The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Debiasing Diffusion Priors via 3D Attention for Consistent Gaussian Splatting

Shilong Jin[1], **Haoran Duan**[2], **Litao Hua**[1], **Wentao Huang**[1], **Yuan Zhou**[1]*
[1]Nanjing University of Information Science and Technology
[2]Tsinghua University
202312490769@nuist.edu.cn, haoran.duan@ieee.org, {202412621441, 202412491490, zhouyuan}@nuist.edu.cn

Abstract

Versatile 3D tasks (e.g., generation or editing) that distill from Text-to-Image (T2I) diffusion models have attracted significant research interest for not relying on extensive 3D training data. However, T2I models exhibit limitations resulting from prior view bias, which produces conflicting appearances between different views of an object. This bias causes subject-words to preferentially activate prior view features during cross-attention (CA) computation, regardless of the target view condition. To overcome this limitation, we conduct a comprehensive mathematical analysis to reveal the root cause of the prior view bias in T2I models. Moreover, we find different UNet layers show different effects of prior view in CA. Therefore, we propose a novel framework, ***TD-Attn***, which addresses multi-view inconsistency via two key components: (1) the 3D-Aware Attention Guidance Module (**3D-AAG**) constructs a view-consistent 3D attention Gaussian for subject-words to enforce spatial consistency across attention-focused regions, thereby compensating for the limited spatial information in 2D individual view CA maps; (2) the Hierarchical Attention Modulation Module (**HAM**) utilizes a Semantic Guidance Tree (SGT) to direct the Semantic Response Profiler (SRP) in localizing and modulating CA layers that are highly responsive to view conditions, where the enhanced CA maps further support the construction of more consistent 3D attention Gaussians. Notably, HAM facilitates semantic-specific interventions, enabling controllable and precise 3D editing. Extensive experiments firmly establish that TD-Attn has the potential to serve as a universal plugin, significantly enhancing multi-view consistency across 3D tasks.

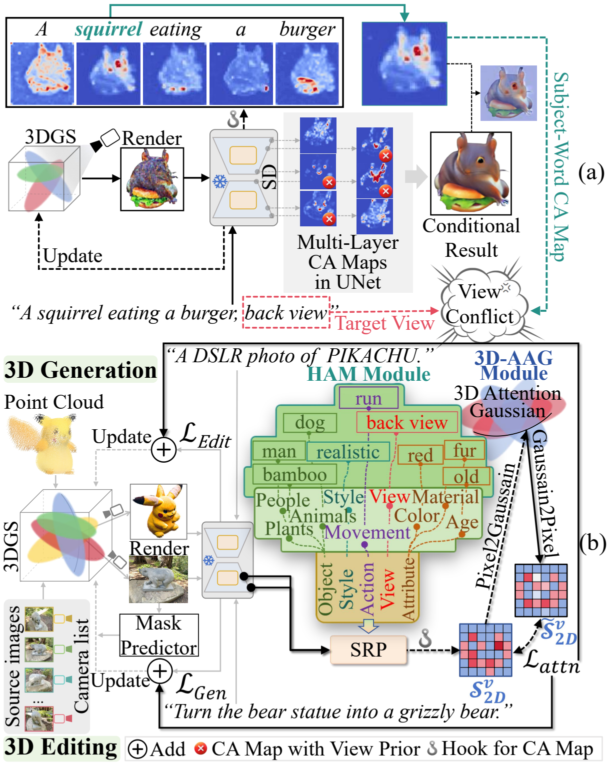

Figure 1: (a) Subject-word CA map anomalies corrupt denoising outputs with layer-varying bias intensity. (b) Overview of our TD-Attn framework integrating HAM and 3D-AAG modules for consistent 3D generation and editing.

Introduction

Text-driven 3D generation (Ding et al. 2024; Hong et al. 2022; Metzer et al. 2023; Michel et al. 2022; Guo et al. 2023) and editing (Wu et al. 2024; Koo, Park, and Sung 2024; Li et al. 2024b) are increasingly applied in industrial design, VR/AR, and digital content production. These technologies allow the creation and modification of 3D assets based on natural language input. Constrained by the complexity of 3D representation learning and the scarcity of 3D data, current text-driven 3D technologies (Liang et al. 2024; Li et al. 2024c,a; Di et al. 2025) typically combine 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) representation with distillation from text-to-image (T2I) models. Specifically, training-free distillation techniques eliminate the dependency on 3D data, while the differentiable 3DGS representation replaces Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) by enabling high-quality rendering with significant acceleration during inference.

However, a fundamental and pervasive challenge persists across existing 3D generation and editing frameworks: the inability to guarantee multi-view consistency. This manifests most prominently as the Janus problem (Armandpour et al. 2023; Hong, Ahn, and Kim 2023; Poole et al. 2022),

---

*Corresponding author.

where 3D objects have conflicting faces or orientations when rendered from different angles. This issue significantly undermines the realism and coherence of the text-to-3D output. Therefore, ensuring view consistency in 3D tasks is crucial for high-quality results, which is the motivation for this paper.

Previous studies (Armandpour et al. 2023; Zhang et al. 2024; Liu et al. 2023; Huang et al. 2024) attribute this inconsistency to the view bias of T2I diffusion models, which arises from training datasets with skewed view distributions and sparse view annotations. We theoretically validate and visualize the impact of prior view bias, as shown in Fig. 1(a). Under a "back view" condition, we input the prompt token with view annotations and its corresponding view-rendered result into the frozen UNet (Ronneberger, Fischer, and Brox 2015). The token-level attention maps show that the subject-word token (e.g., "squirrel") activates frontal-view features, and the spatial distribution of these activations correlates with the distortions observed in the conditional result. This phenomenon demonstrates that the imbalanced training data of T2I models fundamentally shapes the inductive biases of subject-word token cross-attention (CA) scores, which represent the attention regions of words in prompts across spatial features of noisy images. Since high attention scores correspond to regions where generation and editing are concentrated, this bias directly affects the spatial distribution of generated content. Specifically, the abundance of training samples without view annotations paired with prior view images causes subject-words to have higher similarity scores with prior view images, making prior view features activate more easily. When processing prompts without explicit view cues, the network defaults to generating prior view content due to these learned associations. Even when processing prompts with explicit view cues, this prior view preference dilutes or overrides the view conditions, leading to the Janus problem in 3D tasks. Additionally, Fig. 1(a) shows CA maps across multiple UNet layers, which highlights the layer-wise heterogeneity of prior view preference within the UNet.

To address this prior view preference-dominated inconsistency, we first provide a mathematical analysis of how prior view preference in T2I models compromises 3D generation consistency. The analysis reveals that the inherent view bias in T2I models causes conflicts between subject-word and view control conditions, resulting in significant negative gradients during the 3D generation process. Therefore, we propose TD-Attn, a novel framework shown in Fig. 1(b), which comprises two modules: the 3D-Aware Attention Guidance Module (3D-AAG) and the Hierarchical Attention Modulation Module (HAM). First, 3D-AAG constructs a 3D attention Gaussian for subject-word tokens significantly influenced by view preference, which accumulates multi-view 2D attention maps through inverse rendering. The Gaussian effectively dilutes erroneous attention patterns dominated by prior view preference and compensates for the limited spatial information in CA maps from independent 2D denoising processes.

Furthermore, to address the prior view preference differences across different UNet layers, we construct a hierarchical Semantic Guidance Tree (SGT) containing rich semantic words with the assistance of large language models (LLMs). HAM then utilizes the SGT to guide a Semantic Response Profiler (SRP) to localize UNet layers and CA heads with high-response view semantics, and modulates these CA layers to obtain multi-view consistent CA maps, which further support the construction of the 3D attention Gaussian. In particular, HAM not only tracks view semantics, but also enables effective localization and control of other semantics, such as color and material, providing fine-grained control over target semantics in 3D editing tasks. In summary, this paper makes the following contributions:

- We provide a comprehensive mathematical analysis demonstrating how prior view preference affects 3D tasks, with CA mechanism analysis serving as visual validation of this principle. The CA maps reveal that prior view preference from subject-word tokens override the view information from view description tokens, leading us to formulate an optimization objective centered on suppressing prior view preference in subject-word tokens.
- We propose 3D-AAG, which accumulates 2D CA maps from different views to construct a 3D attention Gaussian for subject-word tokens. 3D-AAG then constrains anomalous 2D CA scores to guarantee multi-view consistency.
- We propose HAM, which uses the semantic guidance tree to guide SRP in tracking and weighting CA layers that are highly responsive to view semantics. Then view-enhanced CA maps are further used to construct a more consistent 3D attention Gaussian.
- Proposed HAM enables targeted, semantic-specific interventions during 3D editing, giving precise control over semantics such as color, and material.

Related Work

3D Gaussian Splatting

3D Gaussian Splatting is an explicit and efficient 3D scene representation that models environments using anisotropic 3D Gaussians. Each Gaussian is characterized by its 3D position $\mu \in \mathbb{R}^3$, covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times 3}$, opacity η, and view-dependent color c represented through spherical harmonics coefficients. The representation enables real-time, differentiable rendering (Huang et al. 2025; Lin et al. 2025; Li et al. 2025; Zhou et al. 2025; Qu et al. 2025), through a splatting technique that projects these 3D Gaussians onto 2D image planes while preserving their geometric properties. During optimization, the covariance matrix is factorized into rotation $\boldsymbol{R}$ and scaling $\boldsymbol{S}$ components to ensure numerical stability. A key advantage of 3D Gaussian Splatting is its adaptive densification strategy, where Gaussians dynamically split or clone during training to capture fine scene details while preserving computational efficiency.

Text-driven 3D Generation

Early text-to-3D generation methods (Jain et al. 2022) leverage the cross-modal capabilities of CLIP (Hong et al. 2022) but suffer from limited alignment precision. The emergence

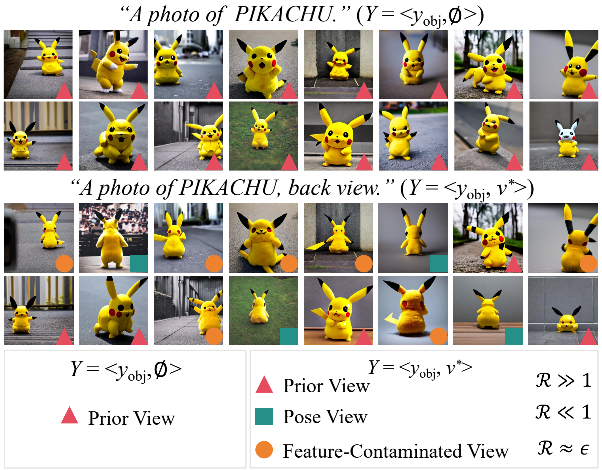

Figure 2: T2I model generation results demonstrating view bias effects. Top: results without view annotations show predominant prior view generation. Bottom: results with "back view" condition exhibit three phenomena: Prior View (red), successful Target View (green), and Feature-Contaminated View (orange).

of diffusion models (Rombach et al. 2022) establishes a new paradigm (Poole et al. 2022; Tang et al. 2023) where 2D diffusion priors guide differentiable 3D representations through distillation techniques, significantly improving generation quality. However, existing methods still face view inconsistency issues where generated 3D objects exhibit the multi-face Janus problem. While some approaches (Shi et al. 2023; Xie et al. 2024; Hong, Ahn, and Kim 2023) fine-tune diffusion models with multi-view datasets or employ geometric constraints and prompt engineering, they either require additional training overhead or lack precise view control.

Method

In this section, we present our innovative framework, TD-Attn, which is developed based on the theoretical analysis of prior view bias in the first section. The pipeline of TD-Attn is illustrated in Fig. 3. In the subsequent sections, we explore the two primary components of our methodology, which effectively guarantee view consistency in 3D tasks.

Theoretical Analysis of View Inconsistency

In T2I diffusion models, image generation can be formulated as the conditional probability distribution $p_\theta(x|Y)$, where θ represents the parameters of the model that govern the generation process, and x represents the generated image given a text prompt Y. The text prompt $Y = \langle y_{\text{obj}}, v^* \rangle$ consists of object description components y_{obj} and view conditions v^*. The training dataset $\mathcal{D} = \{\langle x_i, Y_i \rangle\}_{i=1}^{\mathcal{N}}$ typically exhibits significant view distribution bias and sparse view annotations, which can be formalized as: $p_{\mathcal{D}}(v_{\text{prior}}|y_{\text{obj}}) \gg p_{\mathcal{D}}(v_{\text{other}}|y_{\text{obj}})$ and $p_{\mathcal{D}}(Y_{v^*} = \emptyset) \gg p_{\mathcal{D}}(Y_{v^*} \neq \emptyset)$. During training, when the model learns from text prompts without view annotations, i.e. $Y = \langle y_{\text{obj}}, \emptyset \rangle$, the view information solely derives from the prior view distribution, i.e. $p(v|Y) = p(v|y_{\text{obj}}, \emptyset) \propto p_{\mathcal{D}}(v|y_{\text{obj}})$, this reveals that when processing text without view annotations, T2I models implicitly tend to favor prior view preference generation as the optimal solution. As shown in the upper part of Fig. 2, when generating from prompts such as "A photo of PIKACHU", the model generates images based on the prior view preference.

When processing prompts containing view conditions, the ideal view prior should be $p(v|Y) = \delta_{v,v^*}$, where δ_{v,v^*} is the Dirac delta function, and the model's training objective is $p_\theta^{\text{ideal}}(x|Y) = p_\theta(x|y_{\text{obj}}, v^*)$. However, the learned view distribution deviates from the ideal Dirac distribution due to training data bias, causing subject-word tokens to respond strongly to prior view features. We model this deviation using a bias coefficient ϵ where $0 < \epsilon \ll 1$:

$$p(v|Y) = (1 - \epsilon)\delta_{v,v^*} + \epsilon p_{\mathcal{D}}(v|y_{\text{obj}}), \tag{1}$$

we then analyze the probability ratio by substituting v_{prior} and v^* into $p(v|Y)$:

$$\mathcal{R} = \frac{p(v_{\text{prior}}|Y)}{p(v^*|Y)} = \frac{\epsilon p_{\mathcal{D}}(v_{\text{prior}}|y_{\text{obj}})}{(1 - \epsilon) + \epsilon p_{\mathcal{D}}(v^*|y_{\text{obj}})}, \tag{2}$$

Eq. (1) identifies three distinct generation phenomena: When $\mathcal{R} \ll 1$, the target view v^* dominates, enabling successful "back view" generation (green square marks in Fig. 2); when $\mathcal{R} \gg 1$, prior view bias overwhelms the target condition, resulting in frontal Pikachu generation despite the "back view" condition (red triangle marks); when $\mathcal{R} \approx \epsilon$, comparable view probabilities cause feature contamination, where the model simultaneously responds to both prior and target view conditions, producing images with mixed characteristics such as back-view poses with frontal facial features (orange circle marks).

In 3D tasks, these anomalous CA scores between subject-word tokens and prior view features accumulate across the multi-view optimization process. Specifically, the model optimizes a 3D representation ϕ by leveraging multiple 2D supervision signals from different views. It can be formulated as :

$$\tilde{P}_{3D}(\phi) = \prod_{v^* \in \Lambda} p_{2D}(z_\phi | y_{\text{obj}}, v^*), \tag{3}$$

where Λ represents the set of camera views, and z_ϕ denotes the rendered 2D image from the 3D representation ϕ at view v^*. The gradient with respect to the 3D parameters is:

$$\begin{aligned} &\nabla_\phi \log p_{3D}(\phi) \\ &\quad = |\Lambda| \mathbb{E}_{v^* \in \Lambda} [\nabla_{z_\phi} \log p_{2D}(z_\phi | v^*, y_{\text{obj}}) \frac{\partial z_\phi}{\partial \phi}] \\ &\quad = |\Lambda| \mathbb{E}_{v^* \in \Lambda} [(\nabla_{z_\phi} \log p_{2D}(z_\phi) \\ &\qquad\qquad + \nabla_{z_\phi} \log p_{2D}(v^*, y_{\text{obj}} | z_\phi)) \frac{\partial z_\phi}{\partial \phi}], \end{aligned} \tag{4}$$

to focus on the influence of view priors bias in generation, we can safely neglect the unconditional term $\nabla_{z_\phi} \log p_{2D}(z_\phi)$. While the unconditional term captures the global structure of the generated image, its effect on local

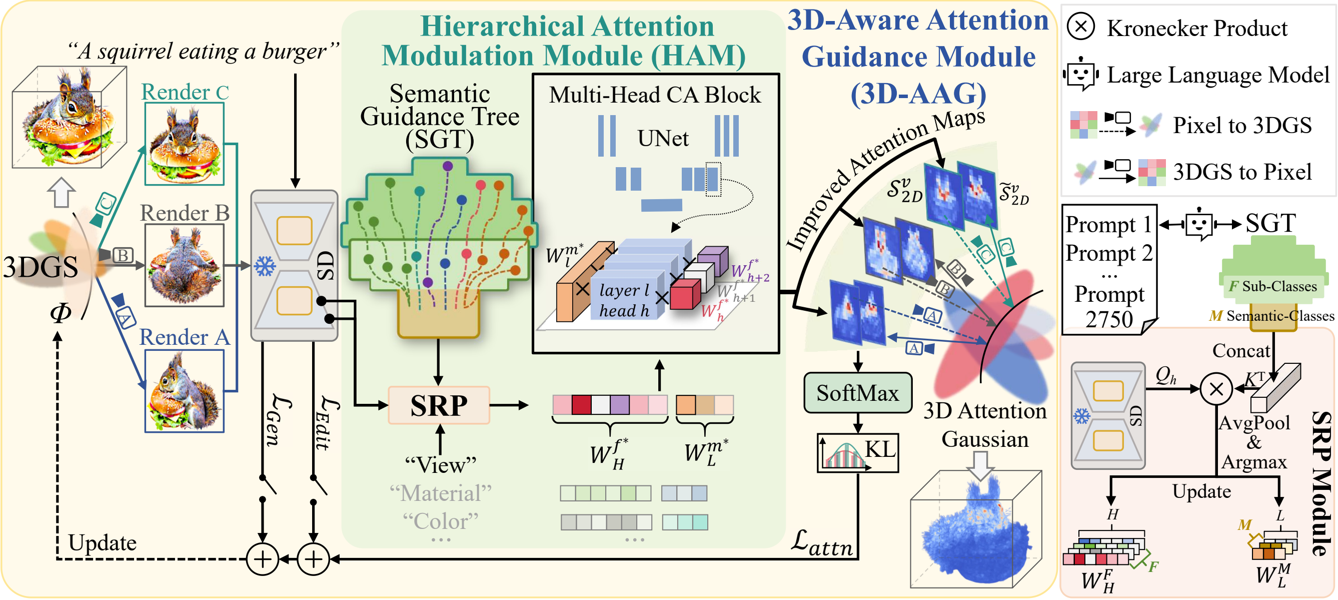

Figure 3: Overview of the TD-Attn framework. HAM guides the CA layers in UNet towards target semantics by weighting them through SRP, which is directed by the SGT. The view enhanced CA maps of subject-word token are then inversely mapped to 3DGS to generate a 3D attention Gaussian with spatial consistency. Finally, the newly rendered subject-word CA maps from the 3D attention Gaussian are used to suppress the prior preference information in the original attention maps.

features such as view is minimal in this context. Consequently, we further decompose the conditional term:

$$\nabla_{z_\phi}\log p_{2D}(v^*, y_{\text{obj}}|z_\phi) = \nabla_{z_\phi}\log p_{2D}(v^*|z_\phi) + \nabla_{z_\phi}\log p_{2D}(y_{\text{obj}}|z_\phi) + \nabla_{z_\phi}\log C, \tag{5}$$

where $C = \frac{p_{2D}(v^*, y_{\text{obj}}|z_\phi)}{p_{2D}(v^*|z_\phi)p_{2D}(y_{\text{obj}}|z_\phi)} = \frac{p_{2D}(v^*|y_{\text{obj}}, z_\phi)}{p_{2D}(v^*|z_\phi)}$. Based on our previous analysis, when certain object descriptions become overly common under prior views, strong dependencies emerge between y_{obj} and v_{prior}. In this case, when v^* approaches v_{prior}, i.e, $C \rightarrow 1$, the term $\nabla_{z_\phi}\log C$ does not significantly interfere with gradient computation. However, when v^* is far from v_{prior}, we have $p_{2D}(v^*|y_{\text{obj}}, z_\phi) \ll p_{2D}(v^*|z_\phi)$, resulting in $C \rightarrow 0$. Consequently, $\nabla_{z_\phi}\log C \ll 0$, resulting in large negative gradient effects that significantly impact the generation process.

Eq. (5) directly explains the empirical observations in Fig. 1(a), where the large negative gradients from the $\nabla_{z_\phi}\log C$ term correspond to the anomalous attention focus regions we observe in practice. This mathematical correspondence reveals why subject-word tokens consistently activate frontal-view features under "back view" conditions, creating the distortions visible in the conditional results. To mitigate this issue, we suppress the prior view preference in subject-word token attention, as shown in Fig. 3.

3D-Aware Attention Guidance Module (3D-AAG)

During each denoising step through CA layers in UNet, 2D CA maps are inversely mapped back to the 3DGS representation (Chen et al. 2024) to accumulate multi-view 2D attention weights. This enables us to construct a view-consistent 3D attention Gaussian, which compensates for the limited spatial information in CA scores computed from independent views. Thanks to the explicit nature of 3DGS, the accumulated attention weight for each Gaussian i can be computed as:

$$w_i = \sum_{v\in\Lambda}\sum_{p\in\mathcal{I}(\mathcal{S}_{2D}^v)}[o_i(p)T_i^v(p)\mathcal{I}(\mathcal{S}(p)_{2D}^v)], \tag{6}$$

where $o_i(p)$ is the opacity of Gaussian i, $T_i^v(p)$ denotes the transmittance along the ray from pixel p to Gaussian i at view v, $\mathcal{I}(\mathcal{S}(p)_{2D}^v)$ is the interpolated attention score of the subject-word token at pixel p, and $\mathcal{S}_{2D}^v$ denotes the CA map of the subject-word token at view v defined as:

$$\mathcal{S}_{2D}^v = \text{Softmax}\left(\frac{Q_v K_{\text{sbj}}^T}{\sqrt{d}}\right), \tag{7}$$

where Q_v is the query matrix projected from the spatial features of the noisy image with dimension d at view v and K_{sbj} is the key corresponding to the subject-word token. Since the 2D attention maps have different resolutions from the rendered images, bilinear interpolation $\mathcal{I}(\cdot)$ is applied to match $\mathcal{S}_{2D}^v$ to the rendering resolution, which ensures that each Gaussian accumulates attention weights from multiple views, effectively diluting the prior view bias and promoting view-consistent 3D parameter updates.

During 3D optimization, 3DGS performs adaptive densification and pruning (Kerbl et al. 2023) to adjust Gaussian distributions. The 3D attention Gaussian synchronizes with these operations, which ensures view-consistent attention patterns throughout optimization while maintaining alignment with the evolving geometric representation.

Since the 3D attention Gaussian integrates CA information from multiple views, it contains more accurate view information compared to any individual 2D CA map obtained from a single optimization step. Therefore, the accumulated 3D attention Gaussian acquires the capability to guide 2D attention maps. Specifically, to emphasize the focus regions of the rendered attention map $\widetilde{\mathcal{S}}^v_{2D}$ at view v, we apply softmax normalization and employ KL divergence loss (Kingma et al. 2014) to constrain the discrete 2D CA maps. The attention loss $\mathcal{L}_{\text{attn}}$ can be formulated as:

$$\mathcal{L}_{\text{attn}} = \text{KL}\left(\text{Softmax}(\widetilde{\mathcal{S}}^v_{2D}) \parallel \mathcal{I}(\mathcal{S}^v_{2D})\right). \tag{8}$$

Hierarchical Attention Modulation Module (HAM)

Eq. (5) demonstrates that 3D tasks suffer from gradient interference due to prior view preference when distilling T2I models, where the conditional gradient term is compromised by anomalous CA scores between subject-word tokens and prior view features. Therefore, we aim to enhance target view semantics in UNet CA layers to minimize the preference ratio $\mathcal{R}$ in Eq. (2), ensuring $\mathcal{R} \ll 1$ and guaranteeing that the posterior probability $p(v|Y)$ in Eq. (1) conforms to the target view v^*.

We adopt a multimodal perspective to examine the interaction patterns between subject-word token keys and image queries within the CA mechanism. As shown in the bottom right of Fig. 3, inspired by HRV (Park et al. 2024), SRP concatenates enriched concept words to form a key matrix K^T, which is then matched against extensive image query matrices Q to identify the highest CA scores among semantic words, thereby defining the primary conceptual focus of each attention head. Unlike HRV, we consider the inherent polysemy in natural language, where concept words often carry multiple semantic meanings. For instance, "match" can refer to a football match, a perfect pairing, or a striking tool for lighting, while "light" can denote physical illumination, object weight, or conceptual simplicity. To mitigate this ambiguity, we construct a semantic guidance tree using LLMs (Brown et al. 2020). As depicted in Fig. 1(b), the SGT employs a three-level hierarchical structure: the root level encompasses M distinct Semantic-Classes (e.g., Object, Attribute); the intermediate level contains F Sub-Classes that belong to each Semantic-Class; and the leaf level comprises F specific Instances, randomly selected from each Sub-Class set. Given the hierarchical relationships between CA heads and UNet layers, we compute separate scores for CA heads and UNet layers to correspond to different semantic granularities in the SGT. To precisely localize layers corresponding to prior view preference, we incorporate view conditions into the majority of these prompts.

For semantic localization, similarity matrices between the obtained query matrix Q and randomly selected instance-words key matrix K^T are computed. For H CA-Heads, the relevance score W^f_h of F Sub-Classes in head h and subclass f is computed as:

$$W^f_h = \text{argmax}_f \left(\text{AvgPool}\left(\frac{Q_h K^T}{\sqrt{d}}\right)\right), \tag{9}$$

and for L UNet layers, the relevance score W^m_l of M Semantic-Classes in layer l and class m is computed as:

$$W^m_l = \text{argmax}_m \left(\sum_{u \in \mathcal{U}_m} \sum_{h \in H_l} \text{AvgPool}\left(\frac{Q_h K^T_u}{\sqrt{d}}\right)\right), \tag{10}$$

where $\mathcal{U}_m$ represents the set of Instances belonging to Semantic-Class m, and H_l denotes the set of attention heads in layer l. This produces one-hot relevance scores that are accumulated into layer-wise weights $W^M_L \in \mathbb{R}^{M \times L}$ and head-wise weights $W^F_H \in \mathbb{R}^{F \times H}$.

The HAM then modulates the UNet's attention computation by applying these learned weights for target semantic f^*:

$$\hat{\mathcal{A}}_h = \lambda W^{m^*}_l W^{f^*}_h \mathcal{A}_h, \tag{11}$$

where $f^* \in m^*$ and $h \in H_l$, λ is the modulation coefficient, $W^{m^*}_l$ represents the layer-wise weight for Semantic-Class m^* at layer l, $W^{f^*}_h$ represents the head-wise weight for Sub-Class f^* at head h, $\mathcal{A}_h$ is the original attention computation at head h, and $\hat{\mathcal{A}}_h$ is the modulated attention computation.

Through this semantic-aware weighting mechanism, the model's sensitivity to target semantics can be selectively enhanced while suppressing prior biases. Notably, the view-enhanced CA maps produced by HAM further support the construction of the more consistent 3D attention Gaussian, and the improved attention consistency helps $\mathcal{L}_{\text{attn}}$ more effectively suppress prior view preference, creating a synergistic effect between our two proposed components that jointly ensures multi-view consistency in 3D tasks. For different 3D tasks, we employ task-specific loss combinations:

$$\mathcal{L} = \mathcal{L}_{\text{Gen}} + \lambda_1 \mathcal{L}_{\text{attn}} \quad \text{for generation tasks}, \tag{12}$$

$$\mathcal{L} = \mathcal{L}_{\text{Edit}} + \lambda_2 \mathcal{L}_{\text{attn}} \quad \text{for editing tasks}, \tag{13}$$

where λ_1 and λ_2 are attention balance coefficients.

Experiments

In this section, we provide a systematic evaluation of TD-Attn through a dual approach integrating qualitative visual analysis and quantitative benchmarking. First, we conduct qualitative comparisons in the first section, followed by comprehensive quantitative analysis in the second section. Finally, we provide ablation experiments to validate the effectiveness of each module in our framework. All experiments were conducted using a single NVIDIA RTX 4090 GPU. Additionally, we employ Stable Diffusion (Rombach et al. 2022) v2.1 base model for distillation in generation tasks, and adopt InstructPix2Pix (Brooks, Holynski, and Efros 2023) based on Stable Diffusion v1.4 to obtain editing results in editing tasks. More implementation details can be found in the appendix.

Qualitative Comparisons

As a universal plugin, TD-Attn can be seamlessly integrated into existing 3D generation and editing frameworks. Fig. 4 demonstrates our method's performance across diverse scenarios. For generation tasks, we integrate TD-Attn into baseline methods including DreamScene (Li et al. 2024a), LucidDreamer (Liang et al. 2024), and GCS-BEG (Li et al.

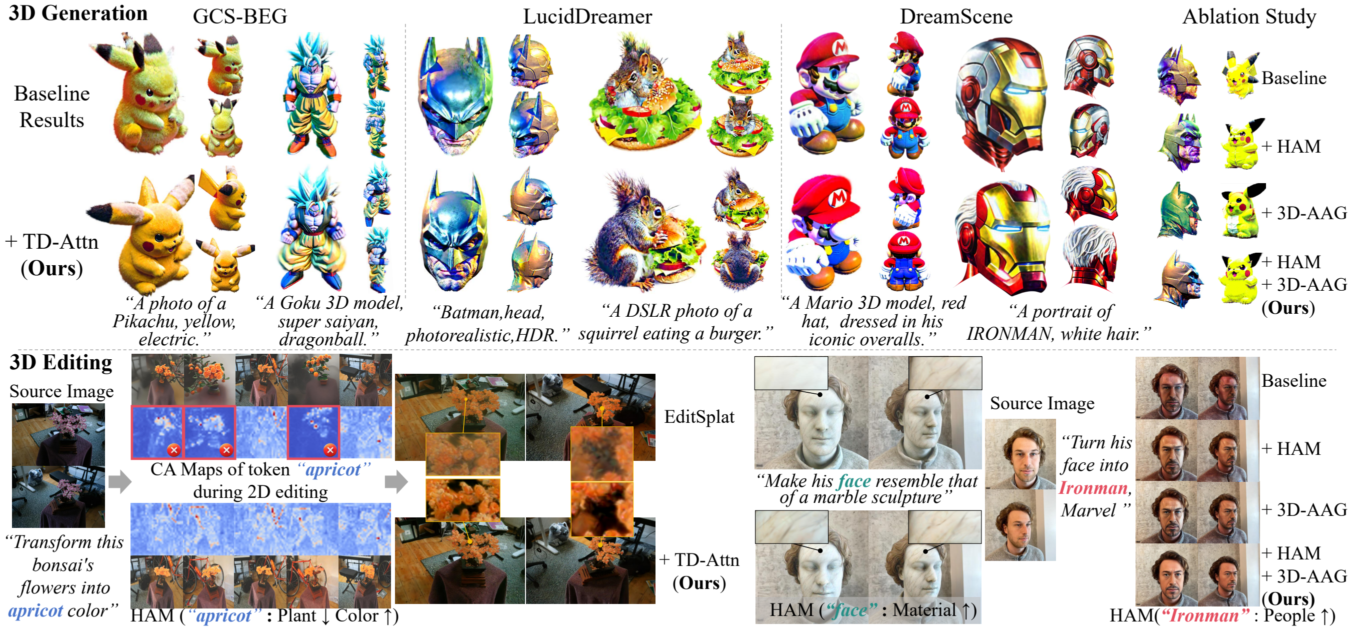

Figure 4: Qualitative comparisons. Top: 3D generation results comparing baseline methods with TD-Attn-enhanced versions, eliminating multi-view inconsistencies and visual artifacts. Right: Ablation study showing progressive improvements. Bottom: 3D editing results demonstrating semantic consistency and multi-view coherence improvements.

Methods	**User Study**		**ImageReward ↑**	**CLIP$_{\text{sim}}$↑**	**Frequency of Inconsistency**	
	Quality ↑	**Consistency ↑**			**f$_{\text{mf}}$(%) ↓**	**f$_{\text{inc}}$(%) ↓**
DreamScene	4.52	3.51	-0.725	0.299	46.7	66.7
DreamScene + Ours	**6.26**	**5.44**	**-0.229**	**0.306**	**33.3**	**53.3**
LucidDreamer	5.34	4.02	-0.386	0.309	26.7	60.0
LucidDreamer + Ours	**7.27**	**6.67**	**0.124**	**0.320**	**13.3**	**33.3**
GCS-BEG	6.13	4.18	0.158	0.312	33.3	60.0
GCS-BEG + Ours	**7.81**	**7.68**	**0.397**	**0.317**	**6.7**	**26.7**

Table 1: Quantitative comparison on 3D generation task.

2024c), with results shown in the second row of Fig. 4. Compared to baseline approaches, our method successfully eliminates extraneous limbs and facial features that plague existing methods, such as additional ears in Pikachu and Batman, or extra faces in Mario, Ironman, and Squirrel.

For 3D editing tasks, we employ state-of-the-art (SOTA) EditSplat (Lee et al. 2025) as our baseline, with improvements demonstrated in the lower portion of Fig. 4. When processing prompts such as "Transform this bonsai's flowers into apricot color", baseline methods often misinterpret "apricot" as a botanical entity rather than a color descriptor. We visualize the CA maps of token "apricot" during the 2D editing process. As shown in Fig. 4, the baseline method's attention maps exhibit erroneous activations on plant-related regions (marked with red crosses), indicating that the model incorrectly associates "apricot" with botanical entities despite the contextual cue "color". This semantic misinterpretation translates to inconsistent editing results across multiple views. Similarly, when handling "Turn his face into Ironman, Marvel", editing results exhibit hybrid states combining human facial features with armor characteristics. TD-Attn provides a way to modulate semantic intensity of specific words within prompts. As visualized in the lower-left portion of Fig. 4, we enhance the "color" semantics of "apricot" while suppressing its "plant" semantics, ensuring semantic consistency of "apricot" throughout the editing process. The HAM module successfully redirects attention toward color-relevant regions, as demonstrated in the corrected CA maps. In contrast, baseline methods frequently exhibit semantic materialization of "apricot" entities, which conflicts with the editing intent. Furthermore, baseline results suffer from multi-view editing inconsistencies, exemplified when processing "Make his face resemble that of a marble sculpture", where different viewpoints exhibit inconsistent facial textures, a phenomenon that TD-Attn successfully mitigates.

Quantitative Comparisons

To intuitively demonstrate TD-Attn's performance, we compare baseline methods with their TD-Attn integrated coun-

Methods	**CLIP$_{\text{sim}}$**↑	**Frequency of Inconsistency**	
		f_{mf}(%)↓	**f_{inc}(%)**↓
Janus issue	**0.318**	100.0	100.0
Baseline	0.307	35.6	62.2
+ HAM	0.311	20.0	57.8
+ 3D-AAG	0.313	24.4	44.4
TD-Attn	0.314	**17.8**	**37.8**

Table 2: Ablation study for generation task.

Methods	**CLIP$_{\text{sim}}$**↑	**CLIP$_{\text{dir}}$**↑	**User Study** ↑
Baseline(EditSplat)	0.253	0.101	4.18
+ HAM	0.272	0.112	5.26
+ 3D-AAG	0.261	0.102	4.73
TD-Attn	**0.277**	**0.114**	**6.34**

Table 3: Ablation study for editing task.

terparts. For generation tasks, we employed standard evaluation metrics ImageReward (Xu et al. 2023) and CLIP$_{\text{sim}}$ (Radford et al. 2021) as objective measures. Additionally, we evaluated the percentage of generated results exhibiting Inconsistency, denoted as f_{inc}. We extended the traditional Janus problem definition to encompass broader inconsistencies including extra limbs and other implausible artifacts. For comprehensive evaluation, we selected diverse prompts involving objects with distinct front-back differences (e.g., portraits, animals, vehicles). The quantitative results are summarized in Tab. 1, where f_{mf} denotes the frequency of Janus problem. As shown in the table, our method achieves the highest scores across all metrics, with particularly notable improvements in mitigating the Janus issue, where our method outperforms the baseline average by approximately 50%.

We conducted a user study comparing our method with the SOTA approaches for subjective evaluation. Using a 1-to-10 scale, participants assessed models based on two key criteria: "Generation Quality" and "3D Consistency." All samples were randomly selected without cherry-picking. As shown in Tab. 1, our method achieves the highest average scores across both evaluation metrics. These results demonstrate that 3D content generated by our approach consistently exhibits superior visual appeal and view consistency.

Ablation Studies

As shown in Fig. 4, both HAM and 3D-AAG effectively improve multi-view consistency, with their combination achieving the best performance. For generation tasks, as shown in Tab. 2, the combination yields the most significant improvements, surpassing baseline methods by 50% and 39.2%.

However, we observed that CLIP$_{\text{sim}}$ metric yields higher scores for generation results with the Janus problem, as shown in Tab. 2. Fig. 5 presents the CLIP similarity distribution between prompts and rendered results for different methods. Under normal conditions, similarity scores decrease progressively from front view to back view, with back views scoring lowest due to reduced information content. However, when the Janus problem occurs, prior preference view features appear across multiple angles, resulting in abnormally high similarity scores. This observation motivates exploring more suitable evaluation methods. Analyzing similarity score distributions across views proves more effective than relying solely on average similarity scores. Results demonstrate that our method produces similarity distributions more closely aligned with the ideal pattern.

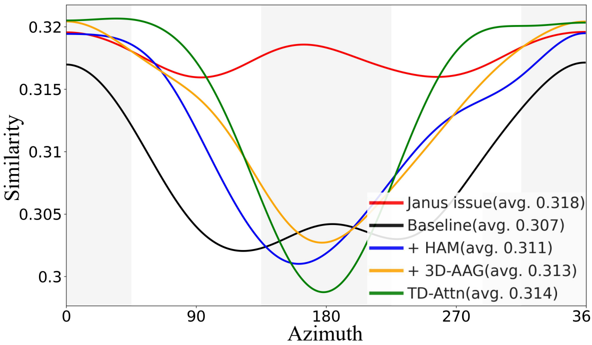

Figure 5: CLIP similarity distribution across different azimuth angles. The plot reveals how multi-view inconsistencies affect similarity scores, where results with the Janus problem exhibit abnormally high scores, while our approach produces a more natural distribution pattern.

For editing tasks, we employ CLIP$_{\text{sim}}$ and CLIP$_{\text{dir}}$ (He et al. 2024; Koo, Park, and Sung 2024; Wu et al. 2024) as objective evaluation metrics, where CLIP$_{\text{dir}}$ measures the consistency between image editing direction and text instruction in CLIP feature space. Tab. 3 demonstrates that our method improves the quality of edited 3DGS. The results of the user study for editing tasks show similarly that our method produces more favorable results. Additional HAM applications in T2I models can be found in the appendix.

Conclusion

We address multi-view inconsistency in 3D tasks by identifying prior view bias as the root cause. Our mathematical analysis reveals that subject-word tokens preferentially activate prior view features, compromising consistency across different perspectives in both generation and editing tasks. TD-Attn provides an effective solution through two key components: a 3D-Aware Attention Guidance Module that constructs view-consistent attention Gaussians by accumulating multi-view cross-attention maps, and a Hierarchical Attention Modulation Module that uses a semantic guidance tree to modulate cross-attention layers for enhanced target semantic awareness. Extensive experiments demonstrate that TD-Attn serves as a universal plugin, significantly reducing inconsistency artifacts in both generation and editing scenarios while maintaining high quality. The framework advances multi-view consistent 3D synthesis and enables fine-grained semantic control across diverse 3D tasks.

References

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*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Y.; Chen, Z.; Zhang, C.; Wang, F.; Yang, X.; Wang, Y.; Cai, Z.; Yang, L.; Liu, H.; and Lin, G. 2024. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21476–21485.

Di, D.; Yang, J.; Luo, C.; Xue, Z.; Chen, W.; Yang, X.; and Gao, Y. 2025. Hyper-3dg: Text-to-3d gaussian generation via hypergraph. *International Journal of Computer Vision*, 133(5): 2886–2909.

Ding, L.; Dong, S.; Huang, Z.; Wang, Z.; Zhang, Y.; Gong, K.; Xu, D.; and Xue, T. 2024. Text-to-3d generation with bidirectional diffusion using both 2d and 3d priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5115–5124.

Guo, P.; Hao, H.; Caccavale, A.; Ren, Z.; Zhang, E.; Shan, Q.; Sankar, A.; Schwing, A. G.; Colburn, A.; and Ma, F. 2023. StableDreamer: Taming Noisy Score Distillation Sampling for Text-to-3D. arXiv:2312.02189.

He, R.; Huang, S.; Nie, X.; Hui, T.; Liu, L.; Dai, J.; Han, J.; Li, G.; and Liu, S. 2024. Customize your nerf: Adaptive source driven 3d scene editing via local-global iterative training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6966–6975.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

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.

Huang, H.; Wu, Y.; Deng, C.; Gao, G.; Gu, M.; and Liu, Y.-S. 2025. FatesGS: Fast and Accurate Sparse-View Surface Reconstruction Using Gaussian Splatting with Depth-Feature Consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4): 3644–3652.

Huang, T.; Zeng, Y.; Zhang, Z.; Xu, W.; Xu, H.; Xu, S.; Lau, R. W.; and Zuo, W. 2024. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5364–5373.

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.

Kingma, D. P.; Rezende, D. J.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Koo, J.; Park, C.; and Sung, M. 2024. Posterior distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13352–13361.

Lee, D. I.; Park, H.; Seo, J.; Park, E.; Park, H.; Baek, H. D.; Shin, S.; Kim, S.; and Kim, S. 2025. Editsplat: Multi-view fusion and attention-guided optimization for view-consistent 3d scene editing with 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11135–11145.

Li, H.; Shi, H.; Zhang, W.; Wu, W.; Liao, Y.; Wang, L.; Lee, L.-h.; and Zhou, P. Y. 2024a. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision*, 214–230. Springer.

Li, Y.; Dou, Y.; Shi, Y.; Lei, Y.; Chen, X.; Zhang, Y.; Zhou, P.; and Ni, B. 2024b. FocalDreamer: Text-Driven 3D Editing via Focal-Fusion Assembly. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4): 3279–3287.

Li, Y.; Lv, C.; Yang, H.; and Huang, D. 2025. Micro-macro Wavelet-based Gaussian Splatting for 3D Reconstruction from Unconstrained Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 5057–5065.

Li, Z.; Hu, M.; Zheng, Q.; and Jiang, X. 2024c. Connecting consistency distillation to score distillation for text-to-3d generation. In *European Conference on Computer Vision*, 274–291. Springer.

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, J.; Li, Z.; Huang, B.; Tang, X.; Liu, J.; Liu, S.; Wu, X.; Song, F.; and Yang, W. 2025. Decoupling Appearance Variations with 3D Consistent Features in Gaussian Splatting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 5236–5244.

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

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.

Michel, O.; Bar-On, R.; Liu, R.; Benaim, S.; and Hanocka, R. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13492–13502.

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, J.; Ko, J.; Byun, D.; Suh, J.; and Rhee, W. 2024. Cross-Attention Head Position Patterns Can Align with Human Visual Concepts in Text-to-Image Generative Models. *arXiv e-prints*, arXiv–2412.

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

Qu, W.; Li, J.; Cheng, J.; Shi, J.; Meng, C.; Ma, C.; Wang, H.; Deng, X.; and Zhang, Y. 2025. HOGSA: Bimanual Hand-Object Interaction Understanding with 3D Gaussian Splatting Based Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6): 6639–6647.

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

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.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

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

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

Wu, J.; Bian, J.-W.; Li, X.; Wang, G.; Reid, I.; Torr, P.; and Prisacariu, V. A. 2024. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *European Conference on Computer Vision*, 55–71. Springer.

Xie, D.; Li, J.; Tan, H.; Sun, X.; Shu, Z.; Zhou, Y.; Bi, S.; Pirk, S.; and Kaufman, A. E. 2024. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6369–6379.

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. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Zhang, Q.; Chen, Z.; Tong, J.; Zhang, J.; Hong, J.; and Li, X. 2024. Viewpoint Consistency in 3D Generation via Attention and CLIP Guidance. *arXiv preprint arXiv:2412.02287*.

Zhou, J.; Fan, L.; Chen, X.; Huang, L.; Liu, S.; and Li, H. 2025. GaussianPainter: Painting Point Cloud into 3D Gaussians with Normal Guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10): 10788–10796.