Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026359369

AI-Driven 3D/4D Content Generation, Editing, and Animation Using Diffusion Models and Gaussian Splatting: A Narrative Review

Yanning Liu*

360 Huntington Avenue, Northeastern University, Boston, MA, USA

**Corresponding author: Yanning Liu.*

Abstract

The convergence of diffusion models and 3D Gaussian Splatting (3D-GS) has catalyzed a paradigm shift in AI-driven 3D/4D content creation, enabling high-quality generation, editing, and animation from natural language instructions. This narrative review synthesizes representative works published between 2023 and 2025 across four interconnected sub-areas: text-to-3D/4D generation, 3D scene editing, dynamic scene animation, and motion and video generation. Through comparative analysis of representation types, supervision strategies, and temporal modeling approaches, we identify convergent design principles—coarse-to-fine optimization in distillation-based methods, explicit representations, regularization against score-distillation variance, and decomposition for controllability—that recur across sub-areas. We catalog six critical limitations and propose concrete directions for improvement for each. This review serves as a foundational reference for researchers entering this rapidly evolving field.

Keywords

3D Gaussian Splatting, diffusion models, text-to-3D generation, 4D generation, 3D editing, neural rendering, score distillation sampling, motion generation, narrative review

1. Introduction

The creation of three-dimensional (3D) and four-dimensional (4D, i.e., dynamic 3D) digital content has traditionally demanded specialized expertise in 3D modeling, animation, and computer graphics, constraining broader adoption across entertainment, gaming, virtual reality, and simulation. Despite decades of advances in graphics tooling, the fundamental bottleneck persists: creating high-quality 3D content requires skilled artists investing hours to days per asset.

The recent convergence of two transformative technologies has begun to alter this landscape fundamentally. Diffusion models [1, 2]—originally developed for 2D image synthesis—have demonstrated generative capabilities extending far beyond their initial scope. The critical breakthrough came with Score Distillation Sampling (SDS) [3], which enables pre-trained 2D diffusion models to serve as differentiable critics for optimizing 3D representations. This mechanism transfers rich visual priors learned from billions of web images

into the 3D domain without requiring large-scale 3D training datasets, spawning a rapidly expanding ecosystem of text-to-3D generation, 3D editing, and 4D generation methods. In parallel, 3D Gaussian Splatting (3D-GS) [4] has emerged as a paradigm-shifting technique for explicit, real-time radiance field rendering. By representing scenes as discrete 3D Gaussian primitives with learnable parameters, 3D-GS addresses two fundamental limitations of Neural Radiance Fields (NeRF) [5]: slow rendering due to extensive neural network evaluation, and the inability to manipulate scene elements entangled in implicit MLP weights. The explicit nature of 3D-GS enables real-time differentiable rendering, direct manipulation of individual primitives, and seamless composition of multiple scenes—properties that align naturally with the requirements of generative and editing pipelines.

The intersection of these technologies has triggered explosive growth in research output, but this rapid expansion has created a fragmented landscape of the literature. Differing terminologies obscure methodological connections across sub-areas. Systematic evaluation standards remain undeveloped. Critical limitations are identified in isolation rather than synthesized into a coherent understanding of the field's bottlenecks, making it difficult for newcomers to discern which methodological choices are fundamental.

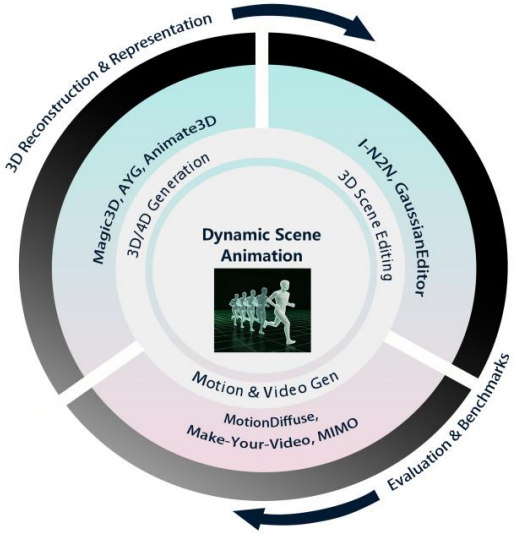
This review addresses these challenges by providing a critical, integrative perspective on AI-driven 3D/4D content creation. Rather than isolated paper summaries, we conduct thematic comparative analysis examining how methods address shared challenges in representation, supervision, and temporal modeling. Through cross-cutting synthesis, we identify convergent design principles and evolutionary trajectories. Building on a systematic cataloging of limitations from the reviewed works, we identify six critical research gaps and propose concrete, actionable directions for each.

1.1 Review Methodology

We searched arXiv, CVF Open Access, and IEEE Xplore (January 2023–March 2025) for works addressing 3D/4D content creation via diffusion model supervision or 3D-GS representations. Papers were included if they appeared in a top-tier venue (CVPR, ICCV, NeurIPS, ECCV, ACM TOG, IEEE TPAMI/TVCG) or had significant citation impact on arXiv, and represented a distinct methodological approach within their sub-area. From approximately 180 candidates, we selected 10 primary works spanning the four sub-areas, emphasizing methodological diversity over exhaustive coverage. Key omitted works are discussed in Section 4.

Figure 1 presents a conceptual framework illustrating the layered dependency structure among the four sub-areas. At the foundation, 3D reconstruction and representation provide the building blocks for all downstream tasks. Generation methods leverage 2D diffusion priors to create content from scratch; editing methods apply similar guidance to modify existing scenes; motion and video generation provide motion signals that drive character animation; and animation methods bridge generation and reconstruction by adding dynamics to static captures. Evaluation infrastructure provides cross-cutting quality assessment across all layers.

Figure 1: Conceptual framework illustrating the layered dependency structure among the four surveyed sub-areas



2. Background and Foundations

2.1 Diffusion Models and Score Distillation Sampling

Diffusion models [1, 2] learn to reverse a gradual noising process: given a data sample $\mathbf{x}_0$, a forward process adds Gaussian noise over T timesteps via $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{(1-\beta_t)}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I})$, and the reverse process learns to denoise via $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t))$. Text-conditioned variants such as Stable Diffusion [18] and Imagen [19] augment this with language encoders trained on billions of image-text pairs.

Score Distillation Sampling (SDS) [3], introduced by DreamFusion, uses a pre-trained diffusion model as a differentiable loss function for 3D optimization. At each step, a random viewpoint renders the 3D representation, noise is added, and the model provides a gradient:

$$\nabla\theta LSDS = Et, \varepsilon[w(t)(\hat{\varepsilon}\varphi(xt; y, t) - \varepsilon)(\partial x)/(\partial\theta)]$$

where $\hat{\varepsilon}_\varphi$ is the predicted noise and $\partial x/\partial\theta$ backpropagates 2D signals into 3D parameters. Classifier Score Distillation (CSD) [20], used by AYG [7], reformulates SDS from a classifier perspective to improve gradient stability. Variational Score Distillation (VSD) [21] (ProlificDreamer) replaces single-point optimization with particle-based variational inference using a LoRA-fine-tuned diffusion model, achieving higher fidelity at greater computational cost. Consensus on which paradigm is preferable remains unresolved, as the fidelity-cost trade-off varies substantially across tasks, and no systematic comparison has been conducted.

2.2 Neural Radiance Fields (NeRF)

NeRF [5] encodes a 3D scene as a continuous volumetric function via an MLP: $F_\Theta: (x, d) \rightarrow (c, \sigma)$, mapping 3D position and viewing direction to color and density. Novel views are synthesized through differentiable volume rendering:

$$C(r) = \sum i = 1NTi(1 - \exp(-\sigma i\delta i))ci, Ti = \exp(-\sum j = 1i - 1\sigma j\delta j)$$

Despite photorealistic novel view synthesis, NeRF is limited by slow rendering (hundreds of MLP evaluations per ray) and entangled implicit representations that prevent direct manipulation of individual objects.

2.3 3D Gaussian Splatting (3D-GS)

3D Gaussian Splatting [4] addresses NeRF's limitations by representing scenes as explicit collections of 3D Gaussian primitives:

$$G(x) = \exp(-(1)/(2)(x - \mu)T\Sigma - 1(x - \mu))$$

Each Gaussian is parameterized by position μ, covariance Σ (decomposed into scaling and rotation), opacity α, and spherical harmonic coefficients for view-dependent color. Differentiable rasterization—projecting 3D Gaussians to 2D and alpha-blending—renders orders of magnitude faster than NeRF's volume rendering. This explicit representation enables real-time rendering, direct manipulation of individual primitives, and trivial composition of independently generated scenes. For temporal content, the dominant paradigm pairs a canonical 3D-GS set with a deformation field $\Phi: (x, t) \rightarrow \Delta x$ that predicts per-Gaussian displacements.

3. Literature Review

This section provides a comparative analysis organized by sub-area, examining how different methods approach shared challenges rather than presenting each paper in isolation.

3.1 Text-to-3D and Text-to-4D Generation

Three representative distillation-based works—Magic3D [6], Align Your Gaussians (AYG) [7], and Animate3D [8]—trace the evolution from static 3D objects to dynamic 4D content, and from pure distillation to data-driven approaches. ProlificDreamer [21] represents the primary VSD-based alternative, while DreamGaussian [22] and LGM [23] exemplify the emerging feed-forward generation paradigm.

Representation. Magic3D [6] employs a hybrid strategy: a hash-grid-encoded NeRF (Instant NGP) in the coarse stage, followed by conversion to a textured mesh (DMTet) for fine-stage refinement—mesh-based SDS optimization outperforming NeRF-based refinement from the same initialization. AYG [7] and Animate3D [8] both adopted 3D Gaussian Splatting with the canonical-GS-plus-deformation-field paradigm, consistent with the field's transition toward explicit representations for real-time tasks. ProlificDreamer [21], by contrast, demonstrates that implicit NeRF representations under VSD can achieve state-of-the-art fidelity, indicating that the choice of representation and the supervision paradigm are coupled. Feed-forward methods DreamGaussian [22] and LGM [23] generate 3D-GS representations in a single forward pass, bypassing per-instance optimization entirely.

Supervision Strategy. The reviewed methods span three paradigms. Magic3D [6] represents pure SDS distillation. Both stages use pre-trained 2D diffusion models (eDiff-I for coarse geometry and Stable Diffusion in latent space for fine texture) as SDS critics, requiring no 3D training data. AYG [7] extends this compositionally by combining gradients from three distinct diffusion model types—MVDream [24] for multi-view consistency, Stable Diffusion for appearance, and a custom text-to-video model for dynamics—within a unified CSD framework. ProlificDreamer [21] introduced VSD: particle-based variational inference with a LoRA-fine-tuned diffusion model, achieving higher fidelity than SDS at greater cost. Animate3D [8] represents the data-driven alternative, training a Multi-View Video Diffusion Model (MV-VDM) from scratch on 1.85 million multi-view videos rendered from 53,340 animated 3D models. This trajectory—zero-shot distillation (SDS/CSD) → enhanced distillation (VSD) → data-driven training—mirrors 2D image generation history and suggests trained generative models will complement distillation-based approaches as 3D/4D datasets grow.

Temporal Modeling. Both AYG and Animate3D employ deformation-based temporal modeling. AYG [7] uses a compact MLP deformation field trained via CSD gradients from the combined diffusion ensemble, generating motion through zero-shot composition. Animate3D [8] uses a HexPlane-structured motion field with L2 reconstruction followed by 4D-SDS refinement, producing higher-fidelity motion constrained by its training distribution.

Limitations. Magic3D is limited to static 3D (40 min / 8 A100 GPUs). AYG cannot handle topological changes and is computationally intensive. Animate3D's quality is bound by its MV-VDM, with synthetic-to-real domain gaps. ProlificDreamer achieves higher quality at substantially greater cost. DreamGaussian and LGM address speed through feed-forward prediction, but trail iterative methods on complex prompts. None of the distillation-based methods operates beyond the object-centric regime.

3.2 3D Scene Editing

The transition from NeRF-based to GS-based editing, exemplified by Instruct-NeRF2NeRF [9] and GaussianEditor [10], constitutes one of the clearest paradigm shifts documented in this review.

The NeRF Editing Baseline. Instruct-NeRF2NeRF [9] introduced language-driven editing of real-captured NeRF scenes via an Iterative Dataset Update scheme that alternates between InstructPix2Pix-based 2D editing and NeRF optimization. While groundbreaking, it inherited NeRF's constraints: ~60 minutes per edit on a high-end GPU, no mechanism for spatially confining edits to target regions, and dependence on pre-existing reconstruction quality. Edits could “bleed” into non-target areas, and large-scale spatial manipulations proved unreliable.

The Gaussian Splatting Turn. GaussianEditor [10] migrated the same conceptual approach—2D diffusion guidance for 3D editing—to a 3D-GS backbone, reducing editing time to 5–10 minutes on comparable hardware. Two key innovations enabled spatial controllability: Gaussian Semantic Tracing unprojects SAM-generated 2D masks into 3D to enable precise, region-specific gradient application, and Hierarchical Gaussian Splatting (HGS) organizes Gaussians by densification age, progressively freezing older generations. In contrast, newer ones remain optimizable—preventing the excessive fluidity that causes vanilla GS to diverge under stochastic SDS guidance. User studies confirmed the magnitude of the improvement: 72.3% preferred GaussianEditor, compared with 15.5% for Instruct-NeRF2NeRF. The improvement cannot be attributed solely to the change in representation, since HGS and semantic tracing co-occurred with it; controlled ablations remain a desideratum.

Broader Significance. This pair demonstrates that identical supervision (InstructPix2Pix-guided iterative updating) performs dramatically better with an explicit representation and appropriate regularization, enabling spatially controlled, interactive-rate 3D editing inconceivable under the implicit paradigm.

3.3 Dynamic Scene Animation

SC-GS [11] and Gaussians-to-Life [12] represent complementary approaches to animating 3D scenes, differentiated primarily by motion source: reconstruction from observation versus generation from text.

Shared Paradigm. Both employ control-point-driven deformation on 3D-GS representations. SC-GS [11] uses ~512 sparse control points with learned time-varying 6-DoF transformations (predicted by an MLP) driving ~100K dense Gaussians through Linear Blend Skinning (LBS) with Gaussian-kernel radial basis function interpolation. Gaussians-to-Life [12] tracks sparse anchor points across frames via CoTracker, lifts 2D trajectories to 3D through monocular depth estimation (UniDepth), and propagates deformation via weighted interpolation or Kabsch rigid estimation. Both employ As-Rigid-As-Possible (ARAP) regularization to encourage locally rigid deformation.

Divergent Motion Sources. SC-GS [11] reconstructs motion from observed multi-view video, achieving high accuracy (PSNR 43.31 on D-NeRF) and uniquely enabling post-reconstruction motion editing by manipulating the learned control graph, but requiring multi-view capture and per-scene optimization. Gaussians-to-Life [12] generates novel motion from text via DynamiCrafter with multi-step denoising and multi-view consistency constraints, enabling animation of static captures without video input—the user specifies “the wind blows through the trees” and provides a bounding box—but cannot add or remove Gaussians, leaving holes where moving objects vacate their original positions. Deformable 3D Gaussians [27] have made partial progress toward this topological limitation through temporally adaptive density control, though large-scale structural changes remain unresolved.

Complementarity. SC-GS excels at high-fidelity reconstruction and editing of observed motion, while Gaussians-to-Life enables creative generation of novel motion from existing captures. A future integration—using generated motion trajectories to initialize control points, then refining them through reconstruction-style optimization—could combine the creative flexibility of text-driven generation with the fidelity of observation-driven reconstruction.

3.4 Enabling Technologies from 2D Domains

Several works in the 2D video and motion domains provide techniques that are transferable to 3D/4D systems. MotionDiffuse [13] applies diffusion models to text-driven human motion generation using skeletal parameters, demonstrating probabilistic motion mapping and fine-grained body-part control for 3D character animation. Make-Your-Video [14] introduces two-stage video training—freezing spatial modules while training lightweight temporal modules with causal attention—a technique adopted in spirit by Gaussians-to-Life's multi-step denoising. MIMO [15] contributes spatially decomposed character synthesis, separating frames into hierarchical layers (human, scene, occlusion) for composable diffusion-based decoding—mirroring SC-GS's motion/appearance split. The common principle—that decomposition enables controllability—recurs throughout the 3D/4D literature. While these methods are not 3D-GS-based and their direct integration into 3D/4D pipelines remains largely aspirational, their design patterns offer transferable architectural insights.

3.5 Evaluation Infrastructure and Foundational Surveys

GPTEval3D [17] leverages GPT-4V as a human-aligned evaluator for text-to-3D generation, achieving strong correlation with expert judgment (Kendall's $\tau = 0.710$) through pairwise comparisons of multi-view renderings. While limited by reliance on a proprietary model (risking benchmark instability if the model is deprecated), $O(n^2)$ pairwise scaling, and restriction to static 3D, it represents the most significant step toward automated 3D evaluation and motivates open-source alternatives. The 3D Gaussian Splatting survey by Fei et al. [16] organizes over 130 papers into six categories, confirming the rapid displacement of NeRF by 3D-GS across generation, editing, reconstruction, and perception tasks.

Table 1 provides a consolidated comparison of all reviewed methods across key design dimensions.

Table 1: Comparative overview of reviewed methods across key design dimensions. "Compute" reports per-instance cost excluding pre-training; GPU architectures vary across rows—direct cost comparison should consider FLOP-equivalent normalization. "Metrics" lists representative quantitative results where reported (N/R = Not Reported). DM = Diffusion Model; SDS = Score Distillation Sampling; CSD = Classifier Score Distillation; VSD = Variational Score Distillation; LBS = Linear Blend Skinning; ARAP = As-Rigid-As-Possible

Method	Representation	Diffusion Prior(s)	Supervision	Temporal Modeling	Reported Metrics	Key Limitation	Compute (per instance)
DreamFusion [3]	NeRF (MLP)	Imagen	SDS	None (static 3D)	User study (preference)	Janus problem; slow	~1.5 hr / 4× TPUv4
Magic3D [6]	Coarse: Hash-grid NeRF; Fine: DMTet mesh	eDiff-I (coarse) + Stable Diffusion latent (fine)	SDS (both stages)	None (static 3D)	User study (preference)	Static only; 40 min	40 min / 8× A100
ProlificDreamer [21]	NeRF (MLP)	Stable Diffusion + LoRA	VSD (particle-based VI)	None (static 3D)	CLIP score, user study	High compute cost	~8 hr / 4× A100
DreamGaussian [22]	3D-GS + mesh refinement	Stable Diffusion + Zero-1-to-3	SDS (single-stage + UV refinement)	None (static 3D)	CLIP score, user study	Quality < iterative methods	~2 min / 1× GPU
AYG [7]	Canonical 3D-GS + MLP deformation field	MVDream + Stable Diffusion + custom T2V DM	CSD (compositional)	Continuous MLP, autoregressive extension	User study (qualitative)	No topological changes	Multi-stage + custom DM training
Animate3D [8]	Canonical 3D-GS + HexPlane motion field	MV-VDM (trained on MV-Video dataset)	L2 recon + 4D-SDS	HexPlane, ARAP regularization	N/R	Synthetic→real domain gap	~40 min / 1× A800 (+ 10 days MV-VDM on 32× A800)
Instruct-NeRF2NeRF [9]	NeRF (MLP)	InstructPix2Pix	Iterative dataset update (2D edit → NeRF optimize)	None (per-frame independent)	User study (preference)	Edit bleeding; 60 min/edit	~60 min / 1× Titan RTX
GaussianEditor [10]	3D-GS + HGS generations	InstructPix2Pix / DDS	Iterative update + DDS	None (per-frame independent)	72.3% user preference vs. 15.5% I-N2N	HGS may lock early errors	5–10 min / 1× RTX A6000
SC-GS [11]	3D-GS + ~512 sparse control points (LBS)	None (observation-driven)	Photometric loss (L1 + D-SSIM)	Control-point MLP + LBS + ARAP	PSNR 43.31 (D-NeRF avg.)	Requires multi-view capture	Per-scene optimization (GPU hours)
Gaussians-to-Life [12]	3D-GS + anchor trajectories	DynamiCrafter (multi-step denoising)	Depth-guided 2D→3D lifting	Anchor interpolation + Kabsch rigid estimation	User study (qualitative)	Cannot add/remove Gaussians	~10 min (optimization-free)
MotionDiffuse [13]	Skeletal motion parameters	T2M diffusion (DDPM)	Diffusion denoising loss	Cross-Modality Linear Transformer	FID, Diversity, Accuracy (HumanML 3D)	No physical plausibility	Standard diffusion training; DDIM inference
Make-Your-Video [14]	Video pixels (latent space)	Stable Diffusion Depth +	Diffusion denoising loss	Pseudo-3D conv + Causal	FVD, CLIPSIM	Structure-conditioned only	Training on WebVid-10M

Method	Representation	Diffusion Prior(s)	Supervision	Temporal Modeling	Reported Metrics	Key Limitation	Compute (per instance)
		temporal modules		Attention Mask	(WebVid-10M)		
MIMO [15]	Spatial layers (human/scene/occlusion)	SD 1.5 + AnimateDiff	Diffusion denoising loss	SMPL-anchored motion codes + temporal attention	FVD, user study	Single-human only	Training on HUD-7K

4. Cross-Cutting Synthesis: Convergent Design Principles and Evolutionary Trajectories

Examining these works collectively reveals deeper patterns that recur across sub-areas. We identify four convergent design principles, then discuss their boundary conditions.

Several principles recur within distillation-based methods. Coarse-to-fine optimization is near-universal under SDS/CSD: Magic3D pioneered the two-stage pipeline (hash-grid NeRF → textured mesh), AYG extended it to 4D (static 3D-GS → deformed dynamic 3D-GS), and Animate3D adapted it for animation (reconstruction → 4D-SDS refinement). Under SDS's high-variance gradients, single-stage optimization from random initialization is prone to instability; the coarse stage provides a well-behaved basin of attraction. Explicit representations dominate for interactive tasks: Magic3D showed mesh refinement outperforms NeRF refinement; AYG and Animate3D independently chose 3D-GS; SC-GS selected GS over 4D NeRF; GaussianEditor migrated from NeRF to GS—in each case, real-time rendering and direct manipulation proved decisive. However, ProlificDreamer [21] demonstrates that implicit representations under VSD achieve state-of-the-art fidelity, and NeRF variants remain competitive on reconstruction benchmarks. Regularization against noisy gradients is a shared necessity under distillation: Magic3D constrains its environment map MLP; AYG uses JSD-based regularization; Animate3D and SC-GS use ARAP losses; GaussianEditor freezes older generations via HGS. The diversity of strategies—statistical, geometric, hierarchical—suggests that the choice of regularization may be as important as the choice of representation. Feed-forward methods (DreamGaussian, LGM) avoid this issue entirely. Decomposition enables controllability: separating motion from appearance (SC-GS), animated from static regions (Gaussians-to-Life), or foreground from background (MIMO) transforms entangled generation into composable sub-problems. This principle carries genuine weight where high-dimensional spatiotemporal content makes entangled optimization particularly brittle.

Boundary Conditions. These principles are conditional generalizations, not universal laws. Coarse-to-fine optimization is specific to distillation-based methods; feed-forward methods (DreamGaussian, LGM) generate in a single pass, and ProlificDreamer's VSD uses particle-based inference rather than noise-level annealing. Explicit representation dominance holds for interactive tasks but is contested where reconstruction fidelity is paramount. Regularization against gradient variance is irrelevant to feed-forward and observation-driven methods. Decomposition for controllability, while useful, is an instantiation of divide-and-conquer; its value lies in the specific choices of decomposition tailored to 3D/4D content.

The field's evolutionary trajectory is equally revealing. The most decisive shift is the transition from implicit to explicit representations for interactive tasks: the transition from Instruct-NeRF2NeRF to GaussianEditor spans roughly six months, and the Fei et al. survey of 130+ GS papers confirms Gaussian representations as the dominant interactive substrate. However, implicit and hybrid representations retain advantages for reconstruction fidelity. A second trajectory leads from score distillation to data-driven and feed-forward generation: Magic3D (pure SDS) → ProlificDreamer (enhanced VSD) → Animate3D (data-driven) → DreamGaussian/LGM (feed-forward, reducing generation from hours to minutes). A third trajectory traces the progression of capability from reconstruction to generation to editing toward composition within a unified 3D-GS representation. A fourth, nascent trajectory leads from object-centric to scene-level generation. Every distillation-based generation paper operates on a single foreground object. Still, animation and video methods have begun to incorporate scene context, and methods such as LucidDreamer [25] have made initial progress in scene-level synthesis.

A conspicuous absence across all reviewed works is standardized evaluation. Most generation papers rely on user studies and qualitative visualization; the most comprehensive survey [16] contains no evaluation section. GPTEval3D addresses text-to-3D but is limited to static content and a proprietary model. Well-known failure modes—the Janus problem (multi-face artifacts), multi-view inconsistency, and floating artifacts—are rarely reported systematically, further complicating cross-method assessment.

5. Research Gaps and Proposed Improvements

Based on the systematic analysis of limitations across the reviewed works, we identify six critical research gaps with evidence synthesized from multiple sources.

Absence of Standardized 3D/4D Evaluation Metrics. Magic3D [6], AYG [7], Animate3D [8], SC-GS [11], and Gaussians-to-Life [12] all rely primarily on user studies and qualitative visualizations. The Fei et al. survey [16] omits evaluation entirely. GPTEval3D [17] addresses text-to-3D evaluation but uses a proprietary model that scales comparisons quadratically. Develop a multi-dimensional automated evaluation framework combining geometric metrics (Chamfer distance, normal consistency), appearance metrics (multi-view FID, LPIPS), motion metrics for 4D (temporal consistency, physical plausibility checks), alignment metrics using open-source VLMs, and efficiency metrics, producing a unified leaderboard analogous to VBench adapted for 3D.

Inability to Handle Topological Changes. AYG [7] “cannot easily produce topological changes”; Gaussians-to-Life [12] “cannot add or remove 3D Gaussians.” The deformation-based paradigm dominating 4D generation is fundamentally topology-preserving. Deformable 3D Gaussians [27] have made initial progress through temporal adaptive density control, but large-scale structural changes—object merging, splitting, or appearing from occlusion—remain unsolved. Develop topology-aware 4D representations through adaptive primitive birth and death in the temporal dimension, hybrid mesh-primitive approaches for large-scale structural changes, and discrete event modeling for topological transitions.

Limited Generalization Capability. Current methods exhibit restricted generalization along two axes. First, generation remains object-centric: Magic3D [6], AYG [7], and Animate3D [8] generate single foreground objects; scene-level methods such as LucidDreamer [25] have made initial progress but do not yet match the quality of object-centric methods. Second, methods trained on synthetic data suffer domain gaps documented by Animate3D [8] and Gaussians-to-Life [12]; the Fei et al. survey [16] identifies few-shot robustness as an open challenge. improvement: Advance scene-level generation through compositional approaches with spatial layout control, and implement domain adaptation via domain randomization, test-time self-supervised adaptation, and multi-domain benchmarks.

Prohibitive Computational Costs. Magic3D [6] requires 40 minutes on 8 A100 GPUs; Animate3D's MV-VDM training consumes 10 days on 32 A800 GPUs; ProlificDreamer [21] requires ~8 hours on 4 A100S; Instruct-NeRF2NeRF [9] takes ~60 minutes per edit. Feed-forward methods (DreamGaussian [22], LGM [23]) have demonstrated viable amortized inference, reducing generation times to minutes, though quality currently trails that of iterative methods. Scale feed-forward generation through larger datasets and improved architectures; develop consistency distillation for few-step inference; and pursue hardware-aware optimization targeting consumer GPU constraints (≤24 GB VRAM).

Limited Motion Controllability. All text-driven methods accept only text prompts. MotionDiffuse [13] provides skeletal-level body-part control; MIMO [15] accepts pose sequences. No method provides fine-grained control over speed, trajectory, amplitude, or style. Beyond the technical input-modality gap, key HCI requirements include: the gulf of evaluation (users cannot efficiently verify motion intent across viewpoints), the absence of iterative refinement, and the need for low-fidelity real-time preview to support creative flow. Proposed improvement: Build hybrid control interfaces combining multi-modal conditioning (text, sketch, example videos, physical parameters) with real-time preview, learnable motion primitives for composable parameterization, and selective re-generation for specific spatial or temporal segments.

Absence of Physical Plausibility Constraints. No reviewed work explicitly models physics beyond ARAP rigidity regularization. PhysGaussian [26] has demonstrated GS kernels as continuum-mechanics particles for simulation, but this capability has not been integrated into generative pipelines. Two complementary strategies: penalty-based methods (soft physics loss terms—collision penalties, balance constraints) for interactive

applications, and fully differentiable simulation (as in PhysGaussian [26]) for offline scenarios demanding physical accuracy. Medium-term goals include learned physics priors for plausible corrections without online simulation costs.

These six gaps form a partially ordered dependency graph. Gap 4 (computational cost) is the primary bottleneck: the prohibitive per-instance optimization expense directly constrains progress on Gap 3 (generalization, since scene-level and cross-domain experiments become economically infeasible) and Gap 6 (physical plausibility, since differentiable simulation adds further computational burden). Gap 1 (evaluation) is the foundational enabler—without standardized metrics, progress on all other gaps is difficult to measure and compare. Gap 5 (limited controllability) and Gap 2 (topological changes) are largely orthogonal but intersect where user-specified topological changes require new control interfaces. This interdependency structure—presented as a hypothesis warranting empirical validation through practitioner surveys or publication-pattern analysis—implies that investment in evaluation infrastructure (Gap 1) and computational efficiency (Gap 4) would have the greatest multiplicative impact, thereby lowering barriers across the entire research frontier.

6. Conclusion

The technologies surveyed carry significant societal implications: text-to-3D/4D combined with character animation could enable photorealistic synthetic media of real individuals; the computational requirements documented in Gap 4 carry non-trivial energy costs; the framing of artist expertise as a "bottleneck" warrants examination given potential labor displacement; and generated content inherits biases from 2D diffusion backbones trained predominantly on Western imagery. We encourage the research community to engage with these concerns alongside technical advances.

This review has examined the rapidly evolving landscape of AI-driven 3D/4D content creation at the intersection of diffusion models and 3D Gaussian Splatting. Through comparative analysis across text-to-3D/4D generation, scene editing, dynamic animation, and motion generation, we have identified convergent design principles—coarse-to-fine optimization within distillation-based methods, explicit representations for interactive tasks, regularization against high-variance score distillation gradients, and decomposition for controllability—each qualified by boundary conditions where counter-examples (feed-forward methods, VSD approaches, implicit representations for reconstruction) demonstrate these principles are conditional rather than absolute. The field's evolutionary trajectory—from implicit to explicit representations, from distillation through enhanced distillation to data-driven and feed-forward generation, and from object-centric to nascent scene-level synthesis—is clear and accelerating.

Six interconnected challenges define the research frontier: the absence of standardized evaluation, inability to handle topological change, limited generalization, prohibitive computational costs, limited motion controllability (encompassing both technical and human-centered dimensions), and the lack of physical plausibility constraints. Addressing these challenges requires cross-pollination across the surveyed sub-areas—evaluation frameworks extended to 4D, control representations adapted for generation, physics simulation integrated with video diffusion, and large-scale 4D datasets enabling the transition from distillation to data-driven and feed-forward methods. At present, these methods are best suited for rapid prototyping of single-object assets; generalization to complex scenes, integration with production pipelines, and validation in non-entertainment domains remain open challenges whose resolution has implications that extend well beyond computer vision and graphics.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[3] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," in *Proc. ICLR*, 2023.

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. ECCV*, 2020.

[6] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3D: High-resolution text-to-3D content creation," in *Proc. IEEE/CVF CVPR*, 2023, pp. 300–310.

[7] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis, "Align your Gaussians: Text-to-4D with dynamic 3D Gaussians and composed diffusion models," in *Proc. IEEE/CVF CVPR*, 2024, pp. 8576–8588.

[8] Y. Jiang, C. Yu, C. Cao, F. Wang, W. Hu, and J. Gao, "Animate3D: Animating any 3D model with multi-view video diffusion," arXiv:2407.11398v2, 2024.

[9] A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-NeRF2NeRF: Editing 3D scenes with instructions," in *Proc. IEEE/CVF ICCV*, 2023, pp. 19683–19693.

[10] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin, "GaussianEditor: Swift and controllable 3D editing with Gaussian splatting," in *Proc. IEEE/CVF CVPR*, 2024, pp. 21476–21485.

[11] Y.-H. Huang, Y.-T. Sun, Z. Yang, X. Lyu, Y.-P. Cao, and X. Qi, "SC-GS: Sparse-controlled Gaussian splatting for editable dynamic scenes," in *Proc. IEEE/CVF CVPR*, 2024, pp. 4220–4230.

[12] T. Wimmer, M. Oechsle, M. Niemeyer, and F. Tombari, "Gaussians-to-Life: Text-driven animation of 3D Gaussian splatting scenes," arXiv:2411.19233v2, Mar. 2025.

[13] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, Jun. 2024.

[14] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, Y. Shan, and T.-T. Wong, "Make-Your-Video: Customized video generation using textual and structural guidance," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 2, pp. 1526–1541, Feb. 2025.

[15] Y. Men, Y. Yao, M. Cui, and L. Bo, "MIMO: Controllable character video synthesis with spatial decomposed modeling," in *Proc. IEEE/CVF CVPR*, 2025, pp. 21181–21191.

[16] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3D Gaussian splatting as a new era: A survey," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 8, pp. 4429–4449, 2024.

[17] T. Wu, G. Yang, Z. Li, K. Zhang, Z. Liu, L. Guibas, D. Lin, and G. Wetzstein, "GPT-4V(ision) is a human-aligned evaluator for text-to-3D generation," in *Proc. IEEE/CVF CVPR*, 2024, pp. 22227–22238.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR*, 2022.

[19] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. NeurIPS*, 2022.

[20] Y. Huang, J. Wang, A. Zeng, H. Cao, X. Qi, Y. Shi, Z.-J. Zha, and L. Zhang, "DreamWaltz: Make a scene with complex 3D animatable avatars," in *Proc. NeurIPS*, 2023.

[21] Z. Wang, C. Lu, Y. Wang, F. Bao, Z. Li, H. Su, and J. Zhu, "ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation," in *Proc. NeurIPS*, 2023.

[22] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "DreamGaussian: Generative Gaussian splatting for efficient 3D content creation," in *Proc. IEEE/CVF CVPR*, 2024, pp. 872–882.

[23] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "LGM: Large multi-view Gaussian model for high-resolution 3D content creation," in *Proc. ECCV*, 2024.

[24] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "MVDream: Multi-view diffusion for 3D generation," in *Proc. ICLR*, 2024.

[25] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, "LucidDreamer: Domain-free generation of 3D Gaussian splatting scenes," arXiv:2311.13384, 2024.

[26] T. Xie, Z. Zong, Y. Qiu, X. Li, Y. Feng, Y. Yang, and C. Jiang, "PhysGaussian: Physics-integrated 3D Gaussians for generative content creation," in *Proc. IEEE/CVF CVPR*, 2024, pp. 6766–6775.

[27] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction," in *Proc. IEEE/CVF CVPR*, 2024, pp. 20341–20351.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access