Make-Your-3D: Fast and Consistent Subject-Driven 3D Content Generation

Fangfu Liu, Hanyang Wang, Weiliang Chen, Haowen Sun, and Yueqi Duan*

Tsinghua University
{liuff23,hanyang-21,chen-wl20,shw22,duanyueqi}@mails.tsinghua.edu.cn

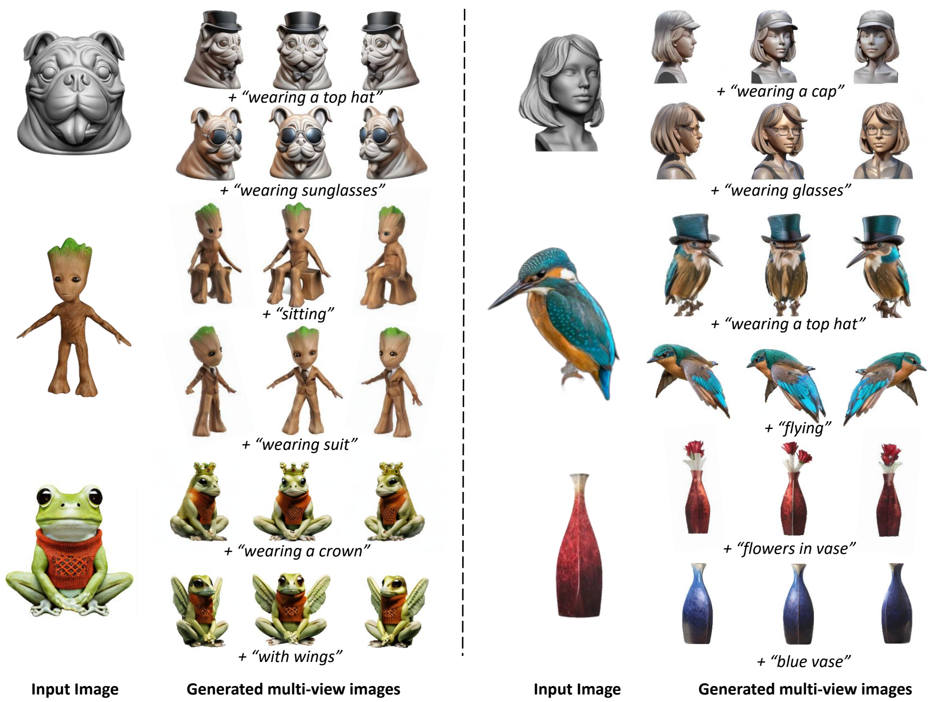

Fig. 1: **Make-Your-3D** can personalize 3D contents from only a **single image** of a subject with text-driven modifications within only **5 minutes**. We show the results from three different perspectives.

Abstract. Recent years have witnessed the strong power of 3D generation models, which offer a new level of creative flexibility by allowing users to guide the 3D content generation process through a single image or natural language. However, it remains challenging for existing 3D generation methods to create subject-driven 3D content across diverse prompts. In this paper, we introduce a novel 3D customization method, dubbed **Make-Your-3D** that can personalize high-fidelity and

* Corresponding author.

consistent 3D content from only a single image of a subject with text description within 5 minutes. Our key insight is to harmonize the distributions of a multi-view diffusion model and an identity-specific 2D generative model, aligning them with the distribution of the desired 3D subject. Specifically, we design a co-evolution framework to reduce the variance of distributions, where each model undergoes a process of learning from the other through identity-aware optimization and subject-prior optimization, respectively. Extensive experiments demonstrate that our method can produce high-quality, consistent, and subject-specific 3D content with text-driven modifications that are unseen in subject image. Project page: `https://liuff19.github.io/Make-Your-3D/`.

Keywords: 3D Generation · Personalization · Fast Speed

1 Introduction

Subject-driven customization has emerged as a prominent aspect within the field of generative models [40, 44, 60, 65, 67], providing a wide range of multimedia applications [61, 69]. It aims to synthesize the individual subject in diverse contexts and styles while retaining high fidelity to subject-specific identities. Despite the great progress of personalization in text-to-image (T2I) [2, 7, 44, 45, 66] and text-to-video (T2V) [32, 42, 58, 60, 62] models, the exploration of customized 3D generation remains relatively limited.

Driven by the advancements in neural 3D representations [20, 27, 34], extensive datasets [4, 8, 9], and the diffusion-based generative models [15, 43], recent works [25, 29, 37, 47, 52, 57] have demonstrated high-quality automated 3D generation from text or image prompt. Although text or image prompts allow for some degree of control over the generated 3D asset, it remains challenging to produce high-fidelity and subject-specific 3D content with text-driven modifications such as novel colors, poses, or attributes that are unseen in any of the input subject images. Enabling the creation of subject-specific 3D assets through flexible text controls would greatly simplify the workflow for artists and streamline 3D acquisition processes. One notable attempt for subject-driven 3D generation is DreamBooth3D [40], which combines a personalizing model (DreamBooth [44]) and a text-to-3D model (DreamFusion [37]) with two-stage tuning for DreamBooth to optimize NeRF [34] representation. However, it has two inherent limitations: (a) time-consuming optimization of NeRF representation with the two fine-tuning stages in DreamBooth and (b) requirements of multiple subject-specific images as input, which significantly limits the range of applications.

In this paper, we propose **Make-Your-3D**, a novel co-evolution framework for fast and consistent subject-driven 3D content generation. Specifically, given only a single casual subject image (without any additional information such as camera pose), we can generate subject-specific 3D assets that align with text-driven modifications as contextualization within 5 minutes, which is **36× faster** than DreamBooth3D [40]. As shown in Fig. 1, we personalize 3D content with geometric consistency and strong appearance identity preservation from given subjects while also respecting the variations (*e.g.*, sitting or wearing suit) provided by input text prompts.

For Make-Your-3D, we draw inspiration from recent advancements in personalized models [66] and multi-view diffusion models [29]. Despite the customization capability of personalized models and the 3D consistency of multi-view diffusion models, there remains a domain gap between the targeted subject and these two models, particularly when the subject is unseen in the training data [44,47]. Therefore, the key idea of our method is to harmonize the distribution of the identity-specific 2D generative model and multi-view diffusion model, aligning them with the distribution of the desired 3D subject. Specifically, we design a co-evolution framework to reduce the variance of distributions, where each model undergoes a process of learning from the other through identity-aware optimization and subject-prior optimization, respectively. Given a casual image captured from a subject, we first lift it to a 3D space through a multi-view diffusion model and capture its multi-views to optimize the 2D personalized model with an identity-enhancement process, which imposes enhanced identity-aware information into the 2D model. Next, we apply the original 2D personalized model to multi-views of the subject with modified text description and obtain more diverse images of the subject. Then we optimize the multi-view diffusion model with such various subject images in the subject-prior optimization process, infusing the subject-specific prior into the multi-view diffusion model. Finally, we cascade the two optimized models to process the single input image of the targeted subject and generate consistent subject-driven 3D results.

We conduct extensive experiments on the dataset used in DreamBooth3D [44] and open-vocabulary wild images captured from a subject with different styles. We also conduct a user study to evaluate the subject and prompt fidelity in our synthesized 3D results. The experiments validate that our Make-Your-3D is capable of producing vivid and high-fidelity 3D assets with strong adherence to the given subject while highly respecting the contextualization in the input text prompts. Compared to DreamBooth3D [40], our method not only surpasses in terms of quality, resolution, and consistency, but also shows a remarkable 36× speed improvement for efficiency. Unlike DreamBooth3D [40], our approach takes only a single wild image as input, eliminating the need for 3-6 carefully selected images of the same subject. Fig. 1 shows sample results of Make-Your-3D on different subjects and contextualizations, indicating that our co-evolution framework promotes the capability of generating subject-specific 3D assets for both models.

2 Related Works

Text-to-3D Generation. With exciting breakthroughs emerging in the image and video generation [3, 16, 41, 63, 70], there has been growing interest in 3D content generation [24, 28], particularly in text-to-3D generation [23]. One approach is to utilize extensive data [4, 8, 9] to train 3D generative models [5, 14, 21, 30, 31, 36, 48, 68, 72], akin to Text-to-Image (T2I) generation. However, due to constraints by the scale and quality of paired text-3D data, these methods are often limited to specific object categories and may exhibit a perceived lack of realism. Another line of text-to-3D is pioneered by DreamFusion [37], which employed a Score Distillation Sampling (SDS) loss to optimize a parametric 3D representation guided by the pre-trained 2D diffusion model [46]. Following works concentrate on improving 3D consistency, novel view quality, and generation speed through strategies such as incorporating 3D priors [26,47,55], crafting a tailored optimization strategy [25, 50, 57], and selecting more expressive and efficient representations [6, 52]. However, only text is not informative enough to express complex scenes or concepts, which can be a hindrance to 3D content creation. Moreover, it is often unattainable to create contextually diverse 3D assets that precisely align with the user-desired objects.

Image-to-3D Generation. Given the rich information embedded in images, numerous studies [10, 29, 33, 52, 54, 59, 64, 71] have explored to generate 3D content from a single image. Early attempts integrated the input image into the optimization pipeline by creating loss based on predicted depth [10,64] or object masks [33], comparing them with the rendered image. Magic123 [38] designs a two-stage coarse-to-fine framework for high-quality image-to-3D generation, employing textual inversion to ensure the generation of object-preserving geometry and textures. DreamGaussian [52] and Repaint123 [71] leverage a more efficient Gaussian splatting representation [20], significantly improving the optimization speed. Wonder3D [29] uses a 2D diffusion model to generate multi-view normal maps with color images and applies a geometry-aware normal fusion algorithm for direct surface extraction. However, despite the capability of these approaches to generate 3D content from a single image, their excessive reliance on images to maintain 3D consistency results in a lack of diversity in the generated 3D content, sometimes resembling more of a 3D reconstruction task. While HarmonyView [59] introduces the concept of harmonizing consistency and diversity, the displayed results often fall short of delivering satisfactory diversity, let alone achieving subject-driven customization. Different from their methods, our work is dedicated to reconstructing the concept of the provided object rather than the input image, thereby preserving the generated diversity.

Subject-Driven Content Creation. An increasing number of works [13,18,19, 22,40,44] are focusing on subject-driven generation, enabling users to personalize the generated content for specific subjects and concepts. Dreambooth [44] finetunes the 2D diffusion model and expands the model's language-vision dictionary with rare tokens using multiple images, achieving personalized text-to-image generation. IP-Adapter [66] realizes controllable generation by incorporating image prompt in the text-to-image models via the design of a lightweight decoupled cross-attention mechanism. VideoBooth [19] injects image prompts into the text-to-video model (T2V) in a coarse-to-fine manner, achieving customized content generation for videos. Despite remarkable success in personalizing T2I and T2V models, they do not generate 3D assets or afford any 3D control. The first at-

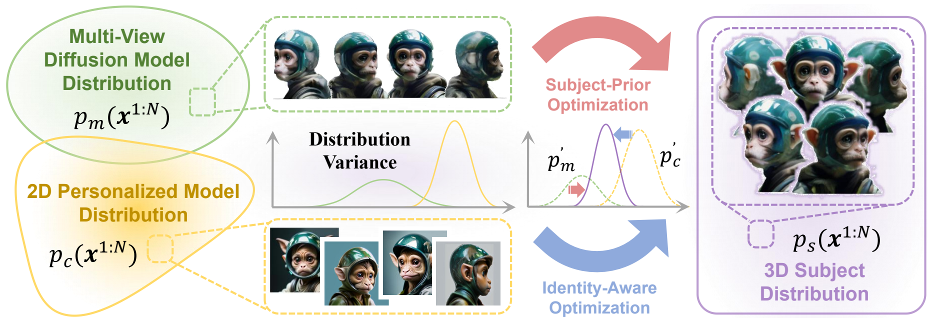

Fig. 2: Distribution variance between the wild subject and pre-trained models. Taking a *monkey* image and text prompt *"with elf ears"* as input, the pre-trained 2D personalized model and multi-view diffusion model generate images out of the distribution of desired ones, *i.e.*, the specific monkey with elf ears. To solve the problem, we carefully design a co-evolution framework including subject-prior and identity-aware optimization to harmonize the distributions and achieves desired 3D assets.

tempt in 3D subject-driven generation is DreamBooth3D [40], which proposes a simple pipeline utilizing DreamBooth [44] for 3D subject-driven generation. However, its generation is constrained by DreamBooth [44], requiring heavy fine-tuning stages with multiple subject images over 3 hours, which limits the range of applications. In contrast, our method can achieve fast subject-driven 3D content generation from only a single image of a subject within 5 minutes.

3 Method

In this section, we introduce our framework, *i.e.*, Make-Your-3D, for fast and consistent subject-driven 3D content generation. Our goal is to align output 3D assets with the distribution of the desired subject. To this end, we first review the scheme of diffusion model (Sec. 3.1), which is the basis of our pre-trained multi-view and personalized models. Then we analyze the distribution variance and optimization target of our method (Sec. 3.2). We further introduce our co-evolution framework, including identity-aware optimization (Sec. 3.3) and subject-prior optimization (Sec. 3.4). Finally, we present our mesh extraction process (Sec. 3.5). An overview of our framework is depicted in Fig. 3.

3.1 Preliminaries

Diffusion Models [15, 49] are probabilistic generative models that comprise two processes: (a) a forward process that gradually adds Gaussian noise to the data following a T steps Markov chain and (b) a denoising process that generates samples from the Gaussian distribution. Let $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ represent the real data under the additional condition $\boldsymbol{c}$, and let $t \in [0, T]$ denote the time step of the

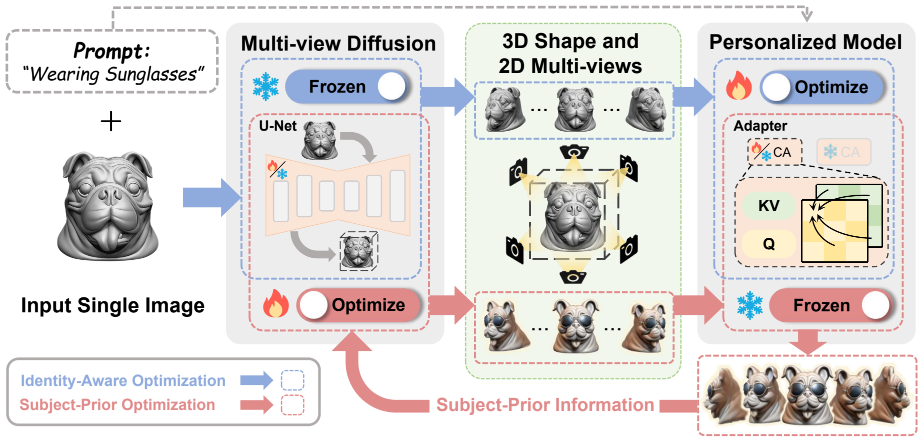

Fig. 3: The overall framework of our proposed Make-Your-3D. Our framework includes identity-aware optimization of 2D personalized model and subject-prior optimization of multi-view diffusion model to approximate subject distribution. The identity-aware optimization (Sec. 3.3) lifts input image to 3D space through a frozen multi-view diffusion model and optimizes the 2D personalized model via multi-views. The subject-prior optimization (Sec. 3.4) adopts diverse images from frozen personalized model to infuse the subject-specific prior into the multi-view diffusion model.

diffusion process. The training objective of the diffusion model $\boldsymbol{\epsilon}_\theta$, which predicts noise, is calculated as the following variational bound:

$$\mathcal{L}_{diff} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{c}, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right) \right\|^2 , \tag{1}$$

where $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$ is the noisy data at step t, and α_t, σ_t are fixed sequence of the noise schedule. For the conditional diffusion models, classifier-free guidance [17] is often employed as a prevailing method. In the sampling stage, the prediction noise is computed as a combination of conditional model $\boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right)$ and unconditional model $\boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c} \right)$:

$$\hat{\boldsymbol{\epsilon}}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right) = w \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right) + (1 - w) \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, t \right) , \tag{2}$$

where w is the guidance scale to adjust the alignment with condition $\boldsymbol{c}$. In our study, we utilize the open-source 2D personalized model [66] and the multi-view diffusion model [29], which are both built upon the Stable-Diffusion model [43], to achieve fast and consistent subject-driven 3D generation.

3.2 The Distribution of 3D Subject

Powered by recent advances in personalizing text-to-image models [44, 45, 66] and image-to-3D models [29, 38, 56], an intuitive idea to achieve customized 3D generation is to naively combine these methods. However, as shown in Fig. 2, it fails to yield satisfactory subject-specific 3D assets due to distribution variance between the wild subject and the pre-trained models [11, 44, 66]. To explore how to approximate the subject domain, we first propose that the distribution of the 3D subject denoted as $q_s(\boldsymbol{z})$, can be modeled as a joint distribution as:

$$q_s(\boldsymbol{z}) = p_s(\boldsymbol{x}^{1:N}|\mathcal{I}, y) = p_s(\boldsymbol{x}^1|\mathcal{I}, y) \cdot p_s(\boldsymbol{x}^{2:N}|\mathcal{I}, y), \tag{3}$$

where $\boldsymbol{x}^{1:N}$ are 2D multi-view color images observed from 3D subject conditioned on subject image $\mathcal{I}$ and text-driven modification y. As we have the pre-trained 2D customized model $p_c(\boldsymbol{x}^1|\mathcal{I}, y)$ and multi-view diffusion model $p_m(\boldsymbol{x}^{1:N}|\mathcal{I}) = \prod_{n=1}^{N} p_m(\boldsymbol{x}^n|\mathcal{I})$, our key insight involves optimizing both p_c and p_m to p'_c and p'_m, which closely align with the distribution of p_s respectively, *i.e.*,

$$p'_c(\boldsymbol{x}^1|\mathcal{I}, y) \approx p_s(\boldsymbol{x}^1|\mathcal{I}, y)\,,\; p'_m(\boldsymbol{x}^{2:N}|\mathcal{I}) \approx p_s(\boldsymbol{x}^{2:N}|\mathcal{I}). \tag{4}$$

Given that the condition y is independent of $\boldsymbol{x}^n$ in the multi-view diffusion model, we can ultimately approximate $q_s(\boldsymbol{z})$ with the optimized models p'_c and p'_m, *i.e.*, $q_s(\boldsymbol{z}) \approx p'_c \cdot p'_m$. Finally, we can formulate the joint distribution as a Markov chain within the diffusion scheme (omit the symbol y and $\mathcal{I}$ for simplicity):

$$p'_c(\boldsymbol{x}_T^1)\, p'_m(\boldsymbol{x}_T^{2:N}) \prod_t p'_m(\boldsymbol{x}_{t-1}^1|\boldsymbol{x}_t^1)\, p'_m(\boldsymbol{x}_{t-1}^{2:N}|\boldsymbol{x}_t^{2:N}), \tag{5}$$

where $p'_c(\boldsymbol{x}_T^1)$ and $p'_m(\boldsymbol{x}_T^{2:N})$ are Gaussian noises. Our key problem is to characterize the distribution $p'_c \to p_s$ and $p'_m \to p_s$ so that we can sample from this Markov chain to generate 3D assets in the subject distribution. Inspired by the derivation above, we carefully design a co-evolution framework, including identity-aware optimization for p_c and subject-prior optimization for p_m.

3.3 Identity-Aware Optimization

To mimic the appearance of subjects from images and synthesize novel renditions of them in different contexts, DreamBooth [44] finetunes all the parameters of the T2I model with 3-6 captured subject images. However, such a tuning strategy that relies on several images, is inefficient and constrained to scenarios where only one image can serve as input. In contrast, we use only a single subject image as input and choose a more efficient adapter-based T2I model (*i.e.*, IP-Adapter [66]) as our 2D personalized model p_c. Despite the user-friendly appeal of using a single input image, the 2D personalized model [66] suffers from distribution variance [11] between the subject and the training data, leading to outputs that do not resemble the subject as shown in Fig. 7, *i.e.*, $p_c(\boldsymbol{x}^1|\mathcal{I}, y) \neq p_s(\boldsymbol{x}^1|\mathcal{I}, y)$. To approximate subject distribution and enhance awareness of identity, we first leverage a multi-view diffusion model [29] p_m to generate the multi-views $\boldsymbol{x}^{(1:N)}$ with view-direction aware prompt $y^{(1:N)}$ given the input subject image $\mathcal{I}$ and text-driven prompt y. Then we apply augmentations to $\boldsymbol{x}^{(1:N)}$ and process them through the pretrained CLIP image encoder [39] $\mathcal{F}$ and get $\mathcal{F}(\boldsymbol{x}^{(1:N)})$. Finally, in the adapter module of the 2D personalized model, we use $\mathcal{F}(\boldsymbol{x}^{(1:N)})$ and $y^{(1:N)}$

to optimize the parameters of image cross-attention layer while freezing the original UNet model and text cross-attention modules. We follow the similar training objective to obtain p'_c in Eq. 1 with the condition $\boldsymbol{c} = \{y, \mathcal{F}(\boldsymbol{x})\}$. The empirical analysis in Sec. 4 demonstrates that our multi-view-based, identity-aware optimization effectively narrows the gap between p_c and the subject domain p_s.

3.4 Subject-Prior Optimization

Powered by the 3D consistency of neural radiance fields, DreamBooth3D [40] distills the fine-tuned DreamBooth to generate 3D assets via score distillation sampling (SDS) [37]. However, this framework suffers from low resolution and time-consuming optimization for per-sample training from scratch, as shown in Fig. 5, 9, limiting the practical usage. In this work, we optimize a more efficient multi-view diffusion framework based on Wonder3D [29] to approximate $p_s(\boldsymbol{x}^{2:N}|\mathcal{I})$ while better achieving fast and high-fidelity personalized 3D generation. Given the multi-views $\boldsymbol{x}^{(1:N)}$ from subject image $\mathcal{I}$ as discussed in Sec. 3.3, we process them through the original 2D personalized model with text-driven modification. Then we obtain diverse outputs $\tilde{\boldsymbol{x}}^{(1:N)}$ from multi-views, which coarsely adhere to the driven text and subject style with strong subject knowledge prior. In addition, we further exploit the subject geometry prior represented by normal maps $\tilde{\boldsymbol{n}}^{(1:N)}$ inferred from $\tilde{\boldsymbol{x}}^{(1:N)}$ by using the off-the-shelf single-view estimator [12]. Finally, we optimize the cross-domain self-attention module in the UNet framework based on multi-view diffusion model [29] to incorporate the subject-specific prior knowledge in the views of 3D distribution. Our objective function consists of two terms: (a) an image diffusion term for subject-prior enhancement and (b) a parameter preservation term for maintaining multi-view ability, which is computed as:

$$\mathcal{L}_{prior} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{n}_0, \boldsymbol{\epsilon}, \boldsymbol{c}_n, \boldsymbol{c}_i, t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}_n, \boldsymbol{c}_i, t \right) \right\|^2 + \lambda \frac{\|\theta - \theta_0\|_1}{N_\theta}, \tag{6}$$

where $\boldsymbol{c}_i, \boldsymbol{c}_n$ are the condition of the subject image with corresponding normal maps, θ_0 is the initial parameter of original multi-view diffusion, N_θ is the number of parameters, and λ is a balancing parameter set to 1. Grounded by visualization studies in Sec. 4, our subject-prior optimization strategy can successfully impose subject prior knowledge into multi-view diffusion model, which aligns $p'_m(\cdot|\mathcal{I}) \to p_s(\cdot|\mathcal{I})$ by Eq. 6 and yields more desired and consistent subject-driven 3D assets.

3.5 Subject-Driven Mesh Extraction

As done in previous co-evolution framework (*i.e.*, identity-aware optimization and subject-prior optimization), we have aligned the 2D personalized model p'_c and the multi-view diffusion model p'_m with the subject distribution p_s. Given the subject image $\mathcal{I}$ and text modification y, we first cascade the two optimized models to process them and obtain subject-driven multiview color images $\hat{\boldsymbol{x}}^{(1:N)}$

with respect to Eq. 3. From $\hat{\boldsymbol{x}}^{(1:N)}$, we apply a recent U-Net based Gaussian model pretrained in LGM [51] to predict 3D Gaussians. Next, we train an efficient NeRF (*i.e.*, Instant-NGP [35]) by using the rendered images from 3D Gaussians, and then convert the NeRF to polygonal meshes [53]. More details can be found in our supplementary materials. With adequately optimized implementation in identity-aware optimization (~ 1 min), subject-driven optimization (~ 3 min), and mesh conversion (~ 1 min), our framework can understand the visual subject in a reference image by approximating the subject distribution, and fast produce high-fidelity, consistent unseen personalized 3D content driven by text modification.

4 Experiments

In this section, we conduct extensive experiments to evaluate our subject-driven 3D content generation framework Make-Your-3D, and show the comparison results against DreamBooth3D [40]. We first present our qualitative results in multi-views and comparisons with baselines [40, 47] in various applications (*e.g.*, stylization, accessorization) (Sec. 4.2). Then we report the quantitative results with a user study (Sec. 4.3). Finally, we carry out more open settings and ablation studies to further verify the efficacy of our framework design (Sec. 4.4). Please refer to the supplementary materials for more results and analysis.

4.1 Experiment Setup

Implementation Details. In our framework implementation, we choose IP-Adapter [66] as our 2D personalized model backbone and apply the learning rates of $1e-4$ with a 0.01 weight decay to the image cross-attention layers with our multi-view based identity-aware optimization. On the other hand, we employ Wonder3D [29] as our multi-view diffusion model and utilize diverse images generated by the original 2D personalized model to subject its U-Net module to subject-prior optimization, with a $5e-5$ learning rates and a $1e-2$ weight decay. Notably, for each subject image, it takes only 5 minutes to complete all optimization stages on a single NVIDIA RTX3090 (24GB) GPU, which is far more efficient than 3 hours tuning on 4 core TPUv4 used in DreamBooth3D [40]. We use a fixed 30 iterations to optimize the personalized model. For the multi-view diffusion model, we use around 100 iterations in subject-prior optimization across different objects. To reconstruct 3D geometry, our method is built on the instant-NGP [35] based Gaussian reconstruction method [53].
Baselines and Metrics. We extensively compare our method with two baselines: DreamBooth3D [40] and an implementation for multi-view dreambooth in MVDream [47]. Since the two baseline methods [40, 47] do not have released related code, their results are obtained by downloading from their project pages. For metrics, we mainly show our results with notable comparisons through visualization. Following [37, 40], we evaluate our approach with the CLIP R-Precision metric in CLIP ViT-B/16, ViT-B/32, and ViT-L-14 models. We also conduct

Table 1: Quantitative comparisons on rendered images with text prompts using different CLIP retrieval models and comparisons on the multi-view consistency, subject fidelity, prompt fidelity, and overall quality score in a user study, rated on a range of 1-10, with higher scores indicating better performance.

Method	ViT B/16	ViT B/32	ViT L/14	Multi-view Consistency	Subject Fidelity	Prompt Fidelity	Overall Quality
DreamBooth3D [40]	0.783	0.710	0.797	6.05	6.42	6.89	5.33
MV DreamBooth [47]	0.805	0.735	0.813	8.76	7.55	6.73	7.59
Make-Your-3D (Ours)	**0.817**	**0.764**	**0.826**	**9.01**	**8.91**	**8.70**	**9.05**

a user study to further demonstrate the subject-driven fidelity, prompt fidelity, consistency, and overall quality of our method.

4.2 Qualitative Results

Visual Results of Make-Your-3D. Fig. 4 shows sample visual results of our method across different subjects with customized text prompts. The results demonstrate high-fidelity and consistent 3D generation with Make-Your-3D for open-vocabulary wild input subject images, achieving faithful alignment respecting the context in the input text prompt.

Qualitative Comparisons. We compare our method with DreamBooth3D [40] in various applications shown in Fig. 5, including color editing, accessorization, stylization, and motion modification. We observe that DreamBooth3D only generates coarse results with multiple images as input over 3 hours. In contrast, our Make-Your-3D produces higher quality subject-driven 3D results with compelling object details from only **a single** subject image within only 5 minutes, which is **36**× faster than DreamBooth3D. We also conduct comparisons with a recent multi-view DreamBooth implementation in MVDream [47] in Fig. 6.

The results further indicate that our method can not only achieve great 3D consistency but also better preserve subject identity without overfitting to data bias in terms of generated styles shown in mutli-view DreamBooth [47].

4.3 Quantitative Results

The middle column of Table 1 shows the average CLIP R-Precision over 160 evenly spaces azimuth renders at a fixed elevation of 40 degrees, following the same setting in DreamBooth3D [40] for fairness. Results clearly demonstrate higher scores for Make-Your-3D, indicating better 3D consistency and text-prompt alignment of our results. For user study, we render 360-degree videos of subject-driven 3D models and show each volunteer with five samples of rendered video from a random method. They can rate in four aspects: 3D consistency, subject fidelity, prompt fidelity, and overall quality on a scale of 1-10, with

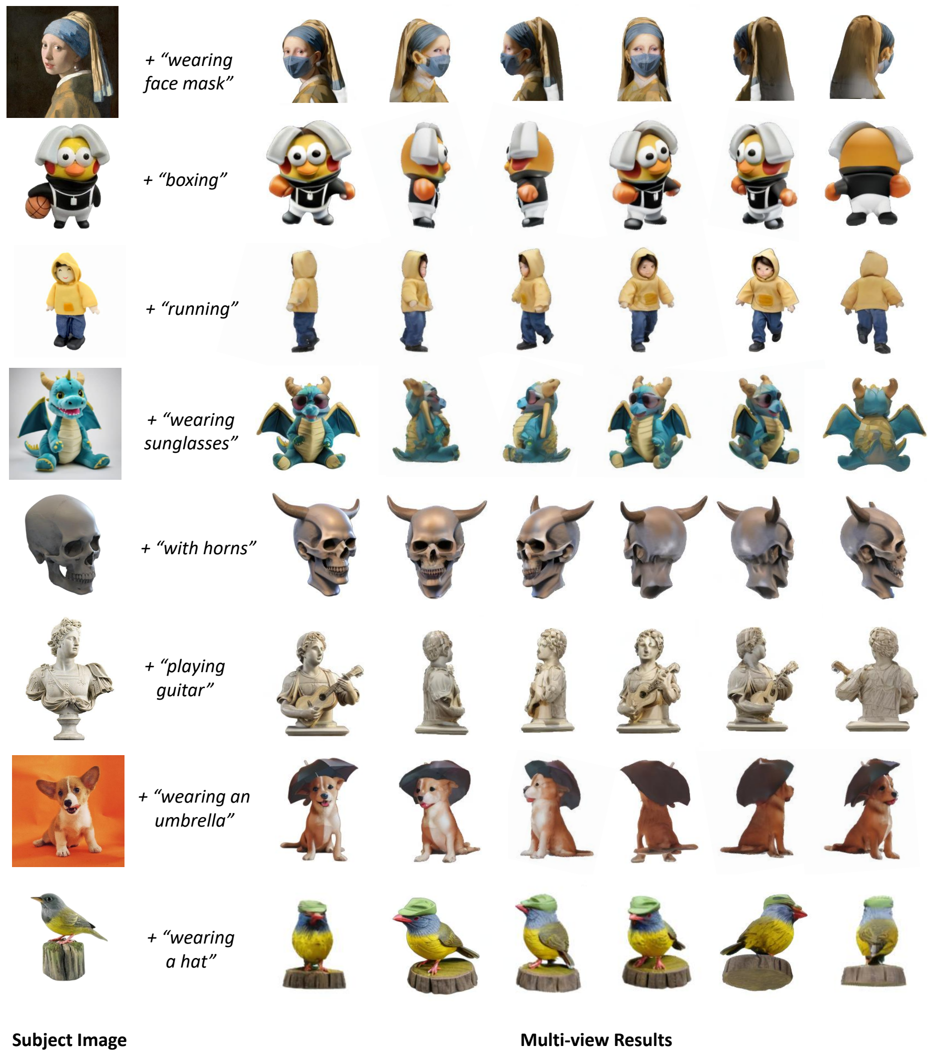

Fig. 4: Visual results of Make-Your-3D on different subjects with customized text inputs. The multi-view results demonstrate that we can generate 3D assets with high-fidelity, 3D consistency, subject preservation, and faithfulness to the text prompts.

higher scores indicating better performance. We collect results from 30 volunteers shown in the right column of Table 1. We find our method is significantly preferred by users over these aspects.

4.4 Ablation Study and Discussion

We carry out ablation studies on the design of Make-Your-3D framework in Fig. 7 to verify the effectiveness of our co-evolution framework. Specifically, we perform

Fig. 5: The qualitative comparisons with DreamBooth3D. We use the same text prompt and only one of the input images as in DreamBooth3D. Notice ours perform better on the object details with less input images.

Fig. 6: The multi-view qualitative comparisons. We are able to generate more realistic objects and achieve better subject preservation with multi-view consistency.

ablation on identity-aware optimization and subject-prior optimization. The results reveal that the omission of any of two elements leads to a degradation in terms of subject-driven fidelity. Notably, the absence of identity-aware optimization leads to worse subject preservation and consistency. The lack of subject-prior optimization results in less plausible multi-view rendering, especially in

Fig. 7: Ablation study of our method. We ablate the design choices of identity-aware optimization and subject-prior optimization.

Fig. 8: More personalization results for humans. Given a customized description and a face image, we can generate high-quality attributes (*e.g.*, beard, clothes) for the 3D character according to various contexts.

cases where the back view lacks informative subject-prior guidance. This illustrates the effectiveness of our overall framework (Fig. 3) that can approximate subject distribution and have great identity-specific preservation. Moreover, our method is robust in various open-vocabulary settings from wild web images and

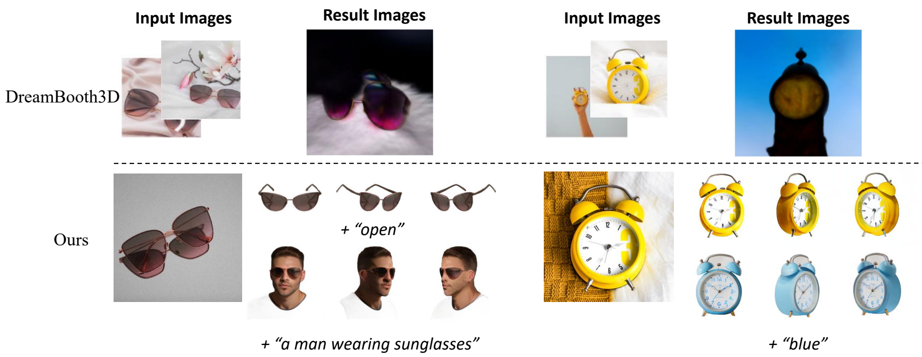

Fig. 9: Comparisons with the failure cases in DreamBooth3D [40]. As Dreambooth3D fails to reconstruct thin object structures like sunglasses and suffers from limited view variation, Our method has made significant improvements in fine details of thin objects and fast 3D personalization from a single subject image.

achieves high-quality results in failure cases of DreamBooth3D [40] shown in Fig. 9. Driven by our co-evolution framework, we can serve more applications such as human personalization shown in Fig. 8, where we can change their attributes like hair, clothes, and more. These surprising results further present the great potential for subject-driven customization with our Make-Your-3D.

5 Conclusion

In this paper, we have proposed Make-Your-3D, a method for fast and consistent subject-driven 3D content generation. To approximate the distribution of the 3D subject, we introduce a novel co-evolution framework. This includes an identity-aware optimization for 2D personalized model and a subject-specific optimization for multi-view diffusion model, through which each model adapts and improves the other's capacity to capture the subject-driven identity. Therefore, our method bridges the distribution variance from the 3D subject, achieving high-fidelity, multi-view coherent, and subject-specific 3D assets that faithfully adhere to the contextualization in text guidance. Notably, we only need a single subject image as input and produce per 3D result within 5 minutes, 36× faster than DreamBooth3D [40]. Extensive qualitative and quantitative experiments verify the effectiveness and efficiency of our co-evolution framework on 3D content personalization and demonstrate the potential for various applications.

Limitations and Future Work. Although our Make-Your-3D allows for high-quality 3D personalization and demonstrates better performance than previous work, the quality still seems to be limited to the backbone itself based on Stable Diffusion v1.5. The larger diffusion model such as SDXL [1] will further improve our performance. In future work, we are interested in exploring the 3D scene-level personalization which is a more challenging and complex task.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62206147

References

1. stable-diffusion-xl-base-1.0. `https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0`, accessed: 2023-08-29
2. Avrahami, O., Hertz, A., Vinker, Y., Arar, M., Fruchter, S., Fried, O., Cohen-Or, D., Lischinski, D.: The chosen one: Consistent characters in text-to-image diffusion models. arXiv preprint arXiv:2311.10093 (2023)
3. Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z.: A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering pp. 1–20 (2024). `https://doi.org/10.1109/TKDE.2024.3361474`
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
5. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction (2023)
6. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation (2023)
7. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36** (2024)
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
10. Deng, C., Jiang, C.M., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors (2022)
11. Du, C., Li, Y., Qiu, Z., Xu, C.: Stable diffusion is unstable. Advances in Neural Information Processing Systems **36** (2024)
12. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2021)
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022)
14. Gupta, A., Xiong, W., Nie, Y., Jones, I., Oğuz, B.: 3dgen: Triplane latent diffusion for textured mesh generation (2023)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

18. Huang, N., Zhang, T., Yuan, Y., Chen, D., Zhang, S.: Customize-it-3d: High-quality 3d creation from a single image using subject-specific knowledge prior (2024)
19. Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: Videobooth: Diffusion-based video generation with image prompts (2023)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
21. Kim, S.W., Brown, B., Yin, K., Kreis, K., Schwarz, K., Li, D., Rombach, R., Torralba, A., Fidler, S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models (2023)
22. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
23. Li, C., Zhang, C., Waghwase, A., Lee, L.H., Rameau, F., Yang, Y., Bae, S.H., Hong, C.S.: Generative ai meets 3d: A survey on text-to-3d in aigc era. arXiv preprint arXiv:2305.06131 (2023)
24. Li, X., Zhang, Q., Kang, D., Cheng, W., Gao, Y., Zhang, J., Liang, Z., Liao, J., Cao, Y.P., Shan, Y.: Advances in 3d generation: A survey (2024)
25. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
26. Liu, F., Wu, D., Wei, Y., Rao, Y., Duan, Y.: Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20763–20774 (2024)
27. Liu, F., Zhang, C., Zheng, Y., Duan, Y.: Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17386–17396 (2023)
28. Liu, J., Huang, X., Huang, T., Chen, L., Hou, Y., Tang, S., Liu, Z., Ouyang, W., Zuo, W., Jiang, J., et al.: A comprehensive survey on 3d content generation. arXiv preprint arXiv:2402.01166 (2024)
29. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
30. Lorraine, J., Xie, K., Zeng, X., Lin, C.H., Takikawa, T., Sharp, N., Lin, T.Y., Liu, M.Y., Fidler, S., Lucas, J.: Att3d: Amortized text-to-3d object synthesis (2023)
31. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation (2021)
32. Ma, Z., Zhou, D., Yeh, C.H., Wang, X.S., Li, X., Yang, H., Dong, Z., Keutzer, K., Feng, J.: Magic-me: Identity-specific video customized diffusion. arXiv preprint arXiv:2402.09368 (2024)
33. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image (2023)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
35. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
36. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts (2022)
37. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion (2022)

38. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., Ghanem, B.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors (2023)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
40. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023)
41. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International conference on machine learning. pp. 1060–1069. PMLR (2016)
42. Ren, Y., Zhou, Y., Yang, J., Shi, J., Liu, D., Liu, F., Kwon, M., Shrivastava, A.: Customize-a-video: One-shot motion customization of text-to-video diffusion models (2024)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
45. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023)
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
47. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation (2023)
48. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion (2022)
49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
50. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior (2023)
51. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation (2024)
52. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
53. Tang, J., Zhou, H., Chen, X., Hu, T., Ding, E., Wang, J., Zeng, G.: Delicate textured mesh recovery from nerf via adaptive surface refinement. arXiv preprint arXiv:2303.02091 (2023)
54. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior (2023)
55. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion (2023)

56. Wang, P., Shi, Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023)
57. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation (2023)
58. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
59. Woo, S., Park, B., Go, H., Kim, J.Y., Kim, C.: Harmonyview: Harmonizing consistency and diversity in one-image-to-3d (2023)
60. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
61. Wu, J., Gan, W., Chen, Z., Wan, S., Lin, H.: Ai-generated content (aigc): A survey. arXiv preprint arXiv:2304.06632 (2023)
62. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. arXiv preprint arXiv:2306.00943 (2023)
63. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models (2023)
64. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views (2023)
65. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)
66. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
67. Zeng, Y., Lu, Y., Ji, X., Yao, Y., Zhu, H., Cao, X.: Avatarbooth: High-quality and customizable 3d human avatar generation. arXiv preprint arXiv:2306.09864 (2023)
68. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models (2023)
69. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023)
70. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion models in generative ai: A survey (2023)
71. Zhang, J., Tang, Z., Pang, Y., Cheng, X., Jin, P., Wei, Y., Ning, M., Yuan, L.: Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting (2023)
72. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., Fu, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation (2023)