CoherenDream: Boosting Holistic Text Coherence in 3D Generation via Multimodal Large Language Models Feedback

Chenhan Jiang[1*] **Yihan Zeng**[2], **Dit-Yan Yeung**[1]

[1]Hong Kong University of Science and Technology
[2]Shanghai Jiao Tong University

Abstract

Score Distillation Sampling (SDS) has achieved remarkable success in text-to-3D content generation. However, SDS-based methods struggle to maintain semantic fidelity for user prompts, particularly when involving multiple objects with intricate interactions. While existing approaches often address 3D consistency through multiview diffusion model fine-tuning on 3D datasets, this strategy inadvertently exacerbates text-3D alignment degradation. The limitation stems from SDS's inherent accumulation of view-independent biases during optimization, which progressively diverges from the ideal text alignment direction. To alleviate this limitation, we propose a novel SDS objective, dubbed as Textual Coherent Score Distillation (TCSD), which integrates alignment feedback from multimodal large language models (MLLMs). Our TCSD leverages cross-modal understanding capabilities of MLLMs to assess and guide the text-3D correspondence during the optimization. We further develop 3DLLaVA-CRITIC - a fine-tuned MLLM specialized for evaluating multiview text alignment in 3D generations. Additionally, we introduce an LLM-layout initialization that significantly accelerates optimization convergence through semantic-aware spatial configuration. Our framework, CoherenDream, achieves consistent improvement across multiple metrics on TIFA subset.As the first study to incorporate MLLMs into SDS optimization, we also conduct extensive ablation studies to explore optimal MLLM adaptations for 3D generation tasks.

Project page — https://chanyn.github.io/CoherenDream

Introduction

3D content generation is essential for diverse applications, including gaming, virtual reality, and robotics simulation. Recently, significant progress has been made in text-to-3D generation through Score Distillation Sampling (SDS) (Poole et al. 2023; Wang et al. 2024; Lin et al. 2023a; Chen et al. 2023; Shi et al. 2024; Metzer et al. 2023). SDS-based methods enable high-quality and diverse 3D generation based on user-provided text inputs, effectively distilling image distributions from 2D diffusion models (Rombach et al. 2022) into parameterized 3D representations.

*Corresponding author: jchcyan@gmail.com

Despite these advances, significant challenges persist in generating 3D content that adheres to user prompts, especially those involving multiple objects and intricate interactions. The limitation stems from SDS's per-view distillation, which lacks global consistency constraints. Independent per-view optimization accumulates biases and progressively deviates from the intended text-to-3D alignment. Furthermore, while recent variants of SDS (Shi et al. 2024; Liu et al. 2024; Li et al. 2024b) adopt fine-tuned diffusion models on specific 3D datasets to enhance 3D consistency, they exacerbate the problem of textual inconsistency in 3D generations (Jiang et al. 2025; Li et al. 2024b), leading to object omissions and physically implausible spatial relationships.

How to develop coherent text-to-3D generation remains relatively unexplored. Recent research efforts, such as JointDreamer (Jiang et al. 2025), emphasize the importance of maintaining textual consistency within the original diffusion model compared to its variants (Shi et al. 2024). DreamView (Yan et al. 2025) trains a view-specific text-to-diffusion model to enhance alignment with specific viewpoints. However, both approaches still face the challenge of bias accumulation during SDS optimization. Another intuitive way to alleviate textual inconsistency in 3D generation is through compositional strategies, such as GALA3D (Zhou et al. 2024) and GraphDreamer (Gao et al. 2024). Nevertheless, GALA3D (Zhou et al. 2024) suffers from non-overlap decomposition, which leads to unnatural spatial relationships. Although GraphDreamer (Gao et al. 2024) advances the modeling of object relationships through scene graphs, it remains limited by compositional optimization, resulting in implausible object combination. In contrast, our work aims to generate coherent results within a holistic 3D representation where object relationships and spatial arrangements maintain textual consistency across all viewpoints.

Recent advances in multimodal large language models (MLLMs) demonstrate powerful cross-modal understanding capabilities, leading to successful integrations in text-to-image generation (Sun et al. 2023; Hu et al. 2024; Feng et al. 2024b; Lian et al. 2023; Cho, Zala, and Bansal 2023). While 2D counterparts are promising, directly adapting MLLMs to 3D generation is not feasible, owing to their limited grasp of 3D representation and spatial relationships. We address these limitations by reformulating text-3D alignment as across-view question-answering tasks. Our key in-

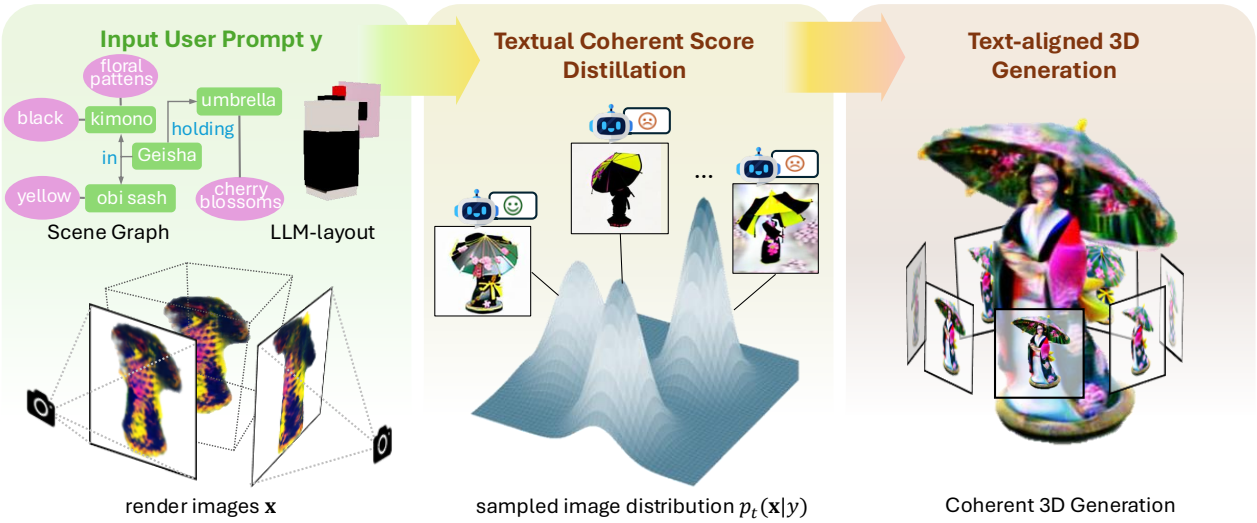

Figure 1: Textual Coherent 3D Generation with CoherenDream. By integrating multimodal LLM feedback into SDS optimization, our Textual Coherent Score Distillation corrects view-bias drift and yields faithful, text-aligned 3D content.

sight is that MLLMs' semantic reasoning capabilities can complement gradient updates from image distribution when properly contextualized within SDS optimization, as shown in Fig 1. To this end, we present **CoherenDream**, which firstly regards MLLMs as dynamic semantic assessor to ensure textual consistency in SDS optimization. We encode the input text and the across-view image distribution from the diffusion model as text-format ground truth, measuring the loss against the text sequences predicted by the MLLM to serve as feedback. This introduces Textual Coherent Score Distillation (TCSD), which adopts MLLM feedback in the SDS optimization to steer the optimization toward a textual-consistent distribution.

Current MLLMs are primarily designed as language assistants, they lack the proficiency to critique 3D generation effectively. To bridge this gap, we further develop a 3DLLaVA-CRITIC based on LLaVA-OneVision 0.5b (Li et al. 2024a) to enhance the quality of the feedback. Specifically, we design a view-aware data collection pipeline to simulate the gradient updates during SDS optimization, ensuring the delivery of accurate feedback. Furthermore, we introduce LLM-layout initialization, which is integrated with TCSD to warm up the 3D representation, thereby enhancing the textual consistency for image distribution from diffusion models. Extensive experimental results demonstrate that CoherenDream not only generates text-coherent 3D content but also outperforms other text-to-3D generation methods in terms of quality and quantitative metrics.

In summary, our contributions are as follows:

- We introduce a novel Textual Coherent Score Distillation (TCSD) for text-coherent 3D generation, guiding optimization with MLLM feedback.
- We propose a view-aware data collection pipeline and fine-tune the 3DLLaVA-CRITIC to provide accurate feedback between text and 3D representations.
- Our CoherenDream establishes a new benchmark in coherent text-to-3D generation, producing 3D content that faithfully reflects user inputs.

Related Works

SDS-based Text-to-3D Generation. The Score Distillation Sampling (SDS) algorithm has achieved surprising results in text-to-3D generation. It pioneers by (Poole et al. 2023), utilizing 2D diffusion model priors (Rombach et al. 2022) to optimize 3D representations. Recent advancements have further refined this technique by enhancing 3D representations (Lin et al. 2023a; Chen et al. 2023; Yi et al. 2024; Tang et al. 2024), improving generation quality (Huang et al. 2023; Wang et al. 2024; Zhu, Zhuang, and Koyejo 2024; Liang et al. 2024), and ensuring 3D consistency (Jiang et al. 2025; Shi et al. 2024; Li et al. 2024b; Seo et al. 2024; Armandpour et al. 2024). Despite these impressive results, these methods still struggle with multi-object prompts (He et al. 2023) and semantic interaction. These challenges often stem from the lack of global textual consistency and accumulation of view-independent biases. To address it, we introduce MLLM feedback into SDS, which dynamically revises the update direction of 3D representations, guiding the score distillation process toward text-3D alignment.

Compositional Text-to-3D Generation. An intuitive approach for text-aligned 3D generation is to decompose holistic representations into individual components for separate optimization before recombination. However, existing methods (Lin et al. 2023b; Bai et al. 2023) suffer from quality degradation due to the challenges in managing layout constraints during NeRF optimization and their dependence on potentially inaccurate predefined layouts. Recently, GALA3D (Zhou et al. 2024) attempts to use 3D

Gaussians representation while dynamically refine layout during generation. However, its non-overlapping constraint leads to unnatural spatial relationships and scale inconsistencies. GraphDreamer (Gao et al. 2024) advances the field by incorporating parsed scene graphs to guide generation. Nevertheless, it inherits the fundamental limitations of compositional approaches, struggling to maintain holistic coherence, particularly in content requiring object interactions. In this work, we warm-up a holistic 3D representation through LLM-layout for SDS optimization. The proposed TCSD further enhance optimization towards text-3D alignment. Our approach can generate holistic 3D content while facilitating more realistic interactions among various objects.

Multimodal Large Language Models. Recent advancements in large language models (LLMs)(Touvron et al. 2023; OpenAI 2023; Anil et al. 2023) have led to increased interest in Multimodal Large Language Models (MLLMs), which combine vision understanding capabilities with LLMs(OpenAI 2023; Li et al. 2024a). Given their promising abilities in visual reasoning and understanding, MLLMs are being explored to enhance 3D generation, particularly in areas such as automatic evaluation (Wu et al. 2024b; He et al. 2023) and data preparation (Sun et al. 2024; Fang et al. 2024). However, existing approaches have not yet integrated MLLMs directly into the 3D generation process. In contrast, our work fine-tunes the LLaVA model (Li et al. 2024a) to provide semantic coherence criticism and view-checking based on multi-view images. We are the first to use MLLM guidance to assist directly in SDS optimization.

Preliminaries

Score Distillation Sampling. Score Distillation Sampling (SDS) employs priors from pre-trained 2D diffusion models to facilitate the generation of 3D content, which is widely adopted in text-to-3D methods (Poole et al. 2023; Lin et al. 2023a; Chen et al. 2023; Wang et al. 2024; Luo et al. 2024). Given a user input prompt y, parameterized 3D representation θ, and pre-trained 2D diffusion model $\Phi(\mathbf{x}_t|y)$ along with noise prediction network $\epsilon_\Phi(\mathbf{x}_t, t, y)$, the objective of SDS is to optimize θ by minimizing the KL-divergence between the rendered image distribution $q_t^\theta(\mathbf{x}_t|c)$ and sampled image distribution $p_t(\mathbf{x}_t|y))$ from diffusion model:

$$\mathcal{L}_{SDS}(\theta) = \min_\theta D_{KL}(q_t^\theta(\mathbf{x}_t|c)||p_t(\mathbf{x}_t|y)). \tag{1}$$

where $\mathbf{x}_t$ represents the noisy rendered image $\mathbf{x} = g(\theta, c)$ at timestamp t, and g is the differentiable rendering function.

Ignoring the UNet Jacobian (Poole et al. 2023), the gradient computation of SDS loss is as follows:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{SDS}(\theta) &\triangleq \mathrm{E}_{t,\mathbf{x}}[w(t)\frac{\sigma_t}{\alpha_t}\nabla_\theta D_{KL}(q_t^\theta(\mathbf{x}_t|c,y)||p_t(\mathbf{x}_t|y))] \\ &\triangleq \mathrm{E}_{t,\epsilon_\Phi}[w(t)(\epsilon_\Phi(\mathbf{x}_t,t,y)-\epsilon)\frac{\partial g(\theta,c)}{\partial\theta}],\end{aligned} \tag{2}$$

α_t and σ_t are hyperparameters of noise schedule, $w(t)$ is the time-dependent weighting function, and $\epsilon_\Phi := (1+s)\epsilon_\Phi(\mathbf{x}_t,t,y) - s\epsilon_\Phi(\mathbf{x}_t,t,\emptyset)$ is the modification of noise prediction with classifier-free guidance (CFG) as s.

Method

In this section, we introduce CoherenDream, a novel text-aligned 3D generation framework as depicted in Fig. 2. We first demonstrate the derivation of Textual Coherent Score Distillation (TCSD), which introduce MLLM feedback for across-view image distribution. Then we introduce a fine-tuned MLLM, 3DLLaVA-CRITIC to better evaluate across-view alignment in 3D generation. Finally, we elaborate on the overall framework CoherenDream, where we integrate three guidance tasks with TCSD and novel LLM-layout initialization technique to further enhance textual consistency.

Textual Coherent Score Distillation (TCSD)

We observe significant limitations in textual understanding and reasoning within the original SDS framework in Eq.(1): 1) $p_t(\mathbf{x}_t|y)$ is constrained by diffusion models, which may diverge considerably from the actual user-prompt distributions (Kirstain et al. 2023; Sun et al. 2023); 2) $p_t(\mathbf{x}_t|y)$ heavily relies on the rendered image $\mathbf{x}$ of last updated 3D representation θ, leading to bias accumulation.

To address these limitations, we introduce informative feedback from MLLMs to $p_t(\mathbf{x}_t|y)$ towards textual consistent distribution. We first define an energy function $\mathcal{E}(\mathbf{x}_t, y)$ to capture text-3D alignment. Then, the ideal textual-consistent distribution $\hat{p}_t(\mathbf{x}_t|y)$ can be formulated as:

$$\hat{p}_t(\mathbf{x}_t|y) \propto p_t(\mathbf{x}_t|y)\cdot\exp(-\mathcal{E}(\mathbf{x}_t,y)) \tag{3}$$

By taking the logarithm of Eq. (3) and then its gradient with respect to $\mathbf{x}_t$, we obtain the guided score function:

$$\nabla_{\mathbf{x}_t}\log(\hat{p}_t(\mathbf{x}_t|y)) = \nabla_{\mathbf{x}_t}\log(p_t(\mathbf{x}_t|y)) - \nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t,y) \tag{4}$$

Based on the connection between diffusion models and score matching (Song and Ermon 2019), ideal text-3d aligned noise prediction can be derived as:

$$\hat{\epsilon}_\Phi(\mathbf{x}_t,t,y) = \epsilon_\Phi(\mathbf{x}_t,t,y) + \sqrt{1-\alpha_t}\cdot\nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t,y)$$

Thus, the textual consistent gradient of Textual Coherent Score Distillation can be formulated by adapting the SDS gradient (Eq.(2)) with $\hat{\epsilon}_\Phi$ as follows:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{TCSD}(\theta) &\triangleq \mathrm{E}_{t,\epsilon_\Phi}[w(t)(\hat{\epsilon}_\Phi(\mathbf{x}_t,t,y)-\epsilon)\frac{\partial g(\theta,c)}{\partial\theta}], \\ &\triangleq \mathrm{E}_{t,\epsilon_\Phi}[w(t)(\epsilon_\Phi(\mathbf{x}_t,t,y)+\lambda\cdot\nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t,y)-\epsilon)\frac{\partial g(\theta,c)}{\partial\theta}],\end{aligned} \tag{5}$$

where $\lambda = \sqrt{1-\alpha_t}$ controls the guidance strength. We define this energy function $\mathcal{E}$ as the autogressive cross-entropy loss between predefined questions/answers $\mathbf{T}$ and the output of an MLLM f_{cr}. The gradient $\nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t,y)$ thus serves as the MLLM guidance δ.

However, computing the MLLM guidance δ directly on the noisy input $\mathbf{x}_t$ is impractical, as the MLLM f_{cr} is trained on clean data and performs poorly on noisy inputs. Therefore, we approximate this gradient by calculating the feedback on the one-step denoised prediction. This final guidance $\delta(\mathbf{T}, \hat{\mathbf{x}}_0)$ for textual consistency is represented as:

$$\delta(\mathbf{T},\hat{\mathbf{x}}_0) = \nabla_{\mathbf{x}_t}\mathcal{L}_{ce}(\mathbf{T}, f_{cr}(\frac{\mathbf{x}_t-\sqrt{1-\alpha_t}\epsilon_\Phi(\mathbf{x}_t,t,y)}{\sqrt{\alpha}})) \tag{6}$$

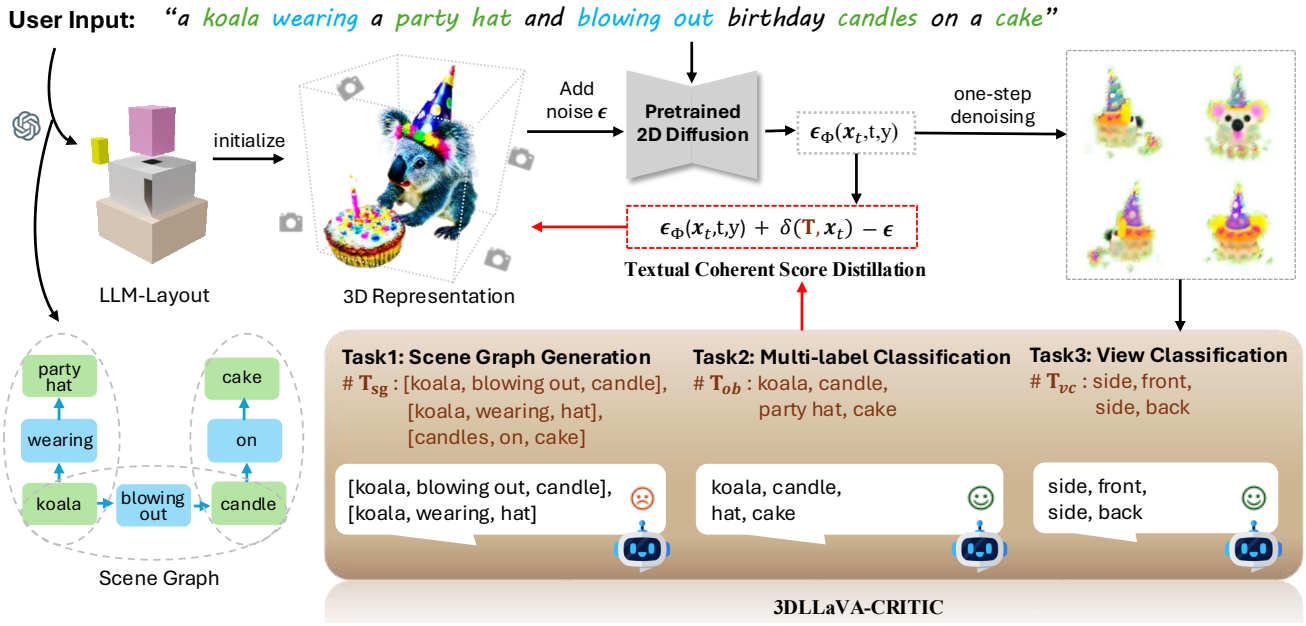

Figure 2: Overview of CoherenDream framework. CoherenDream involves LLM-Layout initialization, textual coherent score distillation and 3DLLaVA-CRITIC with three kinds of guidance tasks, producing text-3D aligned results from MLLM feedback.

3DLLaVA-CRITIC

Original MLLMs are primarily designed as language assistants and cannot assess 3D generation adequately. To bridge this gap, we decompose the evaluation of textual consistency in 3D generation into three question-answering tasks: scene graph generation, multi-label image classification, and view classification. Building on this foundation, we introduce 3DLLaVA-CRITIC, following the architecture of LLaVA-OV (Li et al. 2024a) and fine-tuned using instruct-tuning pairs generated by GPT-4o (OpenAI 2023). Detailed prompts can be found in the the Appendix.

Task Definition

- **Scene graph generation.** We establish this task to enhance global semantic understanding in across-view image distribution. Unlike caption generation, which often includes excessive and irrelevant descriptive details, the scene graph format requires the model to focus on object interactions. Specifically, given across-view observation, the 3DLLaVA-CRITIC produces a set of scene graphs represented as triplets. The form of scene graph triplet is `[subject, relation, object]` or `[object, is, attribute]`. To avoid ambiguity, we define four types of relation: Actions, Spatial Relations, Descriptive Verbs, and Non-specific Connections (e.g., "and").
- **Multi-label classification.** We observe that missing objects are a common phenomenon in textual inconsistency. Therefore, multi-label classification task asks 3DLLaVA-CRITIC to focus on primary objects presented in the user's input. Specifically, 3DLLaVA-CRITIC need to extract objects in given observation.
- **View classification.** Given input images, the 3DLLaVA-CRITIC determines the input camera position from the options: `side, front, back, overhead`. By comparing this classification with the sampled camera positions, the model guides the SDS to generate results that are consistent with the correct viewpoints.

View-aware Data Collection Pipeline Unlike natural images, the sampled image distribution in SDS is conditioned on noisy rendered images from randomly sampled camera poses. To achieve better assessment during SDS optimization, we introduce a view-aware data collection pipeline, as illustrated in Fig. 3. Detailed prompts for each step in the pipeline are available in the Appendix.

Layout condition generation. Inspired by LayoutGPT (Feng et al. 2024a), we utilize GPT-4 (OpenAI 2023) to generate 200 diverse text prompts based on examples in the DreamFusion (Poole et al. 2023) library. Layout is defined as 3D bounding boxes with box center coordinates and dimension: `(x, y, z, h, w, l)`. These layouts are rendered in Blender, and we randomly sample camera poses to produce 32 rendering images per layout.

View-aware image generation. Based on the previously created prompts and layout condition images, we enhance the original prompts by integrating view prompts (e.g., `side, back, front, overhead`). Utilizing DeepFloyd-IF (Shonenkov et al. 2023) and MVDream (Shi et al. 2024), we generate view-aware images by varying random seeds, noise strengths, and denoising steps, in line with the optimization process of CoherenDream. Under each layout condition, we randomly select 4 camera views to create a 2×2 grid image, resulting in a total of 36 grid images.

GPT-4o Annotation. After the last step, we curate 307,409 grid images and then prompt GPT-4o (OpenAI 2023) to generate scene graphs, view prompts, and perform object extraction. These high-quality grid image-text pairs serve as instruction tuning data for proposed 3DLLaVA-CRITIC.

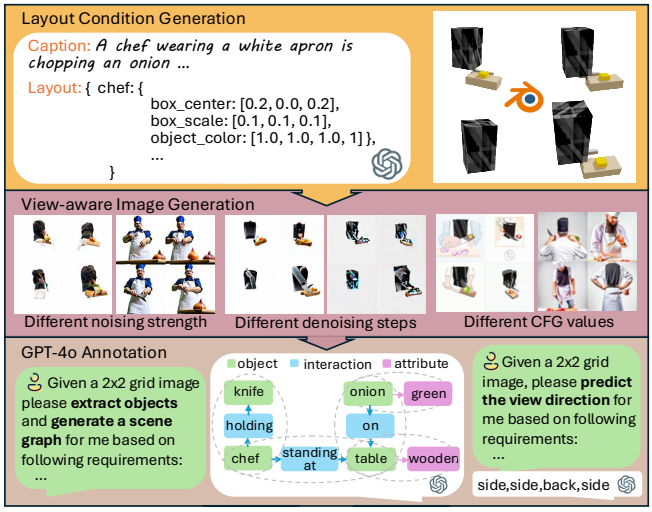

Figure 3: View-aware data collection pipeline for 3DLLaVA-CRITIC that consist of (1) using LLM to generate diverse coarse text prompt and corresponding layout and rendering from random viewpoints in Blender; (2) random sampling images from T2I diffusion model conditioned on layout image and randomly assembling into a 2×2 grid image; (3) employing GPT-4o to extract semantic annotation, including scene graph and view direction.

Framework of CoherenDream

Building upon TCSD optimization, we present the CoherenDream framework that is based on NeRF (Mildenhall et al. 2021) as 3D representation, utilizing Instant-NGP (Müller et al. 2022) with a volume renderer. To ensure coherence with user input, we incorporate three guidance tasks with Eq.(5), which direct sampled image distribution toward the textual consistent distribution. During the optimization process, we employ well-established techniques such as time-annealing (Huang et al. 2023) and resolution scaling-up (Wang et al. 2024). In addition, we introduce a novel technique called LLM-layout initialization, which significantly enhances the quality of the generated 3D content.

Guidance Tasks. We equipe TCSD optimization with three feedbacks from 3DLLaVA-CRITIC: scene graph generation, multi-label classification and view classification. Detailed task descriptions can be found before. For the scene graph generation task, $\mathbf{T}_{sg}$ can be extracted by humans or GPT according to user's input. And answers $\mathbf{T}_{ob}$ of multi-label classification can be directly obtained from $\mathbf{T}_{sg}$. $\mathbf{T}_{view}$ can be decided given the sampled camera pose. Therefore, the overall feedback of CoherenDream is demonstrated as: $\delta_{\text{CoherenDream}} = \lambda_{sg}\delta(\mathbf{T}_{sg}, \hat{\mathbf{x}}_0) + \lambda_{ob}\delta(\mathbf{T}_{ob}, \hat{\mathbf{x}}_0) + \lambda_{view}\delta(\mathbf{T}_{view}, \hat{\mathbf{x}}_0)$.

LLM-layout Initialization. As discussed in Textual Coherent Score Distillation, the rendered images $\mathbf{x}$ of 3D representations significantly influence the sampled image distribution from the diffusion model. The usage of a unit sphere for initialization provides less information condition, leading to the traption of local minimum during the score distillation optimization. It also leads to textual inconsistency. To address this issue, we introduce an LLM-generated layout to warm up CoherenDream generation, which we refer to as LLM-Layout. As illustrated in the "Layout Condition Generation" part of Fig. 3, LLM-Layout consists of a set of cubes in normalized space. The collection of LLM-Layouts mirrors the layout condition generation process detailed in 3DLLaVA-CRITIC. Specifically, we fit the density network of NeRF to align with the occupancy defined by the LLM-Layout. Since the LLM-Layout only provides fundamental placement information, we do not wish to constrain the 3D representation's flexibility regarding shape. To balance this, we implement a decay in importance near the surface, allowing for more freedom in shape representation. The surface weight decay binary cross-entropy loss function is:

$$\mathcal{L}_{\text{LLM-Layout}} = \mathcal{L}_{bce}(\text{occ}_\theta(p), \text{occ}_{\text{LLM-Layout}}(p))\dot{(}1 - e^{\frac{d^2}{2\sigma}})$$

where occ_θ indicates occupancy prediction from density network of NeRF θ, p is the points set sampled from camera view, d denotes the distance of p from the surface, and σ is a hyperparameter that controls the strength of the constraint. During the warming-up phase, which consists of $N = 600$ steps, the noising strength is restricted to the range $[0.4, 0.7]$ to encourage the sampled distribution to adhere more closely to the initialized rendered image. Additionally, the importance of $\mathcal{L}_{\text{LLM-Layout}}$ decays over the warming-up steps. This approach enables us to achieve a better initialization for the SCSD optimization without severely restricting the representational capacity of NeRF.

Experiments

We present the text-to-3D generation results of CoherenDream with qualitative and quantitative evaluations, illustrating state-of-the-art performance. We also make an ablation analysis of the proposed TCSD and 3DLLaVA-CRITIC. More details and experiments can be found in the Appendix.

Textual Consistent 3D Generation

Qualitative Comparison. Fig. 4 shows the comparison with several representative baselines, results produced by official codes. *(i) MVDream (Shi et al. 2024):* While MVDream finetunes multi-view diffusion model on 3D dataset demonstrating heavy textual inconsistency with multi-object prompts, our CoherenDream triggers generalization of multi-view diffusion model guiding by MLLMs feedback. *(ii) DreamView (Yan et al. 2025):* DreamView finetunes on 3D dataset leading to unrealistic texture and implausible interaction (e.g., the mixture of "bear" and "toy car"). Additionally, it need view-specific prompts, which cost human labor. *(iii) JointDreamer (Jiang et al. 2025):* JointDreamer maintains the generalization of original 2D diffusion, but it cannot break through the inherent limitations of diffusion model. It demonstrates object omissions ("child" in the third line) and insufficient semantic interaction understanding (targets for "beside" but shows "on" in the second row). In comparison, our Conherent can produce more text-faithful results.

Quantitative Comparison. We perform quantitative evaluations on a curated 45-prompt subset of TIFA v1.0 (Hu et al.

Figure 4: Qualitative comparison with representative methods. The results indicate that existing text-to-3D generation methods do not produce textual consistent results, involving objects omissions or unnatural interactions (highlighted in red). Conversely, our CoherenDream generates more textually faithful results that benefited from effective MLLM feedback.

Method	TIFA Score ↑	VQAScore ↑	CLIP Score ↑
MVDream (Shi et al. 2024)	77.4	0.73	30.5
DreamView (Yan et al. 2025)	77.9	0.71	31.1
JointDreamer (Jiang et al. 2025)	73.7	0.71	30.8
CoherenDream	**81.4**	**0.79**	**31.4**

Table 1: Comparison on TIFA V1.0 subset.

Method	$\text{Acc}_{\mathbf{T}_{sg}}$ ↑	$\text{Acc}_{\mathbf{T}_{ob}}$ ↑	$\text{Acc}_{\mathbf{T}_{view}}$ ↑
LlaVA-OV-0.5B	0.21	0.34	0.72
3DLlava-CRITIC	0.76	0.89	0.87

Table 2: Quantitative evaluation for 3DLlava-CRITIC.

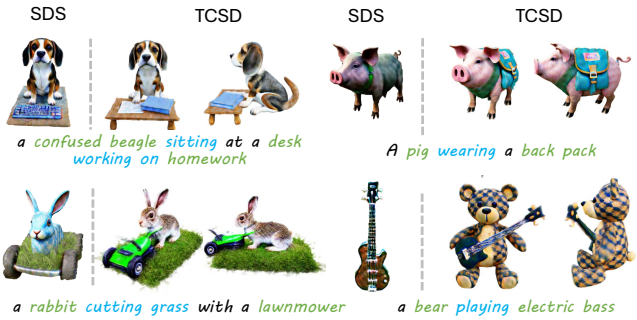

Figure 5: Comparing TCSD with Original SDS. Original SDS exhibits bias accumulation issues resulting in object omissions, while TCSD leverages a dynamic MLLM assessor to produce coherent results in a holistic 3D space.

2023). We adopt a multi-metric evaluation protocol to evaluate textual consistency: TIFA score (Hu et al. 2023), VQAScore (Lin et al. 2024) and CLIP score (Hessel et al. 2021).To ensure comprehensive viewpoint coverage, we uniformly sample 10 azimuth angles around each 3D asset. For VQAScore, we report the maximum score across all viewpoints; For TIFA Score, we compute the intersection of correct answers across all viewpoints. For Clip Score, we adopt the CLIP ViT-B/32 as the feature extractor and calculate the average score across views. Quantitative results in Table 1 demonstrate the consistent superiority of CoherenDream in text-aligned 3D generation across all metrics. Specifically, CoherenDream achieves an improvement of the TIFA score by 4.0% over MVDream, demonstrating its superior corresponding to textual description benefiting from effective feedback from MLLM.

Ablation Study

Ablation on TCSD. To verify the effectiveness of text-to-3D alignment achieved by TCSD, we distill from the multi-view diffusion model used in MVDream (Shi et al. 2024), which is known to overfit on 3D datasets. And we only utilize the original LLaVA-ov-0.5b with scene graph generation task as guidance. The primary distinction between the original SDS and TCSD, is the introduction of an optimization direction δ, which aims for ideal text alignment distribution defined in Eq.(6). Original SDS lacks the understanding and reasoning capabilities necessary for proper alignment, leading to significant misalignment issues. The incorporation of δ serves to mitigate these issues. The results presented in Fig. 5 validate our claims regarding the effectiveness of this approach.

Ablation on LLM-Layout Initialization and 3DLLaVA-CRITIC. We conduct incremental ablations on techniques in CoherenDream, focusing on LLM-Layout Initialization and the fine-tuned 3DLLaVA-CRITIC. As shown in Fig. 6, LLM-Layout Initialization enhances the warm-up process for the 3D representation, effectively activating semantic information within the diffusion model and preventing the op-

Figure 6: Incremental ablations on techniques in CoherenDream framework, which enhances text alignment.

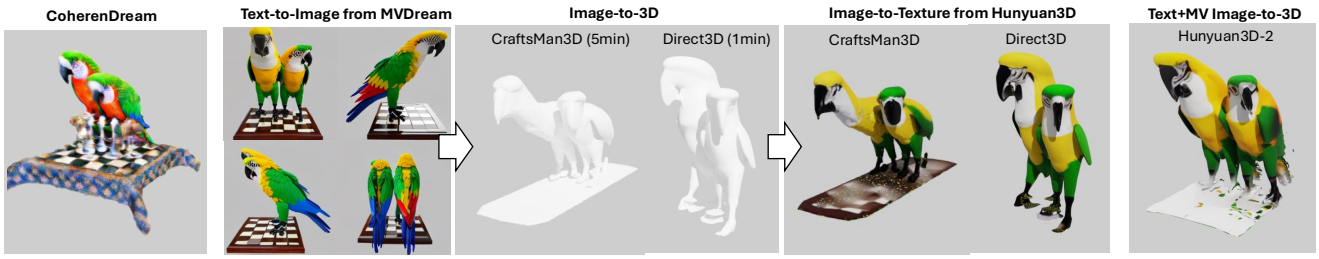

Figure 7: Compare with native 3D model in text-to-image-to-3D pipeline.

timization process from collapsing into local minima. Compared to the scenario without LLM-Layout, this technique helps the model focus on the primary objects and establishes reasonable relative scale relationships. Moreover, the proposed LLM-Layout Initialization strategy does not constrain the creativity of 3D presentations.

However, the domain gap between real images and diffusion-sampled views causes off-the-shelf MLLMs to miss fine attributes and objects (e.g., ignore "a parrot is two" or "tuxedo" as shown in Fig. 6). To address this, we fine-tune 3DLLaVA-CRITIC, which delivers more accurate text-to-3D feedback and yields significantly more faithful 3D outputs. We also quantitatively compare 3DLLaVA-CRITIC against LLaVA-OV-0.5B (Li et al. 2024a) in Table 2. In the absence of a ground-truth dataset, we construct 30 prompts for $\mathbf{T}_{sg}$ and $\mathbf{T}_{ob}$, and randomly choose 10 objects across 10 camera poses for $\mathbf{T}_{view}$. For each validation sample, GPT-4o is given both models' responses alongside the reference answer and tasked with scoring each as correct (1) or incorrect (0). The resulting average accuracies confirm the consistent gains achieved by our fine-tuning.

Compare with native 3D models. Recently, native 3D generation models have attracted attention due to their rapid inference. However, constrained by the limited size and diversity of existing 3D datasets, these native models underperform optimization-based approaches on complex or lengthy textual prompts. Fig. 7 presents a qualitative comparison between our method and several native pipelines, including Craftsman3D (Li et al. 2024c), Direct3D (Wu et al. 2024a), and Hunyuan3D-2 (Zhao et al. 2025). Most native models do not generate 3D geometry directly from text; instead, they employ a multistage text-image-textured-mesh pipeline. As a result, the final mesh quality depends heavily on the intermediate image and often omits fine details. Furthermore, errors introduced during text-to-image synthesis propagate through each stage, leading to significant misalignment between the user's prompt and the native 3D output. These pipelines also tend to produce less realistic textures than SDS-based methods. Finally, although native models offer fast image-to-3D conversion, applying textures to an initially untextured mesh remains a computationally expensive step. In contrast, our CoherenDream maintains higher fidelity to textual prompts and produces more detailed, semantically consistent 3D content.

Conclusion

In this paper, we introduce CoherenDream, a novel text-to-3D framework that leverages a powerful Multimodal Large Language Model (MLLM) to generate 3D results that are faithful to the user's inputs. We demonstrate the framework's effectiveness in optimizing score distillation through semantic feedback derived from the MLLM. Additionally, we incorporate LLM-Layout Initialization to enhance the warm-up process of the 3D representation. Our experiments show that CoherenDream achieves state-of-the-art performance in generating multi-object 3D content, excelling in both visual appearance and alignment with input text.

Acknowledgments

This work has been made possible by a Research Impact Fund project (RIF R6003-21) and a General Research Fund project (GRF 16203224) funded by the Research Grants Council (RGC) of the Hong Kong Government.

References

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Armandpour, M.; Zheng, H.; Sadeghian, A.; Sadeghian, A.; and Zhou, M. 2024. Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond. In *ICLR*.

Bai, H.; Lyu, Y.; Jiang, L.; Li, S.; Lu, H.; Lin, X.; and Wang, L. 2023. CompoNeRF: Text-guided multi-object compositional NeRF with editable 3D scene layout. *arXiv preprint arXiv:2303.13843*.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 22246–22256.

Cho, J.; Zala, A.; and Bansal, M. 2023. Visual programming for step-by-step text-to-image generation and evaluation. *Advances in Neural Information Processing Systems*, 36: 6048–6069.

Fang, Y.; Sun, Z.; Wu, T.; Wang, J.; Liu, Z.; Wetzstein, G.; and Lin, D. 2024. Make-it-Real: Unleashing Large Multimodal Model's Ability for Painting 3D Objects with Realistic Materials. *arXiv preprint arXiv:2404.16829*.

Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2024a. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Feng, Y.; Gong, B.; Chen, D.; Shen, Y.; Liu, Y.; and Zhou, J. 2024b. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4744–4753.

Gao, G.; Liu, W.; Chen, A.; Geiger, A.; and Schölkopf, B. 2024. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *CVPR*, 21295–21304.

He, Y.; Bai, Y.; Lin, M.; Zhao, W.; Hu, Y.; Sheng, J.; Yi, R.; Li, J.; and Liu, Y.-J. 2023. T3Bench: Benchmarking Current Progress in Text-to-3D Generation. *arXiv preprint arXiv:2310.02977*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; and Yu, G. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*.

Huang, Y.; Wang, J.; Shi, Y.; Tang, B.; Qi, X.; and Zhang, L. 2023. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *ICLR*.

Jiang, C.; Zeng, Y.; Hu, T.; Xu, S.; Zhang, W.; Xu, H.; and Yeung, D.-Y. 2025. JointDreamer: Ensuring Geometry Consistency and Text Congruence in Text-to-3D Generation via Joint Score Distillation. In *ECCV*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, W.; Chen, R.; Chen, X.; and Tan, P. 2024b. Sweetdreamer: Aligning Geometric Priors in 2D Diffusion for Consistent Text-to-3D. In *ICLR*.

Li, W.; Liu, J.; Yan, H.; Chen, R.; Liang, Y.; Chen, X.; Tan, P.; and Long, X. 2024c. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. In *CVPR*.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Liang, Y.; Yang, X.; Lin, J.; Li, H.; Xu, X.; and Chen, Y. 2024. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *CVPR*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023a. Magic3D: High-Resolution Text-to-3D Content Creation. In *CVPR*.

Lin, Y.; Wu, H.; Wang, R.; Lu, H.; Lin, X.; Xiong, H.; and Wang, L. 2023b. Towards language-guided interactive 3d generation: Llms as layout interpreter with generative feedback. *arXiv preprint arXiv:2305.15808*.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024. Syncdreamer: Generating multiview-consistent images from a single-view image. In *ICLR*.

Luo, W.; Hu, T.; Zhang, S.; Sun, J.; Li, Z.; and Zhang, Z. 2024. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *NeurIPS*, 36.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12663–12673.

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.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

OpenAI, R. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Seo, J.; Jang, W.; Kwak, M.-S.; Kim, H.; Ko, J.; Kim, J.; Kim, J.-H.; Lee, J.; and Kim, S. 2024. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. In *ICLR*.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2024. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*.

Shonenkov, A.; Konstantinov, M.; Bakshandaeva, D.; Schuhmann, C.; Ivanova, K.; and Klokova, N. 2023. Deepfloyd. https://huggingface.co/DeepFloyd.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Sun, J.; Fu, D.; Hu, Y.; Wang, S.; Rassin, R.; Juan, D.-C.; Alon, D.; Herrmann, C.; van Steenkiste, S.; Krishna, R.; et al. 2023. Dreamsync: Aligning text-to-image generation with image understanding feedback. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Sun, Z.; Wu, T.; Zhang, P.; Zang, Y.; Dong, X.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Bootstrap3D: Improving 3D Content Creation with Synthetic Data. *arXiv preprint arXiv:2406.00093*.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2024. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*.

Wu, S.; Lin, Y.; Zhang, F.; Zeng, Y.; Xu, J.; Torr, P.; Cao, X.; and Yao, Y. 2024a. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37: 121859–121881.

Wu, T.; Yang, G.; Li, Z.; Zhang, K.; Liu, Z.; Guibas, L.; Lin, D.; and Wetzstein, G. 2024b. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22227–22238.

Yan, J.; Gao, Y.; Yang, Q.; Wei, X.; Xie, X.; Wu, A.; and Zheng, W.-S. 2025. DreamView: Injecting View-specific Text Guidance into Text-to-3D Generation. In *ECCV*.

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 *CVPR*.

Zhao, Z.; Lai, Z.; Lin, Q.; Zhao, Y.; Liu, H.; Yang, S.; Feng, Y.; Yang, M.; Zhang, S.; Yang, X.; et al. 2025. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*.

Zhou, X.; Ran, X.; Xiong, Y.; He, J.; Lin, Z.; Wang, Y.; Sun, D.; and Yang, M.-H. 2024. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. *arXiv preprint arXiv:2402.07207*.

Zhu, J.; Zhuang, P.; and Koyejo, S. 2024. HiFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance. In *ICLR*.