УДК 378.004
DOI: 10.25686/978-5-8158-2474-4-2025-113-132

Enhancing 3D Gaussian Splatting with diffusion models: a survey

Maksim Raenchuk
MSU Institute for Artificial Intelligence, Moscow, Russia
Lomonosov Moscow State University, Moscow, Russia

Abstract. 3D Gaussian Splatting (3DGS) has emerged as a technique for real-time novel view synthesis, offering explicit scene representation and efficient rendering. Concurrently, diffusion models have demonstrated unprecedented capabilities in generating and manipulating complex, high-fidelity data distributions. This survey explores the rapidly evolving intersection of these two powerful paradigms: the enhancement of 3DGS using diffusion models. It systematically categorizes recent research that uses diffusion priors to overcome key challenges in 3DGS, specifically examining how these models are integrated into different stages of the pipeline. These integrations include generating Gaussian parameters, providing optimization regularization, refining outputs, and enabling generative capabilities like text-to-3DGS. Existing approaches are analyzed and compared based on their technical innovations, strengths, and limitations. Furthermore, open challenges, such as computational efficiency, multi-view consistency, and controllability, are identified, and promising future research directions are outlined. This survey aims to provide researchers and practitioners with a structured understanding of how diffusion models are advancing the state-of-the-art (SOTA) in 3DGS, fostering further innovation in efficient and generative 3D scene representation.

Keywords: 3D Gaussian Splatting, diffusion models, novel view synthesis, 3D scene representation, generative 3D reconstruction, denoising diffusion probabilistic models, sparse view reconstruction, semantic scene editing, text-to-3D generation.

Introduction

The accelerating demand for immersive technologies–spanning augmented reality (AR), virtual production, and autonomous robotics–has intensified the need for efficient, high-fidelity 3D scene reconstruction and rendering. 3D Gaussian Splatting (3DGS) (Kerbl, et al. 2023) has revolutionized this domain with its explicit, point-based scene representation, enabling real-time photorealistic novel view synthesis while dramatically reducing computational costs compared to neural radiance fields (NeRFs) (Mildenhall, et al. 2020). Despite its impact, 3DGS faces persistent challenges: sensitivity to initialization, artifacts under sparse inputs, and limited geometric coherence. These limitations hinder its deployment in applications requiring robustness to imperfect data or creative control, such as dynamic scene editing and generative content creation.

Concurrently, diffusion models have emerged as a paradigm-shifting force in generative artificial intelligence (AI), demonstrating unprecedented capabilities in synthesizing complex, high-dimensional data distributions through iterative denoising. By learning rich priors from vast datasets, these models excel at hallucinating plausible structures from partial inputs, refining noisy observations, and enabling semantic manipulation via natural language guidance (Rombach, et al. 2022; Blattmann, et al. 2023). Their probabilistic framework offers a natural mechanism to address the ill-posed nature of 3D reconstruction, where multiple valid solutions may explain sparse visual evidence.

Diffusion models enhance 3DGS by providing priors that improve optimization, mitigate artifacts through training regularization, refine results via post-processing, and unlock generative capabilities such as text-to-3D synthesis. This convergence bridges the gap between efficient rendering and generative intelligence, empowering applications from interactive scene editing to on-demand virtual world creation. Yet, research remains fragmented across disparate methodologies with no unified analysis of their trade-offs, scalability, or fundamental limitations.

This survey provides an examination of diffusion-enhanced 3DGS. Cutting-edge approaches are systematized into a coherent taxonomy, their technical innovations analyzed, and performance evaluated across key metrics.

Structure of the Survey

The remainder of this article is organized as follows: Section 2 reviews foundational concepts in 3DGS and diffusion models. Section 3 categorizes diffusion-enhanced 3DGS methodologies by integration strategy (diffusion-guided 3DGS optimization, direct 3DGS generation from latent space, diffusion for consistent novel view synthesis). Section 4 analyzes fundamental challenges and limitations in benchmarking diffusion-enhanced 3DGS methods. Section 5 discusses unresolved challenges and emerging paradigms. Section 6 concludes with reflections on the future of generative 3D scene representation.

Foundational Concepts

This section establishes the mathematical foundations of 3DGS and diffusion models, providing the theoretical underpinnings for their integration. The representation, optimization, and rendering processes of 3DGS are formalized (figure 1 provides an overview of the 3DGS framework), followed by a rigorous treatment of diffusion probabilistic models (the forward and reverse processes of a diffusion model are illustrated in figure 3) and their connection to score-based generative frameworks.

3D Gaussian Splatting (Kerbl, et al. 2023)

Primitive Representation

A 3D scene is fundamentally represented as a collection of anisotropic Gaussian primitives $\mathbf{G} = \{G_i\}_{i=1}^{N}$, where each primitive G_i constitutes a learnable volumetric entity parameterized by its spatial position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$, opacity $\boldsymbol{\alpha}_i \in \mathbb{R} \cap [0,1]$, and view-dependent appearance modeled through spherical harmonics (SH) coefficients $\boldsymbol{\psi}_i \in \mathbb{R}^k$. The Gaussian function defines a radially decaying density field centered at $\boldsymbol{\mu}_i$:

$$G_i(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right).$$

To maintain physical validity during optimization, $\boldsymbol{\Sigma}_i$ is constrained to positive semi-definite configurations through a differentiable factorization into rotational and scaling components:

$$\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top.$$

Here, $\mathbf{R}_i \in \mathrm{SO}(3)$ represents a rotation matrix derived from a normalized quaternion $\mathbf{q}_i \in \mathbb{R}^4$ to avoid gimbal lock, while $\mathbf{S}_i = \mathrm{diag}(\mathbf{s}_i)$ constitutes a scaling matrix with $\mathbf{s}_i \in \mathbb{R}_+^3$ enforcing anisotropic stretching along principal axes. This parameterization ensures both numerical stability and gradient tractability during optimization.

View-Dependent Appearance

The appearance model captures complex light transport effects by encoding color $\mathbf{c}_i$ as a function of viewing direction $\mathbf{d}$ via spherical harmonics basis expansion:

$$\mathbf{c}_i(\mathbf{d}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \boldsymbol{\psi}_i^{(l,m)} Y_l^m(\mathbf{d}),$$

where Y_l^m denotes the spherical harmonics basis function of degree l and order m, and $\boldsymbol{\psi}_i^{(l,m)} \in \mathbb{R}^3$ contains RGB coefficients per band.

The maximum band L governs representational capacity, with higher bands capturing finer specular details at the cost of increased parameter dimensionality. Direction $\mathbf{d}$ is typically derived from the normalized vector between the Gaussian mean $\boldsymbol{\mu}_i$ and the camera origin. This frequency-based decomposition provides rotationally invariant directional representation while maintaining differentiability essential for gradient-based optimization.

Differentiable Rendering

The image synthesis process employs a differentiable rasterizer that aggregates contributions from K depth-ordered Gaussians per pixel $\mathbf{p}$. The composited color follows alpha blending with transmittance accumulation:

$$\mathbf{C}(\mathbf{p}) = \sum_{k=1}^{K} \mathbf{c}_k\, \widehat{\boldsymbol{\alpha}}_k \prod_{j=1}^{k-1}\left(1 - \widehat{\boldsymbol{\alpha}}_j\right),$$

where $\widehat{\boldsymbol{\alpha}}_k = \boldsymbol{\alpha}_k \cdot G_k(\mathbf{x}_k)$ represents the projected opacity modulated by the Gaussian's evaluation at its projected screen-space position $\mathbf{x}_k$.

The elliptical weighted average (EWA) splatting technique projects 3D covariance $\boldsymbol{\Sigma}_i$ to 2D screen space via the Jacobian $\mathbf{J}$ of the projective transformation:

$$\widehat{\boldsymbol{\Sigma}}_i = \mathbf{J}\boldsymbol{\Sigma}_i\mathbf{J}^\top.$$

This projection accounts for perspective distortion during rasterization, ensuring accurate shape preservation across view changes. The sorting operation employs a tile-based parallelization strategy where Gaussians are binned into screen-space tiles prior to depth ordering, enabling efficient GPU implementation.

Optimization

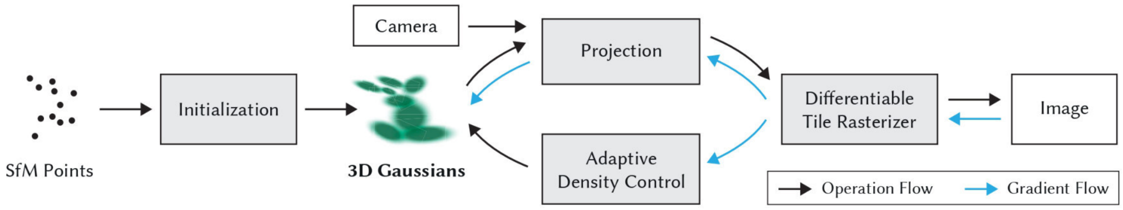


Figure 1. 3DGS framework overview. Optimization starts with the sparse SfM point cloud and creates a set of 3D Gaussians. Subsequent steps optimize and adaptively control the density of this Gaussian set. A fast tile-based renderer is utilized during optimization, enabling competitive training times compared to SOTA fast radiance field methods. After training, the renderer permits real-time navigation across a wide variety of scenes

Parameter optimization minimizes the photometric loss between rendered and ground truth images:

$$\mathcal{L} = \lambda \cdot \mathrm{L1}(\text{render,target}) + (1-\lambda) \cdot \mathrm{DSSIM}(\text{render,target}),$$

where $\mathrm{DSSIM}(\cdot,\cdot)$ enhances structural similarity through a dissimilarity term based on the SSIM metric.

Parameters $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \mathbf{q}_i, \mathbf{s}_i, \boldsymbol{\alpha}_i, \boldsymbol{\psi}_i\}$ are updated via stochastic gradient descent with adaptive learning rates, where positional parameters $\boldsymbol{\mu}_i$ typically employ higher rates than rotational $\mathbf{q}_i$ or appearance $\boldsymbol{\psi}_i$ components.

Adaptive densification dynamically regulates primitive density based on spatial gradient analysis (as shown in figure 2). Regions exhibiting significant view-space position gradients are identified as under-reconstructed, triggering Gaussian cloning with positional perturbation to increase coverage. Conversely, large Gaussians in high-frequency regions are split along principal axes to resolve fine details. Pruning mechanisms periodically remove primitives with opacity $\boldsymbol{\alpha}_i$ below a threshold τ or those residing in low-density regions, maintaining computational efficiency. The optimization alternates between geometric refinement and appearance adjustment, progressively enhancing scene representation fidelity across iterations.

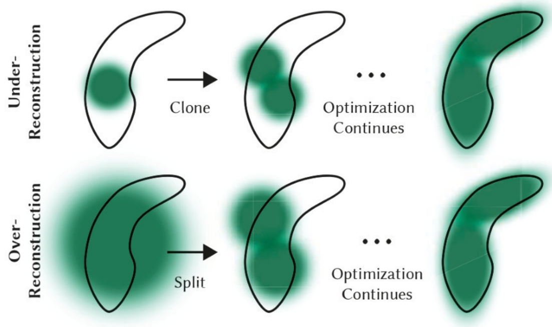


Figure 2. 3DGS framework overview. The adaptive Gaussian densification scheme operates through two primary mechanisms. For under-reconstruction cases shown in the top row, insufficient coverage of small-scale geometry (black outline) triggers cloning of the respective Gaussian. For over-reconstruction cases depicted in the bottom row, small-scale geometry represented by a single large splat undergoes splitting into two components

Recent advancements in adaptive densification (Kheradmand, et al. 2025; Bulò, et al. 2024; Grubert, et al. 2025; Deng, et al. 2025) have demonstrated superior alternatives to conventional heuristic approaches for regulating primitive density. Rather than relying solely on spatial gradient analysis to identify under-reconstructed regions, enhanced methodologies implement structured refinements.

Several studies propose replacing cloning and splitting operations with long-axis splitting strategies, which strategically position new Gaussians to minimize overlap while preserving density distributions. Complementary adaptive pruning techniques significantly reduce redundant primitives by dynamically eliminating low-opacity Gaussians based on iterative opacity thresholds or significance-aware metrics that evaluate rendering impact.

Further improvements incorporate dynamic adjustments to densification thresholds, including exponential scheduling protocols that progressively modulate gradient sensitivity during optimization. To promote efficient resource utilization, regularization terms encourage sparsity in opacity and scale parameters, implicitly pruning underutilized primitives.

Certain frameworks reinterpret Gaussian placement as Markov Chain Monte Carlo (MCMC) sampling, where principled relocation strategies maintain rendering consistency while redistributing Gaussians. The optimization alternates between these enhanced geometric refinements and appearance adjustments, progressively enhancing scene representation fidelity across iterations.

These collective advances address fundamental limitations in gradient-based densification, yielding improved robustness to initialization and more efficient Gaussian utilization without compromising reconstruction quality.

Diffusion Models (Ho, et al. 2020)

Diffusion models constitute a prominent family of generative approaches that learn complex data distributions through a dual process of iterative noise corruption and denoising. These models derive their theoretical foundation from non-equilibrium thermodynamics, where a forward diffusion process progressively injects Gaussian noise into data samples over multiple timesteps, transforming structured data into pure noise. Conversely, a learned reverse process systematically removes noise to recover coherent data structures from random noise inputs. The method's efficacy stems from its stable training dynamics compared to adversarial approaches and its capacity to model complex distributions without restrictive assumptions about data topology.

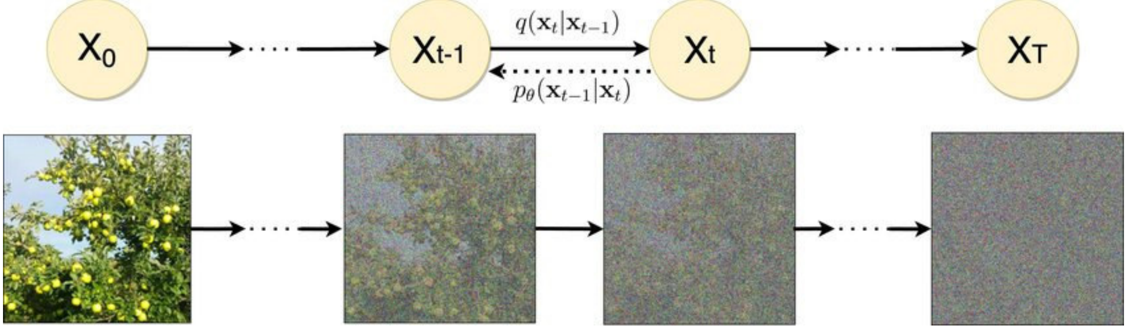

X0
Xt-1
Xt
XT
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$

Figure 3. Diffusion models overview. The graphical illustration of diffusion models with their fixed forward process $\mathrm{q}(\mathbf{x}_\mathrm{t}|\mathbf{x}_{\mathrm{t}-1})$ and learnt backward process $\mathrm{p}_\theta(\mathbf{x}_{\mathrm{t}-1}|\mathbf{x}_\mathrm{t})$

The forward diffusion trajectory operates as a Markov chain that incrementally adds noise to an initial data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. At each timestep t, the conditional distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is governed by a Gaussian transition parameterized by a pre-determined variance schedule $\beta_t \in (0,1)$. This schedule controls the rate of information degradation, typically following a monotonically increasing function that accelerates noise addition in later timesteps. The cumulative effect of T transitions admits a closed-form expression due to the Gaussian nature of the process:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}),$$

where $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

This reparameterization enables efficient sampling of $\mathbf{x}_t$ for arbitrary t without simulating the entire Markov chain, significantly accelerating training.

The generative reverse process is parameterized by a neural network that learns to invert the diffusion trajectory. Starting from isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, the model iteratively refines the sample through transitions:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t,t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t,t)\big).$$

Common implementations fix the covariance $\boldsymbol{\Sigma}_\theta$ to time-dependent constants, while the mean $\boldsymbol{\mu}_\theta$ is reparameterized to predict the injected noise:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\right).$$

Here $\boldsymbol{\epsilon}_\theta$ denotes a deep neural network, typically a U-Net architecture with residual blocks and self-attention mechanisms, conditioned on timestep embeddings. The training objective simplifies to a denoising score matching loss that minimizes the discrepancy between true and predicted noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2],$$

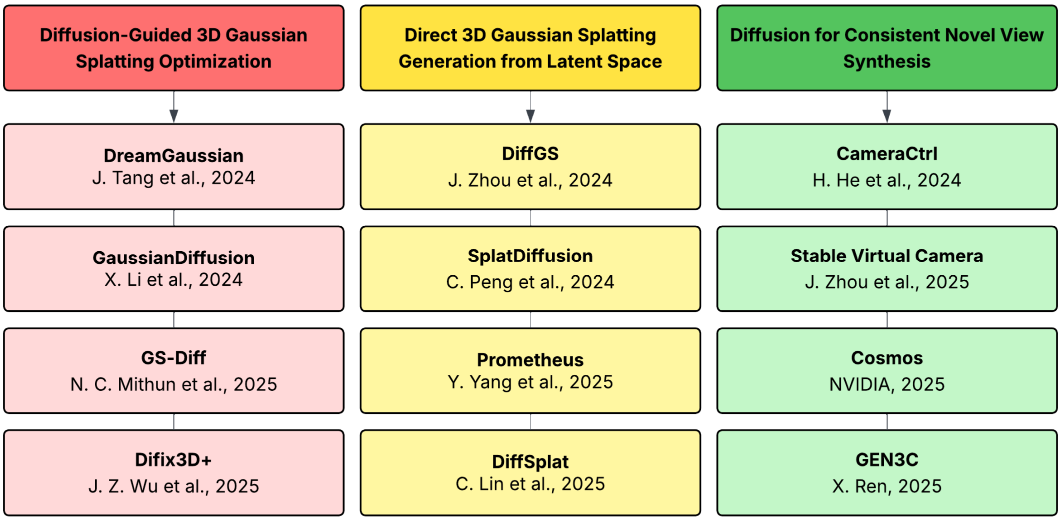
where $t \sim \mathcal{U}(\mathbf{1}, \mathbf{T})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This objective exhibits superior training stability compared to the full variational lower bound, as it circumvents compounding prediction errors across timesteps. During inference, sampling initiates from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applies the learned reverse transitions for T iterations. Advanced samplers such as denoising diffusion implicit models (DDIMs) further accelerate generation by leveraging non-Markovian trajectories that reduce required steps.

Conditional generation extends the framework by augmenting the noise predictor $\boldsymbol{\epsilon}_\theta$ with contextual information $\mathbf{y}$ (e.g., class labels or text embeddings), yielding $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{y})$. Guidance techniques like classifier-free diffusion employ stochastic conditioning to enhance sample fidelity without auxiliary models. Recent innovations integrate diffusion with latent spaces, hierarchical refinement, and hybrid architectures, enabling applications spanning high-resolution image synthesis, molecular design, and video generation. The model's inherent flexibility and strong mode coverage continue to drive empirical advances in generative modeling.

Methodological Taxonomy

This section synthesizes contemporary research integrating diffusion models with 3DGS, categorized by three synergistic paradigms: diffusion-guided 3DGS optimization, direct 3DGS generation from latent space, and diffusion for consistent novel view synthesis. Each approach addresses specific limitations in the 3DGS pipeline–such as sparse view reconstruction, geometric inconsistencies, and generative capability gaps–by leveraging diffusion priors at distinct stages of the scene representation workflow. The three synergistic paradigms discussed in this section are summarized in figure 4.



Diffusion-Guided 3D Gaussian Splatting Optimization	Direct 3D Gaussian Splatting Generation from Latent Space	Diffusion for Consistent Novel View Synthesis
DreamGaussian J. Tang et al., 2024	DiffGS J. Zhou et al., 2024	CameraCtrl H. He et al., 2024
GaussianDiffusion X. Li et al., 2024	SplatDiffusion C. Peng et al., 2024	Stable Virtual Camera J. Zhou et al., 2025
GS-Diff N. C. Mithun et al., 2025	Prometheus Y. Yang et al., 2025	Cosmos NVIDIA, 2025
Difix3D+ J. Z. Wu et al., 2025	DiffSplat C. Lin et al., 2025	GEN3C X. Ren, 2025

Figure 4. Three synergistic paradigms: diffusion-guided 3DGS optimization, direct 3DGS generation from latent space, and diffusion for consistent novel view synthesis

Diffusion-Guided 3D Gaussian Splatting Optimization

DreamGaussian (Tang, et al. 2024)

DreamGaussian pioneers the integration of diffusion priors with 3DGS to address efficiency bottlenecks in optimization-based 3D generation. The framework employs a two-stage pipeline for both image-to-3D and text-to-3D tasks. An overview of the DreamGaussian framework is shown in figure 5. In the first stage, 3D Gaussians are optimized via score distillation sampling (SDS) (Poole, et al. 2022) from 2D diffusion models, leveraging progressive densification to accelerate convergence compared to NeRF-based approaches. This stage generates coarse geometry and appearance within seconds but suffers from texture blurriness due to SDS ambiguity. The second stage addresses this limitation through novel mesh extraction and UV space refinement: an efficient local density query algorithm converts Gaussians to textured meshes, followed by a diffusion-guided fine-tuning that minimizes pixel-wise mean squared error (MSE) losses between rendered views and diffusion-refined images. This refinement explicitly disentangles texture details from geometric ambiguities while avoiding the over-saturation artifacts typical of direct SDS application in UV space.

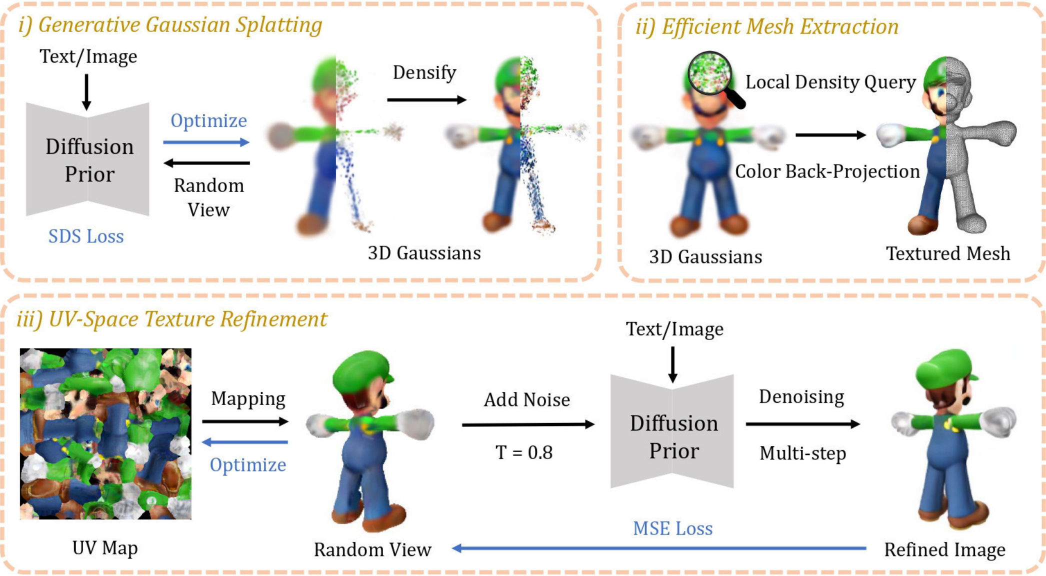


Figure 5. DreamGaussian framework overview

Despite achieving a 10 × speedup over prior methods, DreamGaussian remains susceptible to the multi-face Janus problem (see figure 6 for an example). This artifact, characterized by semantically incoherent surfaces featuring duplicated frontal features (e.g., multiple faces), stems from inconsistent 3D guidance provided by 2D diffusion priors, highlighting inherent cross-view consistency challenges.

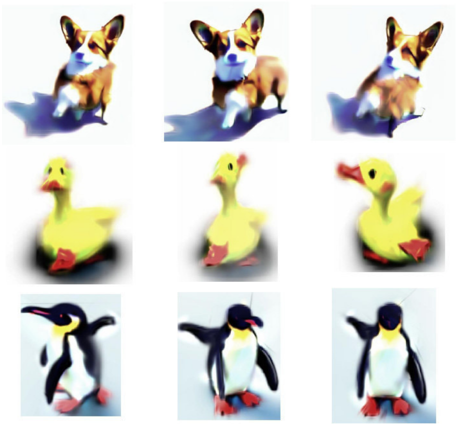


Figure 6. The Janus problem in distilling 3D knowledge from 2D diffusion models refers to a failure mode where generated 3D objects exhibit multiple, inconsistent faces (e.g., two front-facing views) on a single object

The Janus problem–manifesting as geometrically inconsistent surfaces–arises from fundamental limitations in text-to-3D optimization. Distilling knowledge from 2D diffusion models into 3D representations introduces geometric contradictions because these models generate view-specific outputs optimized for local photorealism without explicit cross-view constraints during optimization. Consequently, adjacent viewpoints often synthesize conflicting surface orientations or textures, violating 3D consistency. This issue is exacerbated early in optimization, where high noise levels amplify stochastic variations across views, and in regions with sparse supervision, leading to unconstrained extrapolation and hallucinated structures.

Compounding this issue, the non-convex optimization landscape traps parameters in view-dependent local minima. These minima satisfy individual view losses but collectively produce degenerate geometry incompatible with physical coherence.

Consequently, the Janus artifact represents not merely a rendering flaw but an optimization pathology inherent to disentangled view-wise distillation.

GaussianDiffusion (Li, et al. 2024)

To overcome the core challenge of multi-face Janus artifacts, GaussianDiffusion introduces structured 3D noise injection. Unlike methods sampling independent 2D noise per viewpoint, this approach generates structured noise by rendering randomized 3D Gaussians into view-dependent 2D distributions, ensuring multi-view coherence. Consequently, the denoising process inherently satisfies cross-view geometric constraints without requiring diffusion model fine-tuning, preserving its realism. The GaussianDiffusion framework is depicted in figure 7.

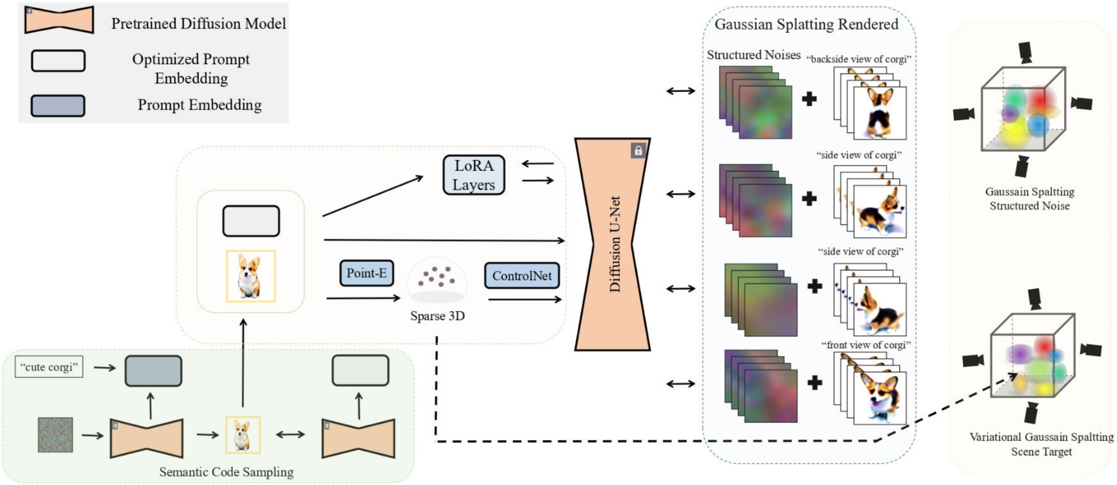


Figure 7. GaussianDiffusion framework overview. A sparse point cloud generated from an image using Point-E undergoes pose projection to create a depth map. This depth map functions as a geometric constraint for ControlNet. Simultaneously, low-rank adaptation (LoRA) provides additional optimization through fine-tuning of the diffusion model. The original sparse point cloud from Point-E serves as the initial input to the 3DGS process. Gradients from the diffusion model are conveyed to 3DGS via SDS. Mitigating issues related to multi-view consistency and artifacts involves the introduction of structured noise and a variational 3DGS approach, resulting in a realistic 3D appearance

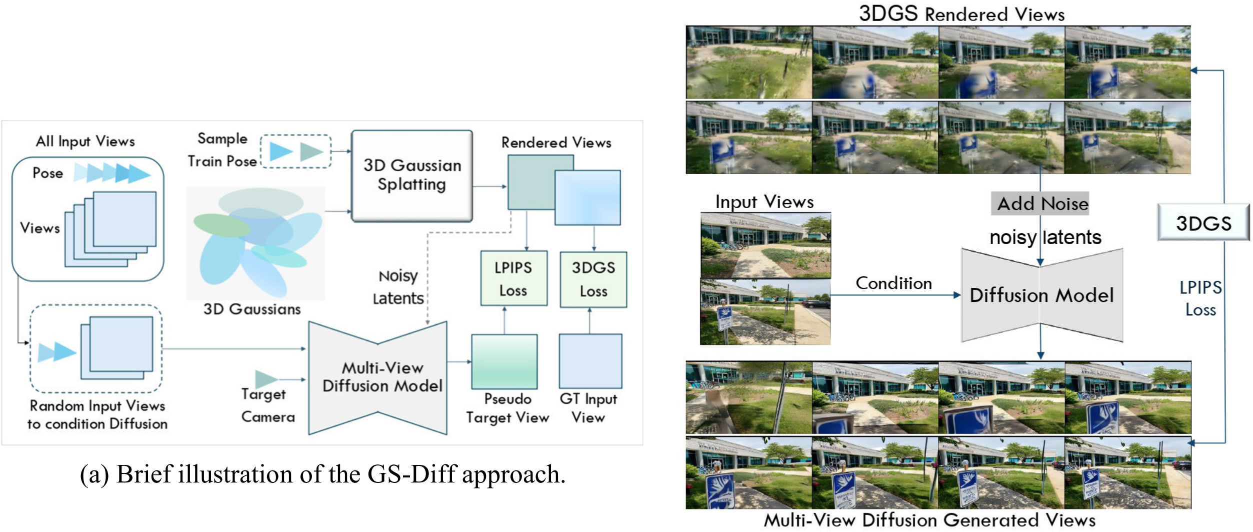
Complementing this, variational 3DGS addresses local minima by modeling Gaussian parameters (e.g., position, scale) as distributions rather than point estimates. Perturbing these parameters with noise scaled to the diffusion timestep encourages coarse-to-fine optimization, expands the convergence domain, and enhances stability.

Quantitatively, GaussianDiffusion demonstrates superior geometric consistency through reduced pose-prediction variance in COLMAP evaluations, though it incurs longer training times than highly optimized baselines like DreamGaussian.

The methodologies diverge in their handling of diffusion priors and geometric constraints. DreamGaussian relies on conventional SDS for initial optimization but circumvents its limitations via separate mesh refinement. GaussianDiffusion rethinks noise perturbation at its core, using shared 3D noise sources and variational distributions to embed multi-view consistency directly into the gradient estimation process. Consequently, DreamGaussian excels in rapid asset generation for downstream applications, while GaussianDiffusion advances robustness and geometric fidelity for complex text prompts, albeit at a higher computational cost.

GS-Diff (Mithun, et al. 2025)

The GS-Diff framework shares foundational principles with prior diffusion-guided 3D reconstruction methods in its utilization of generative priors to regularize underconstrained optimization. Similar to DreamGaussian, GS-Diff leverages diffusion models to mitigate ambiguities arising from sparse or inconsistent inputs. However, while DreamGaussian employs SDS from 2D diffusion models for geometry optimization, GS-Diff diverges significantly by integrating a multi-view diffusion model to synthesize geometrically consistent pseudo-observations. This approach aligns more closely with GaussianDiffusion's philosophy of enforcing cross-view consistency through structured 3D constraints but operates within an explicit reconstruction paradigm rather than generative synthesis. The GS-Diff framework overview is presented in figure 8.



(a) Brief illustration of the GS-Diff approach.

(b) Iterative workflow of the integrated diffusion process with the GS-Diff pipeline.

Figure 8. GS-Diff framework overview: (a) conceptual illustration and (b) iterative diffusion process workflow.

GS-Diff adopts an iterative optimization strategy that interleaves diffusion guidance with 3D Gaussian updates at regular intervals. This stands in contrast to DreamGaussian's sequential pipeline, which separates geometry optimization from texture refinement. By conditioning pseudo-view generation on neighboring input views via camera trajectory interpolation, GS-Diff implicitly enforces multi-view coherence–addressing the Janus problem without requiring explicit 3D noise projection. The method further shares GaussianDiffusion's emphasis on stability through the introduction of an LPIPS (Zhang et al. 2018) loss threshold, which dynamically excludes inconsistent pseudo-views during training. This integration mitigates the risk of hallucination inherent in diffusion models.

Difix3D+ (Wu, et al. 2025)

The Difix3D+ framework shares fundamental operational principles with contemporary approaches that leverage diffusion models for 3D reconstruction enhancement, particularly in its core strategy of distilling diffusion-refined pseudo-views into the underlying 3D representation. Similar to Deceptive-NeRF (Xinhang Liu, et al. 2024) and 3DGS-Enhancer (Xi Liu, et al. 2024), Difix3D+ employs a diffusion model to correct artifacts in rendered views, subsequently using these enhanced images as pseudo-ground truth to update the 3D model parameters. This shared methodology circumvents the computational burden of per-optimization-step diffusion queries by treating the diffusion model as an offline enhancer that generates improved training data. The distillation process aligns with established paradigms where generative priors are transferred to the 3D representation through iterative refinement of pseudo-observations, thereby addressing underconstrained regions while maintaining multi-view consistency. The architecture of the Difix3D+ framework is shown in figure 9.

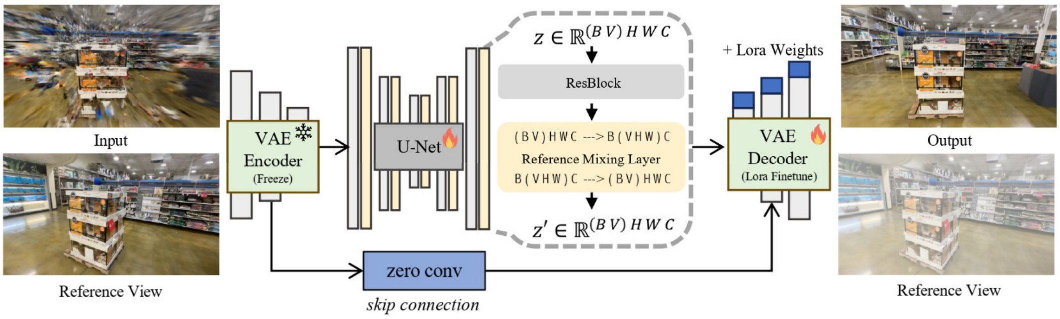


Figure 9. Difix3D+ framework overview. Difix processes a noisy rendered image alongside reference views as input (left), producing an enhanced output image with reduced artifacts (right). The system also generates identical reference views, discarded during practical application and consequently depicted as transparent. The architecture employs a U-Net structure featuring a cross-view reference mixing layer to ensure consistency across reference views. Difix is fine-tuned from SD-Turbo (Sauer et al. 2023, 2024), utilizing a frozen variational autoencoder (VAE) and a LoRA-adapted decoder

Difix3D+ further echoes prior works in its dual-phase utilization of diffusion guidance: during optimization to improve the 3D representation and during inference to refine rendered outputs. This bifurcated approach mirrors architectures like GANeRF (Roessle, et al. 2023) and NeRFLiX (Zhou, Li, Wang, et al. 2023; Zhou, Li, Jiang, et al. 2023), where generative models enhance both the reconstruction process and final rendering quality. The conditioning mechanism on reference views–implemented via cross-attention layers–draws inspiration from multi-view diffusion models (e.g., MVDream (Shi, et al. 2024), SyncDreamer (Y. Liu, et al. 2024)) that aggregate information from input perspectives to maintain contextual coherence. By inheriting these design principles, Difix3D+ operates within the broader trend of integrating 2D generative priors to compensate for 3D reconstruction limitations, particularly in sparse-view scenarios where geometric ambiguities persist.

However, Difix3D+ introduces critical innovations to this shared foundation. The progressive 3D update pipeline–inspired by Instruct-NeRF2NeRF (Haque, et al. 2023)–iteratively refines camera trajectories and augments training data, enabling consistent artifact correction even in extreme novel views. This contrasts with single-step distillation in prior methods, which often struggle with long-range consistency. Additionally, the adoption of a single-step diffusion model (SD-Turbo (Sauer, et al. 2023, 2024)) significantly accelerates both training and inference, addressing efficiency limitations in earlier diffusion-guided frameworks. While maintaining conceptual alignment with existing paradigms, these advancements position Difix3D+ as a scalable solution for large-scale scenes where computational overhead traditionally impedes deployment.

Strong conditioning of the diffusion model on rendered novel views and reference views is crucial for achieving multi-view consistency and high fidelity to input perspectives. When the target trajectory is distant from input views, the conditioning signal weakens, forcing the diffusion model to hallucinate excessively. To mitigate this, an iterative training scheme analogous to Instruct-NeRF2NeRF (Haque, et al. 2023) is adopted, progressively expanding the set of 3D cues rendered to novel viewpoints. Starting with reference views, the 3D representation undergoes optimization with periodic perturbation of ground-truth camera poses toward target views. The resulting novel views are rendered, refined by Difix, and added to the training set for further optimization cycles. This gradual pose perturbation and data augmentation enhance 3D consistency in challenging regions.

Despite distillation, slight multi-view inconsistencies and residual blur persist due to limitations in reconstruction capacity. Difix3D+ addresses this by applying Difix as a final post-processing step during inference, removing artifacts while preserving coherence. Leveraging SD-Turbo's (Sauer, et al. 2023, 2024) single-step architecture, this adds only 76 ms per frame on an NVIDIA A100 GPU–significantly faster than multi-step diffusion models. This dual optimization-inference application of diffusion guidance extends prior frameworks where generative enhancement was typically restricted to training.

Direct 3D Gaussian Splatting Generation from Latent Space

These approaches train diffusion models to directly generate the 3DGS representation itself, often by learning a compact latent space of the Gaussians or functions defining them.

DiffGS (Zhou, et al. 2024)

DiffGS exemplifies this paradigm by reformulating the unstructured nature of 3DGS representation into a continuous functional representation, enabling latent diffusion over Gaussian attributes. Similar to concurrent methods like GaussianCube (Zhang, et al. 2024) and GVGEN (He, et al. 2024), which structure Gaussians into volumetric grids for tractable generation, DiffGS circumvents the discrete challenges of 3DGS representation by introducing three disentangled continuous functions: Gaussian probability function (GauPF), Gaussian color function (GauCF), and Gaussian transform function (GauTF). These functions collectively parameterize the geometry, appearance, and transformations of the Gaussians, effectively distilling the unstructured 3DGS representation into a compact, generative latent space. The DiffGS framework is illustrated in figure 10.

A Gaussian VAE compresses these functions into a low-dimensional latent vector, regularizing the space for stable diffusion training. The latent diffusion model (LDM) then operates in this space, denoising samples to generate novel functional representations conditioned on inputs like text, images, or partial 3DGS. This mirrors the latent-space generation strategies of triplane-based LDMs (e.g., Rodin (Wang, et al. 2022) and 3DGen (Gupta, et al. 2023)), but avoids their reliance on grid structures by directly modeling Gaussian properties through neural predictors. After generation, an octree-guided discretization algorithm extracts discrete Gaussians from the continuous functions. This process optimizes proxy points toward high-probability regions defined by the GauPF, then queries the GauCF and GauTF to assign attributes. Analogous to marching cubes for implicit fields–but tailored to Gaussian primitives–this approach ensures scalable extraction of Gaussians at arbitrary densities, free from voxel-resolution constraints.

While methods like DreamGaussian optimize Gaussians via SDS, DiffGS directly generates the 3DGS representation end-to-end via latent diffusion, eliminating per-scene optimization. This functional approach enhances efficiency and generality, though it inherits challenges in color consistency and geometric fidelity under sparse conditioning, as seen in analogous latent-space generative models.

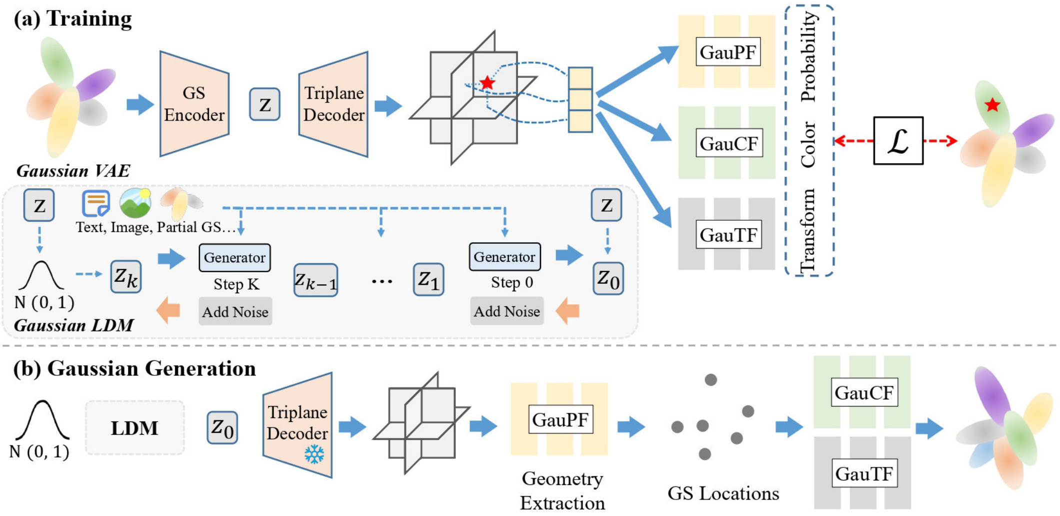


Figure 10. DiffGS framework overview. (a) The fitted 3DGS is disentangled into three 3DGS functions to model Gaussian probabilities (GauPF), Gaussian colors (GauCF), and Gaussian transforms (GauTF). A Gaussian VAE is then trained with a conditional latent diffusion model (LDM) to generate these functions. (b) During generation, Gaussian geometry is first extracted from the generated GauPF, followed by applying GauCF and GauTF to obtain Gaussian attributes

SplatDiffusion (Peng, et al. 2024)

SplatDiffusion addresses the fundamental challenge of modality mismatch in diffusion-based 3D generation by introducing a teacher-guided framework that decouples the denoised signal domain from the supervision domain. Unlike conventional diffusion paradigms requiring aligned 3D supervision, SplatDiffusion leverages pre-trained deterministic predictors (e.g., Splatter Image (Szymanowicz, et al. 2024), Flash3D (Szymanowicz, et al. 2025)) as "noisy teachers" to synthesize corrupted 3DGS samples. This strategy circumvents the scarcity of 3D ground truth by utilizing imperfect teacher predictions as pseudo-targets for diffusion training. Crucially, at noise levels beyond a critical timestep t^*, the distribution of teacher-generated noisy samples approximates that of forward-noised ground truth, inspired by SDEdit (Meng, et al. 2022) principles.

The framework operates in two synergistic stages. The two-stage SplatDiffusion framework is overviewed in figure 11. First, during bootstrapping, the diffusion model learns single-step denoising using direct 3D supervision from the teacher's predictions combined with 2D rendering losses. This initial phase ensures computational efficiency while aligning the model with the teacher's capabilities. Second, multi-step denoising fine-tuning replaces single-step prediction with iterative refinement across multiple timesteps. Here, the model progressively denoises from $t > t^*$ to $t = 0$, supervised exclusively by image reconstruction losses on rendered views. This sequence mirrors inference dynamics, enabling gradient propagation through the full denoising trajectory and significantly enhancing detail recovery.

Additionally, SplatDiffusion integrates a cycle consistency loss, where predicted novel views drive secondary 3D reconstructions that are rendered back to the source view. This regularizes spatial coherence and mitigates geometric drift. For multi-view inputs, the model incorporates guidance gradients during sampling–modulating noise estimates using discrepancies between rendered and target views–which further elevates reconstruction quality without architectural modifications. By operating directly on 3DGS parameters and using only 2D supervision, SplatDiffusion achieves SOTA novel view synthesis while sidestepping the limitations of 3D data dependency. Its flexibility allows seamless integration with diverse teacher models, demonstrating consistent gains across both object-level and scene-level benchmarks.

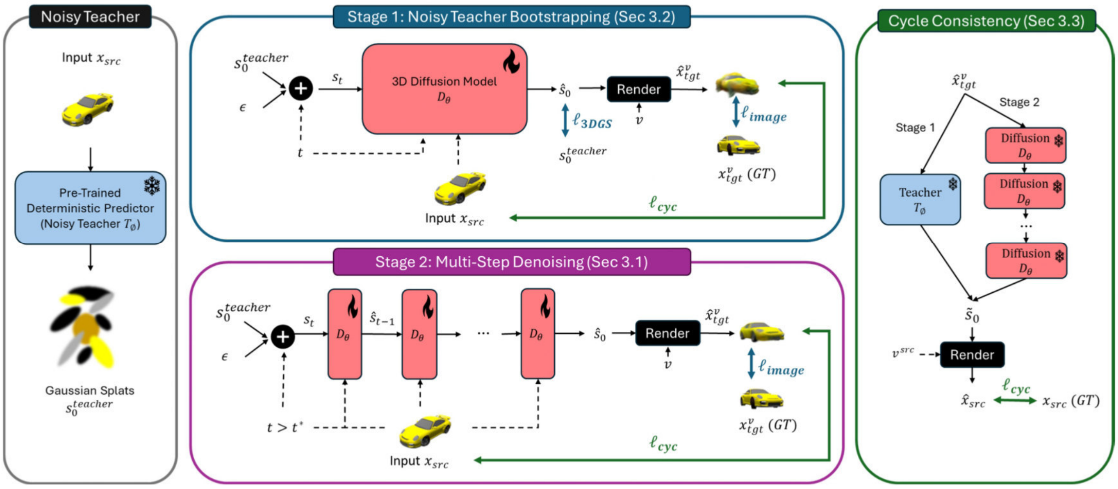


Figure 11. SplatDiffusion framework overview. This framework utilizes a pre-trained deterministic predictor network for 3DGS, termed the "noisy teacher" (left). In stage 1 (top), sampled views are lifted to generate an imperfect 3DGS prediction. This prediction provides noisy samples and supervision for the diffusion denoiser in 3DGS, augmented with additional image supervision. In stage 2 (bottom), the noisy samples are decoupled from supervision. Instead, the noisy teacher generates noisy samples at noise levels $t > t^*$, while a multi-step denoising strategy generates high-quality predictions to facilitate image-only supervision. Both stages incorporate cycle consistency regularization.

Prometheus (Yang, et al. 2025)

Prometheus aligns with emerging paradigms that leverage LDMs (Rombach, et al. 2022) to directly synthesize 3DGS representations, circumventing intermediate optimization or explicit multi-view image synthesis. Similar to DiffGS, Prometheus employs a functional compression strategy: it distills complex 3DGS attributes into a structured latent space amenable to diffusion modeling. However, while DiffGS decomposes Gaussians into geometry, color, and transformation functions, Prometheus adopts a multi-view latent conditioning approach. It encodes RGB-D image observations into a joint latent space using a frozen Stable Diffusion encoder, then decodes these latents into pixel-aligned 3D Gaussians via a modified Stable Diffusion decoder. This formulation implicitly embeds Gaussian properties (position, rotation, scale, opacity, spherical harmonics) within the latent-to-3DGS mapping, effectively structuring the unstructured nature of 3DGS through the lens of 2D priors. The Prometheus training pipeline is shown in figure 12.

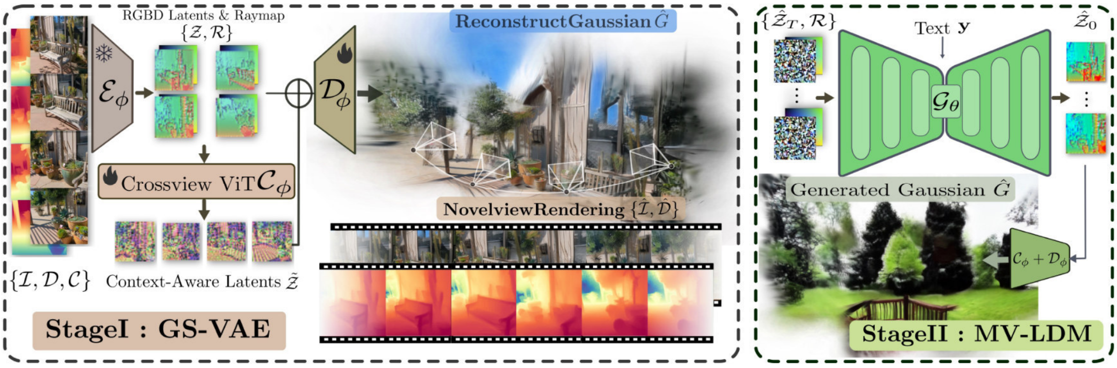

RGBD Latents & Raymap
$\{\mathcal{Z}, \mathcal{R}\}$
$\mathcal{E}_\phi$
$\mathcal{D}_\phi$
Crossview ViT $\mathcal{C}_\phi$
$\{\mathcal{I}, \mathcal{D}, \mathcal{C}\}$
Context-Aware Latents $\tilde{\mathcal{Z}}$
StageI : GS-VAE
Reconstruct Gaussian $\hat{G}$
Novelview Rendering $\{\hat{\mathcal{I}}, \hat{\mathcal{D}}\}$
$\{\mathcal{Z}_T, \mathcal{R}\}$
Text $\mathbf{y}$
$\mathcal{G}_\theta$
$\hat{\mathcal{Z}}_0$
Generated Gaussian $\hat{G}$
$\mathcal{C}_\phi + \mathcal{D}_\phi$
StageII : MV-LDM

Figure 12. Prometheus framework overview. The training process is divided into two stages. In stage 1, the objective is to train a GS-VAE. Utilizing multi-view images along with their corresponding pseudo-depth maps and camera poses, the GS-VAE is designed to encode these multi-view RGB-D images, integrate cross-view information, and ultimately decode them into pixel-aligned 3DGS. In stage 2, the focus shifts to training an MV-LDM. Multi-view RGB-D latents can be generated by sampling from randomly-sampled noise with the trained MV-LDM.

The diffusion process in Prometheus operates on multi-view RGB-D latents rather than explicit Gaussian parameters. MV-LDM denoises these latents conditioned on camera poses and text prompts, analogous to how DiffGS denoises functional representations. Crucially, Prometheus introduces cross-view transformers to fuse latent information across camera views, ensuring spatial coherence in the generated 3DGS–addressing a key challenge in multi-view generative models. This fusion mechanism, coupled with depth-aware latent conditioning, disentangles geometric and appearance attributes, enhancing fidelity. Post-diffusion, the GS-VAE decoder reconstructs the 3DGS scene in a feed-forward manner, paralleling DiffGS's function-to-Gaussian extraction but optimized for scene-level generalization.

While both methods avoid per-scene optimization, Prometheus uniquely harnesses hybrid training data that combines single-view and multi-view inputs, along with a pre-aligned RGB-D latent space, to exploit massive 2D datasets. This contrasts with DiffGS's reliance on explicit 3D function learning. The resulting model achieves rapid generation with robustness to diverse prompts, though it inherits latent diffusion challenges such as sensitivity to noise schedules and guidance trade-offs.

DiffSplat (Lin, et al. 2025)

DiffSplat aligns with emerging methodologies that harness pre-trained 2D diffusion models for direct 3DGS generation, circumventing the limitations of intermediate multi-view synthesis or explicit 3D dataset dependencies. Unlike approaches that structure Gaussians into volumetric grids or functional representations, DiffSplat reinterprets multi-view 3DGS property grids as stylized images, enabling the repurposing of large-scale image diffusion priors. This strategy leverages the inherent capability of image diffusion models to infer 3D geometry–such as depth and surface normals–from web-scale 2D data, thereby bridging the domain gap between image generation and 3D content creation. The DiffSplat framework is summarized in figure 13.

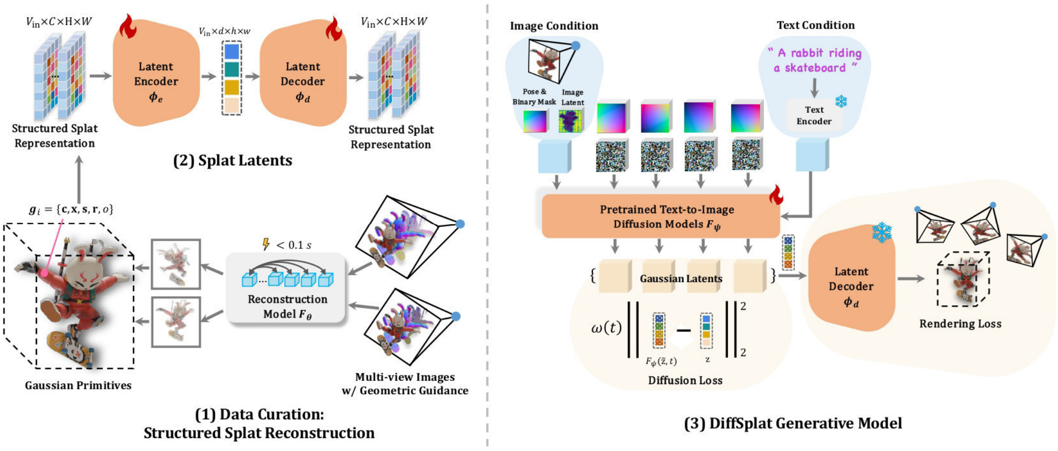

$V_{in}\times C\times H\times W$
Latent Encoder ϕ_e
$V_{in}\times d\times h\times w$
Latent Decoder ϕ_d
$V_{in}\times C\times H\times W$
Structured Splat Representation
Structured Splat Representation
(2) Splat Latents
$g_i = \{\mathbf{c}, \mathbf{x}, \mathbf{s}, \mathbf{r}, o\}$
< 0.1 s
Reconstruction Model F_θ
Gaussian Primitives
Multi-view Images w/ Geometric Guidance
(1) Data Curation: Structured Splat Reconstruction
Image Condition
Pose & Binary Mask
Image Latent
Text Condition
" A rabbit riding a skateboard "
Text Encoder
Pretrained Text-to-Image Diffusion Models F_ψ
Gaussian Latents
$\omega(t)$
$F_\psi(\tilde{z}, t)$
z
Diffusion Loss
Latent Decoder ϕ_d
Rendering Loss
(3) DiffSplat Generative Model

Figure 13. DiffSplat framework overview. (1) A lightweight reconstruction model creates high-quality structured representations for synthetic dataset curation. (2) An image VAE is fine-tuned to encode 3DGS properties into a shared latent space. (3) DiffSplat leverages 2D priors from text-to-image diffusion models to natively generate 3D content from image or text prompts.

Central to DiffSplat is a lightweight reconstruction model that regresses pixel-aligned Gaussian attributes (position, rotation, scale, color, opacity) from multi-view images in under 0.1 seconds. These attributes are structured into 2D grids, analogous to multi-channel images, which serve as scalable pseudo-ground truth data for training. To align these grids with the latent space of image diffusion models, the framework fine-tunes the VAE of pre-trained LDMs (e.g., Stable Diffusion, SDXL), compressing 3DGS property grids into 3DGS latents. This adaptation ensures compatibility with the input expectations of 2D denoising networks while preserving the spatial structure of Gaussian attributes.

The generative backbone of DiffSplat operates by denoising 3DGS latents conditioned on text or single-view images. Crucially, DiffSplat incorporates a 3D rendering loss alongside the standard diffusion loss. This dual-loss mechanism–trained end-to-end–ensures that the model adheres to 3D geometric constraints while leveraging the rich priors embedded in image diffusion architectures. For multi-view generation, DiffSplat explores view-concat and spatial-concat paradigms, where 3DGS latents are fused along view or spatial dimensions, augmented with Plücker embeddings to encode relative camera poses. Minimal architectural modifications (e.g., zero-initialized layers for pose embeddings) ensure seamless integration with diverse base models, enabling techniques like ControlNet to be adapted for controllable 3D generation.

By treating 3DGS property grids as a stylized image modality, DiffSplat effectively "fine-tunes" image diffusion models to generate 3D-consistent Gaussian primitives, eliminating the need for two-stage pipelines or explicit 3D supervision. This approach not only capitalizes on the scalability of 2D diffusion models but also establishes a unified framework for text-to-3D and image-to-3D tasks, achieving SOTA fidelity and generalizability across benchmarks.

Diffusion for Consistent Novel View Synthesis

This approach primarily focuses on generating high-quality, 3D-consistent 2D novel views using diffusion models. While it does not explicitly build or refine a 3DGS representation internally for its primary function, the consistency and quality of its generated views are sufficient to enable the reconstruction of a coherent 3DGS model as a downstream task.

CameraCtrl (He, et al. 2025)

CameraCtrl represents a significant advancement in integrating explicit camera control within video diffusion models, addressing a critical gap in cinematic controllability for generative video synthesis. Unlike conventional approaches that lack precise viewpoint manipulation, CameraCtrl introduces a plug-and-play module that enables accurate trajectory control without modifying core model parameters. The methodology employs Plücker embeddings as the primary camera representation, encoding intrinsic and extrinsic parameters into pixel-wise ray descriptors that provide geometrically interpretable conditioning signals. This formulation captures the epipolar geometry underlying camera movements, offering richer spatial cues than numerical parameter inputs. The camera control module is deliberately decoupled from appearance learning; it processes only Plücker embeddings, ensuring domain-agnostic generalization across diverse visual styles. This design prevents appearance leakage from training data, a common limitation in conditioning frameworks that fuse image latents with control signals. The CameraCtrl framework overview is provided in figure 14.

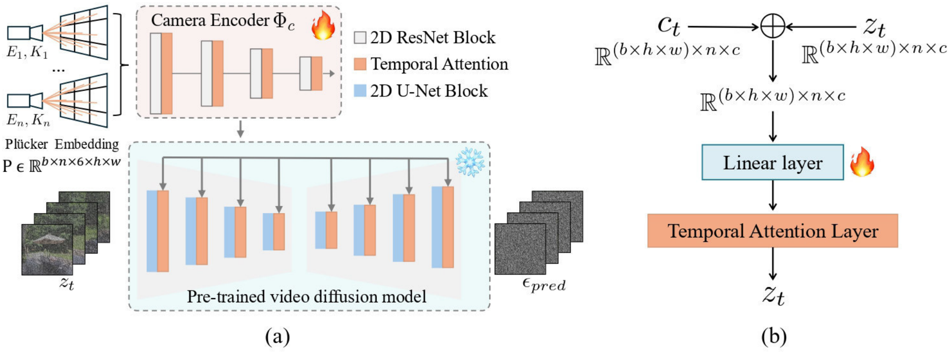

Camera Encoder Φ_c
□ 2D ResNet Block
🟧 Temporal Attention
🟦 2D U-Net Block
E_1, K_1
...
E_n, K_n
Plücker Embedding
$P \in \mathbb{R}^{b\times n\times 6\times h\times w}$
z_t
Pre-trained video diffusion model
ϵ_{pred}
(a)
c_t
z_t
$\mathbb{R}^{(b\times h\times w)\times n\times c}$
$\mathbb{R}^{(b\times h\times w)\times n\times c}$
$\mathbb{R}^{(b\times h\times w)\times n\times c}$
Linear layer
Temporal Attention Layer
z_t
(b)

Figure 14. CameraCtrl framework overview. (a) Given a pre-trained video diffusion model, CameraCtrl trains a camera encoder on it, which takes the Plücker embedding as input and outputs multi-scale camera representations. These features are then integrated into the temporal attention layers of the U-Net at their respective scales to control the video generation process. (b) Details of the camera injection process. The camera features c_t and the latent features z_t are first combined through the element-wise addition. A learnable linear layer is adopted to further fuse two representations which are then fed into the first temporal attention layer of each temporal block

Integration occurs through injection into temporal attention layers of the U-Net backbone, aligning camera dynamics with the sequential nature of video generation. The features derived from Plücker sequences are fused via element-wise addition and linear transformation before modulating temporal self-attention. Empirical analysis reveals that this temporal integration outperforms spatial attention fusion, as camera trajectories inherently exhibit causal dependencies across frames. For training, RealEstate10K is selected due to its diverse camera distributions and appearance alignment with common video diffusion datasets. This choice balances trajectory complexity and visual realism, mitigating geometric ambiguities that arise from synthetic or narrow-distribution data. Quantitative metrics confirm CameraCtrl's superiority in trajectory adherence (RotErr, TransErr) over alternatives like MotionCtrl (Wang, et al. 2024), while maintaining baseline visual quality (FVD, CLIPSim).

The framework demonstrates versatility across text-to-video (e.g., AnimateDiff (Guo, et al. 2024)) and image-to-video (e.g., Stable Video Diffusion (Blattmann, et al. 2023)) pipelines, and seamlessly combines with structural controllers like SparseCtrl (Guo, et al. 2023) for multimodal conditioning. Its plug-and-play nature enables rapid adoption in production workflows, though limitations persist in extreme rotational regimes due to dataset biases. Conceptually, CameraCtrl bridges cinematographic principles with generative AI, enabling dynamic storytelling through controlled viewpoint dynamics.

Stable Virtual Camera (SEVA) (Zhou, et al. 2025)

Similarly, SEVA leverages diffusion models for 3D-consistent view synthesis but operates without explicit intermediate 3D representations. While methods like DreamGaussian and GaussianDiffusion directly optimize 3D Gaussians using diffusion priors, SEVA focuses on synthesizing photorealistic 2D novel views that implicitly satisfy 3D consistency. This approach conditions the diffusion process on input views and their camera poses, utilizing a latent denoising U-Net enhanced with 3D self-attention and temporal operators. The model inherits strong priors from pre-trained 2D diffusion architectures, enabling it to generate large viewpoint changes and smooth interpolations without relying on explicit geometric scaffolds. The SEVA framework is illustrated in figure 15.

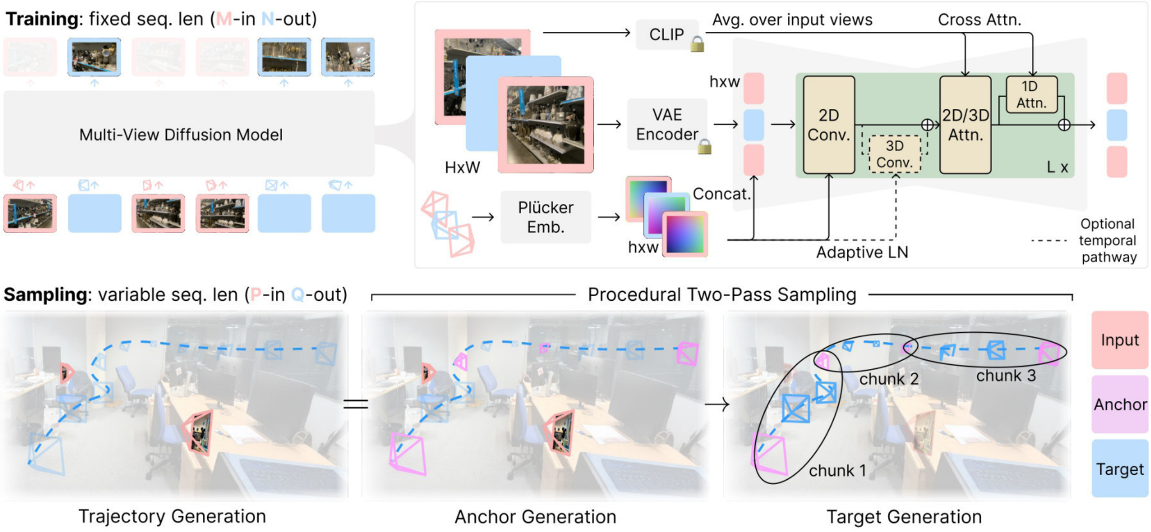


Figure 15. SEVA framework overview. SEVA is trained with a fixed sequence length as an "M-in N-out" multi-view diffusion model using a standard architecture. It conditions on CLIP embeddings, VAE latents of the input views, and their corresponding camera poses. During sampling, SEVA can be cast as a generative "P-in Q-out" renderer operating with variable sequence length, where P and Q need not equal M and N. To enhance temporal and 3D consistency across generated views, particularly when generating along a trajectory, procedural two-pass sampling is presented as a general strategy

Crucially, SEVA's procedural two-pass sampling strategy ensures temporal and spatial coherence across generated views. For long trajectories, a memory bank of anchor views maintains long-range consistency by retrieving spatially nearest neighbors during autoregressive generation. The resulting views exhibit sufficient geometric consistency to enable high-quality 3DGS reconstruction as a downstream task, as validated by distillation experiments. Unlike optimization-based methods, SEVA sidesteps the Janus problem during synthesis by avoiding entangled 3D parameter optimization, instead relying on diffusion's generative capacity to hallucinate plausible yet consistent unseen regions. However, it shares the challenge of handling dynamic textures and ambiguous scenes, where generation uncertainty may necessitate careful guidance scale tuning to balance detail fidelity and coherence. Thus, while SEVA diverges from explicit 3DGS optimization, its output quality and consistency functionally enable comparable downstream 3D reconstruction, streamlining the pipeline through decoupled view synthesis and geometry distillation.

Cosmos (Agarwal, et al. 2025; Alhaija, et al. 2025)

Similarly, the Cosmos world foundation model (WFM) employs diffusion-based architectures for generating 3D-consistent novel views without explicit intermediate geometric representations. Like SEVA, it conditions the diffusion process on camera parameters–specifically leveraging Plücker coordinate embeddings to encode relative camera poses–alongside visual observations from preceding frames. This approach integrates temporal and spatial conditioning through a modified DiT backbone, utilizing hybrid positional embeddings with FPS-aware 3D RoPE (Su, et al. 2023) and learnable absolute embeddings to maintain coherence across arbitrary viewpoints and trajectory lengths. The model's latent space, constructed via a causal tokenizer, compresses high-dimensional visual inputs while preserving critical spatial and temporal relationships, enabling efficient denoising in a reduced-dimensional domain.

For extended sequences, Cosmos adopts a progressive training strategy that incrementally increases resolution and context length, coupled with multi-aspect bucketing to handle diverse scene compositions. This design mirrors SEVA's emphasis on cross-view consistency but extends it through joint image-video training, where domain-specific normalization aligns latent distributions across modalities. The resulting outputs exhibit sufficient geometric fidelity to facilitate downstream 3D reconstruction tasks, as validated by metrics like Sampson error and novel-view synthesis quality. However, Cosmos diverges by directly supporting dynamic perturbations (e.g., robotic actions or autonomous vehicle controls), broadening its applicability beyond static scene synthesis. This flexibility introduces challenges in physics adherence, particularly for contact-rich interactions, which remain an active area of refinement.

GEN3C (Ren, et al. 2025)

GEN3C represents a significant advancement in 3D-consistent video generation by integrating explicit geometric scaffolding into a diffusion-based framework. Unlike purely latent-space approaches, GEN3C constructs a spatiotemporal 3D cache–a collection of colored point clouds generated by unprojecting depth estimates from input images or pre-generated video frames. This cache serves as an approximate 3D scene representation, rendered into 2D videos conditioned on user-specified camera trajectories. The core innovation lies in leveraging these renderings as strong geometric priors for a video diffusion model (e.g., Stable Video Diffusion (Blattmann, et al. 2023) or Cosmos (Agarwal, et al. 2025; Alhaija, et al. 2025)), which is fine-tuned to translate imperfect cache renderings into high-fidelity, temporally coherent videos. The GEN3C framework overview is shown in figure (Ren, et al. 2025).

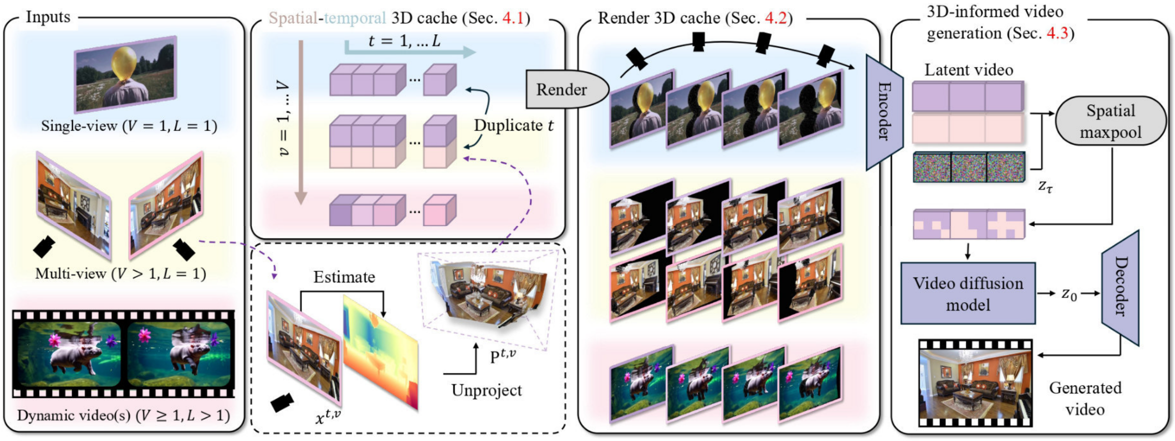

Inputs
Single-view ($V = 1, L = 1$)
Multi-view ($V > 1, L = 1$)
Dynamic video(s) ($V \geq 1, L > 1$)
Spatial-temporal 3D cache (Sec. 4.1)
$t = 1, \ldots L$
$v = 1, \ldots V$
Render
Duplicate t
Estimate
$x^{t,v}$
$p^{t,v}$
Unproject
Render 3D cache (Sec. 4.2)
Encoder
3D-informed video generation (Sec. 4.3)
Latent video
Spatial maxpool
z_τ
Video diffusion model
z_0
Decoder
Generated video

Figure 16. GEN3C framework overview. Given user input–which can be a single-view image, multi-view images, or dynamic videos–a spatiotemporal 3D cache is constructed through depth prediction for each image and subsequent unprojection into 3D. Using camera poses provided by the user, the cache is rendered into videos. These rendered videos are then fed into a video diffusion model to generate a photorealistic video aligned with the target camera poses

The 3D cache dynamically adapts to application contexts: for single-image inputs, it duplicates the initial point cloud across frames; for multi-view static scenes, it aggregates point clouds from each view; and for dynamic scenes, it integrates per-frame depth predictions from source videos. During diffusion, rendered cache frames and disocclusion masks are encoded into latent space, masked to exclude uncovered regions, and fused via viewpoint-invariant max-pooling before conditioning the denoising U-Net. This design ensures the diffusion model focuses generative capacity on disoccluded areas and dynamic refinements rather than inferring scene geometry from camera parameters alone.

GEN3C addresses long-range consistency through autoregressive chunk generation with cache updates. For extended sequences, previously generated frames are depth-estimated, scale-aligned to the existing cache via reprojection error minimization, and unprojected to augment the 3D representation iteratively. This enables applications like driving scene simulation, where horizontal camera shifts beyond training trajectories are synthesized plausibly. Crucially, the explicit cache permits direct 3D editing (e.g., object removal or trajectory modification) by manipulating point clouds before re-rendering and re-generation.

Quantitatively, GEN3C achieves SOTA performance in sparse-view novel view synthesis. Its conditioning strategy yields superior camera controllability over CameraCtrl, particularly under out-of-domain trajectories, while maintaining efficiency. Limitations include dependence on pre-computed dynamics for dynamic scenes and sensitivity to extreme depth noise. By bridging explicit 3D proxies with generative priors, GEN3C establishes a scalable paradigm for cinematic control and 3D-consistent synthesis in sparse-view settings.

Comparative Evaluation Limitations

The rapid proliferation of diffusion-based methodologies for enhancing 3DGS optimization presents significant challenges for systematic benchmarking and comparative evaluation. These limitations stem from fundamental methodological inconsistencies, resource constraints inherent to academic research, and heterogeneous practices in code sharing within the computer vision and graphics communities, as summarized in table 1. This section delineates the structural impediments that preclude direct quantitative comparison across SOTA techniques, contextualizing these challenges within broader research practices while maintaining scholarly rigor.

Table 1. **Code availability of diffusion-enhanced 3DGS methods. The table shows the availability of training and inference code for each method. The table is based on the information provided in the original papers and the availability of the code on the authors' GitHub repositories**

Method	Training code available	Inference code available
Diffusion-Guided 3D Gaussian Splatting Optimization		
DreamGaussian (Tang et al. 2024)	Uses third-party diffusion models	Yes
GaussianDiffusion (Li et al. 2024)	No	No
GS-Diff (Mithun et al. 2025)	No	No
Difix3D+ (Wu et al. 2025)	Yes	Yes
Direct 3D Gaussian Splatting Generation from Latent Space		
DiffGS (Zhou et al. 2024)	Yes	Yes
SplatDiffusion (Peng et al. 2024)	No	No
Prometheus (Yang et al. 2025)	Partially available	Yes
DiffSplat (Lin et al. 2025)	Yes	Yes
Diffusion for Consistent Novel View Synthesis		
CameraCtrl (He et al. 2025)	Yes	Yes
SEVA (Zhou et al. 2025)	No	Yes
Cosmos (NVIDIA, :, Agarwal, et al. 2025; NVIDIA, :, Alhaija, et al. 2025)	Yes	Yes
GEN3C (Ren et al. 2025)	No	Yes

Heterogeneity of Training Data and Experimental Conditions

A primary barrier to comparative analysis lies in the substantial divergence in training datasets and experimental protocols employed across existing studies. Research efforts in diffusion-enhanced 3DGS optimization exhibit pronounced dataset specificity, with methodologies developed and evaluated against bespoke collections of scenes that vary dramatically in content complexity, image resolution, capture conditions, and scene diversity. This dataset fragmentation creates an irreconcilable normalization challenge: performance metrics including PSNR, SSIM, and LPIPS become fundamentally incomparable when derived from non-overlapping visual domains with varying texture richness, lighting conditions, and structural complexity. Consequently, any cross-study quantitative comparison risks confounding methodological efficacy with dataset-specific characteristics. Moreover, the absence of standardized evaluation suites tailored to diffusion-enhanced 3DGS exacerbates this issue, as researchers select validation scenes that optimally demonstrate their method's strengths while potentially obscuring limitations observable under alternative data regimes.

Resource Disparities and Training Inaccessibility

The computational burden associated with retraining diffusion models for 3DGS guidance constitutes a prohibitive barrier to controlled evaluation. Contemporary diffusion frameworks require extensive computational resources–typically hundreds of GPU hours–and specialized hyperparameter tuning to achieve published performance levels. The majority of surveyed works provide only inference code and pretrained checkpoints, omitting critical implementation details necessary for replication of training pipelines. This practice renders method retraining practically infeasible for researchers lacking institutional-scale computational infrastructure, effectively preventing standardized evaluation on common datasets. Even when partial training code exists, documentation gaps regarding data preprocessing, augmentation strategies, and optimization schedules introduce significant reproducibility uncertainties. The challenge intensifies for approaches incorporating custom diffusion architectures or novel conditioning mechanisms, where architectural details are frequently under-specified in publications and absent from repositories. Consequently, the research community faces an asymmetrical landscape where novel methodologies can be empirically validated only against their authors' predetermined baselines under non-reproducible conditions, rather than through independent comparative assessment.

Implementation Inconsistencies and Code Availability Gaps

The absence of standardized implementations across diffusion-enhanced 3DGS frameworks presents another critical barrier to fair benchmarking. Many influential papers in this domain either provide no public implementation at all or release only partial inference codebases lacking training infrastructure. This necessitates arduous reimplementation efforts that introduce significant fidelity risks-particularly when reconstructing complex diffusion-3DGS interaction mechanisms described ambiguously in prose.

Furthermore, the rapid publication tempo in this field outpaces the community's ability to establish standardized evaluation baselines or reproduce existing methods before newer approaches displace them.

Metric Limitations and Qualitative Performance Dimensions

Established quantitative metrics for novel view synthesis inadequately capture the perceptual and structural nuances critical to diffusion-enhanced 3DGS outcomes. Traditional measures like PSNR and SSIM prioritize pixel-level fidelity over geometric coherence, failing to penalize artifacts specific to Gaussian representations such as floaters, over-saturation in high-frequency regions, or topological inconsistencies. These metrics prove particularly inadequate for evaluating diffusion guidance contributions, which often target higher-order scene attributes including material consistency, illumination stability, and view-consistent texture synthesis–dimensions poorly quantified by current measures. The problem intensifies for editing applications, where no standardized metrics exist to assess how diffusion enhancement facilitates semantic manipulation while preserving geometric integrity. Consequently, qualitative assessment remains indispensable yet methodologically problematic: visual comparisons suffer from selective scene presentation and rendering parameter variations across studies. Moreover, the most significant advantages of diffusion-enhanced approaches–such as reduced artifacts in sparse-view reconstruction or enhanced detail recovery in textureless regions–manifest under specific challenging conditions rarely represented in standardized test suites.

These challenges do not diminish the substantive advances in diffusion-enhanced 3DGS optimization but rather highlight the growing pains of an explosively evolving field. As the community coalesces around standardized evaluation practices, the current methodological heterogeneity may ultimately yield robust hybrid approaches synthesizing the strengths of diverse guidance strategies. However, present constraints necessitate scholarly candor regarding the inferential boundaries of cross-method assessment.

Open Challenges and Future Directions

Despite significant advancements in diffusion-enhanced 3DGS, several fundamental challenges persist. Computational efficiency remains a critical bottleneck, as integrating iterative diffusion sampling with 3DGS optimization creates prohibitive training and inference costs–particularly for high-resolution or dynamic scenes. This is exacerbated by the memory-intensive nature of denoising processes, limiting real-time deployment in applications like augmented reality. Multi-view consistency also presents unresolved difficulties, as evidenced by persistent "Janus artifacts" in generative pipelines like DreamGaussian and GS-Diff, where 2D diffusion priors fail to enforce coherent 3D geometry across viewpoints. Such inconsistencies are amplified in sparse-view reconstruction scenarios, where geometric ambiguities lead to artifacts.

Controllability constitutes another significant gap. While methods like CameraCtrl enable trajectory control via Plücker embeddings, they struggle with extreme camera rotations or complex object-centric manipulations. Semantic editing–such as material or lighting changes–often requires per-scene optimization rather than flexible, high-level guidance. Furthermore, dynamic scene modeling remains largely underdeveloped; most frameworks prioritize static scenes, with nascent approaches like GEN3C relying heavily on pre-computed motion data rather than generative dynamics. Physics-aware synthesis (e.g., collision responses or fluid interactions) is almost entirely unexplored within the 3DGS paradigm.

Scalability is equally pressing: current datasets like RealEstate10K or Objaverse lack the diversity needed for robust generalization, particularly for uncommon objects or cultural contexts. This data scarcity intensifies biases and limits cross-domain adaptability.

For dynamic scenes, integrating neural physics engines with 3DGS optimization could synthesize plausible object interactions without pre-specified motion data.

Conclusion

This survey has examined the rapidly evolving synergy between diffusion models and 3DGS, highlighting a paradigm shift in generative 3D scene representation. Diffusion priors effectively address critical limitations inherent in conventional 3DGS pipelines, including sensitivity to initialization, geometric inconsistencies under sparse inputs, artifacts in novel view synthesis, and the lack of semantic controllability. Methodologies integrating these two powerful paradigms have been systematically categorized, revealing distinct yet complementary strategies: diffusion-guided 3DGS optimization leverages 2D or multi-view diffusion models to regularize training and refine geometry with textures; direct latent generation of 3DGS representations utilizes diffusion models operating on compact functional or parameter spaces derived from disentangled Gaussian attributes; and diffusion for consistent novel view synthesis generates multi-view coherent imagery subsequently distilled into high-fidelity 3DGS models.

Key innovations such as Plücker embeddings for geometrically meaningful camera conditioning, variational 3DGS formulations mitigating local minima, structured 3D noise injection ensuring multi-view consistency, and hybrid distillation-rendering losses demonstrate significant advancements. These approaches yield substantial improvements in rendering quality, geometric fidelity, training stability, and generative flexibility across diverse tasks–text-to-3D, image-to-3D, sparse-view reconstruction, and dynamic scene synthesis–while often preserving the real-time rendering advantage of 3DGS. However, fundamental challenges persist, including computational overhead during diffusion-guided optimization, inherent tensions between 2D perceptual quality and strict 3D consistency, limitations in handling complex material properties and lighting, and scalability to large-scale or dynamic scenes.

Future research directions are poised to focus on improving physical plausibility via material-aware and physics-informed modeling, and enabling finer-grained semantic and dynamic control through multimodal conditioning. The convergence of efficient explicit scene representation with powerful generative priors marks a transformative step towards interactive, photorealistic, and controllable 3D content creation, promising significant impact across domains such as augmented reality, virtual production, robotics, and digital twins.

References

1. Blattmann, Andreas, Tim Dockhorn, Sumith Kulal, et al. 2023. *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets*. https://arxiv.org/abs/2311.15127.

2. Bulò, Samuel Rota, Lorenzo Porzi, and Peter Kontschieder. 2024. *Revising Densification in Gaussian Splatting*. https://arxiv.org/abs/2404.06109.

3. Deng, Xiaobin, Changyu Diao, Min Li, Ruohan Yu, and Duanqing Xu. 2025. *Efficient Density Control for 3D Gaussian Splatting*. https://arxiv.org/abs/2411.10133.

4. Grubert, Glenn, Florian Barthel, Anna Hilsmann, and Peter Eisert. 2025. "Improving Adaptive Density Control for 3D Gaussian Splatting." *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 610–21. https://doi.org/10.5220/0013308500003912.

5. Guo, Yuwei, Ceyuan Yang, Anyi Rao, et al. 2024. *AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models Without Specific Tuning*. https://arxiv.org/abs/2307.04725.

6. Guo, Yuwei, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. *SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models*. https://arxiv.org/abs/2311.16933.

7. Gupta, Anchit, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 2023. *3DGen: Triplane Latent Diffusion for Textured Mesh Generation*. https://arxiv.org/abs/2303.05371.

8. Haque, Ayaan, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. *Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions*. https://arxiv.org/abs/2303.12789.

9. He, Hao, Yinghao Xu, Yuwei Guo, et al. 2025. *CameraCtrl: Enabling Camera Control for Text-to-Video Generation*. https://arxiv.org/abs/2404.02101.

10. He, Xianglong, Junyi Chen, Sida Peng, et al. 2024. *GVGEN: Text-to-3D Generation with Volumetric Representation*. https://arxiv.org/abs/2403.12957.

11. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. *Denoising Diffusion Probabilistic Models*. https://arxiv.org/abs/2006.11239.

12. Kerbl, Bernhard, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. *3D Gaussian Splatting for Real-Time Radiance Field Rendering*. https://arxiv.org/abs/2308.04079.

13. Kheradmand, Shakiba, Daniel Rebain, Gopal Sharma, et al. 2025. *3D Gaussian Splatting as Markov Chain Monte Carlo*. https://arxiv.org/abs/2404.09591.

14. Li, Xinhai, Huaibin Wang, and Kuo-Kun Tseng. 2024. *GaussianDiffusion: 3D Gaussian Splatting for Denoising Diffusion Probabilistic Models with Structured Noise*. https://arxiv.org/abs/2311.11221.

15. Lin, Chenguo, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. 2025. *DiffSplat: Repurposing Image Diffusion Models for Scalable Gaussian Splat Generation*. https://arxiv.org/abs/2501.16764.

16. Liu, Xinhang, Jiaben Chen, Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. 2024. *Deceptive-NeRF/3DGS: Diffusion-Generated Pseudo-Observations for High-Quality Sparse-View Reconstruction*. https://arxiv.org/abs/2305.15171.

17. Liu, Xi, Chaoyi Zhou, and Siyu Huang. 2024. *3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-Consistent 2D Diffusion Priors*. https://arxiv.org/abs/2410.16266.

18. Liu, Yuan, Cheng Lin, Zijiao Zeng, et al. 2024. *SyncDreamer: Generating Multiview-Consistent Images from a Single-View Image*. https://arxiv.org/abs/2309.03453.

19. Meng, Chenlin, Yutong He, Yang Song, et al. 2022. *SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations*. https://arxiv.org/abs/2108.01073.

20. Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*. https://arxiv.org/abs/2003.08934.

21. Mithun, Niluthpol Chowdhury, Tuan Pham, Qiao Wang, et al. 2025. *Diffusion-Guided Gaussian Splatting for Large-Scale Unconstrained 3D Reconstruction and Novel View Synthesis*. https://arxiv.org/abs/2504.01960.

22. Niket Agarwal, et al. 2025. *Cosmos World Foundation Model Platform for Physical AI*. https://arxiv.org/abs/2501.03575.

23. Hassan Abu Alhaija, et al. 2025. *Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control*. https://arxiv.org/abs/2503.14492.

24. Peng, Chensheng, Ido Sobol, Masayoshi Tomizuka, Kurt Keutzer, Chenfeng Xu, and Or Litany. 2024. *A Lesson in Splats: Teacher-Guided Diffusion for 3D Gaussian Splats Generation with 2D Supervision*. https://arxiv.org/abs/2412.00623.

25. Poole, Ben, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. *DreamFusion: Text-to-3D Using 2D Diffusion*. https://arxiv.org/abs/2209.14988.

26. Ren, Xuanchi, Tianchang Shen, Jiahui Huang, et al. 2025. *GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control*. https://arxiv.org/abs/2503.03751.

27. Roessle, Barbara, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Niessner. 2023. "GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields." *ACM Transactions on Graphics* 42 (6): 1–14. https://doi.org/10.1145/3618402.

28. Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. *High-Resolution Image Synthesis with Latent Diffusion Models*. https://arxiv.org/abs/2112.10752.

29. Sauer, Axel, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. 2024. *Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation*. https://arxiv.org/abs/2403.12015.

30. Sauer, Axel, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. *Adversarial Diffusion Distillation*. https://arxiv.org/abs/2311.17042.

31. Shi, Yichun, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2024. *MVDream: Multi-View Diffusion for 3D Generation*. https://arxiv.org/abs/2308.16512.

32. Su, Jianlin, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. https://arxiv.org/abs/2104.09864.

33. Szymanowicz, Stanislaw, Eldar Insafutdinov, Chuanxia Zheng, et al. 2025. *Flash3D: Feed-Forward Generalisable 3D Scene Reconstruction from a Single Image*. https://arxiv.org/abs/2406.04343.

34. Szymanowicz, Stanislaw, Christian Rupprecht, and Andrea Vedaldi. 2024. *Splatter Image: Ultra-Fast Single-View 3D Reconstruction*. https://arxiv.org/abs/2312.13150.

35. Tang, Jiaxiang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2024. *DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation*. https://arxiv.org/abs/2309.16653.

36. Wang, Tengfei, Bo Zhang, Ting Zhang, et al. 2022. *Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion*. https://arxiv.org/abs/2212.06135.

37. Wang, Zhouxia, Ziyang Yuan, Xintao Wang, et al. 2024. *MotionCtrl: A Unified and Flexible Motion Controller for Video Generation*. https://arxiv.org/abs/2312.03641.

38. Wu, Jay Zhangjie, Yuxuan Zhang, Haithem Turki, et al. 2025. *Difix3D+: Improving 3D Reconstructions with Single-Step Diffusion Models*. https://arxiv.org/abs/2503.01774.

39. Yang, Yuanbo, Jiahao Shao, Xinyang Li, Yujun Shen, Andreas Geiger, and Yiyi Liao. 2025. *Prometheus: 3D-Aware Latent Diffusion Models for Feed-Forward Text-to-3D Scene Generation*. https://arxiv.org/abs/2412.21117.

40. Zhang, Bowen, Yiji Cheng, Jiaolong Yang, et al. 2024. *GaussianCube: A Structured and Explicit Radiance Representation for 3D Generative Modeling*. https://arxiv.org/abs/2403.19655.

41. Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. https://arxiv.org/abs/1801.03924.

42. Zhou, Jensen, Hang Gao, Vikram Voleti, et al. 2025. *Stable Virtual Camera: Generative View Synthesis with Diffusion Models*. https://arxiv.org/abs/2503.14489.

43. Zhou, Junsheng, Weiqi Zhang, and Yu-Shen Liu. 2024. *DiffGS: Functional Gaussian Splatting Diffusion*. https://arxiv.org/abs/2410.19657.

44. Zhou, Kun, Wenbo Li, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. 2023. *From NeRFLiX to NeRFLiX++: A General NeRF-Agnostic Restorer Paradigm*. https://arxiv.org/abs/2306.06388.

45. Zhou, Kun, Wenbo Li, Yi Wang, et al. 2023. *NeRFLiX: High-Quality Neural View Synthesis by Learning a Degradation-Driven Inter-Viewpoint MiXer*. https://arxiv.org/abs/2303.06919.