CoGrad3D: Spatially-Coupled Timestep Optimization with Orthogonal Gradient Fusion for 3D Generation

Haoyang Tong[1, 2], **Hongbo Wang**[1, 2], **Jin Liu**[1, 3], **Qi Wang**[2], **Jie Cao**[1, 2*], **Ran He**[1, 2]
[1]MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]School of Information Science and Technology, ShanghaiTech University, Shanghai, China
tonghaoyang22@mails.ucas.ac.cn, wanghongbo2024@ia.ac.cn, liujin2@shanghaitech.edu.cn,
wangqi226@mails.ucas.ac.cn, jie.cao@cripac.ia.ac.cn, rhe@nlpr.ia.ac.cn

Abstract

Score Distillation Sampling has driven recent advances in text-to-3D generation. However, current approaches often fail to produce 3D assets that are both rich in detail and consistent across viewpoints. These limitations primarily arise from imbalanced guidance on fine-grained details and an overdependence on single-view optimization—issues exacerbated by the excessive randomness in selecting diffusion timesteps and camera configurations. Such deficiencies commonly lead to blurry textures and inter-view inconsistencies, which degrade visual realism and hinder practical deployment.

To tackle these challenges, we introduce **CoGrad3D**, a unified generative refinement framework that adopts a continuously adaptive optimization strategy. By dynamically modulating the optimization focus based on real-time convergence signals, CoGrad3D ensures balanced progress toward both geometric completeness and high-fidelity detail. Concretely, we propose an adaptive region sampling strategy that emphasizes under-converged viewing areas, promoting stable and uniform optimization. To facilitate the transition from coarse geometry to fine-grained reconstruction, we develop a region-aware temporal scheduling scheme that integrates global training dynamics with local convergence feedback. Furthermore, we introduce a gradient fusion mechanism that consolidates historical gradients from adjacent viewpoints, mitigating view-specific artifacts and promoting the emergence of coherent 3D structures. Extensive experiments demonstrate that CoGrad3D substantially surpasses existing methods in both geometric consistency and texture fidelity, enabling the generation of high-quality, view-consistent 3D models from textual descriptions.

1 Introduction

Text-to-Image (T2I) synthesis has achieved remarkable progress, propelled by the success of large-scale diffusion models like Stable Diffusion (Rombach et al. 2022) and Imagen (Saharia et al. 2022). While originally developed for 2D image synthesis, these powerful models are now being adapted for the more challenging domain of 3D content generation. One dominant strategy involves leveraging pre-trained T2I models as priors to guide per-prompt 3D optimization, a technique pioneered by Score Distillation Sampling (SDS) (Poole et al. 2022; Wang et al. 2023b; Ma et al. 2024; Cheng et al. 2023).

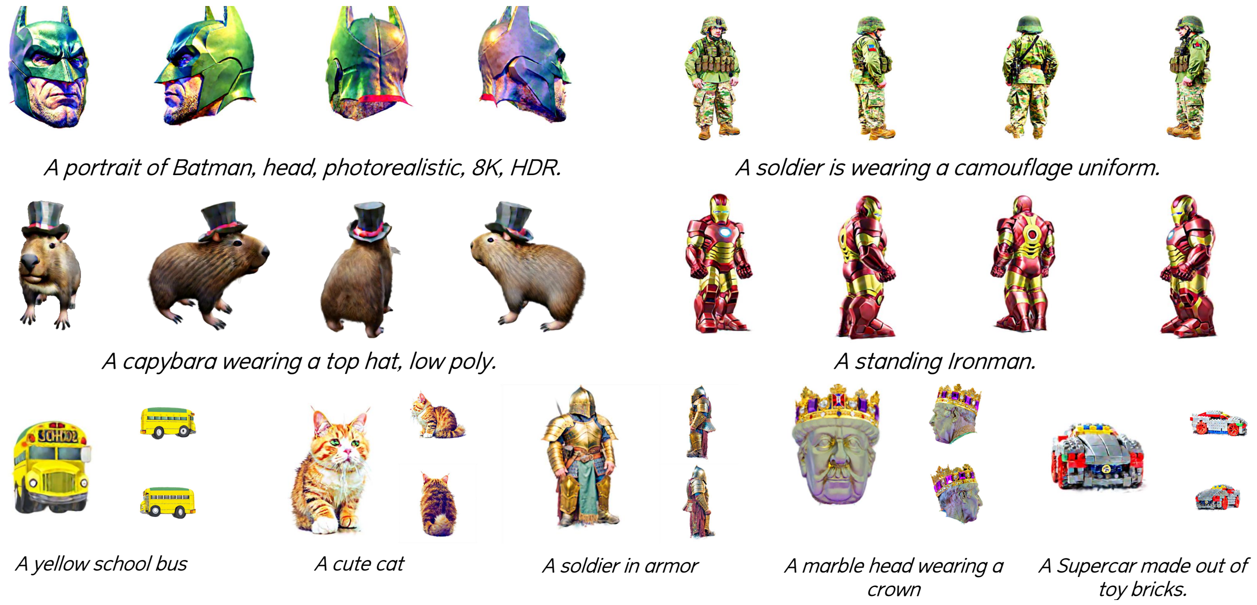
Alongside these optimization-based approaches, data-driven, feed-forward 3D generation has recently emerged as a prominent research direction (Hong et al. 2024; AlBahar et al. 2023; Team 2025). These methods are trained on large-scale 3D datasets and can achieve high inference speeds. However, their performance is often constrained by the diversity of the available 3D data, limiting their generalization to out-of-distribution concepts. In contrast, optimization-based methods like ours require no 3D training data. By leveraging the vast knowledge embedded in 2D diffusion priors, they excel at generating objects with highly detailed textures and demonstrate superior generalization to novel or long-tail concepts. Therefore, addressing the core challenges of optimization-based methods to unlock their full potential remains a critical research endeavor.

However, the per-prompt optimization approach is non-trivial. A fundamental challenge arises from the inherent mismatch between the view-independent nature of 2D diffusion models and the holistic, multi-view consistency required for coherent 3D objects. This deficiency leads to dramatic inconsistencies: features present from one angle may become distorted or vanish entirely from another. This problem is compounded by a reliance on static or coarse-grained timestep schedules, which offer limited control over the optimization process and can lead to overfitting, excessive blurring, or unrealistic geometry (Wang et al. 2024c; Vásquez and Sucar 2011).

To overcome these limitations, we introduce **CoGrad3D**, a novel optimization framework that coordinates the diffusion process across both spatial (viewpoints) and temporal (timesteps) dimensions to enhance multi-view consistency and detail preservation. Our framework integrates three key strategies. First, to combat uneven convergence, a region-aware optimization mechanism directs computational resources to under-optimized areas for focused refinement. Second, to move beyond static schedules, a hybrid timestep scheduling mechanism dynamically adjusts denoising strength based on regional progress, smoothly guiding the synthesis from coarse geometry to fine details. Third, to resolve multi-view inconsistencies, our spatio-temporal

[*]Jie Cao is the corresponding author.

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



A portrait of Batman, head, photorealistic, 8K, HDR.

A soldier is wearing a camouflage uniform.

A capybara wearing a top hat, low poly.

A standing Ironman.

A yellow school bus

A cute cat

A soldier in armor

A marble head wearing a crown

A Supercar made out of toy bricks.

Figure 1: Here we showcase examples from CoGrad3D, our proposed text-to-3D generation framework that creates 3D-consistent and highly detailed content.

gradient fusion method enables direct gradient information sharing between adjacent views, promoting a unified 3D representation. These combined techniques significantly improve the consistency, realism, and detail of the generated 3D content, demonstrating enhanced alignment with text prompts compared to baseline methods.

Our contributions are summarized as follows:

- We propose **CoGrad3D**, a unified Text-to-3D framework that coordinates optimization across spatial and temporal dimensions. Extensive experiments demonstrate that our method outperforms existing approaches in multi-view consistency, geometric fidelity, and alignment with text prompts.
- We introduce a dynamic, region-aware timestep scheduling mechanism that adapts denoising strength based on localized optimization progress, enabling a smooth and stable transition from coarse geometry to fine details.
- We develop a novel spatio-temporal gradient fusion method that explicitly shares information between adjacent views, effectively enforcing geometric coherence and enhancing visual fidelity across different perspectives.

2 Related Work

3D Generation via 2D Priors. To address the scarcity of 3D training data, recent works have leveraged powerful pre-trained 2D generative models for 3D synthesis. Early methods such as Dream Fields utilized CLIP guidance (Jain et al. 2022; Radford et al. 2021a) to capture semantics, though often at the cost of realism. A breakthrough came with SDS (Poole et al. 2022), introduced by DreamFusion, which employs 2D diffusion models as priors. SDS optimizes 3D representations by ensuring their renderings appear realistic under the guidance of the diffusion model, significantly improving fidelity. Subsequent work expanded on this paradigm by exploring diverse 3D representations (Kerbl et al. 2023; Mildenhall et al. 2021; Shen et al. 2021; Loper et al. 2023; Li et al. 2023; Tsalicoglou et al. 2024), refining distillation techniques such as VSD and ASD (Wang et al. 2023b; Ma et al. 2024; Alldieck, Kolotouros, and Sminchisescu 2024; Liang et al. 2024; Wang et al. 2023a), incorporating multi-stage refinement (Qian et al. 2023; Radford et al. 2021b; Feng et al. 2023), and enabling image-conditioned generation (Raj et al. 2023; Liu et al. 2023a). These advances harness the semantic richness of large-scale 2D datasets, enabling open-vocabulary 3D content creation. Our work builds upon this diffusion-guided methodology.

Defect Identification and Correction in 3D Generation. Despite recent progress, 3D models generated from 2D priors often exhibit notable artifacts. Common issues include semantic incompleteness, geometric inconsistencies, unrealistic or blurry textures, and floating artifacts, primarily due to the limited 3D understanding of 2D priors and the nature of the distillation process. These defects are typically identified via multi-view inspection. Various correction strategies have been proposed to mitigate such issues during generation. These include enhancing distillation stability (Hong, Ahn, and Kim 2023; Armandpour et al. 2023), enforcing geometric regularization (Yi et al. 2024b; Wu et al. 2024b; Gao et al. 2024; Dong et al. 2024; Zhou et al. 2024), improving the 3D awareness of 2D priors via fine-tuning or viewpoint conditioning (Shi et al. 2024; Wang and Shi 2023; Hu et al. 2024; Liu et al. 2024b; Seo et al. 2024; Shi et al. 2023; Liu et al. 2024a; Long et al. 2024), and employing targeted losses or sampling strategies to address specific artifacts such as view inconsistency (Huang et al. 2024b; Sun et al. 2024; Ye et al. 2024; Wang et al. 2024a,b; Zhu, Zhuang, and Koyejo 2024; Wu et al. 2024a,c). While post-processing techniques exist, integrated correction during generation is generally preferred for producing coherent and realistic 3D models.

3 Method

In this section, we first review the Diffusion Models and SDS, followed by a detailed explanation of the various components of CoGrad3D. Specifically, CoGrad3D begins by initializing a 3D view space, which is then divided into multiple regional units by uniformly discretizing azimuth, elevation, and camera distance. Each of these units is initialized with sampling weights and a structure designed to record its optimization history. In each iteration, an adaptive strategy is employed to select a target region and its neighborhood, based on accumulated optimization records and their corresponding weights (Sec. 3.2). Subsequently, a camera pose is sampled for the chosen region to render a 2D image projection. Following this, a mixed timestep, denoted as t_{mix}, derived from both local and global optimization progress, guides the diffusion model's image generation, conditioned on noise and a textual prompt (Sec. 3.3). This model generates a primary gradient, $\text{Grad}_{\text{main}}$; simultaneously, auxiliary gradients, Grad_{aux}, are gathered from similar neighboring regions (Sec. 3.4). These gradients are then fused into a mixed gradient, Grad_{mix}, which updates the 3D model via a backpropagation mechanism. Finally, the main gradient and its associated optimization outcomes are stored. This information is utilized to update the sampling weights and to inform subsequent iterations. An overview illustration of our framework is presented in Fig. 2.

3.1 Preliminaries

Diffusion Models Diffusion Models (Ho, Jain, and Abbeel 2020; Song and Ermon 2020) are generative models employing a fixed forward noising process and a learned reverse denoising process. The forward process gradually adds Gaussian noise to data x_0 over T timesteps, defined by variances β_t, with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The noisy sample x_t is drawn from $q(x_t|x_0)$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where x_t converges to standard Gaussian noise $\mathcal{N}(0, \mathbf{I})$ as $t \to T$. The reverse process starts from $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises using a network $\epsilon_\phi(x_t, t, y)$, typically trained to predict the added noise ϵ given x_t, timestep t, and conditioning y. The network parameters ϕ are optimized by minimizing a weighted mean squared error loss:

$$\mathcal{L}(\phi) = \mathbb{E}_{t,x_0,\epsilon}\left[w(t)\|\epsilon - \epsilon_\phi(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, y)\|^2\right], \tag{2}$$

where $w(t)$ is a weighting function. Inference generates samples by iteratively applying the learned ϵ_ϕ starting from noise.

Score Distillation Sampling SDS (Poole et al. 2022) optimizes a differentiable generator $g(\theta)$ by using a pre-trained diffusion model ϵ_ϕ as a prior. It relies on ϵ_ϕ approximating the score function (gradient of the log-density) at noise level t:

$$\nabla_{x_t} \log p(x_t) \approx -\frac{\epsilon_\phi(x_t, t, y)}{\sigma_t}, \quad \text{where } \sigma_t^2 = 1 - \bar{\alpha}_t. \tag{3}$$

SDS updates θ to align generated samples $x = g(\theta)$ with the diffusion prior. A noised version of the generated sample,

$$g(\theta; t) = \sqrt{\bar{\alpha}_t}\, g(\theta) + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

is used to compute the SDS gradient:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \theta) = \mathbb{E}_{t,\epsilon}\left[w(t)\,(\epsilon_\phi(g(\theta;t), t, y) - \epsilon)\,\frac{\partial g(\theta)}{\partial \theta}\right]. \tag{5}$$

Crucially, the diffusion parameters remain fixed, avoiding costly backpropagation through ϵ_ϕ and enabling efficient transfer of 2D priors to tasks such as text-to-3D generation.

3.2 Adaptive Region Sampling

Conventional text-to-3D generation methods typically employ uniform sampling over the view manifold $\mathcal{V}$, ignoring the varying optimization complexities across different viewpoints $v \in \mathcal{V}$. We empirically observe that convergence behavior for geometry and texture differs significantly across views. As a result, uniform sampling may lead to inefficient resource allocation—over-sampling views that converge quickly while under-sampling those with slower optimization dynamics or artifacts. This inefficiency becomes more pronounced for prompts involving complex scenes or multiple objects.

To address this limitation, we introduce an Adaptive Region Sampling strategy. We discretize the continuous view space $\mathcal{V}$ into N disjoint regions $\mathcal{R} = \{R_i\}_{i=1}^{N}$ by quantizing parameters such as elevation, azimuth, and camera distance. Each region R_i is tracked as an independent unit. Rather than sampling uniformly, we introduce a probabilistic scheme that dynamically adjusts sampling based on the optimization history of each region. This approach allows for the dynamic reallocation of computational resources toward regions that present greater optimization challenges, thereby improving overall efficiency.

Specifically, for each region R_i, we maintain its sampling frequency n_i and a sequence of its historical loss values, $\{L_i^k\}_{k=1}^{n_i}$. Let $\bar{L}_i$ be the moving average of the first $n_i - 1$ loss values for that region. At each sampling step, we compute a relative convergence indicator r_i for each region. This indicator is defined piecewise based on the sampling frequency n_i:

$$r_i = \frac{\bar{L}_i \times (n_i - 1) + L_i^{n_i}}{\bar{L}_i \times n_i + \epsilon_{\text{stab}}}\ , \tag{6}$$

when $n_i \geq 2$, where $\epsilon_{\text{stab}} \in \mathbb{R}^+$ is a small positive constant added to ensure numerical stability. Essentially, this indicator r_i evaluates if the most recent loss, $L_i^{n_i}$, is higher than the historical average, $\bar{L}_i$. A value of $r_i > 1$ indicates that the newest loss is greater than the past average, signaling slower convergence or instability in region R_i. For regions in the initial sampling stages ($n_i < 2$), we assign a default indicator value of $r_i = 1$ to encourage initial exploration.

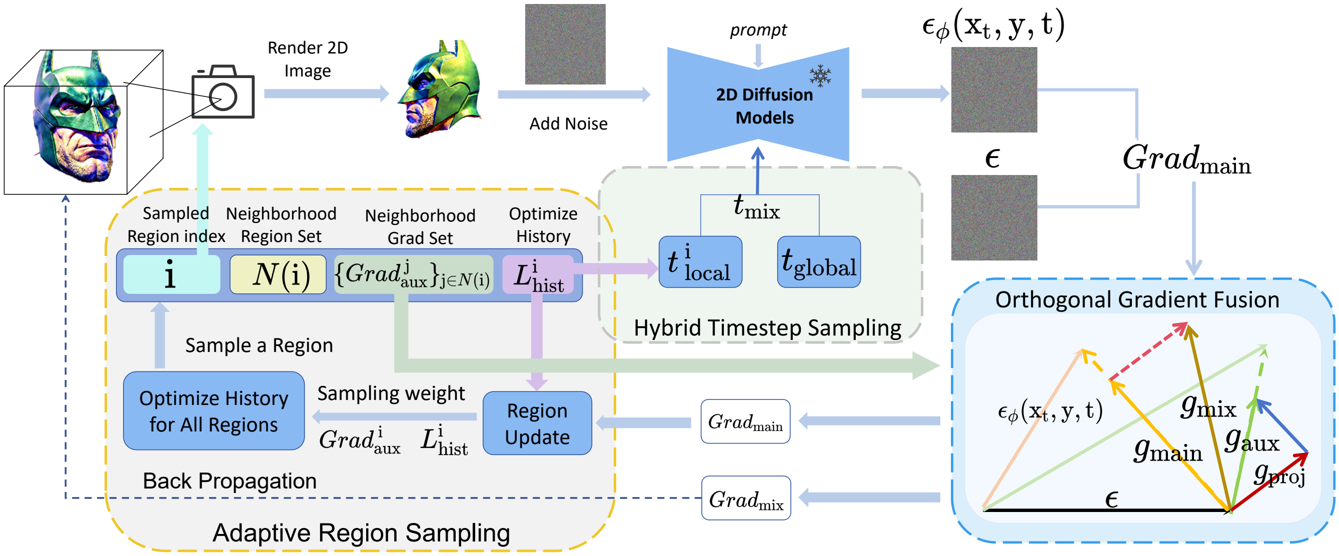


Figure 2: **Pipeline Overview.** We optimize the 3D model using an enhanced SDS framework with three intermingling technical strategies. Adaptive Region Sampling: Selects the next optimization region based on historical progress, efficiently allocating resources. Hybrid Timestep Sampling: Combines global and local loss dynamics to adjust the diffusion timestep t. Orthogonal Gradient Fusion: Enhances gradients by orthogonally projecting historical gradients from neighboring views, improving multi-view consistency while reducing interference.

We then define the sampling probability $P(i)$ for selecting region R_i based on its convergence indicator r_i. We employ an exponential weighting scheme using the Softmax function, which is modulated by a temperature parameter $\tau > 0$. The temperature τ adjusts the sharpness of the distribution: lower values concentrate sampling on poorly-performing regions (those with high r_i), while higher values yield a more uniform distribution. The probability $P(i)$ is calculated as:

$$P(i) = \frac{\exp\left(r_i/\tau\right)}{\sum_{j=1}^{N} \exp\left(r_j/\tau\right)}. \tag{7}$$

This formulation assigns exponentially higher probabilities to regions with larger indicators r_i, thereby prioritizing them for subsequent optimization steps. A region index i is then sampled according to this probability distribution $P = \{P(1), \dots, P(N)\}$. This adaptive mechanism allocates computational resources toward regions that exhibit slower convergence, as indicated by dynamic loss feedback, aiming to accelerate overall convergence and improve the global consistency and fidelity of the generated 3D model.

3.3 Hybrid Adaptive Timestep Sampling

Standard SDS often employs static timestep t sampling, where t is drawn from a fixed probability distribution $p(t)$, such as $t \sim \mathcal{U}(t_{\min}, t_{\max})$. This static strategy neglects the evolving nature of the 3D generation process, which typically transitions from coarse structural formation to fine detail refinement. Furthermore, different view regions often exhibit varying convergence patterns, as indicated by their individual loss histories L_{hist}^i, where i denotes the region index. Consequently, a static sampling strategy may apply suboptimal noise levels at different training stages or for different view regions, thereby impeding both convergence speed and the final model quality.

To address these limitations, we introduce **Hybrid Adaptive Timestep Sampling**, which dynamically selects timestep t based on both global training progress and local region-specific feedback. Let prog denote the global training progress (i.e., the ratio of current iteration to total iterations). We define a global baseline timestep t_{global}, which reflects the appropriate noise level for the current overall training stage. We also maintain an adaptive local adjustment state t_{local}^i for each region i. This local state is updated after each optimization step in region i based on dynamics extracted from its loss history L_{hist}^i. The final timestep sampling for region i combines these global and local signals, potentially also considering information from spatially neighboring regions $N(i)$.

First, the global baseline t_{global} is determined based on the overall progress prog. We employ a strategy, such as a linearly decaying log Signal-to-Noise Ratio (logSNR) schedule, to map progress to a corresponding timestep t_{global}. Concurrently, the local adjustment state t_{local}^i is updated after an optimization step in region i. This update uses features extracted from the recent loss history L_{hist}^i, including the first derivative, the second derivative, or the deviation of the current loss from its exponential moving average L_{ema}^i. A representative update rule could be:

$$\begin{aligned} u_i &:= \text{CLIP}\left(f_u(\nabla L_i, \nabla^2 L_i, (L_i - L_{\text{ema}}^i)), -\Delta t, \Delta t\right) \ , \\ t_{\text{local}}^i &\leftarrow m \cdot t_{\text{local}}^i + u_i \ , \end{aligned} \tag{8}$$

where m is a momentum factor that controls the influence of the previous t_{local}^i, Δt denotes the maximum allowable change in a single update, L_i is the latest loss value recorded in L^i_{hist} and $f_{\text{u}}(\cdot)$ calculates the update value. Specifically, we define f_u as a linear function of its inputs.

To determine the timestep t for the next optimization step in region i, we employ a strategy that incorporates both global and local information. First, a neighborhood-aware adjustment is computed by aggregating the local adjustment values of neighboring regions $j \in N(i)$ using a weighted average, where the weights are typically inversely proportional to the spatial distance between regions. This adjustment reflects the deviation of region i's state from the consensus of its neighboring regions. The local adjustment is then combined with the global baseline t_{global}, scaled by a coefficient C_{local} that controls the influence of neighborhood information. Finally, the resulting value t^i_{mix} is clipped to the permissible range to obtain the base timestep t^i_{base}.

The actual timestep t_i for region i is sampled uniformly from an interval centered at t^i_{base} with half-width Δt, introducing controlled randomness. This yields an adaptive conditional distribution: $t_i \sim p(t \mid L_{\text{hist}}, N(i), \text{prog})$, which integrates global progress, local loss dynamics, spatial context, and controlled randomness to guide timestep selection throughout training.

3.4 Hybrid Orthogonal Gradient Fusion

In the standard SDS framework, the loss L_{SDS} is typically computed based on a single rendered view $x_i = g(\theta, v_i)$ generated from the 3D representation with parameters θ at viewpoint v_i, using the corresponding text prompt y_i. The resulting gradient $g_{\text{main}} = \nabla_\theta L_{\text{SDS}}(\theta, v_i, y_i)$, is used to update θ. However, relying solely on this single-view gradient estimate can be insufficient. The objective, implicitly defined by the diffusion prior ϵ_ϕ, is to optimize θ such that the generated views match the conditional data distribution $p(x|y)$, which is related to the score function $\nabla_{x_t} \log p(x_t|y_i, t)$. Estimating the required multi-view consistent structure and the corresponding score function solely from the gradient derived from a single 2D projection x_i can be challenging. This single-view gradient g_{main} might exhibit high variance or fail to capture the complete geometric and appearance constraints necessary across all viewpoints. Consequently, it may struggle to reliably guide θ towards the desired target distribution over parameters $p(\theta|Y)$ where Y represents the prompt conditioning applied across all relevant views. This limitation is a potential contributor to multi-view inconsistency issues.

To address this issue, we propose **Hybrid Orthogonal Gradient Fusion**, which incorporates information from spatially neighboring views to improve update stability and promote multi-view consistency. The central concept is to fuse historical gradient information obtained from spatially neighboring views v_j when computing the update gradient for the current view v_i. This fusion aims to stabilize the optimization direction and provide a more robust signal for updating θ.

This fusion mechanism is activated conditionally, based on training progress or convergence metrics. When active, for neighboring views v_j belonging to the neighborhood set $\mathcal{N}(v_i)$, we retrieve relevant historical gradients g^j_{hist} which were computed previously when optimizing view v_j at a similar diffusion timestep t. We define the standard SDS gradient for the current view v_i with prompt y_i as the main gradient g_{main}. This gradient is computed using the fixed noise prediction network ϵ_ϕ and the differentiable renderer g:

$$\begin{aligned} g_{\text{main}} &= \nabla_\theta L_{\text{SDS}}(\theta, v_i, y_i) \\ &= \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(\widetilde{x}_i, y_i, t) - \epsilon)\frac{\partial x_i}{\partial \theta}\right], \end{aligned} \tag{9}$$

where $\widetilde{x}_i$ is the noised version of the rendered image $x_i = g(\theta, v_i)$, ϵ is the sampled noise, $w(t)$ is the timestep weight, $\epsilon_\phi(\widetilde{x}_i, y_i, t)$ is the noise estimate predicted by the diffusion model, and $\partial x_i / \partial\theta$ is the Jacobian of the rendering process with respect to θ.

For each selected neighboring area j, we retrieve a stored historical gradient g^j_{hist}, computed at a similar timestep t. We scale each with a similarity-based weight w_{ij} to form an auxiliary gradient $g^j_{\text{aux}} = w_{ij} g^j_{\text{hist}}$. To ensure that this auxiliary information provides complementary guidance and prevents destructive interference with g_{main}, we project each auxiliary gradient g^j_{aux} onto the subspace orthogonal to g_{main}. This projection isolates the component of g^j_{aux} that is orthogonal to g_{main}, denoted as g^j_{proj}:

$$g^j_{\text{proj}} = g^j_{\text{aux}} - \frac{\langle g^j_{\text{aux}}, g_{\text{main}} \rangle}{\|g_{\text{main}}\|^2 + \varepsilon} g_{\text{main}} \; . \tag{10}$$

The final fused gradient g_{fused}, utilized for updating the 3D representation parameters θ, is derived by combining the main gradient with the sum of all projected auxiliary gradients from neighboring views. Notably, while the fused gradient is employed in the current iteration k for the parameter update associated with view v_i, the gradient that is retained in the history for potential future use is the original, unfused main gradient g_{main} computed during this iteration. This avoids uncontrolled accumulation of fused signals and preserves long-term optimization stability.

4 Experiments

This section evaluates the performance of CoGrad3D for text-to-3D content generation. We specifically focus on its capability to generate 3D assets that align well with input text prompts, particularly those used in prior work, while maintaining visual quality and multi-view consistency. More details can be found in Sec. A in Appendix.

4.1 Experimental Setup

Evaluation Metrics. We conduct a comprehensive quantitative evaluation from several perspectives: text-3D alignment, generation quality, 3D consistency, and human preference. For text-3D alignment, we employ CLIP Similarity (Park et al. 2021) based on ViT-B/32 and ViT-L/14 backbones. We report both the mean score for alignment and the standard deviation (std). The std is calculated as the average of standard deviations of scores from results rendered at each angle. A lower value indicates a more stable and robust training process. For generation quality, in

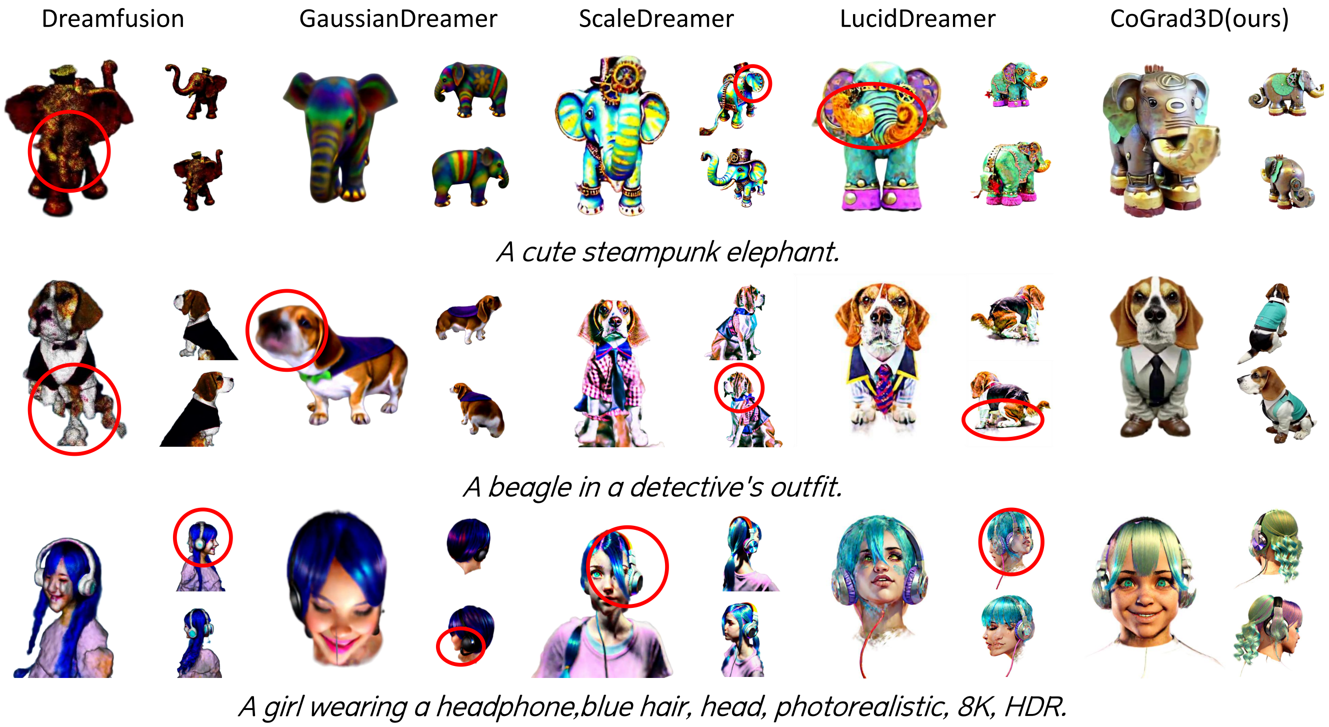

Dreamfusion
GaussianDreamer
ScaleDreamer
LucidDreamer
CoGrad3D(ours)
A cute steampunk elephant.
A beagle in a detective's outfit.
A girl wearing a headphone,blue hair, head, photorealistic, 8K, HDR.

Figure 3: Qualitative comparison between CoGrad3D (Ours) and baseline methods on representative text prompts. Artifacts or inconsistencies are marked with red circles.

the image domain, VQAscore and ImageReward serve as proxies for aesthetic quality, with partial sensitivity to spatial structure. (Huang et al. 2024a; Zhu et al. 2025) Therefore, we utilize the VQAscore (Lin et al. 2024) and the ImageReward (Xu et al. 2023) to judge the aesthetic quality. To evaluate 3D consistency, we measure the Janus Problem Rate (JR), which quantifies the frequency of multi-faced or other spatially inconsistent generations. The detailed measurements can be found in the Appendix A. A.2

User Study. To quantitatively measure human preferences, we conducted a user study involving 25 participants and 40 text prompts. For each prompt, we presented the results generated by all compared methods in a randomized order. Participants were asked to rank the 3D assets based on their adherence to the prompt, visual quality, and overall appeal. To reduce cognitive load and survey duration, each participant was assigned 10 prompt groups for ranking. The "User Rank" metric represents the average rank for each method, where a lower score indicates a higher user preference.

Comparison Targets. We compared CoGrad3D with other text-to-3D methods, many of which are built upon or modify Score Distillation Sampling (SDS). Our comparative analysis includes methods that directly utilize diffusion models for distillation, such as DreamFusion (Poole et al. 2022), TextMesh (Tsalicoglou et al. 2024), and GaussianDreamer (Yi et al. 2024a). Furthermore, we considered methods that improve upon the distillation loss formulation, exemplified by HiFA (Zhu, Zhuang, and Koyejo 2024) and PerpNeg (Armandpour et al. 2023). Additionally, we compared our work with current state-of-the-art methods, primarily including ScaleDreamer (Ma et al. 2024). Consistent3D (Wu et al. 2024c) and LucidDreamer (Liang et al. 2024).

To ensure a fair comparison, we utilized publicly available codebases whenever possible, which are predominantly integrated into the Threestudio framework (Liu et al. 2023b).

4.2 Evaluation of CoGrad3D

Qualitative Results Figure 3 showcases qualitative comparisons on various prompts. Our method, CoGrad3D, consistently generates 3D assets that more accurately adhere to the text prompts compared to the comparison targets. Furthermore, multi-view renderings demonstrate the effectiveness of our approach in producing view-consistent objects with improved geometric and textural quality.

Quantitative Results We present quantitative comparisons on 80 prompts from the DreamFusion gallery (Poole et al. 2022).

The comprehensive evaluation demonstrates that CoGrad3D achieves state-of-the-art results across most metrics, particularly in aesthetic quality. It effectively suppresses the issue of spatial inconsistency. Furthermore, our method stands out as the most popular choice among users in user studies.

Table 1: Quantitative comparison with state-of-the-art text-to-3D generation methods. We report text-3D alignment (CLIP-S), generation quality (VQAscore, ImageReward), 3D consistency (JR), and user preference (User Rank). ↑ indicates higher is better, and ↓ indicates lower is better. Best results are in **bold**.

Method	CLIP-S B/32 (mean ↑)	CLIP-S B/32 (std ↓)	CLIP-S L/14 (mean ↑)	CLIP-S L/14 (std ↓)	VQAscore ↑	ImageReward ↑	JR ↓	User Rank
DreamFusion	24.77	1.54	22.62	1.32	0.4229	-1.4925	0.5250	6.79
Textmesh	23.54	1.06	22.51	1.34	0.4131	-1.3572	0.6125	7.24
GaussianDreamer	27.47	1.22	25.42	1.22	0.4901	0.0021	0.3250	4.57
HiFA	26.72	1.62	24.01	1.49	0.4401	-0.7148	0.4750	6.09
PerpNeg	26.35	1.26	23.70	1.47	0.4071	-1.0281	0.4125	6.22
ScaleDreamer	29.85	1.24	28.08	**1.10**	0.5601	0.3810	0.4375	3.73
Consistent3D	28.76	1.29	26.90	1.23	0.5349	-0.3084	0.4625	5.02
LucidDreamer	30.02	1.19	27.74	1.24	0.5702	0.5022	**0.2000**	3.18
CoGrad3D	**31.42**	**1.02**	**28.91**	1.21	**0.6286**	**0.7449**	0.2250	**2.16**

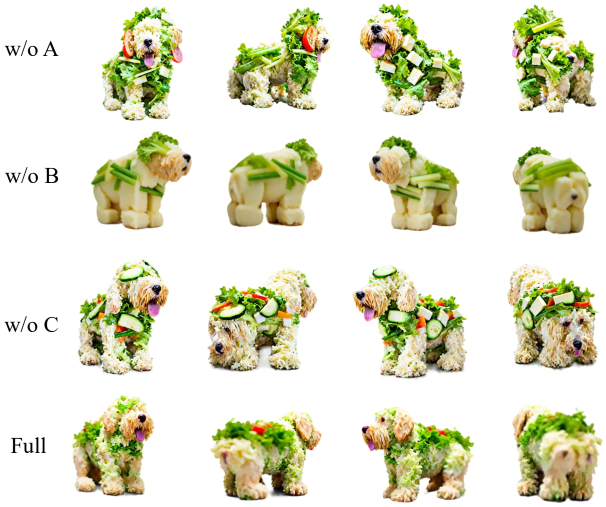


Figure 4: Qualitative results of the ablation study, illustrating the impact of each key module. The prompt used here is “a DSLR photo of a dog made out of salad”.

4.3 Ablation study

To evaluate the effectiveness of each key component in CoGrad3D, we conduct ablation studies by systematically removing each component from the full model. We denote A as Adaptive Region Sampling, B as Hybrid Adaptive Timestep Sampling, and C as Hybrid Orthogonal Gradient Fusion for ease of exposition. Specifically, w/o A replaces the adaptive sampling strategy with uniform random region sampling; w/o B uses random timesteps instead of our hybrid adaptive approach; and w/o C updates the model using only single-view gradients without performing any fusion.

Figure 4 illustrates the qualitative impact of each ablation, where we observe clear visual degradation when any component is removed. These visual results suggest that each module plays a critical role in preserving structural fidelity and semantic alignment. Specifically, the following observations were made:

Table 2: Quantitative Comparison for ablation study.

Method	CLIP-S B/32	CLIP-S L/14	VQA score	Image Reward
w/o A	28.28	24.75	0.4988	-0.0447
w/o B	29.03	24.91	0.5310	-0.4829
w/o C	27.97	25.21	0.5541	0.2448
CoGrad3D	**31.42**	**28.91**	**0.6286**	**0.7449**

- w/o A: Removing this component introduces significant spatial inconsistencies. For example, duplicated facial features (e.g., two dog faces) emerge in the second and third columns, indicating unstable focus on salient regions.
- w/o B: This ablation results in a notable loss of visual diversity. As shown in the figure, most decorative elements (e.g., curly hair, vegetable adornments) are missing, leading to a visibly reduced richness in detail.
- w/o C: Excluding this component further degrades spatial consistency. In particular, duplicated head structures (e.g., two faces appearing at both ends) are observed in the first and fourth columns.

Quantitative results are summarized in Table 2. The full CoGrad3D model consistently outperforms its ablated variants across all evaluation metrics, establishing it as the upper bound. When any of the three modules is removed, the performance drops noticeably, indicating that each component provides a distinct and essential benefit to the final model quality. This confirms the complementary roles of the three modules in enhancing CoGrad3D’s effectiveness.

5 Conclusion

In this work, we introduce CoGrad3D, a novel optimization framework incorporating three key technical strategies. It addresses critical challenges hindering the quality and consistency of Text-to-3D generation methods based on SDS, namely the loss of fine details and pervasive multi-view inconsistencies. We identified the root causes as imbalances in timestep control across generation stages and optimization instability arising from single-view gradient dominance. Extensive experiments demonstrate that CoGrad3D outperforms existing baseline methods.

Acknowledgments

This research is partially funded by several funding sources, including the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. Grant XDA0480302), National Natural Science Foundation of China (Grant Nos. 62425606, 32341009, U21B2045, 62576338), Beijing Natural Science Foundation (Grant No. L252145, QY24383), Young Scientists Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (Grant No. ES2P100116)

References

AlBahar, B.; Saito, S.; Tseng, H.-Y.; Kim, C.; Kopf, J.; and Huang, J.-B. 2023. Single-Image 3D Human Digitization with Shape-Guided Diffusion. In *SIGGRAPH*.

Alldieck, T.; Kolotouros, N.; and Sminchisescu, C. 2024. Score distillation sampling with learned manifold corrective. In *European Conference on Computer Vision*, 1–18. Springer.

Armandpour, M.; Sadeghian, A.; Zheng, H.; Sadeghian, A.; and Zhou, M. 2023. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*.

Cheng, X.; Yang, T.; Wang, J.; Li, Y.; Zhang, L.; Zhang, J.; and Yuan, L. 2023. Progressive3D: Progressively Local Editing for Text-to-3D Content Creation with Complex Semantic Prompts. In *The Twelfth International Conference on Learning Representations*.

Dong, S.; Ding, L.; Huang, Z.; Wang, Z.; Xue, T.; and Xu, D. 2024. Interactive3d: Create what you want by interactive 3d generation. 4999–5008.

Feng, L.; Wang, M.; Wang, M.; Xu, K.; and Liu, X. 2023. MetaDreamer: Efficient Text-to-3D Creation With Disentangling Geometry and Texture. *arXiv preprint arXiv:2311.10123*.

Gao, G.; Liu, W.; Chen, A.; Geiger, A.; and Schölkopf, B. 2024. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21295–21304.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, S.; Ahn, D.; and Kim, S. 2023. Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation. *Advances in Neural Information Processing Systems*, 36: 11970–11987.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2024. LRM: Large Reconstruction Model for Single Image to 3D. In *The Twelfth International Conference on Learning Representations*.

Hu, Z.; Zhao, M.; Zhao, C.; Liang, X.; Li, L.; Zhao, Z.; Fan, C.; Zhou, X.; and Yu, X. 2024. Efficientdreamer: high-fidelity and robust 3d creation via orthogonal-view diffusion priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4949–4958.

Huang, S.; Sun, S.; Wang, Z.; Qin, X.; Xiong, Y.; Zhang, Y.; Wan, P.; Zhang, D.; and Jia, J. 2024a. Placiddreamer: Advancing harmony in text-to-3d generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6880–6889.

Huang, Y.; Wang, J.; Shi, Y.; Tang, B.; Qi, X.; and Zhang, L. 2024b. Dreamtime: An Improved Optimization Strategy for Diffusion-Guided 3d Generation. In *ICLR*.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 867–876.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Li, M.; Duan, Y.; Zhou, J.; and Lu, J. 2023. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12642–12651.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6517–6526.

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. *arXiv preprint arXiv:2404.01291*.

Liu, M.; Shi, R.; Chen, L.; Zhang, Z.; Xu, C.; Wei, X.; Chen, H.; Zeng, C.; Gu, J.; and Su, H. 2024a. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10072–10083.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024b. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *ICLR*.

Liu, Y.-T.; Guo, Y.-C.; Voleti, V.; Shao, R.; Chen, C.-H.; Luo, G.; Zou, Z.; Wang, C.; Laforte, C.; Cao, Y.-P.; et al. 2023b. threestudio: a modular framework for diffusion-guided 3D generation. *cg. cs. tsinghua. edu. cn*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024.

Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9970–9980.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. 851–866.

Ma, Z.; Wei, Y.; Zhang, Y.; Zhu, X.; Lei, Z.; and Zhang, L. 2024. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *European Conference on Computer Vision*, 1–19. Springer.

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.

Park, D. H.; Azadi, S.; Liu, X.; Darrell, T.; and Rohrbach, A. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D Using 2D Diffusion. In *ICLR*.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; and Ghanem, B. 2023. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raj, A.; Kaza, S.; Poole, B.; Niemeyer, M.; Ruiz, N.; Mildenhall, B.; Zada, S.; Aberman, K.; Rubinstein, M.; Barron, J.; et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2349–2359.

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.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Seo, J.; Jang, W.; Kwak, M.; Kim, I. 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*.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: A Single Image to Consistent Multi-view Diffusion Base Model. *arXiv:2310.15110*.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *ICLR*.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

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 e-prints*, arXiv–2406.

Team, T. H. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. arXiv:2501.12202.

Tsalicoglou, C.; Manhardt, F.; Tonioni, A.; Niemeyer, M.; and Tombari, F. 2024. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*, 1554–1563. IEEE.

Vásquez, J. I.; and Sucar, L. E. 2011. Next-best-view planning for 3d object reconstruction under positioning error. In *Advances in Artificial Intelligence: 10th Mexican International Conference on Artificial Intelligence, MICAI 2011, Puebla, Mexico, November 26-December 4, 2011, Proceedings, Part I 10*, 429–442. Springer.

Wang, H.; Cao, J.; Liu, J.; Zhou, X.; Huang, H.; and He, R. 2024a. Hallo3D: Multi-modal hallucination detection and mitigation for consistent 3D content generation. *Advances in Neural Information Processing Systems*, 37: 118883–118906.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023a. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12619–12629.

Wang, P.; Fan, Z.; Xu, D.; Wang, D.; Mohan, S.; Iandola, F.; Ranjan, R.; LI, Y.; Wang, Z.; and Chandra, V. 2024b. SteinDreamer: Variance Reduction for Text-to-3D Score Distillation via Stein Identity. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Wang, P.; and Shi, Y. 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*.

Wang, P.; Xu, D.; Fan, Z.; Wang, D.; Mohan, S.; Iandola, F.; Ranjan, R.; Li, Y.; Liu, Q.; Wang, Z.; et al. 2024c. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9037–9047.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36: 8406–8441.

Wu, J.; Gao, X.; Liu, X.; Shen, Z.; Zhao, C.; Feng, H.; Liu, J.; and Ding, E. 2024a. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3202–3211.

Wu, K.; Liu, F.; Cai, Z.; Yan, R.; Wang, H.; Hu, Y.; Duan, Y.; and Ma, K. 2024b. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wu, Z.; Zhou, P.; Yi, X.; Yuan, X.; and Zhang, H. 2024c. Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9892–9902.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 15903–15935.

Ye, J.; Liu, F.; Li, Q.; Wang, Z.; Wang, Y.; Wang, X.; Duan, Y.; and Zhu, J. 2024. Dreamreward: Text-to-3d generation with human preference. 259–276.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024a. 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.

Yi, T.; Fang, J.; Zhou, Z.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Wang, X.; and Tian, Q. 2024b. GaussianDreamerPro: Text to Manipulable 3D Gaussians with Highly Enhanced Quality. *CoRR*.

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. In *International Conference on Machine Learning*, 62108–62118. PMLR.

Zhu, J.; Chen, Z.; Wang, G.; Xie, X.; and Zhou, Y. 2025. SegmentDreamer: Towards High-fidelity Text-to-3D Synthesis with Segmented Consistency Trajectory Distillation. arXiv:2507.05256.

Zhu, J.; Zhuang, P.; and Koyejo, S. 2024. HIFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance. In *The Twelfth International Conference on Learning Representations*.

A Implementation details

A.1 Core Framework and Diffusion Prior.

CoGrad3D is built upon the official implementation of LucidDreamer (Liang et al. 2024). We use stable-diffusion-2-1-base (Rombach et al. 2022) as the diffusion prior. All experiments were conducted on NVIDIA V100 GPUs.

A.2 The specific implementation method of metric.

Specifically, the metrics involved in this paper are calculated as follows:

- **CLIP Similarity:** We measure the cosine similarity between the CLIP(Radford et al. 2021a) embeddings (using specified backbones like ViT-B/32 and ViT-L/14) of rendered images (averaged across 120 views in a 360° orbit) and the CLIP embedding of the corresponding text prompt. For the quantitative evaluation, we compute this similarity for each prompt in the dataset, and then report the mean of these per-prompt similarity scores across all evaluation prompts. We may also report the average variance on the same view of these scores to indicate the stability of performance across different prompts.
- **VQAscore and ImageReward:** We selected six points at intervals of 60 degrees on the circumference as representatives of this generation, and evaluated the scores of these points using the pre-trained "clip-flant5-xl" and "ImageReward-v1.0" models, and averaged the scores of the six points. Each model was experimented on 80 prompts, and the average score of each prompt was the overall score of the model.
- **Janus Problem Rate:** We tested the results produced by different models under 80 prompts, rendered them as videos, and counted the frequency of artifact occurrence or 3d inconsistency.

A.3 hyperparameters involved in in experiment

The hyperparameters involved in Adaptive Region Sampling. The continuous observation space V is discretized into a finite set of regions. This is achieved by quantizing key camera parameters: camera elevation is discretized into 3 levels, covering the range from $-30°$ to $60°$. Azimuth is divided into 30 uniform sectors over the full $360°$. Camera distance is quantized into 2 levels; for the object-centric scenes under consideration, these levels are sampled from a distribution $U(1.5, 2.0)$. The total number of distinct regions, is the product of the number of discrete levels chosen for each parameter, resulting in 180 regions in this configuration. For each region R_i, its sampling frequency n_i and historical average loss $L_{i,k}$ are maintained. To compute the relative convergence indicator r_i , a small stability constant ϵ_{stab} is used, set to 1×10^{-6}. The sampling probability $P(i)$ for selecting region R_i employs an exponential weighting scheme, where the temperature parameter τ is set to 0.1. A lower τ value results in a stronger prioritization of regions with higher r_i (indicating slower convergence), whereas a higher τ leads to more uniform sampling across regions.

The hyperparameters involved in Hybrid Adaptive Timestep Sampling. The global reference timestep t_{global} is determined by a linearly decaying schedule based on the logSNR. Specifically, as the global training progress prog increases from 0.2 to 0.8, the logSNR is linearly interpolated between its value at $t = 1000$ and its value at $t = 1$, where t=1000 corresponds to a progress of 0, and t=1 corresponds to a progress of 1.

Each region maintains a local adjustment state t^i_{local}, which is updated based on its recent loss history $L^{\text{i}}_{\text{hist}}$. The update rule incorporates a momentum factor $m = 0.9$ and leverages the function $f_{\text{u}}(\cdot)$, which computes the update as a weighted combination of the current loss L^i_{current}, its first-order derivative ∇L_i, and the deviation δL_i from its exponential moving average L^i_{ema}:

$$f_{\text{u}}(\nabla L_i, \nabla^2 L_i, \delta L_i) = c_1 \nabla L_i + c_2 \nabla^2 L_i + c_3 \delta L_i,$$

where the weights are set to $c_1 = 0.5$, $c_2 = 0.3$, and $c_3 = 0.2$. To ensure stability, the resulting update is clipped to the range $[-0.05, 0.05]$, preventing abrupt changes in the local timestep.

To incorporate spatial coherence, a neighborhood-aware adjustment is applied using inverse-distance weighting. For each region i, the local states t^j_{local} from its neighboring regions $j \in N(i)$ are aggregated, where the neighborhood $N(i)$ includes all regions whose L1 distance from region i is less than 2. The influence of each neighboring region is scaled by a coefficient

$$C_{\text{local}} = 0.2 \times 2^{-d_{ij}},$$

with d_{ij} denoting the L1 distance between regions i and j.

The aggregated result yields a base timestep t^i_{base}, which is then clipped to the valid range $[100, 900]$. Finally, the actual timestep t_i used during optimization is sampled uniformly from the interval

$$t_i \sim \mathcal{U}(t^{\text{i}}_{\text{base}} - \Delta t,\, t^{\text{i}}_{\text{base}} + \Delta t),$$

where $\Delta t = 100$ is chosen to introduce stochasticity and encourage robustness.

The hyperparameters involved in Hybrid Orthogonal Gradient Fusion. This fusion mechanism is activated between 20% and 75% of the total optimization steps to prevent overly uniform generation in the early stages and blurriness in the later stages. For a selected neighboring historical gradient g^j_{hist}: we first select candidate gradients from regions with an L1 norm distance less than 2 from the current view v_i. From these candidates, we further select historical gradients whose optimization step difference from the current total optimization steps is less than 10% of the current total optimization steps, and whose recorded optimization timestep difference is less than 50. The weighting factor w_{ij} is set to 2^{-d_j}, where d_j is the L1 norm distance between v_i and v_j.The projection of g^j_{aux} onto the subspace orthogonal to g_{main} uses a small constant $\delta = 1 \times 10^{-8}$ in the denominator for numerical stability.The original, unfused main gradient g_{main} is stored in the history for future use, preventing uncontrolled accumulation of fused information.

The hyperparameters involved in the training process We use 3d gaussian as the 3D scene representation, optimized using the AdamW optimizer. For each experiment, we train for 6,000 iterations.

B Discussions and Limitation

B.1 Discussions with Data-Driven Methods

Recently, data-driven, feed-forward 3D content generation has become a research hotspot. However, its reliance on 3D datasets limits its generalization ability and richness. Unlike existing data-driven methods, our approach does not require any 3D dataset to train a 3D generator. On the one hand, for test text prompts that fall within the distribution of the training data, these supervised data-driven methods may generate higher-quality outputs than our unsupervised method. However, by leveraging more diverse and high-quality 2D diffusion models trained on richer datasets for 3D generation, our model is capable of generating textures with extremely high fidelity and detail. Furthermore, because its training is not confined to specific 3D assets, it exhibits excellent generalization potential when handling out-of-distribution and long-tail concepts.

B.2 Limitation

Our approach relies on a pretrained diffusion model without any 3D prior knowledge. On the one hand, it benefits from the model's ability to generate high-quality and diverse outputs; on the other hand, it may sometimes yield suboptimal results, particularly in complex 3D modeling scenarios. These dual aspects highlight key directions for the future development of 3D generation. First, it is essential to develop generative models trained specifically on datasets centered around diverse and high-quality 3D assets, which would better capture the inherent complexity and subtle variations of 3D structures. Second, to leverage richer 2D datasets, effective strategies must be designed to identify and mitigate biases transferred from pretrained models. Addressing these challenges is crucial for enhancing the accuracy and reliability of 3D generation and remains an important area for continued research.