ImageBind3D: Image as Binding Step for Controllable 3D Generation

Zhenqiang Li
Zhengzhou University
Zhengzhou, China
lizhenqiang@gs.zzu.edu.cn

Jie Li*
Shanghai Jiao Tong University
Shanghai, China
lijiecs@sjtu.edu.cn

Yangjie Cao
Zhengzhou University
Zhengzhou, China
caoyj@zzu.edu.cn

Jiayi Wang
Zhengzhou University
Zhengzhou, China
wangjy@gs.zzu.edu.cn

Runfeng Lv
Zhengzhou University
Zhengzhou, China
lv1023785937@gs.zzu.edu.cn

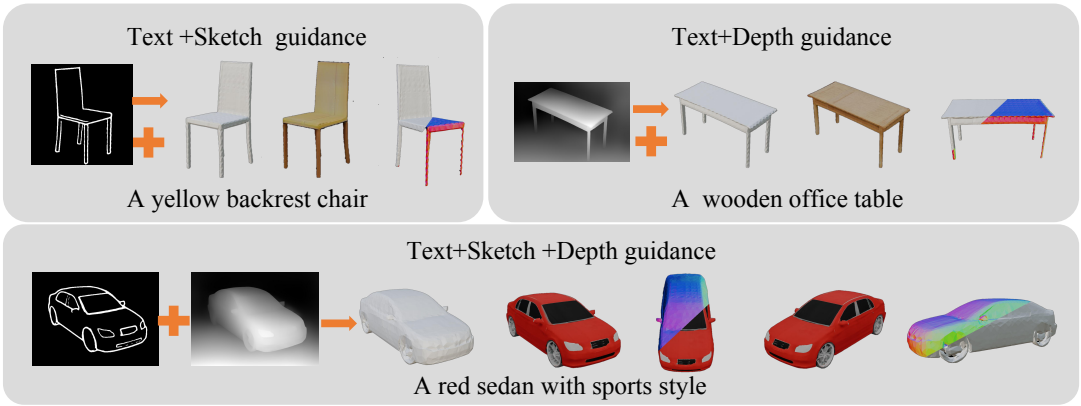

Figure 1: We propose ImageBind3D, a simple but effective approach that can offer guidance in multiple forms to feed-forward 3D generative models, while not affecting the original network architectures, generation capacity, and efficiency. Thanks to ImageBind3D, we can achieve more controllable outcomes, as opposed to the random results generated by GAN-based models or optimization-based techniques (e.g., GET3D and Dreamfusion). Furthermore, ImageBind3D can generate 3D objects with composable guidance.

Abstract

Recent advancements in 3D generation have garnered considerable interest due to their potential applications. Despite these advancements, the field faces persistent challenges in multi-conditional control, primarily due to the lack of paired datasets and the inherent complexity of 3D structures. To address these challenges, we introduce ImageBind3D, a novel framework for controllable 3D generation that integrates text, hand-drawn sketches, and depth maps to enhance user controllability. Our innovative contribution is adopting an inversion-align strategy, facilitating controllable 3D generation without requiring paired datasets. Firstly, utilizing GET3D as a baseline, our method innovates a 3D inversion technique that synchronizes 2D images with 3D shapes within the latent space of 3D GAN. Subsequently, we leverage images as intermediaries to facilitate pseudo-pairing between the shapes and various modalities. Moreover, our multi-modal diffusion model design strategically aligns external control signals with the generative model's latent knowledge, enabling precise and controllable 3D generation. Extensive experiments validate that ImageBind3D surpasses existing state-of-the-art methods in both fidelity and controllability. Additionally, our approach can offer composable guidance for any feed-forward 3D generative models, significantly enhancing their controllability. The code will be available at https://imagebind-3d.github.io/imagebind3d/.

*Corresponding author

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680845

CCS Concepts

• **Computing methodologies** → **Appearance and texture representations**; **Mesh models**.

Keywords

3D object generation, Conditional generation, Multimodal diffusion model

ACM Reference Format:
Zhenqiang Li, Jie Li, Yangjie Cao, Jiayi Wang, and Runfeng Lv. 2024. ImageBind3D: Image as Binding Step for Controllable 3D Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680845

1 Introduction

High-quality 3D objects are becoming increasingly important in various applications, e.g., metaverse, film special effects, and social platforms. However, the manual creation of 3D assets is very slow and tedious and requires specific technical knowledge and refined artistic skills. To accelerate this process, numerous studies have explored the use of generative models[4, 15, 23, 45] for 3D generation, yielding significant advancements. However, these approaches still lack the capability for multi-conditional control.

Existing methods[11, 24, 25, 40] typically require large-scale and paired shape data to enable effective training in the field of single-condition-guided 3D generation. Benefiting from 3D supervision, the generated results exhibit commendable geometric fidelity and consistency across different viewpoints. However, these methods exhibit a notable limitation in preserving fine-grained appearance details and ensuring robust controllability. They are constrained to textual inputs and can't incorporate additional guidances, such as sketches or depth maps. Yet, the collection and annotation of 3D data present substantial difficulties.

Recently, NeRF[36, 39] and 3D Gaussian Splatting[12, 27, 29, 41], have attracted considerable attention in the field of view synthesis, owing to their remarkable ability to represent complex scenes and produce high-fidelity rendering results. Numerous studies [8, 23, 31, 45] have employed NeRF and GS as 3D representations for text-based 3D generation tasks. Dream Fields[23], Dreamfusion[45], and GSGEN [8] were introduced to mitigate the constraints posed by limited datasets. By harnessing the capabilities of pre-trained vision-language models for guidance, these approaches extract 3D insights from 2D models. These techniques demonstrate excellent performance in generating high-fidelity and coherent 3D objects, meeting various textual prompts provided by the users. Due to the lack of paired textual and shape data, the task of generating 3D shapes from text is highly challenging for the following reasons. On one hand, the 2D diffusion model will introduce biases from the internet dataset into 3D generation. On the other hand, the absence of 3D priors in 2D models gives rise to problems of geometric inconsistency and discontinuity. Additionally, due to the multi-step iterative nature of diffusion models, the generation process often necessitates several tens of minutes.

While these methods can achieve promising generative quality, they notably lack the flexibility in user control capability to accurately guide the generation of 3D objects according to users' specific ideas. Specifically, the absence of multi-conditional control capability results in generated outputs that are usually uncontrolled and unstable. For instance, recent methods like GET3D[15] and Dreamfusion[45] fail to achieve accurate control over generated outputs through the combination of different conditions, e.g., combining text with sketches or depth maps, as shown in Figure.1. This paper tries to dig out the control capabilities that 3D generation models have implicitly learned.

Going beyond existing approaches, we introduce a novel 3D generation method named ImageBind3D, which enables multi-conditional 3D generation without the need for matched 3D datasets. Our ImageBind3D methodology adopts an inversion-align two-stage approach that effectively exploits the control capabilities offered by diffusion models, facilitating multi-conditional 3D generation. Inspired by ImageBind[16] and LanguageBind[65], we employ images as an intermediary representation to connect 3D shapes with text, sketch, and depth maps. In the first stage, employing GET3D as a baseline, we design an encoder-based 3D inversion algorithm that aligns images and shapes in latent space, as shown in Figure 2: Stage 1. Next, we extract multi-modal information from images to serve as pseudo-labels for 3D objects. In the second stage, We design a 3D multi-modal diffusion model in the latent space of the 3D GAN and inject additional guiding information into the diffusion model using decoupled attention, as shown in Figure 2: Stage 2. Utilizing a 3D multi-modal diffusion model, ImageBind3D can generate accurate 3D objects under multiple conditions. We summarize our main contributions as follows:

- We design a 3D multi-modal diffusion model that enables accurate control for generating high-quality 3D objects, while also supporting multi-conditional guidance.
- We introduce an encoder-based 3D inversion method to align images and 3D shapes in latent space.
- Employing images as an intermediary, we develop a pseudo-label generation strategy between shapes and various modalities, thus eliminating the necessity for matched 3D datasets.

2 Related Work

GAN-based models. Researchers have explored various methods for generating different 3D representations, including 3D voxel grids [14, 17, 34, 53], clouds [1, 37, 59, 63], implicit models [9, 35, 42, 62], octrees [13, 22], and meshes[3, 15, 21, 30]. However, the primary emphasis of these approaches lies in 3D content generation, with limited attention paid to controllability aspects. Employing semantic or edge maps, pix2pix3D [10] and SofGAN [6] facilitate new view synthesis, performing admirably for views closely aligned with the input. Yet, when generating views distant from the original conditions, they exhibit degraded quality with rough geometric and lack of fine details. Closely related to our work, TAPS3D [58] and ISS [32] establish a relationship between the input text and the latent space to achieve text-guided 3D generation. However, they only support text guidance and do not allow for more refined constraints on shape and appearance.

Diffusion-based generative models. Diffusion model[20, 54, 55] have recently achieved state-of-the-art performance in multiple generative tasks, such as text-to-image [44, 47, 49, 50], text-to-video[2, 28, 46, 52] and text-to-3D[8, 45, 51]. Dreamfusion [45] and SJC [57] employ Neural Radiance Fields (NeRF) to represent 3D structures, and subsequently utilize Score Distillation Sampling (SDS) for optimizing the rendering of new perspective images. These methods facilitate zero-shot text-to-3D generation, however, they are constrained by their low-resolution output, slow generation process, over-smoothing, over-saturating, and multi-faceted issues. Concurrently related to our method, HOLODIFFUSION [25]

introduces Warp-Conditioned-Embedding [18] as a pseudo-3D representation for 3D diffusion, which is constructed from multi-view features. Subsequently, this representation is rendered into 2D space and further optimized with the aid of 2D diffusion models. It should be noted that their pseudo-3D representation has inspired the 3D representation of our ImageBind3D. Control3D[7] enables the generation of primary views constrained by text and sketches, thus offering a degree of controllability with the SDS strategy. However, it still faces issues such as inconsistencies in geometry and views, and slow generation speeds.

3 Background

3D genetative model. GET3D [15] is a novel approach for 3D object generation, capable of producing high-fidelity textured 3D shapes through multi-view supervision. Specifically, GET3D maps noise vectors $z \in \mathrm{N}(0, \mathrm{I})$ to a textured mesh. The generation process includes the geometry branch and texture branch. The geometry branch is responsible for the differentiable generation of a surface mesh. Additionally, the texture branch generates a texture field, allowing for color queries to be performed at surface points. Following the design of StyleGAN [26] and PTI [48], they map $z1$ and $z2$ to intermediate latent spaces $w1$ and $w2$. By leveraging the differentiable render, the complete procedure is fully differentiable. The adversarial objective is defined as follows:

$$\begin{aligned} L(D_x, G) =& \mathbb{E}_{z\in N, c\in C}[g(D_x(R(G(z), c)))] \\ &+ \mathbb{E}_{I_x \in p_x}[g(-D_x(I_x))] \\ &+ \lambda ||\nabla D_x(I_x)||_2^2, \end{aligned} \tag{1}$$

where $g(u) = -\log(1 + \exp(-u))$, p_x represents the distribution of real images, R stands for rendering, and λ is a hyperparameter.

Challenges: This method enables the rapid generation of high-quality 3D objects. However, it still lacks multi-conditional control capability.

Diffusion model for Image Synthesis. Latent Diffusion Models[49] achieved significant advancements in the realm of text-to-image synthesis. T2I-adapter[38] and ControlNet[61] dig out the hidden abilities of T2I models, and then explicitly use them to control the generation, including text, semantic maps, and sketches. LDM represents a two-stage diffusion model comprising an autoencoder and a UNet-based denoiser. The optimization process can be expressed by the following formulation:

$$\mathcal{L} = \mathbb{E}_{Z_t, C_{\epsilon,t}}(\|\epsilon - \epsilon_\theta(Z_t, C)\|_2^2), \tag{2}$$

$Z_t = \sqrt{\overline{\alpha}_t} Z_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I)$ is the noised feature map at step t, as a combination of a scaled initial feature map Z_0 and scaled noise ϵ, where ϵ is drawn from a standard normal distribution. C represents the conditional information. ϵ_θ is a U-Net-based denoising architecture. Following T iterative steps, the final artifact $\hat{Z}_0$ is propagated into the decoder phase of the autoencoder to perform image generation. They utilize the cross-attention model to incorporate text into the denoising framework, which could be defined as follows:

$$Z' = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \tag{3}$$

where c_t is text features, $Q = ZW_q, K = c_t W_k$ and $V = c_t W_v$ represent the query, key, and values from text features. W_q, W_k and W_v represent the weight matrices.

Challenges: The primary advantage of this method is its capability for fine-grained control, yet it necessitates a significant investment in computational power and expansive training data. In the process of designing a 3D diffusion model, this issue becomes more evident.

4 Method

To achieve the best of both worlds, we propose an inversion-align approach named ImageBind3D, as shown in Figure 2. Our method harnesses the multi-condition guidance capability of diffusion models for Controllable 3D generation. To address computational challenges, we propose denoising in the intermediate latent space of 3D GAN. In response to the absence of paired datasets, we adopt a pseudo-labeling strategy to generate labels for 3D objects. In subsequent sections, we first introduce our 3d inversion method, which aligns images and 3D objects in the latent space (Section 4.1). Subsequently, we discuss our pseudo-label generation method, which forms connections between shapes and multiple modalities, with images serving as the central component (Section 4.2). Lastly, we propose our 3D multi-modal diffusion model, which can generate accurate 3D objects according to different input conditions (Section 4.3).

4.1 Image-based 3D Inversion

Controllable 3D generation encounters dual challenges: the absence of paired datasets and the substantial computational resources required. To address these two challenges, we design a 3D inversion method. Similar to the encoding-decoding architecture of SD[49], we adopt latent codes as 3D representations to simplify computational complexity. Utilizing the 3D inversion method, we establish a mapping relationship between images and 3D representations, which provides us with a benchmark for aligning 3D objects across different modalities. Employing GET3D as the foundational baseline, we develop an encoder-based 3D GAN inversion approach, which adopts an encoder-decoder architecture, as shown in Figure 2: Stage 1. Our encoder architecture comprises a VAE encoder and an MLP feature mapping layer. While the decoder follows the generator of GET3D architecture with frozen parameters. We utilize MLP to map image features from various hierarchical levels to the geometric latent variables $w1$ and appearance latent variables $w2$. By constraining the weights of this generator, our model can concentrate on achieving semantic congruence between the input images and the generated 3D objects. Hence, in the training phase, optimization is solely focused on the parameters of VAE and MLP. In the inversion process, we aim to find an intermediate latent variable to minimize the disparity in reconstruction loss between input images and their 3D renderings. It can be defined as follows:

$$\min_E \sum_{i=1}^{N} L(x(i), R(G(E(x(i)), \theta))), \tag{4}$$

where $G(w; \theta)$ is the 3d object generated by GET3D, which parameterized by weights θ, R is a differentiable renderer, E is a VAE encoder.

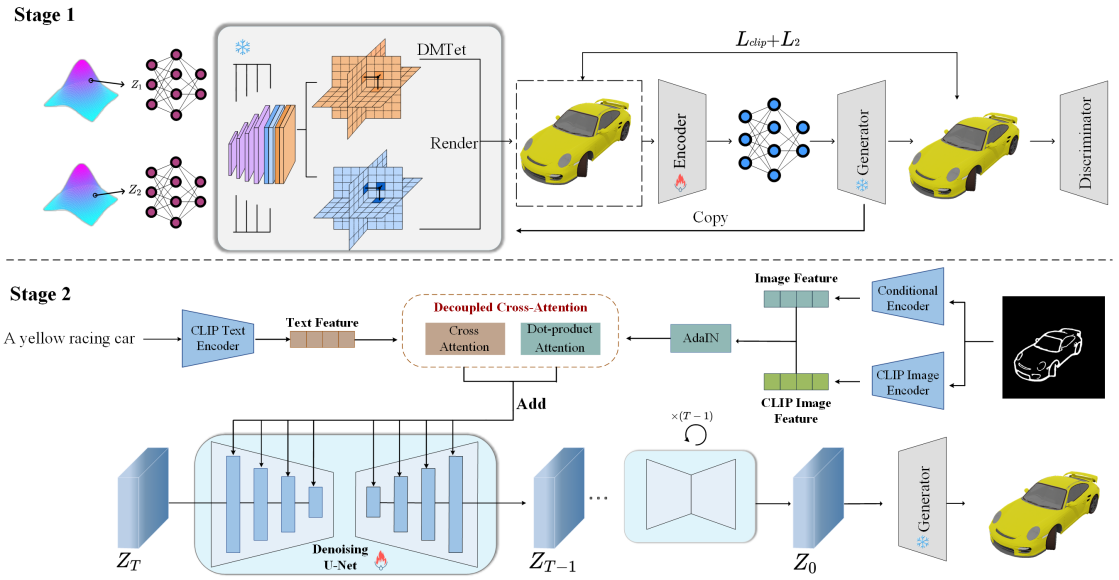

Figure 2: Our ImageBind3D is a two-stage approach for multi-conditional 3D generation. In the first stage, we employ a 3D inversion technique to align images and shapes within the GAN's latent space. Next, we generate a pseudo-label centered around the images. In the second stage, we introduce our 3D multi-modal diffusion model for multi-conditional 3D generation.

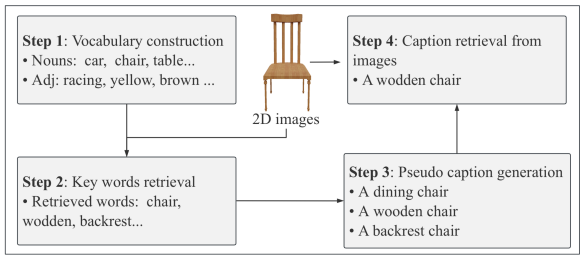

Figure 3: We demonstrate our pseudo caption generation module, which is inspired by TAPS3D[58]. Within this module, there are four distinct steps: vocabulary construction, keyword retrieval, pseudo caption generation, and caption retrieval from images.

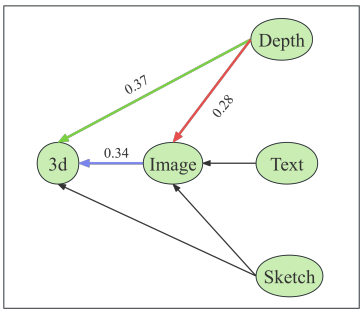

Figure 4: We calculated the average distance between different modalities using examples from two classes: Car and Chair.

It is observed that using $L(D_x, G)$ only allows the model to create plausible geometry corresponding to the input image. However, the generated appearance is unnatural and blurry. The primary challenge we face is the discordance between the adversarial loss and our objective of achieving a direct one-to-one mapping during the inversion process. To address this problem, we introduce two additional losses: image similarity loss, and pixel-wise loss. The image similarity loss is defined as follows:

$$L_{\text{CLIP}} = 1 - \cos(\text{E}_i(I_x), \text{E}_i(I_{gt})), \tag{5}$$

where E_i is the image encoder of CLIP, I_x and I_{gt} present the rendered images and input images. The pixel-wise loss is L2 loss, which signifies the Euclidean norm between the input and rendered images. The overall loss is obtained by blending these three components. It can be defined as follows:

$$L = \lambda_1 L(D_x, G) + \lambda_2 L_2 + \lambda_3 L_{\text{CLIP}}, \tag{6}$$

4.2 Pseudo Label Generation

The most recent 3D shape generation models are primarily fueled by data. Benefiting from large-scale training data, their performance enhancement is notable. However, when we aim to train a multi-condition 3D generation model, we are confined to a restricted set of multi-modal conditioned 3D objects. To address this challenge, we follow the methodology of [58, 64] and propose generating pseudo-labels for 3D objects centered around images. Our pseudo-label strategy is predominantly focused on text descriptions, sketches, and depth maps. For pseudo captions, we adopt a four-step pseudo caption generation method from TAPS3D[58], as shown in Figure 3. Initially, we construct a vocabulary using ShapeNet-related[5] nouns and adjectives found in the CLIP vocabulary. Next, we gather multiple words based on the 2D-rendered images. Subsequently,

candidate captions are generated using the retrieved words. Lastly, we choose a caption by assessing text-image similarities computed with the CLIP model. Following the paradigm of ControlNet[61], we've incorporated sketch and depth estimation models to predict sketches and depth maps.

As shown in Figure 4, through an image-centric pseudo-label strategy, we can efficiently establish connections between multi-modal data. Taking depth, image, and 3D as examples, it is observed that the direct mapping generation of 3D objects results in smaller intervals between modalities. The distance between modalities is obtained using equation 5. Considering the semantic gap, we opted not to employ a strategy of controlling images with different conditions and then using images to control 3D generation. On the contrary, our objective is to directly establish mapping relationships between different modalities and shapes, without intermediaries.

4.3 Multimodal Diffusion for Controllable 3D Generation

Due to the absence of multi-condition guidance in GAN frameworks, we present a 3D multimodal diffusion model for controllable 3D generation, as shown in Figure 2: Stage 2. We have adopted U-Net architecture and decoupled the cross-attention module. The multimodal diffusion model is designed to generate diverse latent codes, each corresponding to distinct input conditions. The generated latent codes are inputted as control signals into the generator, thereby enabling controllable 3D generation.

This innovation draws inspiration from the methodologies SD[49] and HoloDiffusion[25]. Specifically, HoloDiffusion leverages multi-view approaches to forge a comprehensive 3D representation suitable for diffusion. In our denoising architecture, z_t is constructed by concatenating geometric latent variables and appearance latent variables, with dimensions of 512*31. Within the latent space, 512×22 dimensions are allocated to present geometric attributes, leaving the remaining 512×9 dimensions to present appearance characteristics.

As illustrated in Equation 3, the original SD model utilizes the cross-attention mechanism to incorporate text into the denoising framework. To achieve multi-conditional control, one straightforward approach is to concatenate the features from disparate conditions and subsequently feed them collectively into the cross-attention layers. However, our findings indicated that this methodology fell short of efficacy. Inspired by Ip-adapter[60], we propose our decoupled cross-attention mechanism, comprising both cross-attention and dot-product attention modules. The output of the cross-attention Z' can be computed by Equation 3. Our dot-product attention module consists of three components: the CLIP encoder, VAE encoder, and AdaIN feature fusion module. These two encoders are employed to extract visual prompt features at semantic and geometric levels. Utilizing these image features, we apply Adaptive Instance Normalization (AdaIN) to normalize two features c_s and c_g. It can be defined as follows:

$$\hat{Q}_s = \text{AdaIN}(Q_s, Q_g), \tag{7}$$

$$\hat{K}_s = \text{AdaIN}(K_s, K_g), \tag{8}$$

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y), \tag{9}$$

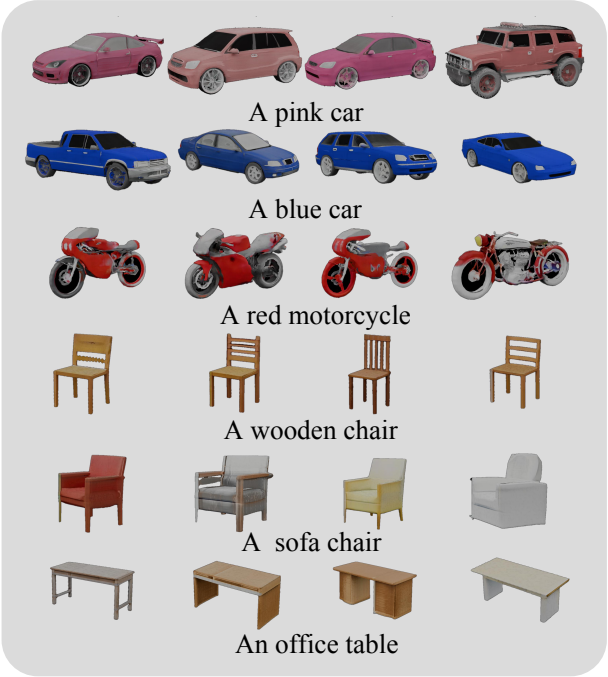

Figure 5: We demonstrate the generative diversity and textual guidance of our method. Each row employs the text prompt with different samples of random noise as input.

where x, y present CLIP and VAE feature, μ, σ present the mean and standard deviation of features. We concatenate K_n and $\hat{K_m}$, as well as V_n and V_m, respectively, to obtain K_{nm} and V_{nm}. Our dot-product attention can be defined as:

$$Z'' = \text{Attention}(\hat{Q}_m, K_{nm}^T, V_{nm}) = \text{Softmax}\left(\frac{\hat{Q}_m K_{nm}^T}{\sqrt{d}}\right) V_{nm}, \tag{10}$$

Where $\hat{Q}_m$, K_{nm} and V_{nm} represent the query, key and value. Subsequently, the output of condition dot-product attention is added to the output of text cross-attention. The decoupled cross-attention is specified as follows:

$$\begin{aligned} Z^{\text{new}} &= \text{Attention}(Q, K, V) + \alpha * \text{Attention}(\hat{Q}_m, \hat{K}_{nm}^T, V_{nm}) \\ &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V + \alpha * \text{softmax}\left(\frac{\hat{Q}_m \hat{K}_{nm}^T}{\sqrt{d}}\right) \cdot V_{nm} \end{aligned} \tag{11}$$

where α is weight factor, if $\alpha = 0$, the model become the original text-guided diffusion model. Specifically, our additional control signals include sketches or depth maps, as well as combinations of various conditions.

5 Experiment

5.1 Implementation Details and Metrics

We conduct training and evaluation on ShapeNet[5]. Our experimental evaluations are performed on four complex geometric categories, including Car, Table, Chair, and Motorbike. We utilize the GET3D model as our 3D generator. In the inversion stage, our inversion experiments are performed with a batch size of 16 and executed on 2 Nvidia 3090Ti-24G GPUs. It costs 15 hours for 3D

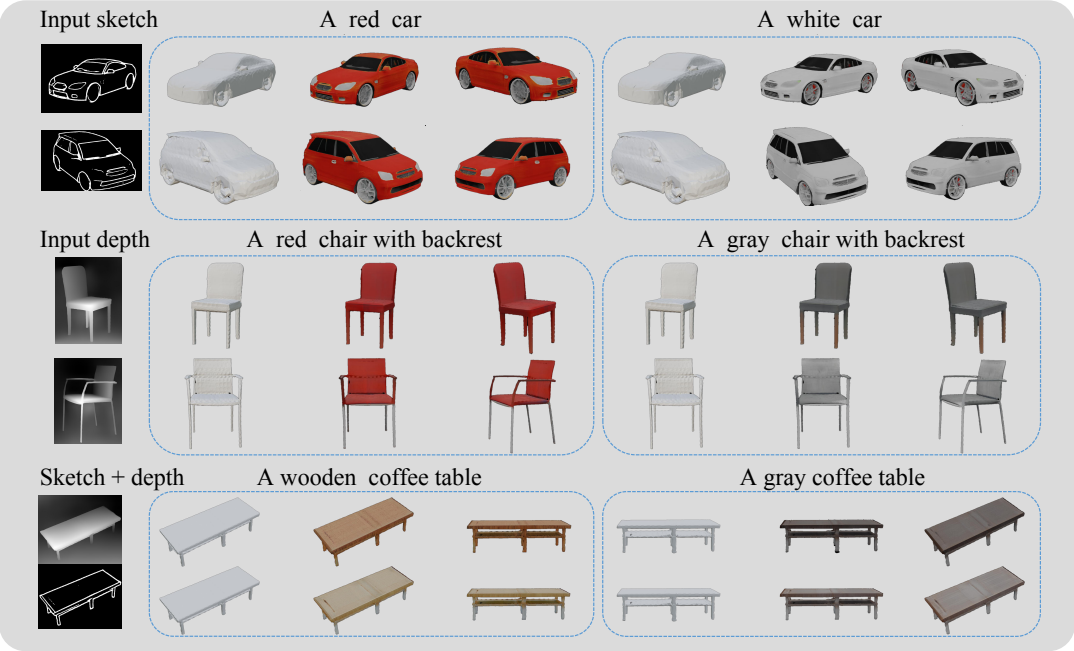

Figure 6: We show the multi-conditional control capabilities of our method. Each column takes the text prompt and additional conditions as the model input.

inversion training. In the alignment stage, we set the batch size to 64. The training process begins with the text-to-3D diffusion model, followed by the training of the multi-modal diffusion model. The training sessions lasted 15 and 6 hours for the respective stages.

5.2 Qualitative Results

In Figure 5, we illustrate our method's generative diversity and textual control efficacy across four object classes. In each row, we sample random noises alongside the provided textual prompts as inputs to generate textured 3D shapes. We observe that the rendered 2D images possess semantic coherence with the provided prompts. Furthermore, the 3D results present substantial diversity in texture and geometric structures, even when generated from the same textual inputs. Additionally, the Motorbike, Table, and Chair exhibit more complex geometric structures compared to the Car. However, we still effectively control the texture and geometric details of these challenging categories, demonstrating the robustness of our approach. In Figure 6, we show the multi-conditional control capabilities of our method across different classes. In the initial four lines, we present our ability to generate 3D objects by combining text with sketches or depth maps. Our approach enables the generation of multiple appearances for 3D objects using the same shape prompts and different text inputs. Furthermore, our method can generate diverse 3D objects based on the same text and varying visual prompts. In the last two lines, we present the results of 3D generation based on multiple conditions such as text, sketches, and depth maps. We observed that when we introduce a concept with text and then further refine the object with sketches or depth maps, it leads to a more realistic and controllable generation. This allows ordinary users to rapidly create 3D objects as they imagine, enhancing the user-friendly design experience.

5.3 Comparisons with State-of-the-art Methods

We compared our method with existing works through qualitative and quantitative analyses. To demonstrate the quality of our 3D generation, we compared it with existing works using three metrics: FID[19], R-precision[43], and FPD[33]. They are employed individually to assess the quality of rendered 2D images, the distance between textual and image representations, and the geometric fidelity of 3D models.

Qualitative Comparisons. Due to the current research methods mainly supporting single-condition guided 3D generation, we performed comparative experiments under different conditions. Our comparative analysis encompasses the following methods: TAPS3D[58], ISS[32], and pix2pix3d[10]. TAPS3D and ISS only support text-guided 3D generation, whereas pix2pix3d enables sketch-guided 3D generation. For a fair comparison, we conducted retraining of pix2pix3d and ISS at a resolution of 1024×1024. In Figure 7, we present the results of qualitative comparisons. We show text-guided, sketch-guided, and combined text and sketch-guided 3D generation separately. We compare across two categories, showcasing the generated results for each category from two different views. For example, with the text "a wooden backrest chair", our ImageBind3D can generate richer details than TAPS3D and ISS. Compared to pix2pix3D, our method can generate results that better match the sketch description, as shown in the sixth column of Figure 7. It can be observed that our method outperforms TAPS3D, ISS, and pix2pix3d in generating 3D textured shapes.

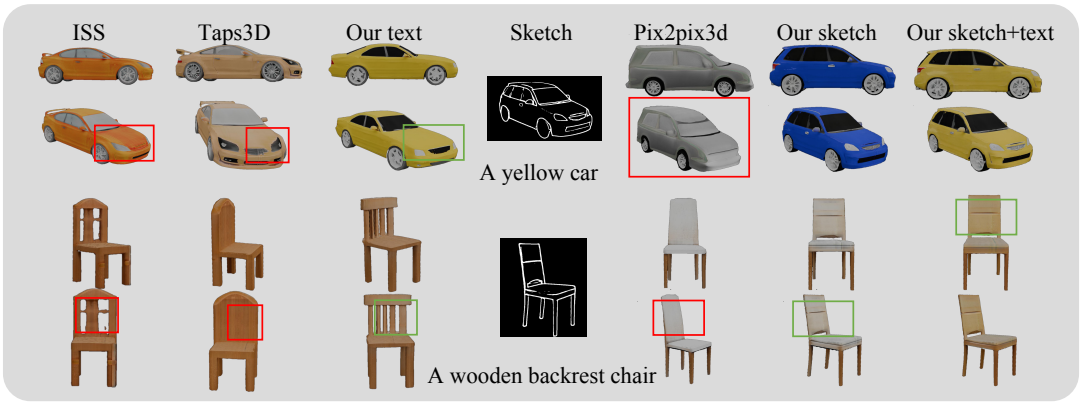

Figure 7: Our comparative experiments are conducted on two classes: Car and Chair. For each object, we display rendered images from two perspectives. We compare the text-guided 3D generation results of TAPS3D[58] and ISS[32] separately and contrast them with pix2pix3D[10] for sketch-guided generation results.

Table 1: Evaluation is performed on the Car and Chair categories using the FID, R-Precision, and FPD metrics. As the resolution generated by 3DFuse is 256×256, we downsample the generated results for comparison. Ablation-1 and Ablation-2 represent the experimental results of our ablation study.

Method	Car			Chair		
	FID (↓)	R-Precision(R=1)(↑)	FPD (↓)	FID (↓)	R-Precision(R=1)(↑)	FPD (↓)
3DFuse[51]	65.23	59.32 ± 2.15	N/A	94.75	67.55 ± 2.47	N/A
TAPS3D[58]	34.62	62.36 ± 1.93	337.67	44.83	60.19 ± 1.89	342.23
ISS[32]	37.18	60.36 ± 2.03	364.93	44.96	58.72 ± 2.23	585.79
Ablation-1	538.61	56.18 ± 1.98	1786.42	673.18	54.07 ± 2.01	2487.52
Ablation-2	668.35	57.13 ± 2.17	1875.68	797.31	55.75 ± 2.04	2613.57
Ablation-3	35.18	61.94 ± 2.06	N/A	50.28	59.87 ± 1.97	N/A
Ablation-4	35.57	62.37 ± 2.04	N/A	59.69	19.52 ± 1.96	N/A
Our-256*256	**30.46**	**64.05±1.87**	N/A	**40.27**	**64.41±2.05**	N/A
Our	**29.04**	**64.57±1.91**	**305.14**	**38.36**	**64.79±1.94**	**322.85**
Our+sketch	28.93	64.68 ± 1.90	N/A	37.41	64.92 ± 1.93	N/A
Our+depth	28.16	64.49 ± 1.93	N/A	37.62	64.73 ± 1.94	N/A

Quantitative Comparisons. To ensure fairness in comparison, we test on the same dataset using the official codes from GitHub. Table 1 provides quantitative comparisons. In particular, we downsampled our results to match the resolutions of 3DFuse [51]. 3DFuse adopts NeRF for 3D representation and utilizes SDS techniques, thereby imposing limitations on both the resolution and speed of the resultant outputs. In contrast, our backbone GET3D model boasts a larger capacity, enabling higher resolutions. Experimental findings demonstrate that our approach outperforms 3DFuse in text-guided generation quality for specific categories. Comparing the existing works with ours in Table 1, we can observe that our method outperforms the three state-of-the-art works across all three evaluation metrics.

In Table 2, we compare our method with others in terms of inference speed. Optimization-based techniques, such as Dreamfusion[45], 3DFuse[51], and GSGEN[8] demand several tens of minutes. Although DreamGaussian[56] reduces optimization time to 2 minutes, the resulting resolution is only 256*256 and the geometric quality is poor. However, it is important to emphasize that this comparison is not entirely fair for these optimization-based algorithms, as they are designed for open-world 3D generation. ISS, utilizing optimization strategies for each object, takes approximately 10 minutes. TAPS3D, employing a direct mapping text feature to latent space, operates in 6.5 seconds. Our approach delivers rendering results in 0.32 seconds and requires 1.09 seconds for mesh generation.

5.4 Ablation Study

3D Inversion Module. In ablation-1, we removed L_{clip} and L_2 during 3D inversion training and conducted experiments following the original inversion-align strategy. All other settings remain unchanged, yet the generated results underperform our original ImageBind3D across three evaluation metrics, as shown in Table 1. We execute ablative analyses on our methodology under textual guidance, encompassing "a red car" and "a chair with a backrest". The experimental results of ablation-1 are visualized in the second and fifth rows of Figure 8. **Pseudo Label Module.** In ablation-2 of

Table 2: We compare the inference times of various methods across distinct prompts. The inference times for Dreamfusion[45] and GSGEN[8] were obtained from their papers. For 3DFuse[51], DreamGaussian[56], and ISS[32], we computed the average time from 50 sample sets. For TAPS3D[58] and our method, the average time was derived from 500 sample sets.

Method	Device	Output	Time
Dreamfusion[45]	TPUv4	Rendering	90 min
3DFuse[51]	3090Ti	Rendering	30 min
GSGEN[8]	11G	Mesh	100 min
DreamGaussian[56]	3090Ti	Mesh	2 min
ISS[32]	3090Ti-24G	Mesh	10 min
TAPS3D[58]	3090Ti-24G	Mesh	6.5 sec
Ours-text	3090Ti-24G	Rendering	0.32 sec
Ours-text	3090Ti-24G	Mesh	1.07 sec
Ours-text+sketch	3090Ti-24G	Mesh	1.09 sec
Ours-text+depth	3090Ti-24G	Mesh	1.08 sec

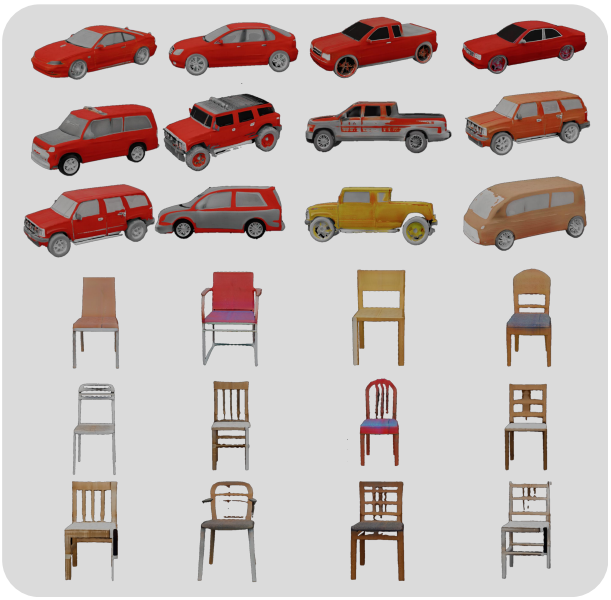

Figure 8: We perform ablation studies in our method, with two different text prompts "a chair with a backrest" and "a red car". The first row of experimental results corresponds to our complete model, followed by the results of the ablation-1 experiment in the second row, and those of the ablation-2 experiment in the third row.

Table 1, we directly removed the pseudo-label module and utilized a diffusion model to control the image directly for 3D generation. The experimental results are depicted within the third and sixth rows of Figure 8. Experimental results indicate that pseudo pseudo-label module plays a significant role in our method. Furthermore, by comparing ablation-1, ablation-2, and the original ImageBind3D, we observe that our inversion-based alignment mechanism contributes more substantially to the overall effectiveness.

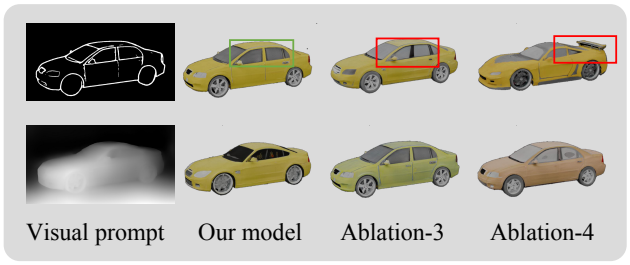

Figure 9: We show 3D results using various fusion mechanisms. All generated outputs are derived from identical text and visual prompts.

Decoupled Attention Module. To validate the effectiveness of our decoupled attention module, we conduct ablation-3 and ablation-4. In ablation-3, we adopt the decoupled attention mechanism of IP-Adapter[60]. In ablation-4, we directly add additional information to text, then input them into the cross-attention module. We compared the performance of these two approaches on FID and $R-P$ evaluation metrics, as shown in Table 1. Additionally, qualitative comparisons are illustrated in Figure 9. We merge the textual description "a yellow car" with different visual prompts. In the first row, the generated result from ablation-3 does not match the sketch conditions for the window part, and in ablation-4, the generated result for the car's tail part is also inconsistent with the input sketch. The experimental results indicate that our decoupled attention module effectively and precisely incorporates different guiding information into 3D generation.

6 Conclusion

We propose a novel 3D generation framework that enables controllable and high-quality 3D generation with multi-conditional guidance. Initially, we introduce a 3D inversion approach to establish correspondences between images and 3D objects, and then employ the latent codes as 3D representation. Next, we generate pseudo-labels to facilitate model training. Finally, we design a 3D multi-modal diffusion model to control the generation of 3D objects. During the inference stage, our method does not require additional optimization steps. Our generation method enables regular users to generate controllable and high-quality 3D objects within acceptable processing times.

Limitations. The primary limitation of our method lies in its generation capability, which is constrained by the original generative model. This issue could be addressed by adopting a more extensive and powerful generative model. Besides, we cannot produce different fine-grained details for different object components.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302458 and the Collaborative Innovation Major Project of Zhengzhou (20XTZX06013).

References

[1] 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.

[2] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23040–23050.

[3] 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.

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

[5] 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).

[6] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. 2022. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)* 41, 1 (2022), 1–26.

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

[8] Zilong Chen, Feng Wang, and Huaping Liu. 2023. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585* (2023).

[9] Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5939–5948.

[10] 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.

[11] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. 2023. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015* (2023).

[12] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. 2023. LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS. *arXiv preprint arXiv:2311.17245* (2023).

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5501–5510.

[14] Matheus Gadelha, Subhransu Maji, and Rui Wang. 2017. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 402–411.

[15] 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. In *Advances In Neural Information Processing Systems*.

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[17] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. 2019. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9984–9993.

[18] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. 2021. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4700–4709.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[21] Tao Hu, Liwei Wang, Xiaogang Xu, Shu Liu, and Jiaya Jia. 2021. Self-supervised 3D mesh reconstruction from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6002–6011.

[22] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. 2023. Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2697–2706.

[23] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 857–866. https://doi.org/10.1109/CVPR52688.2022.00094

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

[25] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. 2023. Holodiffusion: Training a 3D diffusion model using 2D images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18423–18433.

[26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[28] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023).

[29] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. 2023. Compact 3D Gaussian Representation for Radiance Field. *arXiv preprint arXiv:2311.13681* (2023).

[30] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2020. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 677–693.

[31] 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.

[32] Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. 2022. Iss: Image as stetting stone for text-guided 3d shape generation. *arXiv preprint arXiv:2209.04145* (2022).

[33] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. 2022. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17896–17906.

[34] Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. 2020. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674* (2020).

[35] 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.

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

[37] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. 2019. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575* (2019).

[38] 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).

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[40] 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).

[41] Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. 2023. Compressed 3D Gaussian Splatting for Accelerated Novel View Synthesis. *arXiv preprint arXiv:2401.02436* (2023).

[42] Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.

[43] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[45] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

[46] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. 2023. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483* (2023).

[47] 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.

[48] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)* 42, 1 (2022), 1–13.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[50] 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.

[51] 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).

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

[53] Edward J Smith and David Meger. 2017. Improved adversarial systems for 3d object generation and reconstruction. In *Conference on Robot Learning*. PMLR, 87–96.

[54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[55] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

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

[57] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023. 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.

[58] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. 2023. TAPS3D: Text-Guided 3D Textured Shape Generation from Pseudo Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16805–16815.

[59] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4541–4550.

[60] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

[61] 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.

[62] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. 2022. SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 52–63.

[63] Linqi Zhou, Yilun Du, and Jiajun Wu. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5826–5835.

[64] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. 2022. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124* (2022).

[65] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852* (2023).