CyC3D: Fine-grained Controllable 3D Generation via Cycle Consistency Regularization

Hongbin Xu[1], Chaohui Yu[2], Feng Xiao[1], Jiazheng Xing[3], Hai Ci[3], Weitao Chen[4], Fan Wang[2], Ming Li[5*]

[1]South China University of Technology
[2]Alibaba Group
[3]National University of Singapore
[4]Fudan University
[5]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
liming@gml.ac.cn

Abstract

Despite the remarkable progress of 3D generation, achieving controllability, i.e., ensuring consistency between generated 3D content and input conditions like edge and depth, remains a significant challenge. Existing approaches often struggle to maintain accurate alignment, leading to noticeable discrepancies. To address this issue, we propose CyC3D, a new framework designed to enhance controllable 3D generation by explicitly encouraging cyclic consistency during training between the second-order 3D content, generated based on extracted signals from the first-order generation, and its original input controls. Specifically, we employ an efficient feed-forward backbone that can generate a 3D object from an input condition and a text prompt. Given an initial viewpoint and a control signal, a novel view is rendered from the generated 3D content, from which the extracted condition is used to regenerate the 3D content. This re-generated output is then rendered back to the initial viewpoint, followed by another round of control signal extraction, forming a cyclic process with two consistency constraints. *View Cycle Consistency* ensures coherence between the two generated 3D objects, measured by semantic similarity to accommodate generative diversity. *Condition Cycle Consistency* aligns the final extracted signal with the original input control, preserving structural or geometric details throughout the process. Extensive experiments on zero-shot GSO/ABO benchmarks demonstrate that CyC3D significantly improves controllability, especially for fine-grained details, outperforming existing methods across various conditions (e.g., +14.17% PSNR for edge, +6.26% PSNR for sketch).

Code — https://toughstonex.github.io/cyc3d.github.io/
Extended version — https://arxiv.org/abs/2504.14975

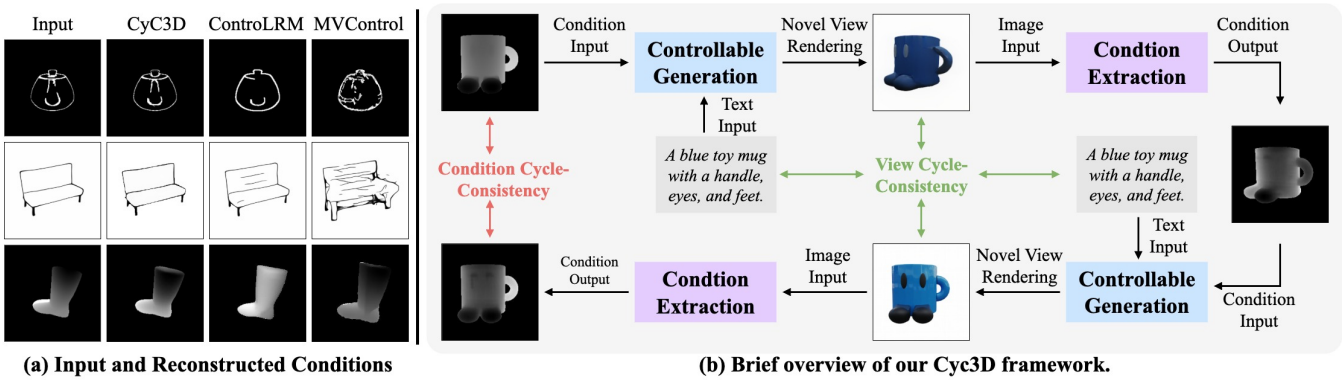
1 Introduction

In generative multimodal learning era (Brown et al. 2020; Qwen et al. 2025; Li et al. 2025a; Zhao et al. 2025; Liu et al. 2025, 2024; Li et al. 2024a; Miao et al. 2025; Shi et al. 2025; Su et al. 2025), the creation of 3D models from text descriptions (text-to-3D) or image collections (image-to-3D) is a fundamental task in computer graphics, attracting growing interest from a wide range of domains, including VR/AR (Behravan 2025), digital game design (Xu et al. 2024b), and filmmaking (Wang et al. 2025). The mainstream approaches can be broadly categorized into two groups: optimization-based methods and feed-forward methods. Optimization-based 3D generation methods leverage Score Distillation Sampling (SDS) (Poole et al. 2022) to iteratively refine 3D objects based on textual or visual inputs, effectively distilling the rich prior knowledge embedded in large pre-trained diffusion models (Rombach et al. 2022). In contrast, feed-forward 3D generation approaches (Nichol et al. 2022; Tang et al. 2023b; Hong et al. 2024, 2023; Zou et al. 2024; Tochilkin et al. 2024; Chen et al. 2025a) generate 3D content in a single inference pass, significantly improving efficiency and making real-time applications more feasible.

While significant progress has been made in 3D generation research, a critical yet under-explored challenge remains—*the controllability of 3D generation*. Controllable generation seeks to integrate precise control signals, such as edges, sketches, depth maps, surface normals, and text prompts, as conditioning inputs. By incorporating these structured constraints, controllable 3D generation enhances the flexibility and reliability of the synthesis process, enabling more precise manipulation of the generated content.

To advance research in controllable 3D generation, MVControl (Li et al. 2024b) extends ControlNet, originally designed for conditioned 2D image generation, to a multi-view diffusion model (Shi et al. 2023). It first predicts four-view images, which are then processed by a multi-view Gaussian reconstruction model (Tang et al. 2025a) to generate the final 3D representation. In contrast, ControLRM (Xu et al. 2024a) aims to develop an end-to-end controllable 3D generation framework by replacing the original image encoder of a large reconstruction model (LRM) (Hong et al. 2023) with a conditional encoder, enabling more precise conditioning. The inference pipeline of these controllable 3D generation works is summarized in Fig. 1 (a). While these studies have demonstrated a degree of controllability, a significant challenge remains, *i.e.,* , achieving precise and fine-grained conditional control. As illustrated in Fig. 1 (b), the extracted signals from the 3D models generated by MVControl and

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Input and Reconstructed Conditions **(b) Brief overview of our Cyc3D framework.**

Figure 1: Brief overview of our proposed **CyC3D**.

ControLRM exhibit noticeable discrepancies from the original input controls, underscoring the need for more refined and accurate controllable 3D generation.

To address this challenge, we propose CyC3D, a novel and general training framework that enhances the controllability of 3D generation. Our approach extends a feed-forward 3D generation network into a cyclic paradigm, enabling more precise and consistent adherence to input control signals, as illustrated in Fig. 1 (c). Specifically, given an input control condition on a reference view, the generated 3D content is first rendered from randomly sampled new viewpoints. From these rendered images, the corresponding control conditions are extracted and fed back into the generation model as inputs, ensuring that the model learns to maintain control consistency across views. Subsequently, the model regenerates the 3D content, which is then rendered back to the initial reference viewpoint, where the extracted control conditions should match the original input condition. To further refine controllability, we introduce two complementary consistency constraints: *View Cycle Consistency*, which ensures that the two generated contents have similar semantics in embedded space since they are generated with the same prompt and highly coupled conditions, and *Condition Cycle Consistency*, which enforces strict alignment between the extracted control signals from the final rendered images and the original input condition. We compare our CyC3D against state-of-the-art approaches on GSO and ABO datasets (Xu et al. 2024a) for zero-shot evaluation. Experimental results demonstrate that our framework significantly enhances controllability across all benchmarks while preserving competitive generation quality compared to existing methods.

We highlight the contributions of this work as follows: (1) *New Insight and Framework:* We reveal that existing efforts in controllable 3D generation still perform poorly on fine-grained controllability, and propose a novel framework (CyC3D) to handle the problem, which extends the original single-stage novel view synthesis pipeline to a double-stage one with cycle consistency. (2) *Cyclic Consistency Feedback:* To refine controllability, we propose two complementary consistency constraints: condition cycle consistency and view cycle consistency. The former regularizes the consistency between input and reconstructed conditions, the latter enhances the semantic consistency of different stages. (3) *Evaluation and Promising Results:* We provide a comprehensive zero-shot evaluation of the controllability on GSO/ABO benchmarks, and demonstrate that CyC3D comprehensively outperforms existing methods.

2 Related Work

Text-/Image-to-3D Generation. 3D generation falls into two paradigms: (1) Optimization-based methods (e.g., DreamFusion (Poole et al. 2022), Styleme3d (Zhuang et al. 2025), Inter3D (Chen et al. 2025b)) use SDS loss with diffusion priors to optimize neural fields without large text-to-3D datasets, but suffer from the Janus problem due to 2D–3D inconsistency. (2) Feed-forward models (e.g., Instant3D (Li et al. 2024a), LRM (Hong et al. 2023), TripoSR (Tochilkin et al. 2024), TextSplat (Wu et al. 2025)) leverage large 3D datasets and transformers to predict NeRFs efficiently; multi-view variants like SyncDreamer (Liu et al. 2023) and MVDream (Shi et al. 2023) further improve quality via multi-view diffusion. While Cycle3D (Tang et al. 2025b) implements cycle consistency in inference, our approach embeds it into training to enable real-time feed-forward generation.

Controllable 3D Generation. Controllable 3D generation remains a key challenge, drawing inspiration from controllable 2D methods. MVControl (Li et al. 2024b), inspired by ControlNet (Zhang, Rao, and Agrawala 2023), integrates a trainable control network with multi-view diffusion but suffers from limited generalization and procedural complexity. ControLRM (Xu et al. 2024a) overcomes these issues with an end-to-end feed-forward design for fast inference and precise control. Building on ControlNet++ (Li et al. 2025b) and ControLRM, we enhance conditional controls to enable more sophisticated and nuanced manipulation of 3D assets.

3 Method

3.1 Generation Backbone

For controllable 3D generation, our CyC3D can be combined with arbitrary feed-forward backbones (e.g. ControLRM (Xu et al. 2024a), MVControl (Li et al. 2024b)).

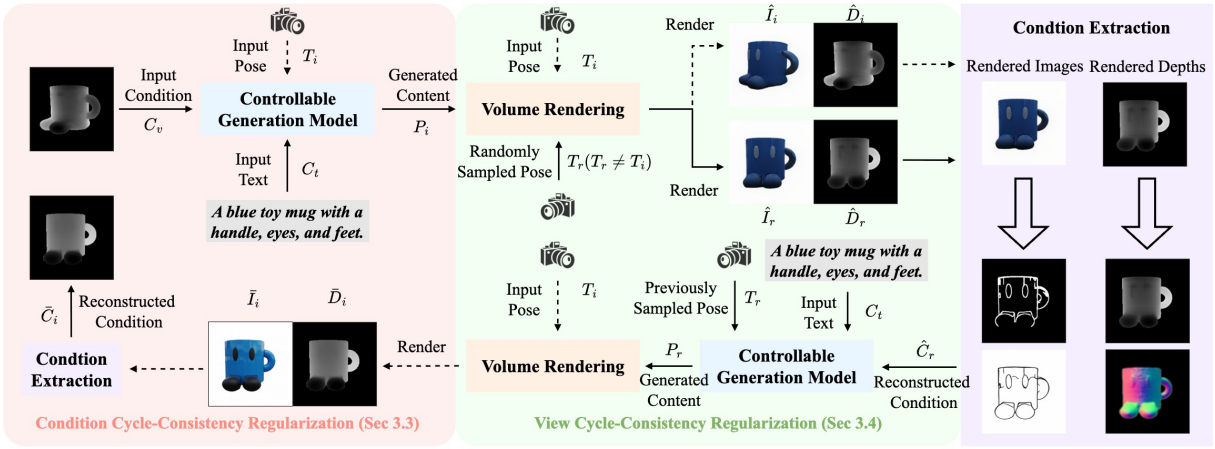


Figure 2: Training framework of our proposed CyC3D. It contains two kinds of regularization: condition cycle consistency regularization (Sec. 3.3) and view cycle consistency regularization (Sec. 3.4).

In default, we select ControLRM (Xu et al. 2024a) as our backbone model. ControLRM contains a 2D condition extractor and a 3D triplane decoder. The conditional generator in ControLRM is designed to take both text and 2D visual conditions as inputs. This generator can utilize either a transformer or diffusion backbone to create initial 2D latent representations, which are subsequently transformed into 3D models. The 3D triplane decoder is inherited from a pre-trained large reconstruction model (Hong et al. 2023). The triplane transformer decoder employs a unique representation of 3D data by decomposing it into three orthogonal planes (X, Y, and Z). Each plane is processed using transformer layers, which utilize self-attention mechanisms to capture spatial relationships and dependencies within the data.

3.2 Condition Extraction

To ensure the gradient back-propagation of cycle consistency in our CyC3D framework during training, the utilized condition extractor should be fully differentiable. The design of the condition extractor for each condition control is shown as follows:

Canny Condition. We extract edge maps using the Canny algorithm (Canny 1986). Following (Li et al. 2025b), a differentiable Canny extractor can be implemented based on Kornia (Riba et al. 2020) to extract edge maps.

Sketch Condition. Following ControlNet (Zhang, Rao, and Agrawala 2023), we utilize a pre-trained CNN network to extract sketch conditions from images. Since the CNN network is differentiable, the gradients can be passed backward through the network.

Depth Condition. In analogy with the sketch condition, we also utilize a pre-trained foundation model (Depth Anything (Yang et al. 2024)) to extract the depth condition from rendered images. As an alternative, we can also normalize the rendered depth map via volume rendering to extract the geometric prior from the 3D content.

Normal Condition. In practice, we find that the existing normal estimation models (Bae, Budvytis, and Cipolla 2021) utilized in ControlNet (Zhang, Rao, and Agrawala 2023) tend to predict over-smoothed normal maps. It may result in a large gap between the exact 3D shape of the rendered multi-view images and the estimated normal maps. Consequently, we adopt an alternative solution to predict the depth map first and then convert the depth map to a normal map via a differentiable module (Huang et al. 2021). As an alternative, we also extract the normal from the rendered depth maps via volume rendering.

3.3 Condition Cycle Consistency Regularization

Denote the input edge condition as $C_e \in \mathbb{R}^{H\times W}$, the sketch condition as $C_s \in \mathbb{R}^{H\times W}$, the depth condition as $C_d \in \mathbb{R}^{H\times W}$, and the normal condition as $C_n \in \mathbb{R}^{H\times W\times 3}$. The input visual condition is noted as $C_v \in \{C_e, C_s, C_d, C_n\}$. Given the text prompt C_t and the input visual condition as C_v, the generation model outputs the triplane representation $P \in \mathbb{R}^{3\times C_p\times H_p\times W_p}$ corresponding to the input conditions.

$$P_i = f_{\text{gen}}(C_v, C_t, T_i), \tag{1}$$

where T_i is the camera pose of the input condition, which is usually assumed to be an identity matrix after normalizing all the camera extrinsics to a frontal view. f_{gen} is the backbone generation model for controllable 3D generation.

Given a 3D point $x \in \mathbb{R}^3$, the color and density filed (c, σ) can be obtained via sampling the corresponding features on the triplane P_i: $(c, \sigma) = f_{\text{mlp}}(P_i, x)$. By sampling the points along the viewpoint $T_r \in \mathbb{R}^{4\times 4}$, we can render the image on view T_r via volume rendering (Mildenhall et al. 2021):

$$\hat{I}_r, \hat{D}_r = f_{\text{render}}(P_i, T_r), \tag{2}$$

where T_r is the randomly sampled viewpoints as shown in Fig. 2. $\hat{I}_r$ is the rendered image on view T_r, and $\hat{D}_r$ is the rendered depth map. Then we can use the corresponding condition extractor f_{cond} introduced in Sec. 3.2 and Fig. 2 (b) to extract the condition map on the sampled view T_r:

$$\hat{C}_r = f_{\text{cond}}(\hat{I}_r), \tag{3}$$

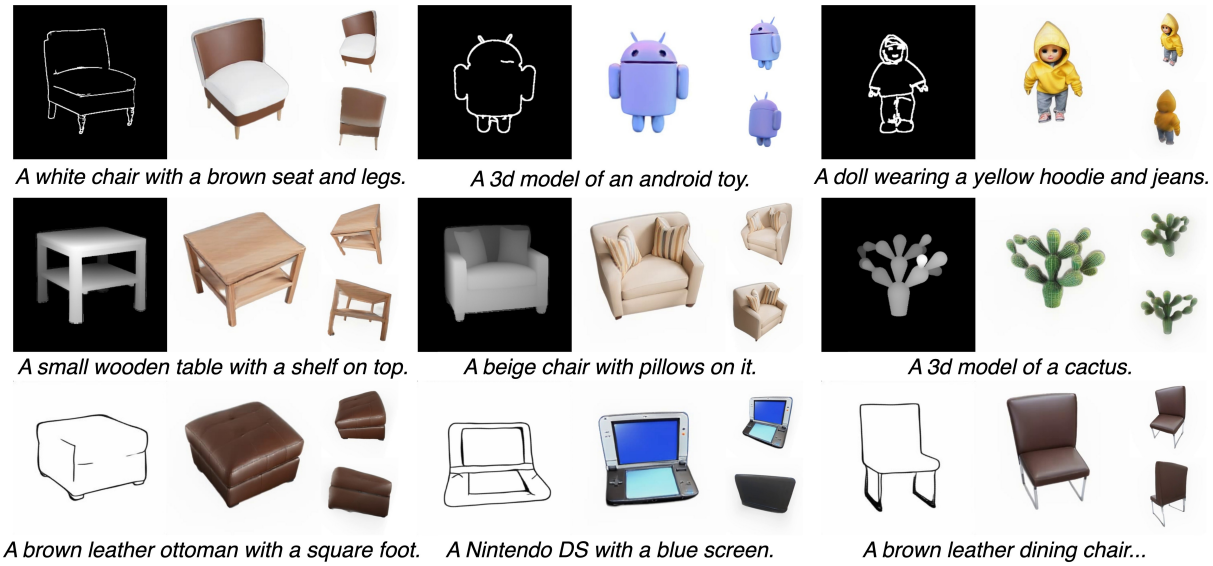


Figure 3: Visualization results of our CyC3D. The input conditions and the rendered images at different views are visualized. Our CyC3D can generate high-quality results with consistent conditional controls.

where $\hat{C}_r$ is the extracted condition map. Note that the condition extractor f_{cond} is differentiable.

In a second time, we can feed the obtained condition map $\hat{C}_r$ on novel view T_r back to the generation model f_{gen} again:

$$P_r = f_{\text{gen}}(\hat{C}_r, C_t, T_r), \tag{4}$$

where the predicted triplane P_r can further be used to render on the input reference view T_i:

$$\overline{I_i}, \overline{D_i} = f_{\text{render}}(P_r, T_i), \tag{5}$$

where $\overline{I_i}$ is the rendered image on the input reference view, and $\overline{D_i}$ is the rendered depth map. Then we can extract the condition map from the rendered image $\overline{I_i}$:

$$\overline{C_i} = f_{\text{cond}}(\overline{I_i}), \tag{6}$$

where $\overline{C_i}$ is the generated condition map back on the input reference view.

As shown in Fig. 2, we can calculate the difference between the reconstructed condition and the input condition to regularize the fine-grained controls. Depending on the type of conditional controls, we also propose two different kinds of conditional cycle-consistency feedback as follows:

2D Conditional Control Feedback. If we only consider 2D conditional controls extracted from the rendered images (e.g., Canny map extracted from RGB image), we can use a condition extractor to reconstruct the control signal. The generated condition map $\overline{C_i}$ via Eq. 6 should be consistent with the original condition map C_v on the input reference view.

$$L_{\text{cond}} = \sum_{T_r} \|\overline{C_i} - C_v\|_2^2. \tag{7}$$

If the randomly sampled camera pose T_r equals the input viewpoint T_i, the two-step inference in cycle consistency will be degraded to a single one (as shown in Fig. 2 (c)), saving the computation cost:

$$\hat{I}_i, \hat{D}_i = f_{\text{render}}(f_{\text{gen}}(C_v, C_t, T_i), T_i), \tag{8}$$

$$\hat{C}_i = f_{\text{cond}}(\hat{I}_i), \tag{9}$$

where $\hat{C}_i$ is the extracted condition map on the input reference view. In this way, we can modify Eq. 7 as:

$$L_{\text{cond}} = \sum_{T_r \neq T_i} \|\overline{C_i} - C_v\|_2^2 + \|\hat{C}_i - C_v\|_2^2. \tag{10}$$

3D Conditional Control Feedback. If we consider 3D conditional controls, such as depth or normal conditions, the 3D shape of the generated objects should also be regularized for fine-grained control.

(1) If the input condition is depth map ($C_v = C_d$), the explicit consistency on the rendered depth map and the input conditional depth map can be calculated as follows:

$$L_{\text{cond-d}} = \sum_{T_r} \|f_{\text{norm}}(\overline{D_i}) - C_d\|_2^2, \tag{11}$$

where f_{norm} is the depth normalizing function. Since the rendered depth $\overline{D_i}$ and the input depth C_d have a gap in the scale of depth, we utilize f_{norm} to normalize them to the same scale and calculate a scale-agnostic loss following (Ranftl et al. 2020).

In analogy with Eq. 10, when the randomly sampled camera pose T_r equals the input viewpoint T_i, there is no need to run the inference twice during propagation. We can modify Eq. 11 as follows:

$$L_{\text{cond-d}} = \sum_{T_r \neq T_i} \|f_{\text{norm}}(\overline{D_i}) - C_d\|_2^2 + \|f_{\text{norm}}(\hat{D}_i) - C_d\|_2^2, \tag{12}$$

where $\hat{D}_i$ can be obtained from Eq. 8.

Figure 4: Qualitative comparison among our CyC3D and the other state-of-the-art controllable 3D generation methods. In the 1st row, the baseline methods fail to generate consistent content with the text prompts. In the 2nd row, ControLRM generates a fainter color compared with our CyC3D.

(2) If the input condition is a normal map ($C_v = C_n$), the explicit depth-normal consistency can be further calculated as follows:

$$L_{\text{cond-n}} = \sum_{T_r} \|f_{\text{d2n}}(\overline{D_i}) - C_n\|_2^2, \tag{13}$$

where f_{d2n} is a differentiable module (Huang et al. 2021) that can convert the depth map to a unified normal map.

In analogy with Eq. 10 and 12, when the randomly sampled camera pose T_r equals the input viewpoint T_i, there is no need to run the inference twice during propagation. We can thus modify Eq. 13 as follows:

$$L_{\text{cond-n}} = \sum_{T_r \neq T_i} \|f_{\text{d2n}}(\overline{D_i}) - C_n\|_2^2 + \|f_{\text{d2n}}(\hat{D}_i) - C_n\|_2^2, \tag{14}$$

where $\hat{D}_i$ can be obtained from Eq. 8.

3.4 View Cycle Consistency Regularization

Since the condition cycle consistency regularization (Sec. 3.3) is a strong constraint that enforces the reconstructed and the input conditions to be as similar as possible, only using this loss might lead to overfitting towards the input viewpoint. If the generated 3D content from the initial view fails, the remaining processes might be misled to a wrong cross-view mapping. Only using the condition cycle consistency on the input view can not handle this issue, and the semantic multi-view consistency might be ignored during optimization. Consequently, we propose View Cycle Consistency Regularization to ensure convergence.

Rendering Regularization Loss. To regularize the possible failures in generation, the image reconstruction loss between the rendered images from generated 3D contents and the ground truth images on different views is used to supervise the generation model f_{gen}:

$$L_{\text{render}} = \sum_{T_l} \|f_{\text{render}}(f_{\text{gen}}(C_v, C_i, T_i), T_l) - I_l^{\text{gt}}\|_2^2, \tag{15}$$

where T_l means the pre-defined ground truth viewpoints in the multi-view dataset, and I_l^{gt} is the corresponding ground truth image. Note that this constraint can only be used in the first-time feedforward computation of the generation model. Because the randomness of the generation model might be depressed if enforcing the first-time and second-time generated 3D contents to be the same.

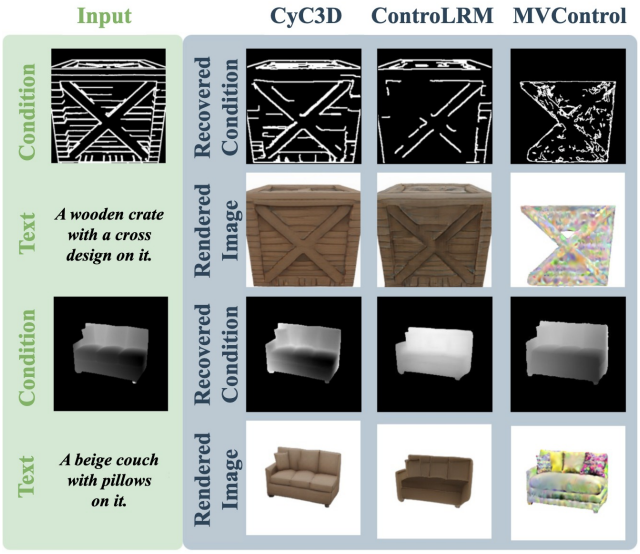


Figure 5: Qualitative comparison of controllability among our CyC3D and the existing state-of-the-art methods in controllable 3D generation.

Semantic Consistency Loss: We utilize the text prompt C_t to regularize the semantic consistency of different stages.

$$\begin{aligned} L_{\text{clip}} = \sum_{T_r} & (1 - \cos(f_{\text{clip-t}}(C_t), f_{\text{clip-i}}(\hat{I}_r))) + \\ & (1 - \cos(f_{\text{clip-t}}(C_t), f_{\text{clip-i}}(\overline{I}_i))). \end{aligned} \tag{16}$$

View Consistency Regularization: The overall view cycle consistency loss can be computed as follows:

$$L_{\text{view}} = L_{\text{render}} + \alpha \cdot L_{\text{clip}}, \tag{17}$$

Dataset	Methods	Edge			Sketch			Depth			Normal	
		PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	M-MSE↓	Z-MSE↓	R-MSE↓	NB-MSE↓	DN-CON↓
GSO	**GSGEN**	13.01	0.780	0.0526	15.07	0.772	0.0322	0.1277	0.1268	0.0311	0.0230	0.0497
	GaussianDreamer	12.41	0.766	0.0597	14.58	0.792	0.0362	0.0740	0.0947	0.0391	0.0225	0.0366
	DreamGaussian	9.94	0.686	0.1048	14.42	0.765	0.0380	0.0980	0.0916	0.0406	0.0216	0.0335
	VolumeDiffusion	13.25	0.818	0.0497	16.50	0.836	0.0231	0.1202	0.0953	0.0531	0.0190	0.0172
	3DTopia	9.63	0.704	0.1134	15.54	0.801	0.0287	0.1193	0.1064	0.0420	0.0268	0.0372
	MVControl	10.44	0.723	0.0952	14.51	0.746	0.0370	0.0555	0.0645	0.0160	0.0209	0.0315
	ControLRM-T	12.43	0.836	0.0607	17.17	0.836	0.0202	0.0567	0.0713	0.0599	0.0120	0.0359
	ControLRM-D	12.47	0.837	0.0600	17.23	0.837	0.0197	0.0383	0.0546	0.0650	0.0119	0.0357
	CyC3D-T (Ours)	**15.17**	**0.885**	**0.0342**	18.31	0.889	0.0159	**0.0051**	**0.0037**	0.0065	**0.0027**	0.0062
	CyC3D-D (Ours)	15.09	0.885	0.0347	**18.38**	**0.890**	**0.0156**	**0.0051**	0.0041	**0.0060**	0.0028	**0.0060**
ABO	**GSGEN**	12.75	0.737	0.0614	14.25	0.744	0.0384	0.1004	0.0967	0.0463	0.0171	0.0479
	GaussianDreamer	11.92	0.736	0.0699	13.75	0.734	0.0439	0.0809	0.1061	0.0521	0.0188	0.0463
	DreamGaussian	9.38	0.683	0.1178	12.90	0.716	0.0523	0.1267	0.0762	0.0433	0.0188	0.0349
	VolumeDiffusion	12.58	0.734	0.0619	15.03	0.780	0.0321	0.1007	0.1198	0.0671	0.0188	0.0264
	3DTopia	8.48	0.633	0.1469	14.41	0.741	0.0372	0.1916	0.1844	0.0559	0.0346	0.0418
	MVControl	9.52	0.656	0.1161	13.34	0.686	0.0479	0.0454	0.0467	0.0181	0.0133	0.0350
	ControLRM-T	10.49	0.552	0.0909	15.77	0.777	0.0272	0.0514	0.0497	0.0700	0.0136	0.0572
	ControLRM-D	11.51	0.592	0.0725	15.80	0.778	0.0269	0.0493	0.0545	0.0661	0.0158	0.0548
	CyC3D-T (Ours)	12.98	**0.756**	0.0540	**16.55**	**0.847**	**0.0225**	0.0080	**0.0078**	0.0102	**0.0029**	**0.0102**
	CyC3D-D (Ours)	**12.99**	0.755	**0.0539**	16.45	0.845	0.0231	**0.0075**	0.0087	**0.0091**	0.0030	0.0109

Table 1: Quantitative comparison of controllability with state-of-the-art controllable 3D generation methods on **GSO** and **ABO** benchmarks. ↑ denotes higher result is better, while ↓ means lower is better. The best results are highlighted in **bold**. We provide the results of four different conditions: edge, sketch, depth, and normal.

where α is the set to 5.0 following (Xu et al. 2024a) to balance the scale of different terms as default.

3.5 Overall Loss

Finally, the overall loss is the combination of the aforementioned losses:

$$L_{\text{total}} = \begin{cases} L_{\text{view}} + \lambda \cdot L_{\text{cond}} + \beta \cdot L_{\text{cond-d}}, & \text{if } C_v = C_d \\ L_{\text{view}} + \lambda \cdot L_{\text{cond}} + \beta \cdot L_{\text{cond-n}}, & \text{if } C_v = C_n \\ L_{\text{view}} + \lambda \cdot L_{\text{cond}}, & \text{otherwise} \end{cases} \tag{18}$$

where λ is set to 1.0 in default, and β is set to 0.1 for regularization.

4 Experiment

4.1 Implementation Details

Datasets. Our training dataset is the training split of the G-Objaverse (**GOBJ**) dataset (Qiu et al. 2024), a sub-set of Objaverse (Deitke et al. 2023). Following the selection principle of (Li et al. 2024b), we utilize the test samples collected by (Xu et al. 2024a) gathered from 2 different datasets for zero-shot evaluation: 80 samples from Google Scanned Objects dataset (**GSO**) (Downs et al. 2022), and 80 samples from Amazon Berkeley Objects dataset (**ABO**) (Collins et al. 2022).

Training Details. In analogy with (Xu et al. 2024a), we initialize our network with the weights from pretrained OpenLRM-base (Hong et al. 2023). For CyC3D-T, the conditional backbone is a transformer-based network; For CyC3D-D, the conditional backbone is a diffusion-based network. The training might cost 2 days for CyC3D-T and 3 days for CyC3D-D on 32 Nvidia V100-32G GPUs.

Methods	GOBJ→GSO			GOBJ→ABO		
	FID↓	CLIP-I↑	CLIP-T↑	FID↓	CLIP-I↑	CLIP-T↑
GSGEN	344.61	0.740	0.289	366.47	0.669	0.259
GaussianDreamer	278.70	0.810	0.300	225.38	0.787	0.277
DreamGaussian	359.65	0.760	0.279	392.95	0.723	0.247
VolumeDiffusion	299.61	0.715	0.259	350.46	0.679	0.242
3DTopia	331.39	0.727	0.283	231.55	0.751	0.272
MVControl	278.08	0.816	0.288	217.97	0.802	0.291
ControLRM-T	260.75	**0.846**	0.289	202.14	0.827	0.282
ControLRM-D	171.13	0.838	0.302	181.84	0.836	0.292
CyC3D-T (Ours)	255.30	0.840	0.295	196.55	0.836	0.291
CyC3D-D (Ours)	**168.71**	**0.846**	**0.303**	**180.34**	**0.843**	**0.294**

Table 2: Generation quality (**FID**) and CLIP score (**CLIP-I**, **CLIP-T**) comparison with state-of-the-art controllable 3D generation methods **GSO**, and **ABO** benchmarks. ↑ means higher result is better, while ↓ means lower is better. The best results are emphasized in **bold**. We provide the average results of 4 different conditions (edge, sketch, depth, and normal) in the table.

Evaluation and Metrics. We evaluate controllable 3D generation under four conditions: edge, sketch, depth, and normal. Following (Xu et al. 2024a), controllability is measured by comparing input conditions with those extracted from generated 3D content. For edge/sketch, we use PSNR, SSIM, and MSE; for depth/normal, we use MSE-based metrics. Specifically: (1) M-MSE/Z-MSE: depth similar-

Methods	Control				Quality		
	Edge	Sketch	Depth	Normal	FID	CLIP-I	CLIP-T
OpenLRM-S	0.107	0.046	0.133	0.026	268.7	0.807	0.292
OpenLRM-B	0.090	0.045	0.131	0.021	258.0	0.822	0.293
OpenLRM-L	0.079	0.045	0.138	0.023	256.4	0.824	0.294
TGS	0.052	0.042	0.140	0.021	258.9	0.841	0.295
TripoSR	0.098	0.042	0.119	0.024	261.8	0.835	0.289
MVControl	0.095	0.037	0.016	0.033	278.1	0.816	0.288
CyC3D-T	**0.034**	**0.016**	0.007	0.006	255.3	0.840	0.295
CyC3D-D	0.035	**0.016**	**0.006**	**0.006**	**168.7**	**0.846**	**0.303**

Table 3: Comparison with ControlNet-based baselines (ControlNet + Image-to-3D methods) on **GSO** benchmark.

Loss	Edge			Sketch		
	PSNR	SSIM	MSE	PSNR	SSIM	MSE
w/o L_{cond}	12.87	0.841	0.0570	17.52	0.888	0.0187
w L_{cond}	15.09	0.885	0.0347	18.38	0.890	0.0156

Table 4: Ablation experiments of 2D condition cycle consistency.

ity using Midas (Ranftl et al. 2020) or ZoeDepth (Bhat et al. 2023); (2) R-MSE: rendered vs. input depth consistency in 3D space (Xu et al. 2024a); (3) NB-MSE: normal consistency via Normal-BAE (Bae, Budvytis, and Cipolla 2021) (Li et al. 2025b); (4) DN-CON: depth-normal consistency between rendered depth and input normals (Xu et al. 2024a).

4.2 Experimental Results

Qualitative Results. Fig. 3 shows qualitative results of our CyC3D, which uses ControLRM's backbone (Xu et al. 2024a) with a pre-trained ControlNet replacing the diffusion model. A comparison with state-of-the-art methods (ControLRM (Xu et al. 2024a) and MVControl (Li et al. 2024b)) is provided in Fig. 4. Our method generates high-quality 3D content while faithfully preserving conditional details.

Controllability Evaluation. Following (Xu et al. 2024a), the comparisons of controllability with state-of-the-art controllable 3D generation methods on **GSO/ABO** benchmarks are presented in Tab. 1. From the tables, we can find that our CyC3D can achieve state-of-the-art performance in the quantitative comparison with other baselines, including GSGEN(Chen et al. 2024), GaussianDreamer(Yi et al. 2024), DreamGaussian(Tang et al. 2023a), VolumeDiffusion(Tang et al. 2023b), 3DTopia(Hong et al. 2024), MVControl(Li et al. 2024b), and ControLRM-T/D(Xu et al. 2024a). For example, the R-MSE of our CyC3D-D is 0.0060, which is much lower than the baselines of ControLRM-D (0.0650 R-MSE). Furthermore, we also provide a qualitative comparison of controllability in Fig. 5.

Generation Evaluation. To investigate whether improving controllability compromises generation quality, we provide the quantitative results of Fréchet Inception Distance (FID) (Heusel et al. 2017) and CLIP-score (Mohammad Khalid et al. 2022) following (Li et al. 2024b) in Tab. 2. Furthermore, the qualitative results are provided in Fig. 4. We report the CLIP-I metric to quantify image similarity between reference and rendered images, and the CLIP-T metric to evaluate text-image alignment using textual annotations.

Loss	Depth			Normal	
	M-MSE	Z-MSE	R-MSE	NB-MSE	DN-CON
w/o L_{cond}	0.0383	0.0546	0.0650	0.0034	0.0205
w L_{cond}	0.0192	0.0241	0.0321	0.0023	0.0187
w L_{cond}+$L_{\text{cond-d/n}}$	0.0051	0.0041	0.0060	0.0011	0.0056

Table 5: Ablation experiments of 3D condition cycle consistency.

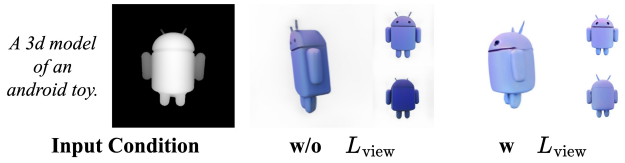


Figure 6: Ablation experiments of view cycle consistency.

Comparison with ControlNet-based Baselines. We evaluate ControlNet-based baselines combined with state-of-the-art image-to-3D models (OpenLRM (Hong et al. 2023), TGS (Zou et al. 2024), TripoSR (Tochilkin et al. 2024)), where ControlNet-generated images are converted to 3D assets using foreground masks from Rembg for background removal. Results on GSO are reported in Tab. 3.

4.3 Ablation Study

Effect of Condition Cycle Consistency. We evaluate the effectiveness of 2D (e.g., Eq. 10 for edge/sketch) and 3D (Eqs. 12, 14 for depth/normal) conditional control feedback via ablation studies in Tabs. 4 and 5. The results show that each condition cycle consistency term improves controllability across different modalities.

Effect of View Cycle Consistency. To evaluate the effectiveness of view cycle consistency feedback (e.g., Eq. 17), we provide qualitative ablation results in Fig. 6. As the figure shows, without view cycle consistency regularization, the generated 3D contents might be overfitting to the input viewpoint. Though the rendered results in the input view are reasonable, the lack of regularization on the remaining views might lead to a low-quality 3D shape.

5 Conclusion

We propose CyC3D, a novel framework for improving the controllability during 3D generation guided by different conditional controls (edge/sketch/depth/normal). The core idea of cycle-view consistency aligns input conditions with generated models via multi-view rendering and condition re-extraction feedback. The quantitative and qualitative evaluation results on GSO and ABO benchmarks show the state-of-the-art controllability of CyC3D.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62502317).

References

Bae, G.; Budvytis, I.; and Cipolla, R. 2021. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13137–13146.

Behravan, M. 2025. Generative AI Framework for 3D Object Generation in Augmented Reality. *arXiv preprint arXiv:2502.15869*.

Bhat, S. F.; Birkl, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chen, A.; Xu, H.; Esposito, S.; Tang, S.; and Geiger, A. 2025a. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, 338–355. Springer.

Chen, G.; He, Y.; Yu, M.; Yu, F. R.; Xu, G.; Ma, F.; Li, M.; and Zhou, G. 2025b. Inter3D: A Benchmark and Strong Baseline for Human-Interactive 3D Object Reconstruction. *arXiv preprint arXiv:2502.14004*.

Chen, Z.; Wang, F.; Wang, Y.; and Liu, H. 2024. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21401–21412.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21126–21136.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2553–2560. IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hong, F.; Tang, J.; Cao, Z.; Shi, M.; Wu, T.; Chen, Z.; Yang, S.; Wang, T.; Pan, L.; Lin, D.; et al. 2024. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Huang, B.; Yi, H.; Huang, C.; He, Y.; Liu, J.; and Liu, X. 2021. M3VSNet: Unsupervised multi-metric multi-view stereo network. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3163–3167. IEEE.

Li, J.; Qiu, X.; Xu, L.; Guo, L.; Qu, D.; Long, T.; Fan, C.; and Li, M. 2025a. UniF2ace: Fine-grained Face Understanding and Generation with Unified Multimodal Models. *arXiv preprint arXiv:2503.08120*.

Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; and Chen, C. 2025b. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision*, 129–147. Springer.

Li, M.; Zhou, P.; Liu, J.-W.; Keppo, J.; Lin, M.; Yan, S.; and Xu, X. 2024a. Instant3d: instant text-to-3d generation. *IJCV*.

Li, Z.; Chen, Y.; Zhao, L.; and Liu, P. 2024b. Controllable Text-to-3D Generation via Surface-Aligned Gaussian Splatting. *arXiv preprint arXiv:2403.09981*.

Liu, S.; Li, J.; Zhao, G.; Zhang, Y.; Meng, X.; Yu, F. R.; Ji, X.; and Li, M. 2025. EventGPT: Event Stream Understanding with Multimodal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 29139–29149.

Liu, Y.; An, J.; Zhang, W.; Li, M.; Wu, D.; Gu, J.; Lin, Z.; and Wang, W. 2024. RealEra: Semantic-level Concept Erasure via Neighbor-Concept Mining. *arXiv preprint arXiv:2410.09140*.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Miao, C.; Chang, T.; Wu, M.; Xu, H.; Li, C.; Li, M.; and Wang, X. 2025. FedVLA: Federated Vision-Language-Action Learning with Dual Gating Mixture-of-Experts for Robotic Manipulation. *arXiv preprint arXiv:2508.02190*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mohammad Khalid, N.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, 1–8.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qiu, L.; Chen, G.; Gu, X.; Zuo, Q.; Xu, M.; Wu, Y.; Yuan, W.; Dong, Z.; Bo, L.; and Han, X. 2024. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9914–9925.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.

Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; and Bradski, G. 2020. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3674–3683.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Shi, Y.; Yan, W.; Xu, G.; Li, Y.; Chen, Y.; Li, Z.; Yu, F. R.; Li, M.; and Yeo, S. Y. 2025. Pvchat: Personalized video chat with one-shot learning. *arXiv preprint arXiv:2503.17069*.

Su, Z.; Qiu, X.; Xu, H.; Jiang, T.; Zhuang, J.; Yuan, C.; Li, M.; He, S.; and Yu, F. R. 2025. Safe-Sora: Safe Text-to-Video Generation via Graphical Watermarking. *arXiv preprint arXiv:2505.12667*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2025a. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, 1–18. Springer.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023a. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Tang, Z.; Gu, S.; Wang, C.; Zhang, T.; Bao, J.; Chen, D.; and Guo, B. 2023b. Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. *arXiv preprint arXiv:2312.11459*.

Tang, Z.; Zhang, J.; Cheng, X.; Yu, W.; Feng, C.; Pang, Y.; Lin, B.; and Yuan, L. 2025b. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7320–7328.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*.

Wang, Q.; Luo, Y.; Shi, X.; Jia, X.; Lu, H.; Xue, T.; Wang, X.; Wan, P.; Zhang, D.; and Gai, K. 2025. CineMaster: A 3D-Aware and Controllable Framework for Cinematic Text-to-Video Generation. *arXiv preprint arXiv:2502.08639*.

Wu, Z.; Xu, H.; Xu, G.; Nie, P.; Yan, Z.; Zheng, J.; Qu, L.; Li, M.; and Nie, L. 2025. TextSplat: Text-Guided Semantic Fusion for Generalizable Gaussian Splatting. *arXiv preprint arXiv:2504.09588*.

Xu, H.; Chen, W.; Zhou, Z.; Xiao, F.; Sun, B.; Shou, M. Z.; and Kang, W. 2024a. ControLRM: Fast and Controllable 3D Generation via Large Reconstruction Model. *arXiv preprint arXiv:2410.09592*.

Xu, Y.; Ng, Y.; Wang, Y.; Sa, I.; Duan, Y.; Li, Y.; Ji, P.; and Li, H. 2024b. Sketch2Scene: Automatic Generation of Interactive 3D Game Scenes from User's Casual Sketches. *arXiv preprint arXiv:2408.04567*.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6796–6807.

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

Zhao, F.; Li, M.; Xu, L.; Jiang, W.; Gao, J.; and Yan, D. 2025. FaVChat: Unlocking Fine-Grained Facial Video Understanding with Multimodal Large Language Models. *arXiv preprint arXiv:2503.09158*.

Zhuang, C.; Hu, Y.; Zhang, X.; Cheng, W.; Bao, J.; Liu, S.; Yang, Y.; Zeng, X.; Yu, G.; and Li, M. 2025. Styleme3d: Stylization with disentangled priors by multiple encoders on 3d gaussians. *arXiv preprint arXiv:2504.15281*.

Zou, Z.-X.; Yu, Z.; Guo, Y.-C.; Li, Y.; Liang, D.; Cao, Y.-P.; and Zhang, S.-H. 2024. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10324–10335.