\theta(\mathbf{x}_i^t, t, c(I, F_C^t, \mathbf{c}_i))\|, \tag{2}$$

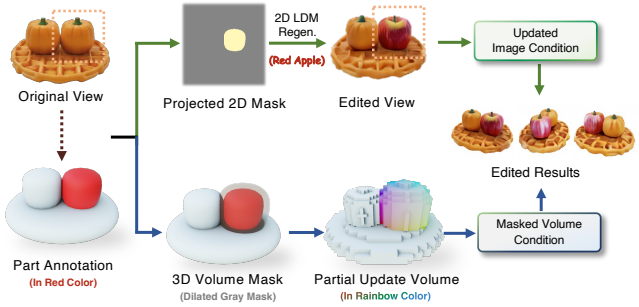

Figure 3: Proxy-bounded part editing. We update the 2D image condition and 3D control volume with masks from users' part annotation of the proxy.

where ϵ_θ is the model predicted noise, $c(I, F_C^t, \mathbf{c}_i)$ is the conditioned embedding of candidate image I, 3D control volume F_C^t and camera view $\mathbf{c}_i$. During the training procedure, we use zero convolution [Zhang et al. 2023] for the proxy feature volume convolution UNet f_{VP} while freezing other layers, which enables manipulating control strength during the generation.

3.2 Interactive Generation Workflow

In 3D modeling tasks, artists are likely to adjust the target object back and forth, and progressively edit the local part for satisfactory results. However, the interactive generation and previewing for 3D objects remains an open problem due to the lack of fine-grained controlling ability and slow reconstruction speed [Cheng et al. 2023b; Li et al. 2023a]. Hence, we develop a novel interactive and responsive generation workflow upon the Coin3D framework, which fully leverages the piecewise proxies of the condition for easy and precise part editing, and reuses 3D control volume for interactive previewing.

Proxy-bounded part editing. As the coarse proxies are mainly constructed with basic shape elements, we design an interactive local part editing workflow based on the elements in the proxy. Specifically, users can specify a certain piece from the basic shapes, and regenerate the piece content. For example, we can regenerate one of the pumpkins into a red apple by designating the sphere on the plate, as shown in Fig. 3. However, because the multiview diffusion model is both conditioned on 3D volume and 2D images, it is not trivial to realize the editing regardless of the complete conditions. Therefore, we propose a two-pathway condition editing scheme that considers both 2D and 3D conditions, as illustrated in Fig. 3. For 2D conditions, we construct a 2D mask by projecting masked proxies at the desired editing view and perform diffusion-based 2D regenerating (a.k.a. masked image-to-image inpainting) [Meng et al. 2021; Zhou et al. 2023] with the mask. We then use the edited image as the image condition for the denoising steps. For 3D conditions, we first construct a 3D feature mask by slightly dilating the masked proxy, which ensures seamless fusion of the newly generated content. Then, during each denoising step, we reuse the cached original 3D control volume and only partially update the unmasked volume according to the feature mask M, as:

$$\hat{F}_C^t = (1 - M)F_C^t + M\tilde{F}_C^t, \tag{3}$$

where $\hat{F}_C^t$ is the updated volume by fusing cached volume F_C^t and predicted volume $\tilde{F}_C^t$ at t. By enabling proxy-bounded masks on both 2D and 3D conditions, we can precisely edit the local part at the original object while preserving other parts unchanged.

Interactive preview with progressive volume caching. To ensure a smooth experience for interactive generation, we want to preview the editing results in a few seconds and inspect the edited effect from arbitrary viewpoints. Hence, we design a progressive volume caching mechanism, which memorizes the latest 3D control volume for each timestamp t. Then, during the preview stage, we transfer the user's viewpoint spanning poses $\mathbf{c}'$ inside the modeling software to the viewpoint condition and volume projection in multiview diffusion. To make the preview responsive, we use the cached 3D control volume without re-running the 3D adapter, and instantly decode [Bohan 2023] the preview image for each step.

3.3 Volume-Conditioned Reconstruction

The outcome of the diffusion model is a set of multiview images of the object, so we need to reconstruct it to 3D representation (e.g., using NeuS [Wang et al. 2021]) for CG applications. However, naïvely reconstructing with multiview images is sub-optimal and might result in unexpected geometry due to limited viewpoints (see Sec. 4.4). Therefore, we integrate 3D-aware context from the 3D control volume F_C^t to the reconstruction stage, which improves the reconstruction quality. Specifically, we propose a volumetric-based score distillation sampling, called volume-SDS, which integrates the 3D control prior from the voxelized feature F_C^t to the field's back-propagation as the following:

$$\Delta_x L_{V-SDS} = w(t)(\epsilon_\theta(\mathbf{x}_t, t, c(I, F_C^t, \mathbf{c})) - \epsilon), \tag{4}$$

where $w(t)$ is the weighting function [Poole et al. 2022]. In this way, the reconstruction can be decently supervised by 3D control signals to achieve better mesh quality (see Sec. 4.4). Please refer to the the supplementary material for more details of the reconstruction.

4 EXPERIMENTS

We first compare our method with image-based 3D generation in Sec. 4.1, and compare with controllable 3D object generation methods in Sec. 4.2. Then, we show the interactive generation applicability with designated part editing in Sec. 4.3. Finally, we perform ablation studies to analyze the design of our framework in Sec. 4.4.

4.1 Comparison on Proxy-based and Image-based 3D Generation

So far, the most stable 3D object generation pipelines are mainly image-based, i.e., giving a single image as a conditioning input, and then generating multiview images for reconstruction [Liu et al. 2023a; Long et al. 2023; Shi et al. 2023a] or direct 3D representations [Hong et al. 2023]. Unlike these methods, we use a coarse shape proxy as a guidance through the entire interactive generation pipeline. Since all these methods use image conditions to bootstrap

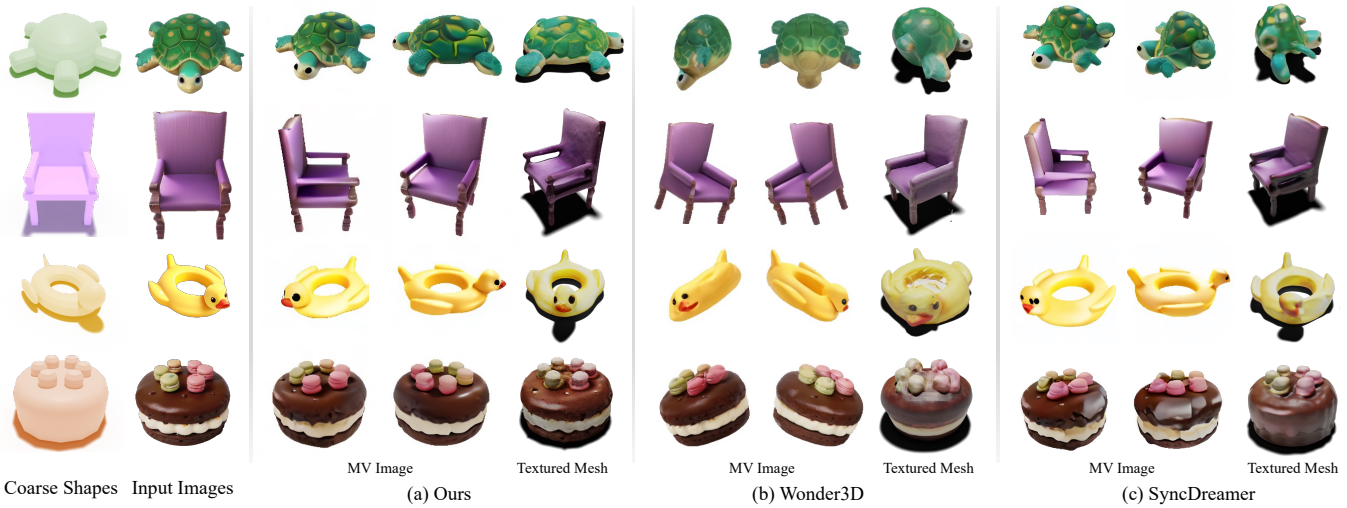

Figure 4: We compare our proxy-based generation method with image-based methods (i.e., Wonder3D [Long et al. 2023] and SyncDreamer [Liu et al. 2023a]) on the generated multiview images and reconstructed textured mesh.

Methods	Quantitative Metrics			User Study	
	CLIP Score ↑	ImageReward ↑	GPTEvals3D ↑	Matching Degree ↑	Recon. Quality ↑
Proxy-Based vs. Image-Based 3D Generation					
Wonder3D [Long et al. 2023]	0.251	-0.557	980	1.613	1.770
SyncDreamer [Liu et al. 2023a]	0.260	-0.152	962	1.654	1.594
Ours	**0.266**	**0.026**	**1035**	**2.733**	**2.634**
Controllable 3D Object Generation					
Fantasia3D [Chen et al. 2023a]	0.212	-1.597	810	1.267	1.273
Latent-NeRF [Metzer et al. 2023]	0.246	-1.188	1146	1.930	1.918
Ours	**0.249**	**-0.749**	**1204**	**2.801**	**2.809**

Table 1: We perform quantitative evaluation and user studies on the 3D generation task.

the diffusion model, we first compare our method with SOTA image-based generation methods (i.e., Wonder3D [Long et al. 2023] and SyncDreamer [Liu et al. 2023a]) using the same image candidates, where our method also add extra coarse shapes as conditioning.

Qualitative comparison. We show the multiview images and the reconstructed textured meshes in Fig. 4. As shown in Fig. 4, the predicted views and the textured meshes from Wonder3D and SyncDreamer both have some artifacts (e.g., distorted green turtle and yellow swimming ring at the first and third row in Fig. 4 (b) (c), missing hollowed handrail and short legs at the second row in Fig. 4 (b), missing white creamy middle layer at the fourth row in Fig. 4 (c)). Thanks to the proxy-guided conditioning and volume-conditioned reconstruction, our method can synthesize multiview images free of single view ambiguity by complementing 3D context from the proxy (e.g., complete chairs with correct hollowed handrail in Fig. 4 (a)), and also consistently reconstruct 3D objects with intact shape and vivid appearance.

Quantitative comparison. We use CLIP score [Radford et al. 2021] to evaluate the text-object matching degree, and ImageReward [He et al. 2023; Xu et al. 2023a] and GPTEvals3D [Wu et al. 2024] to evaluate the perceptual quality of the predicted multiview images. As presented in Table 1, our method achieves the overall best metrics, demonstrating that adding proxy-based conditioning can improve the quality of 3D generation tasks. Note that Wonder3D's ImageReward score is lower than SyncDreamer's due to the evaluator's bias of orthogonal image views, while their Elo scores [Elo 1967] evaluated by GPTEvals3D are comparable.

User study. We also conduct a user study to compare our method with others. Following TEXTure [Richardson et al. 2023], we ask 30 users to sort 35 testing examples in random order based on the perceptual quality and content matching degree (w.r.t the given image or text prompts), and assign the scores by their ranking (i.e., with a score of 3 for the ordered best one and a score of 1 for the last one). As reported in Table 1, our method achieves the best score among all the methods. More details can be found in the supplementary material.

4.2 Comparison on Controllable 3D Object Generation

We compare our method with Controllable 3D object generation methods, including Latent-NeRF [Metzer et al. 2023] and Fantasia3D [Chen et al. 2023a]. Latent-NeRF introduces a sketch-shape guided loss, which constrains the density field close to the surface of the shape proxy, while Fantasia3D uses coarse shapes as the geometry initialization of DMTet [Shen et al. 2021]. In the experiment, we give all the methods the same coarse shape and the text prompts as a guidance and output the neural reconstructions' rendered views for comparison (since Latent-NeRF does not officially provide mesh extraction). As shown in Fig. 4, Latent-NeRF generally obtains plausible results but fails to produce tiny shape and appearance details (e.g., blurry textured sofa, fox missing clear eyes and right arms and missing sunflower in Fig. 5 (b)), which indicates that directly applying 3D control to the 3D representation is sub-optimal since it might not work smoothly with the generic SDS loss. For Fantasia3D, since it only uses the 3D shape for initialization rather than supervision, it often generates overgrowth results that do not

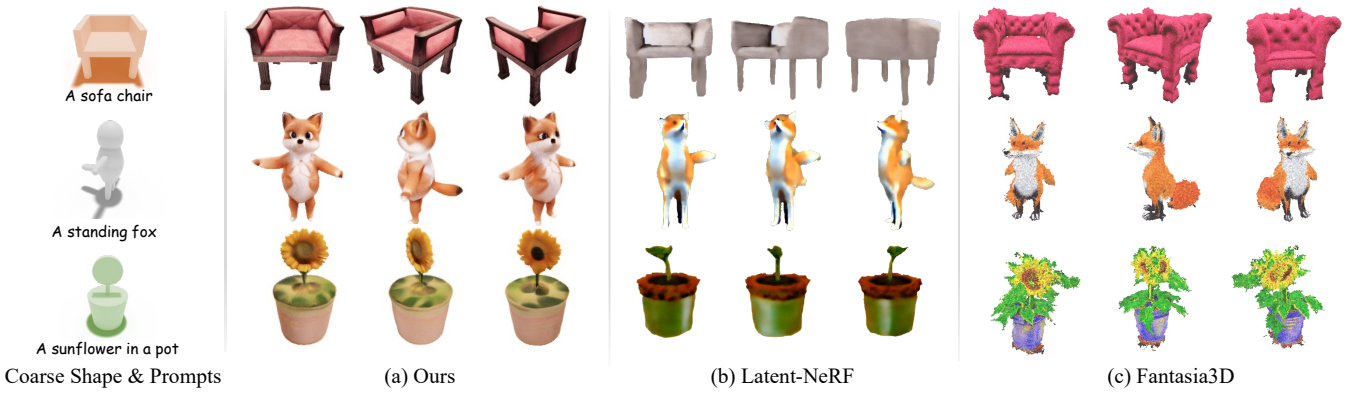

Figure 5: We compare the Controllable 3D generation with Latent-NeRF [Metzer et al. 2023] and Fantasia3D [Chen et al. 2023a].

Figure 6: We conduct interactive generation with part editing on several basic shape proxies.

follow the given shape (e.g., inflate sofa, the fox with incomplete downward arms, pot with many leaves and a broken sunflower in Fig. 5 (c)) and also slightly suffers from "multi-face Janus problem" (e.g., multi-face fox in Fig. 5 (c)). Since both Latent-NeRF and Fantasia3D use vanilla 2D diffusion model as prior while being agnostic to the multi-view correlations, their results are sensitive to the initialization and hyperparameter settings. In contrast, our method directly adds 3D-aware control to the diffusion process, which essentially controls the supervisory of reconstruction's 2D diffusion prior and consistently achieves high fidelity generation following users' shape guidance. It is also noteworthy that, both Latent-NeRF and Fantasia3D require a long period of reconstruction (e.g., dozens of minutes) to give an impression of what the object might look like, making it unusable for interactive modeling, while our framework bypasses the reconstruction stage and allows to preview the 3D object in only a few seconds.

4.3 Interactive Generation with Part Editing

We now present examples of interactive generation with progressive part editing. As shown in Fig. 6, users can first generate a basic instance (e.g., a base of cake, a teddy bear, or a penguin) with shape proxy, and then progressively add new shape blocks with changed text prompts (e.g., adding a small cake with candles, a green hat and red scarf, or even progressively add a torch and a red backpack from the back view), which seamlessly enrich the content of the instance while maintaining other parts unchanged. Notably, all these editing operations can be finished in roughly 5~10 seconds, which then allows interactive previewing of the edited 3D results. Please refer to the supplementary video for the demonstration.

Figure 7: We inspect the efficacy of volume-conditioned reconstruction.

4.4 Ablation Studies

Volume-conditioned reconstruction. We then inspect the efficacy of the volume-SDS loss by ablating during the shape reconstruction. As shown in Fig. 7, by adding volume-SDS loss, we achieve better geometry reconstruction (e.g., less floater and more reasonable chair bottom) than naïve training on fixed multiview images [Liu et al. 2023a; Long et al. 2023].

Proxy-bounded part editing. We finally analyze the proxy-bounded part editing by ablating proxy conditioning and mask dilation strategy in Fig. 8. Specifically, we choose a multi-step editing example,

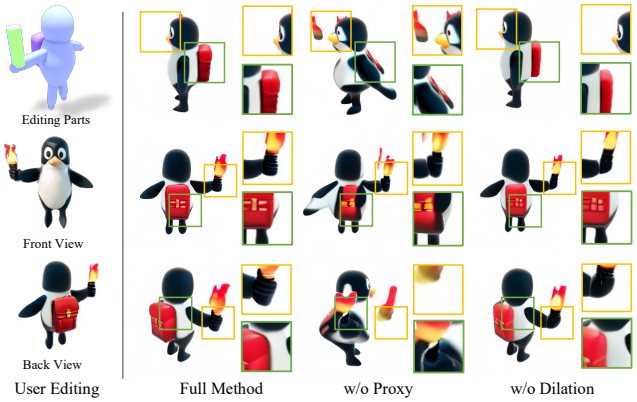

Figure 8: We analyze the importance of proxy guidance and 3D mask dilation in proxy-bounded part editing.

where the backpack should be edited from the back view. We merge the front and edited back image conditions using mixed denoising [Bar-Tal et al. 2023]. As shown in Fig. 8, the editing without the proxy condition would result in broken shapes (e.g., dangling flames and distorted body), while disabling mask dilation would also make the editing less natural (e.g., tightly fused red bag and broken hands). By equipping with the full strategies, we achieve a seamless editing effect while preserving other content unchanged.

More ablation studies can be found in the supplementary material.

5 CONCLUSIONS

We have proposed a novel 3D assets generation pipeline, named Coin3D. Our method successfully integrates 3D-aware control from coarse shape proxies to the 3D object generation task and enables an interactive generation workflow, where users can freely alter prompts/shapes or regenerate designated local parts, and inspect the changes with interactive 3D preview in a few seconds.

Limitations and future works. First, our workflow starts from synthesizing 2D image candidates, which provides users with quick preview and selection but requires prompt engineering to obtain a clean and satisfactory result without complex background textures. In the future, we can finetune a 2D diffusion model with object-centric data [Deitke et al. 2023] and introduce LLM-based prompt enhancement [Gustavosta 2023] to handle this issue. Second, due to the limited resolution of the base diffusion model [Liu et al. 2023c], our method cannot produce fine-level details (e.g., complex fur textures or wrinkled surface), which can be further improved by adopting better backbones [Shi et al. 2023a] or taking the refinement stage with high-resolution optimization [Lin et al. 2023; Tang et al. 2023a]. Third, our reconstructed texture meshes already baked lighting effects while lacking PBR materials for modern rendering pipelines. In future work, we can train a material-disentangled diffusion model to enable generating objects with PBR materials.

ACKNOWLEDGMENTS

We would like to acknowledge the support from the NSFC (No. 62102356), Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. We also express our gratitude to all the anonymous reviewers for their professional and constructive comments. The authors from Zhejiang University are also affiliated with the State Key Laboratory of CAD&CG.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. 2018. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*. PMLR, 40–49.

Chong Bao, Yinda Zhang, Yuan Li, Xiyu Zhang, Bangbang Yang, Hujun Bao, Marc Pollefeys, Guofeng Zhang, and Zhaopeng Cui. 2024. GeneAvatar: Generic Expression-Aware Volumetric Head Avatar Editing from a Single Image. *arXiv preprint arXiv:2404.02152* (2024).

Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2023. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20919–20929.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. (2023).

Shariq Farooq Bhat, Niloy J Mitra, and Peter Wonka. 2023. LooseControl: Lifting ControlNet for Generalized Depth Conditioning. *arXiv preprint arXiv:2312.03079* (2023).

Ollin Boer Bohan. 2023. Tiny AutoEncoder for Stable Diffusion. https://github.com/madebyollin/taesd. Accessed: 2023-10-03.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16123–16133.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5799–5809.

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

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873* (2023).

Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. 2023b. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1148–1156.

Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2022. Sem2nerf: Converting single-view semantic masks to neural radiance fields. In *European Conference on Computer Vision*. Springer, 730–748.

Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. 2023b. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784* (2023).

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. 2023a. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4456–4465.

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 628–644.

Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. 2023. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2920–2929.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.

Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 2023. 3d-aware conditional image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4434–4445.

Anastasia Dubrovina, Fei Xia, Panos Achlioptas, Mira Shalah, Raphaël Groscot, and Leonidas J Guibas. 2019. Composite shape modeling via latent space factorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8140–8149.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life* 22, 8 (1967), 242–247.

Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A Efros. 2022. Blobgan: Spatially disentangled scene representations. In *European Conference on Computer Vision*. Springer, 616–635.

Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 605–613.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems* 35 (2022), 31841–31854.

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 216–224.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

Gustavosta. 2023. MagicPrompt. https://huggingface.co/Gustavosta/MagicPrompt-Stable-Diffusion. Accessed: 2023-10-03.

Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. 2021. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14072–14082.

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789* (2023).

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. 2023. T^3Bench: Benchmarking Current Progress in Text-to-3D Generation. arXiv:2310.02977 [cs.CV]

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400* (2023).

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–11.

Heewoo Jun and Alex Nichol. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463* (2023).

Hiromichi Kamata, Yuiko Sakuma, Akio Hayakawa, Masato Ishii, and Takuya Narihira. 2023. Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion. *arXiv preprint arXiv:2303.15780* (2023).

Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 371–386.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Florian Kluger, Hanno Ackermann, Eric Brachmann, Michael Ying Yang, and Bodo Rosenhahn. 2021. Cuboids revisited: Learning robust 3d shape fitting to single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13070–13079.

Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. 2023b. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12642–12651.

Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. 2023a. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608* (2023).

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2023b. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885* (2023).

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

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023c. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *arXiv preprint arXiv:2309.03453* (2023).

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2023. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008* (2023).

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8446–8455.

Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makaravets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. 2024. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12663–12673.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023).

Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. 2020. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning*. PMLR, 7220–7229.

Swiya Nath and Dénes Szücs. 2014. Construction play and cognitive skills associated with the development of mathematical abilities in 7-year-old children. *Learning and instruction* 32 (2014), 73–80.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022).

Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy Mitra. 2023. Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D. *arXiv preprint arXiv:2312.02190* (2023).

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508* (2023).

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721* (2023).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18603–18613.

Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. 2023. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937* (2023).

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. 2023a. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110* (2023).

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2023b. MVDream: Multi-view Diffusion for 3D Generation. *arXiv:2308.16512* (2023).

Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. 2022. Epigraf: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems* 35 (2022), 24487–24501.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023a. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653* (2023).

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023b. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184* (2023).

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023a. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12619–12629.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*. 52–67.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. 2023b. PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction. *arXiv preprint arXiv:2311.12024* (2023).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. 2017. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems* 30 (2017).

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. 2024. GPT-4V (ision) is a Human-Aligned Evaluator for Text-to-3D Generation. *arXiv preprint arXiv:2401.04092* (2024).

Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2690–2698.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023a. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977 [cs.CV]

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023c. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20908–20918.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. 2023b. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217* (2023).

Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. 2022a. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*. Springer, 597–614.

Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. 2024. Dreamspace: Dreaming your room space with text-driven panoramic texture propagation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 650–660.

Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. 2022b. Neural rendering in a room: amodal 3d understanding and free-viewpoint rendering for the closed scene composed of pre-captured objects. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–10.

Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2021. Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13779–13788.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*. 767–783.

Wang Yu, Xuelin Qian, Jingyang Huo, Tiejun Huang, Bo Zhao, and Yanwei Fu. 2023. Pushing the Limits of 3D Shape Generation at Scale. *arXiv preprint arXiv:2306.11510* (2023).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Xingchen Zhou, Ying He, F Richard Yu, Jianqiang Li, and You Li. 2023. RePaint-NeRF: NeRF Editting via Semantic Masks and Diffusion Models. *arXiv preprint arXiv:2306.05668* (2023).

Supplementary Material

In this supplementary material, we describe more details of our method in Sec. A. Besides, we also conduct more experiments in Sec. C. More qualitative results can be found in our supplementary video, and the source code will be released upon the acceptance of this paper.

A IMPLEMENTATION DETAILS

A.1 Dataset Preparation

In our experiment, we use the LVIS subset of Objaverse [Deitke et al. 2023] to train the model, which contains 28,000+ objects after a heuristic cleanup process following Long et al. [Long et al. 2023]. For training view rendering, we set up 16 image views with -30° pitch and evenly facing towards the object from 360°.

A.2 Evaluation Data Preparation and User Study

For quantitative comparison, we produced 30 testing examples (coarse shapes and users' prompts) for each experiment (Sec. 4.1 and Sec. 4.2). Then, for each example, we generate four images at four poses $\{\mathbf{c}_i | i = 1, 5, 9, 13\}$ from 16 evenly distributed viewpoints, and calculates CLIP score [Radford et al. 2021], ImageReward [He et al. 2023; Xu et al. 2023a] and GPTEvals3D [Wu et al. 2024] average Elo scores for each standalone view. For user studies, we prepare 35 examples and merge the output images of each method into one image. Then we ask 30 participants to sort the merged images. In the comparison on proxy-based and image-based 3D generation, we merged four multiview images and four textured mesh rendering images into one. In the comparison on controllable 3D object generation, we merge four rendering images into one, since LatentNeRF is difficult to extract the textured mesh.

A.3 Training and Network Details

During the initialization of the training, we follow Zhanget al. [Zhang et al. 2023] to keep the multiview convolution network weights f^{VM} fixed, and use a dual 3D UNet structure to implement the 3D feature Adapter with trainable copy initialization strategy. Specifically, our proxy is first voxelized at a resolution of $32 \times 32 \times 32$, where the value of each voxel would be assigned to 1 if there is any occupied 3D point inside the voxel. After that, the features will be up-convolution to 64 channels through two layers of 3D convolution. For the training of the 3D adapter, we sample 256 points on each object surface as a coarse proxy, and train the model at 256×256 resolution. The learning rate is 0.00005 and the batch size is set to 8, with 100K training iterations. The total training process of the 3D Adapter takes about two days on an Nvidia A100-80G graphic card.

During the textured mesh reconstruction stage, we use NeuS [Wang et al. 2021] as the neural representation. The total loss function is defined as the following:

$$L = L_{rgb} + L_{V-SDS} + L_{mask} + R_{eik} + R_{sparse} + R_{smooth} \quad (5)$$

where L_{rgb} is the L_2 loss between the generated multiview images $\mathbf{x}_{(i:N_v)}$ and images rendered under the camera poses $\mathbf{c}_{(i:N_v)}$. L_{V-SDS} is the volume-SDS loss proposed in the main paper. L_{mask} is the BCE mask loss of the rendered opacity, where the mask M is obtained with existing methods [Kirillov et al. 2023]. R_{eik} is the Eikonal loss that regularize the magnitude of the SDF gradients of each sample point to be unit length. R_{sparse} and R_{smooth} are respectively the sparse term used to reduce the floater of SDF and the smooth term used to smooth the 3D surface. For the 3D reconstruction of each instance, we takes about 5 minutes on a single NVidia A100-80G graphic card.

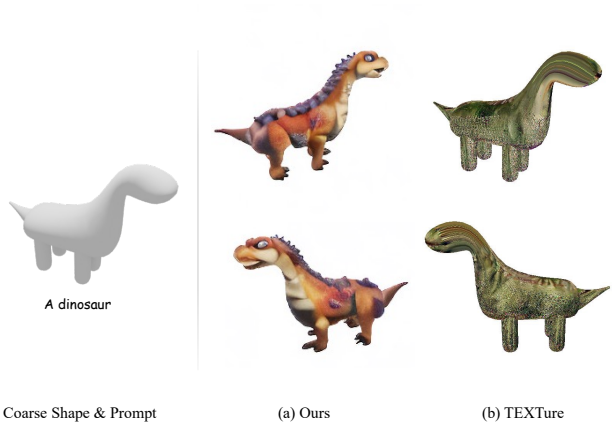

Figure I: We compare our method with the texture synthesis method TEXTure [Richardson et al. 2023].

Figure J: We compare our proxy-bounded part editing with Fantasia3D [Chen et al. 2023a] fine-tuning.

B MORE DISCUSSIONS

B.1 Necessity of Proxy-guided 3D Generation

For personalized generation demands, we think only using text / images is insufficient and also unintuitive for expressing 3D structures of objects and their spatial relationships. Hence, granting

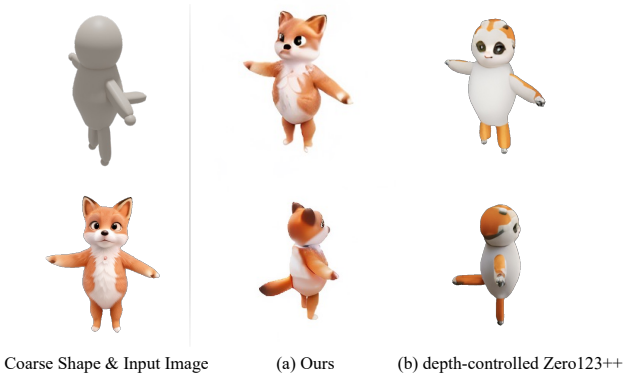

Figure K: We compare our method with the depth-controlled 3D generation pipeline (Zero123++ [Shi et al. 2023a]).

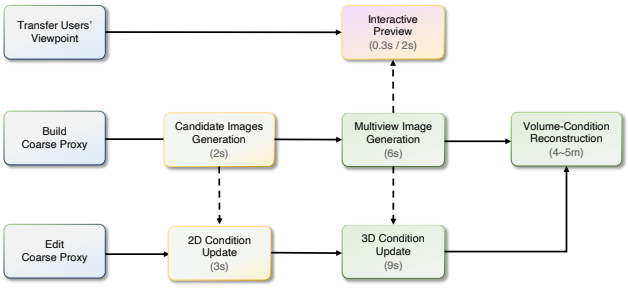

Figure L: Runtime overview of interactive generation.

system 3D-aware controllability with 3D proxy is necessary for 3D generation. As for the acquisition of 3D proxies, we believe this is not an obstacle for target users, as it can be assembled easily using kids' software like Tinkercad, taken from 3D modeling games from SteamVR, or using LLM+procedural modeling instructions. Similarly, ControlNet uses control images from raw sketches to delicate line art, which also requires basic painting skills.

B.2 More Limitations

First, the resolution of 3D-aware control is bounded by the size of the proxy feature volume, which cannot fully leverage control from complex high-poly models. For example, we cannot generate a large-scale urban scene with satisfactory building details. Second, our method requires manual tuning control strength to balance between over-constrained and under-constrained, which is also similar to ControlNet [Zhang et al. 2023] as the control strength mainly depends on the creators' aesthetic choices.

C MORE EXPERIMENTS

C.1 Runtime Evaluation

As shown in Fig. L, given a 3D proxy and text prompt, Coin3D first takes 2s to generate candidate images and 6s to generate 3D-aware conditioned multiview images. During previewing, Coin3D

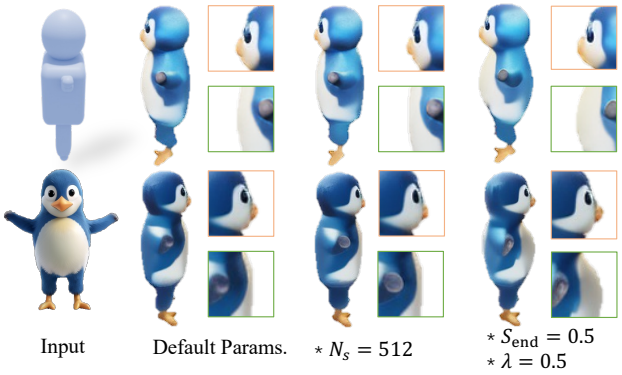

Figure M: We analyze the 3D-aware controlling strength with different control parameters. As shown above, increasing proxy samples would add more shape constraints, while setting lower weights would give more freedom of generation.

responds to the camera rotating instantly (<0.3s) and takes 2s for convergence (baseline without volume-cache requires 6s). For interactive editing, Coin3D takes 3s to update the 2D condition and 9s to update the feature volume, which is then ready for previewing. As a comparison, existing editable 3D generation methods take much longer for feedback, e.g., ~1h for Progressive3D and FocalDreamer, ~0.5h for Fantasia3D. Finally, Coin3D takes 4~5m (600 iterations) to reconstruct and export the textured mesh (see Fig. 1 and Fig. 4 from the main paper for "proxies vs. textured meshes").

C.2 3D-Aware Controlling Strength

We analyze the adjustable 3D-aware controlling strength with different control parameters under the fixed seed in Fig. M and Fig. M. For the default parameters, we set the number of proxy samples $N_s = 256$ and fully unlocked weight control ($\lambda = 1.0$ and $S_{\text{end}} = 1.0$) through the whole diffusion process. As a comparison, we set the $N_s = 512$ or set partial weight control ($\lambda = 0.5$ and $S_{\text{end}} = 0.5$, i.e., only half weight applied and disable weights for the last half of the denoising steps), where λ is the weight when adding intermediate outputs of f_{VP} to f_{VM}, and S_{end} is the ending step of 3D-aware controlling. We found that increasing the number of proxy samples would add more constraints to make it close to the given shape, while setting lower weights would give more freedom to the network to predict a curved shape.

C.3 Interactive Editing vs. Fine-Tuning Fantasia3D

Since Fantasia3D supports fine-tuning on the pre-trained mesh, it can also conduct interactive editing by updating text prompts. Therefore, we compare our interactive editing method with Fantasia3D fine-tuning. We use the user-guided generation method mentioned in Fantasia3D. The first stage performs geometry optimization and texture optimization based on the initial shape, and the second stage uses the first stage's optimized object but modifies the text prompt during optimization. During the experiment, we first use a flat cylinder as the initial input shape, and use text

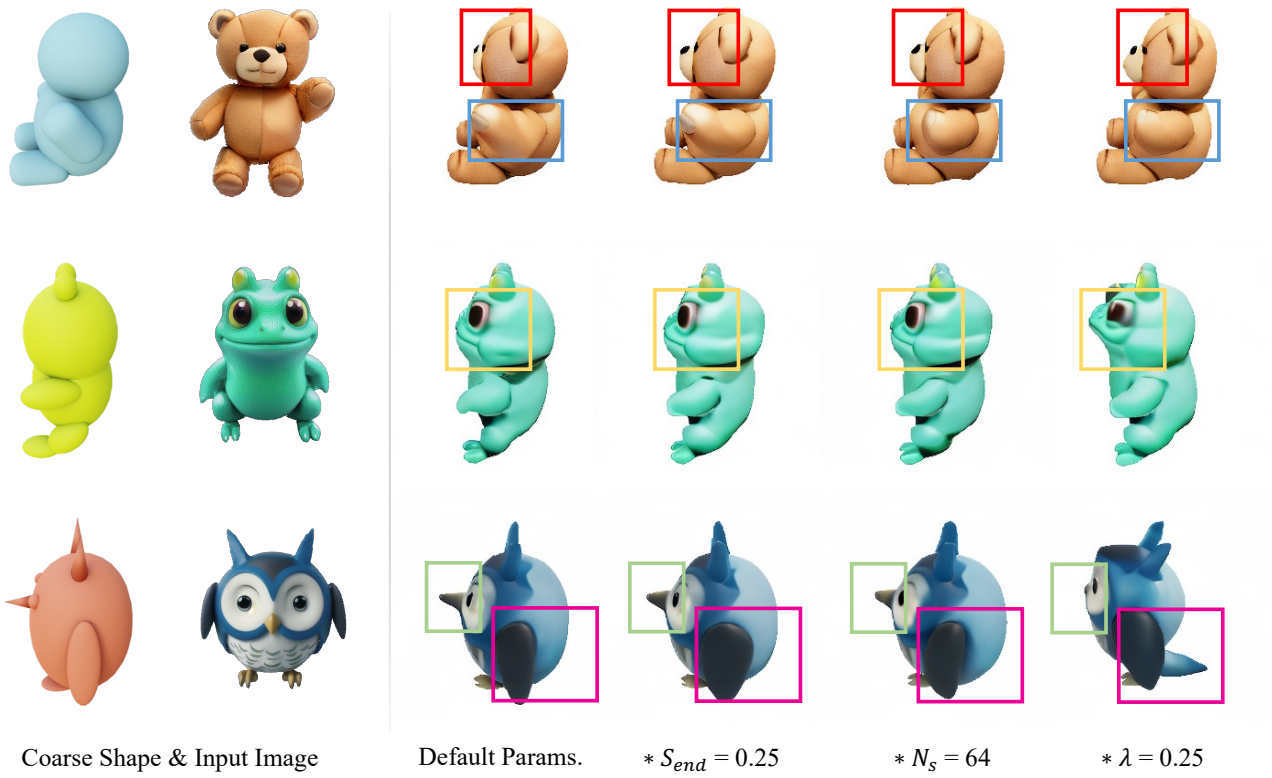

Figure M (cont.)**: We analyze the 3D-aware controlling strength with different control parameters. As shown above, setting lower weights would give more freedom of generation.**

prompt "a birthday cake base" to guide the optimization. Second, we add candles by modifying the proxy shape and updating the text prompt to "a birthday cake with three candles on top" and As shown in Fig. J, our method successfully generates appealing cakes with candles, while Fantasia3D fails to achieve a reasonable result (e.g., almost no complete candle on the top).

C.4 Proxy-Based 3D Generation vs. Texture Synthesis

We also compare our method with texture synthesis work TEXTure [Richardson et al. 2023], which generates UV textures given the corresponding geometry. As shown in Fig. I, when given the same coarse proxy for generation, TEXTure tends to generate tightly bounded textures on the given mesh, resulting in blurry appearances and invisible facial features of the dinosaur. In contrast, our method allows a certain degree of freedom during the generation, which gracefully synthesizes the dinosaur with vivid facial details. The experiment demonstrates that the proxy-based 3D generation is far beyond the texture synthesis task, as it requires the method to generate more details upon the coarse proxy shape.

C.5 Proxy-Based 3D Generation vs. Depth-Controlled 3D Generation

We compare our proxy-based 3D generation with the depth-controlled 3D generation pipeline from Zero123++ [Shi et al. 2023a], where we feed the Zero123++ with the multiview depth maps rendered from the coarse shape proxy. As shown in Fig. K, even only given a coarse shape of the animal with no ears, our method still generates cute animal ears upon the simple shape, while Zero123++ can only synthesize novel views of the object that tightly fit to the coarse shape proxy with poor facial details. This demonstrates that simply using 2D depth control as a condition in multiview generation cannot achieve the ideal coarse 3D control ability like ours, which further proves the value of adding 3D-aware control in a 3D manner.

C.6 User Studies of Ablation Studies

We selected 10 examples of the ablation studies in Sec. 4.4 and asked 24 users to judge whether the proposed strategies improve the results. The statistic shows 78% of users believe volume-SDS improves the quality, 75% of users think 3D mask dilation makes editing more natural, 96% of users find the proxy helpful in maintaining shape integrity.

C.7 User Study of 3D Interaction of Proxy-based 3D Generation

We also conducted a user study of proxy-based 3D generation. We show users the process of making coarse shapes in Blender, as well as the generated multi-view images and 3D reconstruction results. We asked each participant to rate three questions: (a) the difficulty of using 3D modeling tools; (b) overall satisfaction with the effectiveness of our approach; (c) willingness to use our methods, on a scale of 1 to 5, where 1 in (a) means easy to use, 5 in (b) means satisfied with the effectiveness, and 5 in (c) means willing to use. The score of (a) is 2.38, (b) is 4.62, and (c) 4.46, which indicates that most of the participants consider the difficulty of coarse shape modeling is acceptable and are willing to use our method.

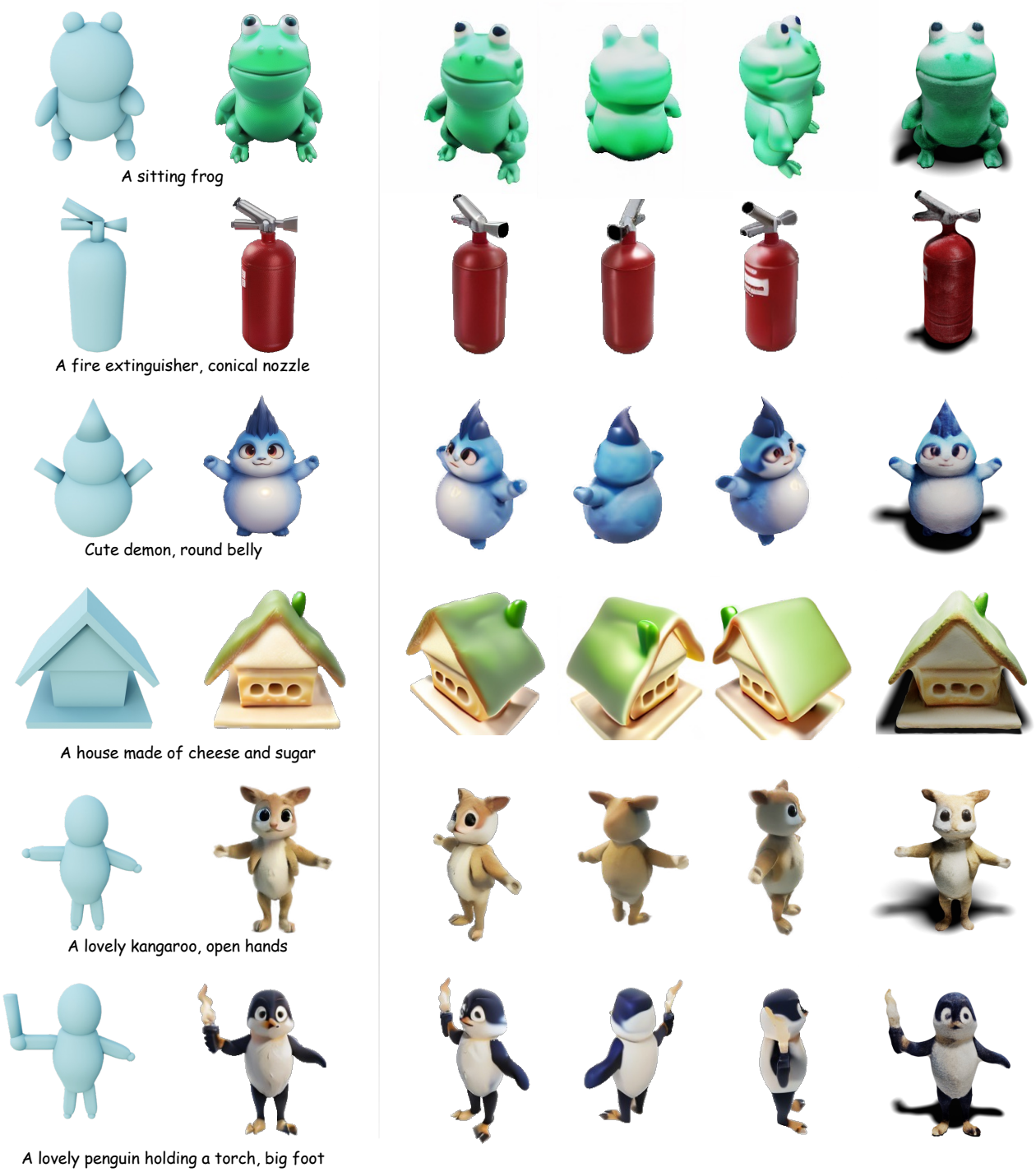

Figure N: More examples of controllable 3D object generation.

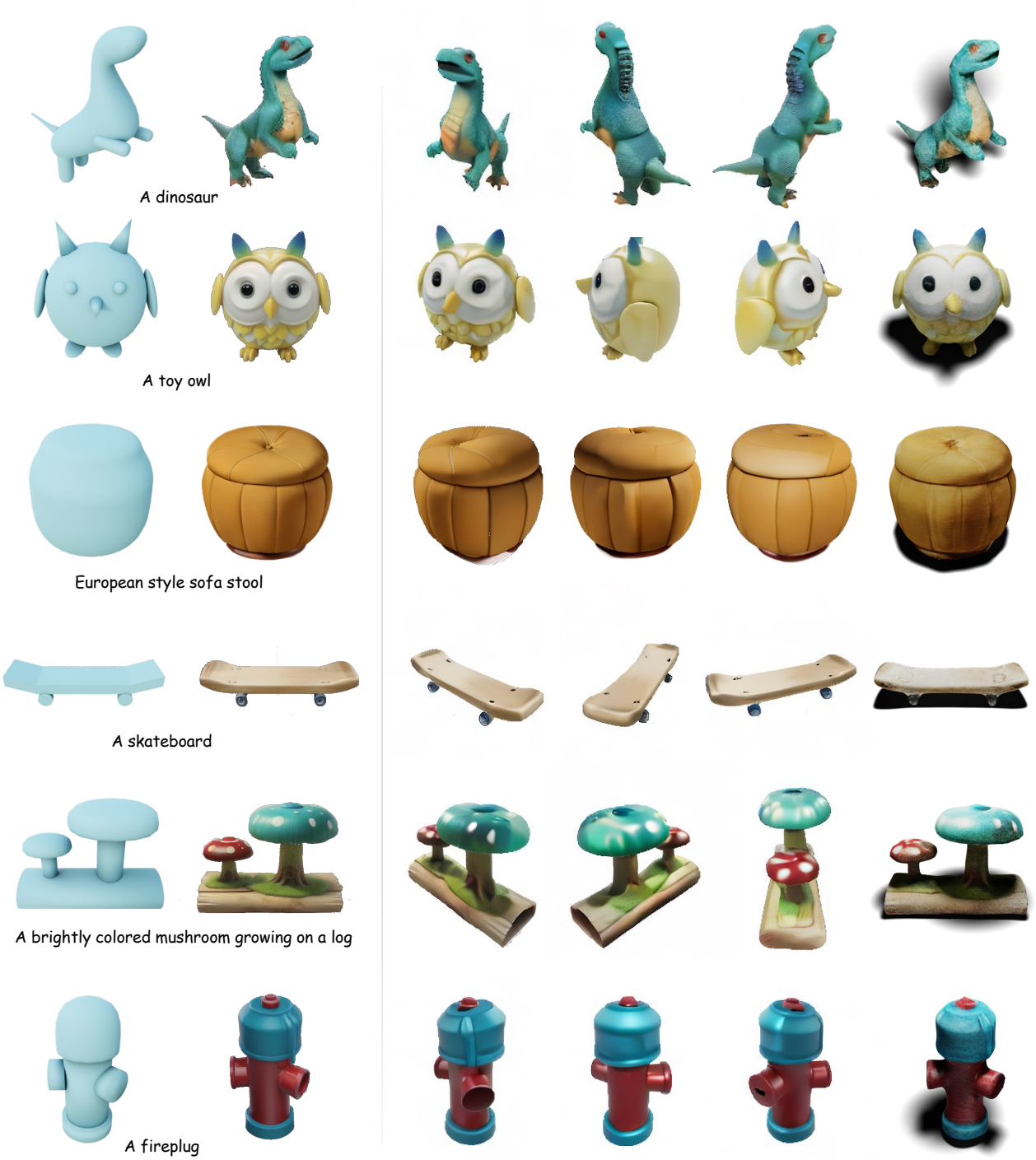

Figure N (cont.)**: More examples of controllable 3D object generation.**