# DreamVGGT: VGGT-Guided Generative Gaussian Splatting For Scene Customization

Anonymous ICME submission

***Abstract*—VGGT has demonstrated powerful performance in multi-view synthesis. However, existing diffusion-based 3D generation frameworks underutilize the explicit geometric structure provided by VGGT, leaving the generated 3D models without effective grounding. To address these challenges, we propose DreamVGGT, a unified framework that synergizes diffusion-based generation with explicit geometric–semantic guidance to achieve high-fidelity 3D model synthesis and editing from a single image. Specifically, we employ VGGT-guided initialization to ground the Score Distillation Sampling (SDS) process, followed by our proposed Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO) to strictly align Gaussians via adaptive uncertainty and boundary synergy. This yields a semantically explicit 3DGS representation that enables controllable object editing, including insertion, removal, and deformation. Experiments confirm effective object editing, enabling customizable simulation and training environments.**

***Index Terms*—3D Gaussian Splatting, Diffusion Models, Image-to-3D Generation, 3D Scene Editing**

## I. Introduction

Generating realistic, editable, and structurally coherent 3D scenes is essential for advancing digital content creation, immersive virtual reality, and autonomous robotic systems. However, constructing high-quality 3D models remains a significant challenge. Although differentiable rendering techniques such as NeRF [1] and 3DGS [2] have demonstrated impressive geometric and photometric fidelity in multi-view reconstruction, these methods inherently depend on dense viewpoint coverage. Under single-view or highly sparse-view conditions, recovering 3D structure from a 2D image remains a fundamentally ill-posed problem: depth suffers from scale ambiguity, occluded regions lack geometric constraints, texture boundaries tend to degrade, and global consistency is difficult to enforce. Meanwhile, although 2D diffusion models [3] excel at generating high-quality images, their direct use for 3D generation remains problematic. Lacking true 3D geometric priors and multi-view consistency, diffusion-based 3D structures are often unstable, semantically ambiguous, noisy in density distribution, and prone to deformation or drift during optimization, making them unsuitable for realistic simulation scenarios. Recently, direct 3D inference models such as VGGT [4] have attracted considerable attention, particularly for their ability to infer depth and point clouds in an unsupervised manner from a single view. However, simply concatenating VGGT's geometric priors with the generative capability of diffusion models often degenerates into feature-level or output-level stacking, without establishing effective mutual constraints during optimization. Consequently, diffusion models cannot leverage VGGT's structural stability, and the geometric priors from VGGT are not effectively refined by diffusion's appearance-consistent gradients, leading to unstable depth, loose structural coherence, and blurred semantic boundaries—ultimately failing to address the fundamental challenges of single-view 3D reconstruction. To overcome these challenges, we introduce DreamVGGT, a unified framework for generating, initializing, and editing high-quality 3D representations from a single image.

Fig. 1. DreamVGGT enables scene customization by generating semantically explicit 3D models from single image and supporting object-level editing operations, including insertion, removal, and deformation.

At the core of DreamVGGT lies a synergistic workflow that bridges the gap between explicit geometric priors and generative diffusion guidance. First, we employ a VGGT-guided initialization strategy (Sec. II-A), utilizing explicit depth and point cloud priors to physically ground the 3D Gaussians, preventing the geometric collapse often seen in random initialization. Second, to rigorously align the generative appearance with these structural priors, we propose an Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO) mechanism (Sec. II-B). This mechanism adaptively balances Score Distillation Sampling (SDS) gradients with geometric constraints via uncertainty gating, ensuring that high-confidence geometric regions are preserved while low-confidence regions are refined by diffusion. Finally, to enable downstream utility, DreamVGGT establishes a semantically explicit 3DGS representation (Sec. II-C). By enforcing boundary synergy between semantic labels and geometric discontinuities, our framework naturally supports object-level editing operations, including insertion, removal, and deformation, through direct parameter manipulation. By unifying diffusion-based generation with explicit geometric–semantic grounding, DreamVGGT introduces a robust solution to the longstanding challenge of single-view 3D generation. It greatly enhances the realism, controllability, and scalability of synthetic environments. As illustrated in Fig. 1, our method enables high-fidelity editing operations

To summarize, DreamVGGT offers the following key contributions:

- We introduce a VGGT-guided initialization strategy that leverages depth and point priors to provide grounded initialization, resolving scale ambiguity and preventing geometric collapse at the beginning.
- We propose an uncertainty-aware optimization mechanism (U-DSPO) that synergizes SDS supervision with geometric and semantic priors via uncertainty gating and boundary constraints, ensuring structural accuracy and sharp semantic delineation.
- We construct a semantically explicit 3D Gaussian representation that supports consistent object-level editing operations, including insertion, removal, and deformation, enhancing the utility of generated models for practical applications.

## II. Methods

As shown in Fig. 2, the framework visualizes the process of producing 3D results from 2D image inputs; for text-based inputs, the system first synthesizes a 2D representation before lifting it into 3D. Sec. II-A introduces a VGGT-guided diffusion-based single image to 3DGS Initialization module that grounds 3D Gaussians from a single image using depth and point cloud priors. Sec. II-B presents Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO), which refines Gaussians through the adaptive optimization of diffusion and geometric-semantic constraints. Sec. II-C describes the resulting semantically explicit representation that enables direct object-level scene editing operations, including removal, insertion, and deformation.

### *A. Diffusion-Based Single Image to 3DGS Initialization Guided by VGGT*

Traditional geometry-based point cloud initialization methods, such as COLMAP, struggle with sparse or single-view inputs, leading to insufficient spatial anchoring for 3D generation. To address this limitation, we introduce Visual Geometry Grounded Transformer (VGGT) [4] as a single-view structural prior provider that offers depth, semantics, and point-level geometry from a single input image. Our framework accepts either a single image or a text prompt as input. For text input, we first synthesize a 2D image via Stable Diffusion Model before processing it as image input. Given the input image (either directly provided or text-synthesized), we leverage the VGGT to extract four structural priors: semantic map $\hat{\mathcal{S}}(\mathbf{x})$, depth uncertainty maps $\mathbf{U}(\mathbf{x})$, depth map $\mathcal{D}_{\text{raw}}$, and a sparse 3D point cloud $\mathcal{P}_{\text{raw}}$. Since monocular depth estimation produces only relative depth without absolute scale, we normalize both the depth map and point cloud to a canonical range $[0, 1]$ via min-max normalization, obtaining the normalized depth $\tilde{\mathcal{D}}$ (from $\mathcal{D}_{\text{raw}}$) and normalized point cloud $\tilde{\mathcal{P}}_{\text{vggt}}$ (from $\mathcal{P}_{\text{raw}}$). This normalization ensures numerical stability and consistent spatial initialization across diverse scenes. These priors establish a physics-grounded initialization for the 3D Gaussian field, preventing the structural collapse, scale drift, and shape distortion that commonly arise when relying solely on Score Distillation Sampling (SDS) without explicit geometric constraints. Specifically, the normalized priors serve as spatial anchors that constrain the diffusion-based optimization toward structurally plausible configurations, while SDS refines appearance and promotes multi-view consistency.

To better align the Gaussian distribution with the coarse-to-fine nature of the diffusion process, we adopt an adaptive densification strategy to progressively refine the Gaussian representation and capture fine-grained geometric details. We initialize Gaussian means $\mu_i$ by randomly sampling points from $\tilde{\mathcal{P}}_{\text{vggt}}$, assign initial opacity $\alpha_i^0 = 0.1$, and set isotropic covariances $\Sigma_i$ according to local point density (k-nearest neighbors with $k = 20$). Every 100 iterations, Gaussians with accumulated gradients exceeding a threshold are densified by splitting (for large Gaussians) or cloning (for small ones), and pruned when opacity drops below 0.005 or scale becomes excessively large, ensuring efficient representation and preventing over-parameterization.

To distill 2D diffusion priors into the 3DGS parameter space, we employ Score Distillation Sampling (SDS) [5]. At each iteration, we render the 3D scene from a randomly sampled viewpoint $p \sim \mathcal{P}_{\text{cam}}$ (uniform spherical sampling) and compute:

$$\nabla_\Theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,p,\epsilon}\left[w(t)\left(\epsilon_\phi(z_t; t, c) - \epsilon\right)\frac{\partial z_t}{\partial \Theta}\right], \qquad (1)$$

where $\Theta$ denotes the 3DGS parameters, $t \sim \mathcal{U}(0.02, 0.98)$ is the diffusion timestep, $z_t$ is the noised latent encoded by a VAE $\mathcal{E}$, $\epsilon_\phi$ is the denoising network conditioned on the CLIP embedding $c$ of the input image, and $w(t) = \sigma_t^2/\bar{\alpha}_t$ follows the DDPM noise schedule.

This SDS-based optimization refines the appearance of the 3D Gaussians to align with the learned 2D diffusion prior across multiple viewpoints. To ensure fidelity to the input view, we enforce reference-view consistency during the initialization stage via a photometric loss. Specifically, at each iteration $k$, we compute the MSE between the rendered RGB and the input image, weighted by $\lambda_{\text{rgb}}^{(k)}$, which linearly increases from 0 to $10^4$ over the first 3000 iterations. This schedule is empirically designed to balance the gradient magnitudes between the SDS guidance and the photometric constraints. While this initialization stage provides coarse geometry and appearance consistency, fine-grained geometric and semantic alignment with VGGT predictions remains necessary. We therefore introduce an Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO) stage (Sec. II-B) to refine geometry, semantics, and spatial coherence under adaptive cross-modal constraints.

### *B. Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO)*

Although diffusion-based initialization establishes a coarse global structure, the resulting 3D Gaussians frequently exhibit geometric misalignment and semantic inconsistencies due to the scale ambiguity inherent in monocular priors. To mitigate these issues, we propose Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO). Unlike standard linear fusion strategies, U-DSPO incorporates an adaptive uncertainty

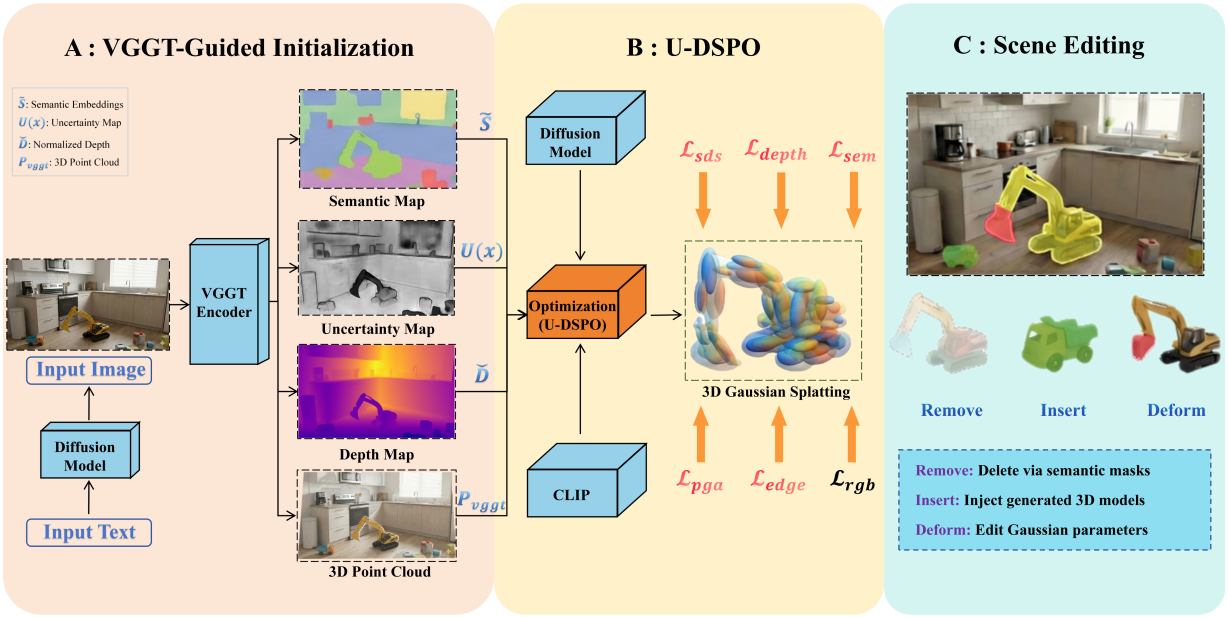


Fig. 2. Overview of our framework. In part A, the system accepts a single image and utilizes the VGGT encoder to extract structural priors, including semantics, uncertainty map, depth, and point cloud for 3D Gaussian initialization. In part B, these priors guide the Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO), where the 3D Gaussian Splatting is refined through CLIP-conditioned diffusion, enforced by adaptive SDS, depth, and semantic constraints. In part C, the resulting semantically explicit representation enables object-level editing.

mechanism and a boundary synergy constraint to balance VGGT-derived structural priors with diffusion-based visual consistency.

We first address the reliability of geometric priors. Monocular depth predictions from VGGT exhibit non-uniform reliability. To address this, we model the pixel-wise uncertainty $\mathbf{U}(\mathbf{x}) \in [0,1]$ derived from the predictive variance. We define the geometric confidence as $\mathbf{C}(\mathbf{x}) = 1 - \mathbf{U}(\mathbf{x})$ and formulate an uncertainty-aware depth consistency loss:

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \mathbf{C}(\mathbf{x}) \cdot \left\| \mathcal{D}^{\text{ren}}(\mathbf{x}) - \tilde{\mathcal{D}}(\mathbf{x}) \right\|_1, \tag{2}$$

where $\mathcal{D}^{\text{ren}}$ is the rendered depth and $\tilde{\mathcal{D}}$ is the normalized VGGT depth. Functionally, $\mathbf{C}(\mathbf{x})$ serves as an adaptive gating mechanism. In regions of high uncertainty (low confidence), the penalty for depth mismatch is attenuated, thereby shifting the optimization focus to the SDS loss to recover plausible structures via multi-view consistency. In contrast to conventional monocular depth estimation approaches that typically output relative depth maps suffering from scale ambiguity, VGGT provides metric depth and explicit point cloud priors. This absolute spatial grounding is critical for initializing 3D Gaussians with physical consistency, preventing the geometry collapse often observed when lifting 2D diffusion priors directly.

Simultaneously, we introduce a constraint to resolve semantic inconsistencies where labels do not align with geometric discontinuities. We assume that significant transitions in semantic categories should spatially coincide with depth discontinuities. To enforce this structural coherence, we implement a boundary synergy objective via an edge-aware smoothness loss:

$$\mathcal{L}_{\text{edge}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \|\nabla \mathcal{D}^{\text{ren}}(\mathbf{x})\|_1 \cdot \exp\left(-\beta \left\| \nabla \hat{\mathcal{S}}(\mathbf{x}) \right\|_2\right), \tag{3}$$

where $\nabla$ is the spatial gradient operator and $\beta$ is a scaling factor. Here, the term $\exp(-\beta\|\nabla\hat{\mathcal{S}}\|)$ acts as a semantic edge indicator. Within a semantic object, the loss enforces surface smoothness, whereas at semantic boundaries, the penalty is relaxed to permit geometric transitions. This ensures that the optimized 3D Gaussians maintain geometric silhouettes consistent with the semantic layout.

Finally, to prevent deviation from the latent scene manifold during stochastic SDS optimization, we utilize the sparse point cloud $\tilde{\mathcal{P}}_{\text{vggt}}$ as physical anchors. We refine the global structural alignment using a normal-weighted Chamfer distance:

$$\mathcal{L}_{\text{pga}} = \sum_{i=1}^{N} \eta_i \min_{\mathbf{q} \in \tilde{\mathcal{P}}_{\text{vggt}}} \|\mu_i - \mathbf{q}\|_2, \tag{4}$$

where $\eta_i = \max(0, \mathbf{n}_i \cdot \mathbf{n}_{\mathbf{q}})$ weights the alignment based on the local normal consistency between the Gaussian normal $\mathbf{n}_i$ and the normal $\mathbf{n}_{\mathbf{q}}$ of the nearest point $\mathbf{q}$.

To ensure semantic consistency, we define the semantic segmentation loss as:

$$\mathcal{L}_{\text{sem}} = -\frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \sum_{c=1}^{C} \hat{\mathcal{S}}_c(\mathbf{x}) \log \mathcal{S}_c^{\text{ren}}(\mathbf{x}), \tag{5}$$

where $\hat{\mathcal{S}}_c(\mathbf{x})$ is the VGGT-predicted semantic probability for class $c$ at pixel $\mathbf{x}$, $\mathcal{S}_c^{\text{ren}}(\mathbf{x})$ is the rendered semantic probability,

and $C$ is the number of semantic classes. Each Gaussian $g_i$ is assigned a learnable semantic embedding that is rendered via weighted splatting to produce per-pixel semantic logits.

The total objective integrates diffusion supervision with the proposed U-DSPO constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SDS}} + \lambda_{\text{rgb}}^{(k)} \mathcal{L}_{\text{rgb}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_s (\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{edge}}) + \lambda_p \mathcal{L}_{\text{pga}}, \tag{6}$$

where $\mathcal{L}_{\text{rgb}} = \|I^{\text{ren}}(\Theta; p^{\text{ref}}) - I^{\text{input}}\|_2^2$ enforces reference-view consistency with time-dependent weight $\lambda_{\text{rgb}}^{(k)}$ (warm-up strategy detailed in Sec. II-A). Here, $p^{\text{ref}}$ is the input camera pose and $I^{\text{ren}}$ is the rendered RGB from the reference viewpoint. After warm-up (first 3000 iterations), this weight remains constant throughout the U-DSPO stage. This unified strategy ensures that the generated 3D models possess accurate geometry, stable semantic boundaries, and realistic appearance.

We empirically tuned the hyperparameters $\lambda_d$, $\lambda_s$, and $\lambda_p$ to balance the gradient magnitudes between the SDS guidance and geometric priors. The RGB weight $\lambda_{\text{rgb}}^{(k)}$ follows a linear warm-up schedule to gradually enforce photometric consistency. For the Gaussian optimization, we adopt standard densification and pruning strategies following DreamGaussian [6], ensuring training stability and geometric convergence. We set $\lambda_d = 2.0$, $\lambda_s = 1.5$, $\lambda_p = 0.8$ and train with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\text{lr} = 2 \times 10^{-3}$). Densification and pruning occur every 100 iterations with gradient threshold 0.005.

### C. VGGT-Grounded Scene Editing

Unlike conventional post-processing, DreamVGGT integrates editing directly into the U-DSPO, leveraging VGGT priors to ensure geometric consistency.

For object removal, we avoid artifact-prone deletion by refining boundaries via the boundary synergy objective (Eq. 3). Before pruning, we briefly fine-tune the scene with increased semantic weight to sharpen the geometric transition between the target object and background, enabling clean separation based on the optimized semantic probability mask.

Regarding object insertion, we utilize the VGGT canonical depth map $\tilde{\mathcal{D}}$ to resolve scale ambiguity and floating artifacts. Specifically, the user specifies a 2D insertion point $(u, v)$ on the screen space. To ensure physical plausibility, we explicitly anchor the object by unprojecting this coordinate to 3D space using the absolute depth value $d = \tilde{\mathcal{D}}(u, v)$, automatically snapping the model to the underlying surface geometry. New models are initialized at physical coordinates derived from $\tilde{\mathcal{D}}(u, v)$. We then apply a localized U-DSPO, enforcing the uncertainty-aware depth loss (Eq. 2) to align the object's contact surface with the scene geometry, while SDS harmonizes global illumination.

Finally, to enable object deformation, we treat the sparse cloud $\tilde{\mathcal{P}}_{\text{vggt}}$ as a structural skeleton. Instead of disrupting Gaussian parameters directly, we apply a control field to the corresponding VGGT points and utilize the point-alignment loss $\mathcal{L}_{\text{pga}}$ (Eq. 4) to drive the Gaussian swarm. This optimization forces the geometry to flow naturally with the anchors, preserving topological integrity during deformation.

## III. Experimental Results

### A. Implementation Details

For scene synthesis, we adhere to the standard configuration of 3D Gaussian Splatting [2], employing 30,000 steps to reconstruct the scene and using the Adam optimizer to train the parameters of the Gaussians. For object generation via text or visual prompts, we utilize the settings described in DreamGaussian [6]. Here, Gaussians are initially set with an opacity of 0.1 and a grey hue, arranged within a sphere of 0.5 radius. We enhance the rendering resolution for Gaussian splatting from 64 to 512 and adjust mesh sampling between 128 and 1024. The RGB loss weight linearly increases from 0 to $10^4$ over the first 3000 iterations as described in Sec. II-A. All experimental procedures are carried out on a single NVIDIA A100-80GB GPU. We conducted experiments on the Mip-NeRF

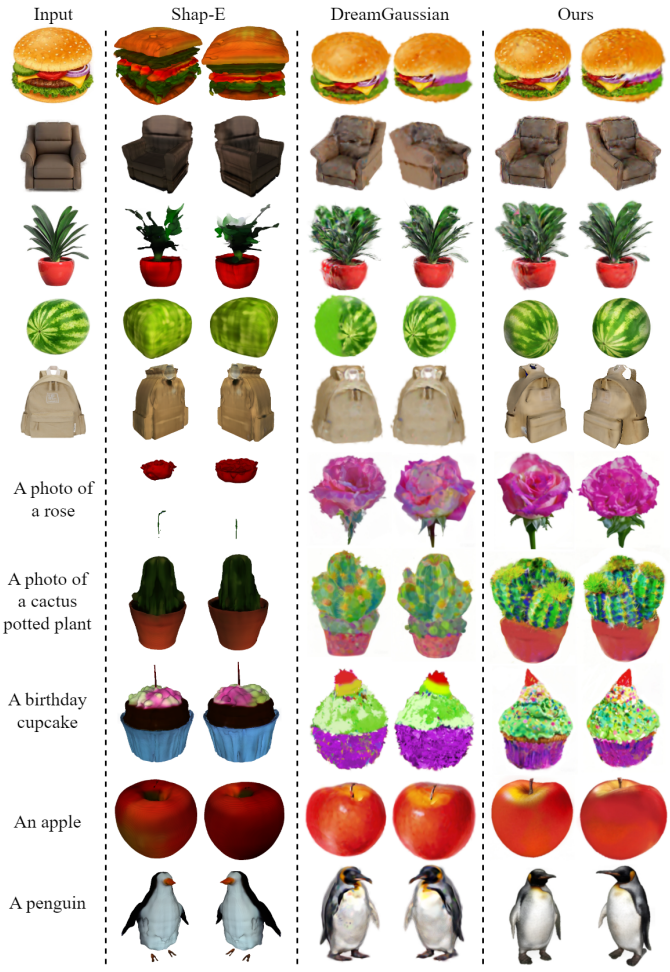


Fig. 3. Comparisons of Image-to-3D and Text-to-3D generation. Our method maintains superior structural fidelity. By anchoring the optimization with VGGT-guided initialization, we effectively prevent the geometric collapse and shape distortion that commonly affect purely SDS-driven baselines.

360 dataset [7] for object removal, insertion and customization. We quantify the image-to-3D generation quality using the CLIP similarity metric [8]. Following the evaluation protocol established in [8], we utilize their benchmark dataset of 30

images. We compute the average CLIP score between the rendered novel views and the reference image to report the final metric.

| Methods | Representation | CLIP-Similarity | User Study |
|---|---|---|---|
| One-2-3-45 [9] | NeRF | 0.578 | 4.85 |
| Zero-1-to-3 [10] | NeRF | 0.697 | 6.13 |
| Shap-E [11] | NeRF | 0.574 | 4.37 |
| DreamGaussian [6] | 3DGS | 0.689 | 6.35 |
| GALA3D [12] | 3DGS | 0.716 | 7.62 |
| **Ours** | **3DGS** | **0.725** | **7.86** |

TABLE I
**QUANTITATIVE COMPARISONS** OF GENERATION QUALITY AND HUMAN EVALUATION FOR OBJECT-CENTRIC 3D GENERATION ARE PRESENTED IN SEPARATE COLUMNS.

### *B. Object-centric Generation*

We quantitatively assessed the quality of objects generated by various image-to-3D methods [13], [14] by employing the average CLIP similarity metric, comparing the generation results derived from identical text prompts or reference images. Fig. 3 visualizes some of the generation results. To further validate these findings, we conducted a comprehensive human evaluation with 40 participants, including digital artists and UX designers, who rated the generation quality. Participants used a 10-point scale, where higher scores indicate better perceived quality. The results of this evaluation are summarized in Table I.

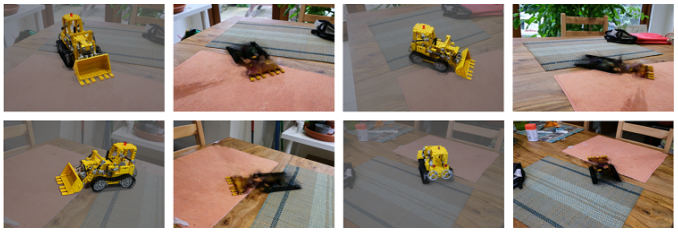

Fig. 4. Render the image after removing the Gaussians of the semantic part of the Lego tank corresponding to the kitchen scene in the Mip-Nerf 360 Dataset [7].

### *C. Scene Customization*

In our experiments, we demonstrate the capabilities of our model with a focused study on scene customization. We manipulate the semantic Gaussians of the Kitchen scene from the Mip-NeRF 360 dataset [7], effectively removing most of the structure. Crucially, the clean separation observed in Fig. 4 results from our boundary synergy constraint (Eq. 3). This mechanism enforces strict alignment between semantic and geometric edges, avoiding the jagged artifacts typically seen when pruning unconstrained neural fields.

Fig. 5 demonstrates diverse editing capabilities, including object replacement, depth-anchored insertion using $\tilde{\mathcal{D}}(u, v)$ to resolve scale ambiguity, color modification, and precise removal. Fig. 6 shows sequential insertion of multiple objects without disrupting existing content, demonstrating seamless integration for simulation and virtual reality applications.


Fig. 5. Visualization of insertion, removal, and customization of the synthesized scenes in the counter, kitchen, and room scenes of the Mip-Nerf 360 dataset [7].

Fig. 6. Our method adds different 3D objects to the scene one by one to obtain rendering images. In the first row, which renders the counter scene, an apple and blueberries are added to the tray in the first scene, a tomato is added in the second scene, an orange is added in the third scene, and a cup is added in the fourth scene. In the second row, which renders the room scene, an Ultraman figure and a notebook are added in the first scene, a game controller is added in the second scene, a mouse is added in the third scene, and a green plant is added in the fourth scene.

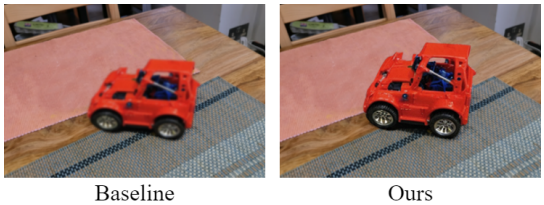


Fig. 7. The visualization of substituting the original LEGO excavator with a new truck object, generated by DreamGaussian baseline [6] and our method incorporating depth constraints, in the Kitchen scene of the Mip-NeRF 360 dataset [7].

### D. Ablation Study

We also report PSNR/LPIPS/mIoU in ablation (Table II), confirming consistent improvements across photometric and semantic metrics. Given the absence of a standardized metric for evaluating the quality of customization or scene editing, we assessed the performance of our method by integrating a new object into the Kitchen scene, comparing it against the baseline DreamGaussian [6]. As depicted in Fig. 7, the 3D objects created using our method, which incorporates depth constraints, show markedly improved geometry, texture, and clarity over the baseline. The sharper rendering of the objects significantly enhances visual quality, resulting in crisper and more detailed imagery.

For each component of DreamVGGT, the effectiveness of the depth prior in enhancing geometric clarity and the semantic prior in segmenting 3D objects has been sufficiently demonstrated in Fig. 7 and Fig. 4. The uncertainty gating mechanism in U-DSPO plays a crucial role in stabilizing geometry–appearance alignment. To further validate its contribution, we include a supplementary ablation: removing uncertainty gating decreases PSNR by 0.8 and mIoU by 1.4, confirming its effectiveness in adaptive supervision.

TABLE II
ABLATION STUDY OF LOSS COMPONENTS IN DREAMVGGT. WE PROGRESSIVELY INCORPORATE EACH COMPONENT TO VALIDATE ITS CONTRIBUTION TO VISUAL QUALITY (PSNR, LPIPS) AND SEMANTIC ACCURACY (MIOU).

| Loss Components | | | | | | Metrics | | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{\text{SDS}}$ | $\mathcal{L}_{\text{rgb}}$ | $\mathcal{L}_{\text{depth}}$ | $\mathcal{L}_{\text{pga}}$ | $\mathcal{L}_{\text{sem}}$ | $\mathcal{L}_{\text{edge}}$ | **PSNR** ↑ | **LPIPS** ↓ | **mIoU** ↑ |
| ✓ | | | | | | 23.15 | 0.185 | 81.2 |
| ✓ | ✓ | | | | | 25.40 | 0.162 | 82.5 |
| ✓ | ✓ | ✓ | | | | 27.12 | 0.141 | 83.8 |
| ✓ | ✓ | ✓ | ✓ | | | 28.55 | 0.125 | 84.9 |
| ✓ | ✓ | ✓ | ✓ | ✓ | | 29.80 | 0.108 | 86.4 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **31.25** | **0.092** | **88.5** |

Discussion: Table II demonstrates that each proposed component contributes to the overall quality. The uncertainty-aware depth term $\mathcal{L}_{\text{depth}}$ improves geometric consistency, boosting PSNR from 25.40 to 27.12. Furthermore, even with a strong semantic baseline, the inclusion of the edge synergy loss $\mathcal{L}_{\text{edge}}$ further sharpens the semantic boundaries, increasing the mIoU by 2.1 points (from 86.4 to 88.5). These results suggest that incorporating cross-modal structural consistency contributes to enhancing both 3D generation fidelity and editing precision. We also evaluated the geometric fidelity of the utilized priors against standard relative depth estimation baselines on 100 randomly sampled images from the Mip-NeRF 360 dataset scenes. The results demonstrate that VGGT priors significantly enhance structural accuracy, reducing depth RMSE by $12.4\%$ and improving boundary mIoU by 5.2 points. Additional quantitative evaluations are provided in the supplementary material.

## IV. CONCLUSION

In this work, we present DreamVGGT, a unified 3D generation framework that enhances the structural fidelity and controllability of single-view scene synthesis. We integrate VGGT-guided initialization with our proposed Uncertainty-aware Depth-Semantic-Point Optimization (U-DSPO). This synergy grounds diffusion-based generation with explicit geometric priors, ensuring physical plausibility alongside visual realism. Furthermore, by enforcing boundary synergy within the 3D Gaussian Splatting [2], our approach achieves a semantically explicit scene structure. This allows for object-level manipulation, including insertion, removal, and deformation. Ultimately, DreamVGGT not only supports advancements in robotic simulation environments but also extends the potential for augmented and virtual reality, providing adaptive, high-fidelity, and interactable 3D models for immersive experiences.

## REFERENCES


[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[4] Facebook Research and Others, "Vggt: Visual geometry grounded transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, Best Paper Award.

[5] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

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

[7] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469, 2021.

[8] Luke Melas-Kyriazi, Iro Laina, C. Rupprecht, and Andrea Vedaldi, "Realfusion 360° reconstruction of any object from a single image," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8446–8455, 2023.

[9] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," *arXiv preprint arXiv:2303.11328*, 2023.

[11] Heewoo Jun and Alex Nichol, "Shap-e: Generating conditional 3d implicit functions," 2023.

[12] Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang, "Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting," *arXiv preprint arXiv:2402.07207*, 2024.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR.

[14] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.