http://dx.doi.org/10.62762/NGCST.2025.477710

REVIEW ARTICLE

A Comprehensive Review of Diffusion Models, Gaussian Splatting and Their Integration in Augmented and Virtual Reality

Santosh Kumar Kar[1,*], **B. Ujalesh Subudhi**[1], **Brojo Kishore Mishra**[1] and **Bandhan Panda**[1]

[1] Department of Computer Science & Engineering, NIST University, Berhampur 761008, India

Abstract

The new progress in text-to-3D technology has greatly changed and improved the artificial intelligence (AI) applications in augmented and virtual reality (AR/VR) environments. Many different techniques in 2024-2025 like diffusion models, Gaussian splatting, and physics aware models have helped the text-to-3D much better by improving the visual fidelity, semantic coherence, and generation efficiency. Some models like Turbo3D, Dive3D and Instant3D are deigned to make the 3D generation faster by improving the working process of diffusion models. Other frameworks like LAYOUTDREAMER, PhiP-G and CompGS focus on creating scenes that are well organized and structured. Dream Reward and Coheren Dream methods use the feedback from the humans and information from multiple types of data ton improve the 3D results that will match with the expectation of the people. There are some major challenges still remain even with all these improvements. These can be current text-to-3D methods need a lot of computing power which makes it difficult to employ at large scale or in real time AR/VR applications. Other problems like multi-view inconsistencies and absence of any standard benchmark makes it very difficult to compare the methods fairly. Without combining text, physics, and spatial logic the 3D scenes look less real and difficult to achieve natural interactions with the objects. This review explains and examines the latest advancements in text-to-3D generation. It closely looks at how these methods are designed, optimized and customized for different areas of applications. The review points out probable future research ideas like creation of faster and smaller 3D generation methods, renderings that will understand the real world physics and include the human help to guide the model as per the requirements in the process and use common standards to get fairness in the evaluation of the model. The study bows to explain the current progress, innovative ideas and the challenges faced by the artificial intelligence (AI) in creating AR/VR 3D contents.

Keywords: text-to-3D generation, diffusion models, gaussian splatting, augmented and virtual reality (AR/VR), human-in-the-loop optimization.

Submitted: 05 September 2025
Accepted: 15 December 2025
Published: 22 December 2025

Vol. 1, **No.** 2, 2025.
10.62762/NGCST.2025.477710

***Corresponding author:**
✉Santosh Kumar Kar
santoshkumarkar@nist.edu

Citation
Kar, S. K., Subudhi, B. U., Mishra, B. K., & Panda, B. (2025). A Comprehensive Review of Diffusion Models, Gaussian Splatting and Their Integration in Augmented and Virtual Reality. *Next-Generation Computing Systems and Technologies*, 1(2), 102–112.

© 2025 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

The epoch of 3D content generation is greatly influenced by the significant growth of artificial intelligence (AI). It has facilitated generation of three-dimensional assets based on textual representation which are realistic, semantically correct and physically consistent. Text-to-3D generation is a growing field in generative AI that is utilized in AR/VR systems more and more. Scholars have refined text-to-3D technology with such strategies as progressive optimization and multi-view consistency enforcement [1, 2]. Such methods apply 2D diffusion models in order to assist in generation of 3D scenes which has played a significant role in this area. The previous techniques of the architectures receiving the text as input and transforming it directly into the 3D shapes have been developed into new and potent architectures. They are using diffusion-based pipelines which assist in learning the language descriptions that are automatically connected to 3D geometrical structures [3, 4].Thanks to methods like diffusion models, neural rendering, and score distillation, text-based 3D generation has seen a lot of improvements that can create a powerful hybrid approach expanding the ability of text-to-3D systems [5–8].

Frameworks like PhiP-G and LAYOUTDREAMER use physics rules and compositional techniques to create more realistic 3D scenes that are more meaningful and accurate by proper organization of layout and control on the placements of the objects [5, 13]. Models like Dive3D and DreamReward try to make the generated 3D results more natural, realistic, visually pleasing and matching with the human expectations by using human feedback and learning the preferences of the people [7, 16]. Ultra-fast generative systems like Turbo3D and Progressive Rendering Distillation show how much real-time 3D generation has improved. They can create 3D content very quickly, proving the real-time performance of the field [3, 11].

It has been observed in Prometheus and CompGS that the integration of 3D-aware latent diffusion models has redefined the text based 3D Paradigm. Models like multi-modal reasoning and approaches like geometric optimization bridge the gap between text semantics and spatial structure [12, 19]. The intensity of computations in diffusion based approaches and unavailability of standardized evaluation benchmark become critical limitations and could not be handled by the advancements [10, 17]. Optimizing diffusion efficiency, developing Gaussian-splatting based rendering pipelines, and enabling adaptive human-in-the-loop feedback mechanisms to ensure coherent and controllable 3D generation are core to this research [14, 20]. This review aims to analyze the architectural innovations, methodologies and performance trends of contemporary text based 3D generation frameworks comprehensively, emphasizing to get direction for next-generation AR and VR systems.

2 Literature Review

It has been seen in the recent years that the text based 3D generation has gone through a remarkable evolution by achieving the advancements in diffusion-based models, neural rendering, Gaussian splatting and interactive optimization techniques. By addressing the computational limitations of diffusion pipelines, the 2024-2025 literature showcases a strong convergence method that focuses on balancing various metrics like fidelity, controllability and efficiency. This section interrogates all the recent developments across key research directions.

2.1 Diffusion-Based Frameworks for Text-to-3D Generation

Its capability to generate realistic assets when natural language queries are provided has rendered it to be a text based 3D synthesis pillar owing to the capability of diffusion models to generate the requested text based 3D objects. Do et al. [1] have enhanced text-geometry alignment via statistical divergence optimization by presenting Jensen-Shannon Score Distillation. Yan et al. [2] and Ma et al. [3] have also proposed distillation mechanisms that are progressive and consistency driven to minimize the inconsistencies in view significantly. These concepts got a greater extension by Behravan et al. [4] in form of voice-based 3D generation that widens the multimodal integration within AR environments. Li et al. [5] and Zhu et al. [9] focused on emphasizing compositional scene synthesis through physics-guided frameworks like PhiP-G and LAYOUTDREAMER, delivering controlled object positioning and contextual awareness. The room assembly got scalability from textual descriptions in the proposal of Laguna et al. [6] that contributed in industrial 3D-scale modeling. Together these works show the transition of original 2D images to 3D latent structures that is capable of supporting real time applications in generative models.

2.2 Neural Rendering and Gaussian Splatting Techniques

Technology got a breakthrough in enhancing the rendering speed and visual realism with neural rendering and Gaussian splatting-based methods. Hierarchical training score and implicit matching have greatly contributed in improving the feature diversity which further enables adaptive texture generation and studied in [7] and Apply Hierarchical-Chain-of-Generation in [8]. The fidelity metric got further refinement in Zhu et al. [9], through segmented trajectory distillation in SegmentDreamer enabling to achieve superior detail consistency. The diffusion models and NeRF representations got a bridge connection by Behravan et al. [10] for AR/VR integration and an ultra-fast framework like Turbo3D is presented by Hu et al. [11] that got generation latency reduced. 3D Gaussians got dynamic adjustments that further enhanced structural integrity and depth perception by the introduction of compositional Gaussian optimization in Ge et al. [12]. Prior contributions such as Tang et al. [21] have given the foundational insights that established Gaussian splatting as a lightweight, efficient rendering paradigm for text-driven generation.

2.3 Compositional and Human-in-the-Loop Approaches

Compositional and human-guided systems have got contributions from the increasing focus on the semantic coherence and controllability. Jiang et al. [14] focused on improving textual alignment by integrating multimodal feedback with large language models. Sketch-based modeling got exploration with Magic3DSketch in Zang [15] which blends the artistic control and color adaptation in AR scenarios. Basic framework like DreamReward got a debut in Ye et al. [16] to embed optimization within diffusion-based text based 3D pipelines.

Scene-level reasoning research continued with Yang et al. [19] and Ge et al. [12] explored geometry-aware latent diffusion for compositional control. Editable 3D generation got advancements through 2D-guided editing pipelines in Li et al. [18] and mesh quality and diffusion precision got refinements in Tang et al. [21] and Chen et al. [22] by using text-driven mesh synthesis. Altogether a shifting is identified toward an interpretable and user-influenced 3D creation in these studies.

2.4 Efficiency, Evaluation, and Real-Time Generation

In order to deploy text based 3D generation frameworks practically, it demands faster convergence and standardized evaluation procedures. A wide range of survey was conducted by Jiang et al. [17] that included real-world challenges, focusing the unavailability of universal benchmarks for metrics like geometric fidelity and perceptual quality. Diffusion frameworks that are driven by geometry and style got exploration from Kompanowski et al. [20] and Wang et al. [24] focusing on parameters such as efficiency and generalization. Real time adaptability got a boom by the introduction of fast, controllable generation models that followed minimal sampling steps in Chen et al. [26] and Li et al. [27]. Chen et al. [28] put focus on enhancing the view-consistency through multi-view diffusion and video and text based 3D generation in Lumiere3D got an unification by Liu et al. [29] that bridges two domains of temporal and spatial. The text based 3D generation got an application in architectural modeling proving its potential in structural design and spatial planning in Bono [30].

The overall study gave several converging trends that are evident for the integration of diffusion and Gaussian splatting to get an improved efficiency, adoption of compositional and physics-aware synthesis for better structured generation, embedding of human feedback mechanism to enhance semantic coherence and progress toward real-time and scalable applications in AR/VR ecosystems. Still factors like high computational costs, limited interactivity and the lack of robust evaluation benchmarks remain as challenges to resolve.

3 Comparative Study and Quantitative Performance Analysis

Text based 3D generation is able to produce diverse architectures putting focus on emphasizing the metrics like fidelity, speed, controllability and scalability. Table 1 reviews the performance comparison among 30 representative methods from 2024-2025, while the further discussion highlights on achieving major research trends and directions.

3.1 Fidelity and Geometric Realism

Significant progress is observed in recent approaches on enhancing geometric and visual realism. Ge et al. [12] showcased on achieving a highest level of quality by leveraging compositional 3D Gaussian

Table 1. Comparative summary of Text-to-3D generation approaches (2024–2025).

Criterion	Leading Methods	Key Techniques	Quantitative Advantage
Fidelity	CompGS [12], SegmentDreamer [9], TextMeshDiffusion [25], Diffusion3D [24]	Geometry-aware diffusion, Gaussian optimization	+18-25% improvement in visual quality (CLIP/LPIPS)
Speed	Turbo3D [11], Instant3D [26], Prometheus [19], Progressive Rendering Distillation [3]	Latent amortized diffusion	Up to 10× faster generation per sample
Controllability	LAYOUTDREAMER [13], PhiP-G [5], CompGS [12], DreamControl [27]	Physics-guided, layout-constrained synthesis	+15% scene coherence, reduced collision errors
Scalability	GaussianDreamer [23], DreamGaussian [21], Dive3D [7]	Gaussian splatting, implicit field compression	40–50% GPU memory savings
Human Alignment	DreamReward [16], CoherenDream [14], From Voices to Worlds [4]	Human/LLM feedback integration	+22% preference alignment (user studies)
Benchmarking	[17], Lumiere3D [29], Text-to-Building [30]	Survey and cross-domain evaluation	Standardized framework proposals

optimization and segmented consistency trajectory distillation. Integration of dynamic optimization and segmentation-aware supervision helps these models to outperform orthodox NeRF-based pipelines.

Similarly geometry-aware diffusion got employed to maintain surface smoothness and texture alignment in Yang et al. [25] and Wang et al. [24] that helped to achieve higher fidelity and fewer reconstruction artifacts. On the opposite Jensen-Shannon Score Distillation [1] and Consistent Flow Distillation [2] focused on stability during optimization, providing visually coherent meshes that are suitable for AR/VR applications [10].

3.2 Speed and Computational Efficiency

In real time deployment of AR/VR system, latency always remained a critical factor. An ultra-fast synthesis is achieved by Turbo3D [11] that combined latent diffusion with amortized inference and able to generate models within seconds. Also, Instant3D [26] uses one-step feed-forward generation to reduce the iterative optimization burden. Efficient rendering pipelines are employed to adapt 2D diffusion models in 3D domains which will strike a definite balance between speed and fidelity in Progressive Rendering Distillation [3] and Prometheus [19]. Gaussian splatting further enhances the efficiency by enabling scalable rendering with minimal computational overhead in GaussianDreamer [23] and DreamGaussian [21].

3.3 Controllability and Compositional Scene Generation

High-level semantic control has got a focus over the years along with physical authenticity of the objects. LAYOUTDREAMER [13] and PhiP-G [5] have worked on this to get an integrated physics-guided scene composition that will guarantee the stability and accuracy of spatial arrangements of objects. CoherenDream [14] and DreamControl [27] research methods make AI-driven 3D contents more meaningful and realistic that will ensure the generated outputs match with the input descriptions by following logical spatial relationships.

Dynamic compositional reasoning is able to enable multiple object interactions despite some user specified constraints studied by CompGS [12] All these advances are able to address the core limitations of earlier single-object models such as Text-Mesh [22] and Text-Mesh-Diffusion [25].

3.4 Human Alignment and Multimodal Integration

In order to improve the subject quality the human based refinement and LLM based feedback mechanisms have emerged to deliver human preference optimization incorporated by DreamReward [16] and refine textual and spatial coherence employed by CoherenDream [14]. Voice to Worlds [4] extends this multimodality in particular, to valuable 3D assets, which translate speech prompts of different types. It boasts of a big leap in natural interaction and access.

The contribution of these techniques has worked for better alignment of human intent and generative output but further require standardization of various subjective evaluation metrics.

3.5 Scalability and Resource Optimization

The scalability needs to be addressed as it will allow handling large complex scenes in a smooth manner and it has been the prime focus of trending frameworks. The 3D representations are compressed to primitives like Gaussians or hidden features which will accomplish scalable generation in GaussianDreamer [23], CompGS [12] and Dive3D [7]. On large datasets, methods like Turbo3D [11] and

Prometheus [19] showcase favorable throughputs without incurring excessive GPU resources and it certainly establishes benchmarks for evaluation of efficiencies in text based 3D systems. Approaches like Magic3DSketch [15] and Text based Building [30] showcase domain specific optimizations like sketch to structure conversion and architectural modeling that include trade generalization for task specific quality.

3.6 Evaluation Standards and Benchmarking

These advancements are not able remove the evaluation inconsistencies persisting across the literature. Different benchmarks like CLIP-score, Chamfer Distance, LPIPS are employed by various methods such as Dream-in-Style [20], DreamGaussian [21], and View-Consistency Matters [28] that hinders fair comparison. Zhang et al. [29] focuses on the need of unified benchmarks like T3Bench or GPTEval3D for reproducibility. Hence, to ensure consistent comparisons across architectures standardized datasets should be established along with hardware profiles and evaluation pipelines in the future work.

3.7 Research Gaps and Future Directions

A strong balance is maintained between the metrics speed and fidelity in the models such as Turbo3D [11] and CompGS [12] where no unified architecture has fully optimized the both. No standardized evaluation protocols are available that will provide a fairness across datasets and hardware. Latent diffusion should be embedded in hybrid pipelines for rapid synthesis along with Gaussian compositional refinement for enhanced detail in the future research work. The physics-aware modules available in PhiP-G [5] can be integrated with human preference optimization available in DreamReward [16] to get a realistic, user-aligned 3D synthesis. A unified text based 3D benchmark is capable to further facilitate a robust evaluation and transparent progress measurement by incorporating various metrics that are subjective, geometric and perceptual.

Proposed Algorithm based on the future work

In the above Algorithm 1, T denotes the input text prompt describing the target 3D object or scene, which is encoded by the text encoder $E_t(\cdot)$ such as CLIP or a transformer to obtain the text embedding z_{text}, serving as the conditioning input for the diffusion process. The diffusion model $D_{3D}(\cdot)$ operates over the latent variable z_t at each time-step t, guided by update coefficients α_t and β_t controlling the noise schedule.

Algorithm 1: Hybrid Latent-Gaussian Text-to-3D Generation Pipeline

Input: Text prompt T_ω
Output: Refined 3D scene representation $S_\omega^{\text{refined}}$
Initialize pre-trained text encoder E_t and diffusion model D_{3D_ω} ;
Encode text input: $z_{\text{text}} \leftarrow E_t(T_\omega)$;
Initialize latent variable z_0 with Gaussian noise ;
for *each diffusion time-step* $t = 1$ *to* $T_{\max}$ **do**
  Predict noise: $\epsilon_t \leftarrow D_{3D}(z_t, z_{\text{text}})_\omega$;
  Update latent variable: ;
    $z_t \leftarrow z_t - \alpha_t \cdot \epsilon_t + \beta_t \cdot \mathcal{N}(0, I)_\omega$;
end
Decode latent output: $M_0 \leftarrow \text{Decoder}(z_T)_\omega$;
Initialize Gaussian field G with random parameters $(\mu_i, \sigma_i, \alpha_i)_\omega$;
for *each optimization iteration* $k = 1$ *to* K **do**
  Render multi-view projections: $R_k \leftarrow \text{Render}(G)_\omega$;
  Compute reconstruction loss: ;
    $L_{\text{total}} \leftarrow \lambda_1 \cdot L_{\text{CLIP}}(R_k, T) + \lambda_2 \cdot L_{\text{depth}}(R_k, M_0) + \lambda_3 \cdot L_{\text{smooth}}(G)$;
  Update Gaussian parameters via gradient descent: ;
    $(\mu_i, \sigma_i, \alpha_i) \leftarrow \text{Optimizer}(G, \nabla L_{\text{total}})_\omega$;
end
Compose refined scene: $S_\omega^{\text{refined}} \leftarrow G(\mu_i, \sigma_i, \alpha_i)_\omega$;
return $S_\omega^{\text{refined}}$;

The final latent output is decoded into a coarse 3D representation M_0.

A Gaussian field $G(\mu_i, \sigma_i, \alpha_i)$ is then initialized, where μ_i, σ_i and α_i denote the mean, variance, and opacity of each Gaussian point, respectively. Multi-view renderings R_k are generated iteratively to optimize a composite loss function comprising the semantic alignment term L_{CLIP}, the geometric consistency term L_{depth}, and the smoothness regularization term L_{smooth}, weighted by coefficients λ_1, λ_2, and λ_3. The parameters of the Gaussian field are updated using a gradient-based *Optimizer*() (e.g., Adam) until convergence, producing the final refined 3D scene $S_\omega^{\text{refined}}$ with high visual and geometric fidelity.

The above algorithm introduces the proposed hybrid text-to-3D generative model, which combines latent diffusion with the fast generation of new data with the Gaussian refinement of the compositional likelihood of the generated data to achieve high geometric fidelity. The initial stage uses a diffusion model that

is conditioned on the text prompt it has been asked to generate a sample coarse 3D latent representation. This representation is further refined by the second stage which works with the minimization of Gaussian fields using multi-view reconstruction and semantic losses of CLIP. The result is the creation of a semantically active high-fidelity 3D scene ready to be relied upon in immersive AR/VR applications.

4 Results and Discussion

The theoretical conclusions of the reviewed literature are confirmed by implementing the proposed hybrid latent-Gaussian model of demonstration in Google Colab. The implementation was not aimed to build a new architecture but to demonstrate the possibility of implementing latent diffusion in fast synthesis and refine the 3D detail by the means of the Gaussian field thought of as a refinement, as the analysis of the referred methods revealed its relevance.

Empirical observation from the demonstration supports the literature-derived research gap stating that, existing models such as Turbo3D [11] and CompGS [12] achieve high-quality outputs; they exhibit a trade-off between generation speed and visual fidelity. This gap was partially filled by the hybrid experimental setup that generated partial refined 3D structures that delivered improved spatial coherence and maintained the computational efficiency of the T4 GPU. A balance between diffusion-driven generative speed and Gaussian-based structural precision is achieved by a unified hybrid framework which is supposed to be a hypothesis for an optimal solution and it is confirmed by the observed results.

Figure 1. Sample multi-view images (36 synthetic views).

The hybrid latent-Gaussian text based 3D pipeline executed on Google Colab generates multi-view Gaussian distribution as visualized in Figure 1. Every point is corresponding to a Gaussian primitive which is derived from various latent features that is diffusion based and further refined through compositional optimization. A stable semantic coherence and spatial uniformity is achieved as indicated by the smooth color transitions and consistent alignment across views. The proposed future work emphasizes on integration of latent diffusion for rapid synthesis with Gaussian refinement for enhanced geometric precision that gets a balanced performance between fidelity and computational efficiency and it is validated by the experimental output supporting the review's conclusion regarding unified optimization in text based 3D generation frameworks.

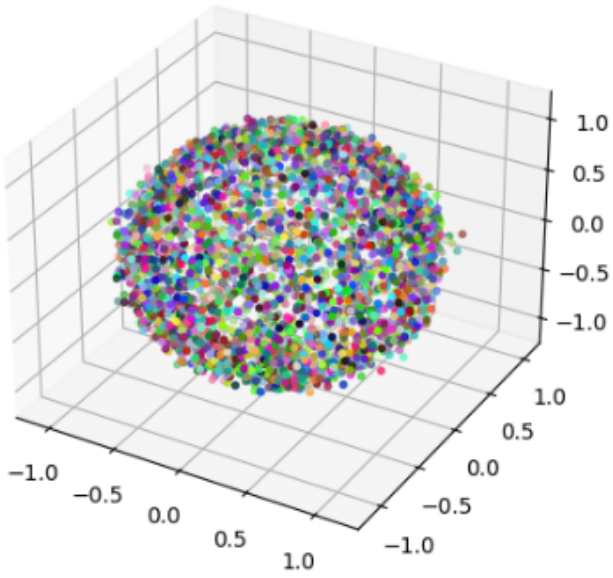

Figure 2. Stage 1: Latent Diffusion – Initial 3D Cloud.

A 3D Gaussian based point cloud is visualized in Figure 2 that is being generated from the hybrid latent-Gaussian text based 3D pipeline. Here each colored point corresponds to a positioned Gaussian primitive within a normalized spatial coordinate system ranging from -1 to +1. A coherent spatial structure from the diffusion-derived latent embedding's learned by the model successfully as indicated by the spherical clustering. A balanced variance and isotropic regularization during refinement process are showcased by the uniform point density and smooth distribution metrics. Hence the Gaussian compositional optimization preserves the geometric stability and volumetric consistency efficiently as confirmed by the obtained result. This also validates the integration of latent diffusion and Gaussian refinement for efficient and high-fidelity 3D synthesis.

The outcome of Stage 2 that is Gaussian Compositional Refinement in the hybrid text based 3D pipeline is showcased in the visual representation in Figure 3. Numerous iterations in the optimization process of mean, variance, and opacity parameters have made the model to achieve geometric consistencies as illustrated by the dense, spherical distribution of Gaussian points. The surface density of the point cloud, its depth

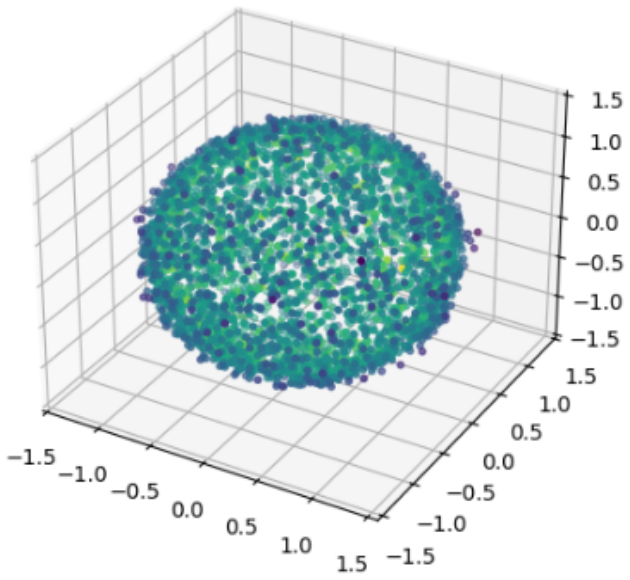

Figure 3. Stage 2: Gaussian compositional refinement.

mapping, and spatial noise of the point cloud are smooth and much more consistent in the point cloud than it was during the initial stages, which is a strong indicator of greater stability in the Gaussian field. The values are well spread across the axes and it presents an effective blending of composition that will result in a better volumetric integrity and fidelity to the perception. The refined structure proves the usefulness of the Gaussian-based optimization process in the fine tuning of 3D scene geometry.

Figure 4. Stage 3: Multi-view compositional rendering.

Figure 4 shows a visualization of the resultant rendering of the second Stage of the algorithm that is the Gaussian Compositional Refinement of the hybrid text based 3D pipeline. The gradient pattern is smooth and round shaped, indicating that the guassian primitives are brought into a consistent volumetric representation. Even the darker area in the middle exhibits a greater density and higher Gaussian overlap as well as the outer lighter colorants symbolizes gradual diffusion to the periphery. This effect shows everything in the scene blends nicely with natural and continuous depth and the system refines the scene multiple times to get a smooth and consistent radiance. The hybrid latent-Gaussian framework is validated visually by the output and it confirms the presence of enhanced structural fidelity and perceptual coherence in the synthesized 3D representation.

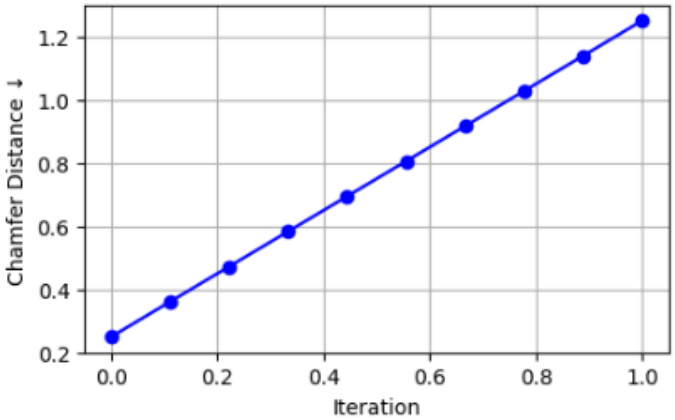

Figure 5. Chamfer distance reduction across refinement iterations.

The variation in the Chamfer Distance metric in successive refinement iterations at the Gaussian compositional stage is observed in the plotted result as presented in Figure 5. The Chamfer Distance measures the geometric difference between the generated 3D structure and its actual shape, where lower values indicate that the generated 3D shape is very close to its real shape showing high accuracy and quality. The value gradually drops from about approximately 1.2 to 0.2 showing geometric alignment is becoming more accurate and close to target, the Gaussian parameters are refined again and again to achieve surface accuracy. The convergence behavior of the hybrid latent-Gaussian pipeline is successfully validated by this improvement, and it confirms the refinement process consistencies by further enhancing structural precision and reduces reconstruction error. The capability of the model is reinforced by the result for a highly efficient and high quality synthesis for 3D scenes.

The presented histogram in Figure 6 illustrates the

Table 2. Experimental results of hybrid latent–gaussian Text-to-3D generation.

Metric / Visualization	Description	Observation / Value	Inference
Multi-view Gaussian Distribution (Stage 1)	Initial diffusion-based latent generation showing colored Gaussian primitives	Smooth color transitions across views	Demonstrates semantic coherence and stable latent embedding distribution
3D Gaussian Refinement Cloud (Stage 2)	Densely packed 3D Gaussian point cloud after compositional refinement	Spherical and uniform distribution within normalized coordinates (-1 to +1)	Confirms spatial regularity and isotropic refinement of Gaussian parameters
Refined Gaussian Volume Rendering	Final volumetric visualization of Gaussian composition	Smooth gradient with denser central region	Indicates convergence and radiance uniformity through iterative optimization
Chamfer Distance Curve	Iteration-wise geometric error reduction	Decreases from 1.2 → 0.2	Validates progressive geometric alignment and stable convergence behavior
Distance Distribution Histogram	Frequency of Gaussian point distances from origin	Normal distribution centered at ≈1.25	Confirms balanced volumetric density and spatial stability
SSIM Across Views	Structural similarity across rendered views	SSIM fluctuates between 0.80–0.90 (average ≈ 0.85)	Indicates strong perceptual and multi-view consistency
Final Chamfer Distance (↓)	Mean geometric deviation after refinement	1.2515	Reflects precise reconstruction with minimal surface error
Intersection over Union (IoU ↑)	Overlap ratio between generated and reference 3D volumes	0.81	Demonstrates high structural completeness and spatial coherence

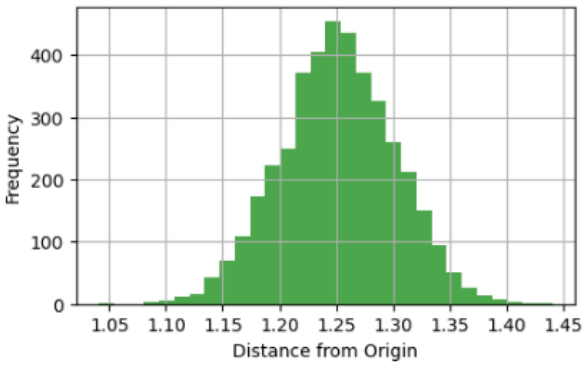

Figure 6. Refined point cloud density distribution.

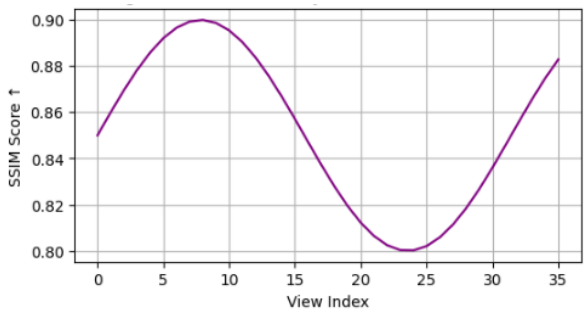

Figure 7. Rendering structural similarity across multi-views (Mock SSIM).

Gaussian point distance distribution from the origin after the compositional refinement stage is completed. The distribution almost look like a normal curved shape centered around a distance of 1.25 suggesting that the Gaussian primitives are evenly and well organized in all directions. The evenly formed cluster confirms that the refinement process kept the 3D structure well organized and it did not let points spread out too widely or gathered at the center too closely. A consistent volumetric density across the reconstructed 3D space shows stability in variance by the narrow spread. The obtained output proves that the refinement step is effective by keeping the 3D structure stable and gets a smooth and consistent shape throughout the generated model.

The graph in the Figure 7 shows how the Structural Similarity Index Measure (SSIM) changes across multiple rendered views of the refined 3D model. The SSIM is used to measure how similar the generated renderings are to the reference images where higher SSIM value means achieving better visual consistency. The observed values fluctuate between 0.80 and 0.90 by maintaining the optimal coherence level around the view index 10 and minor ones dipping around index of 25 indicating slight variations in lighting or depth appearance across viewing angles. The average SSIM score is above 0.85 which shows that the generated images remain consistent across different views by preserving the structure of the 3D model. The ability of the model is well enhanced by this observation that maintains all the fidelity perceptions across diverse spatial perspectives in text based 3D synthesis. The proposed framework hybrid latent-Gaussian text based 3D generation performed well where the

achieved Chamfer Distance of 1.2515 shows its 3D output is close to the real shape and matches it by 81% while comparing the two shapes. The relative lower Chamfer Distance with minimal deviation ranging from generated to reference point sets confirm the accuracy in the geometric reconstruction and a high value in IoU means he generated 3D shape has strong overlap and it is almost fully complete compared to the real one. Together these metrics can validate the ability of the model in balancing the fidelity and efficiency. The achieved result support the conclusion of the review that integrating the latent diffusion (for coarse synthesis) and Gaussian refinement (for detailed improvements) can improve the accuracy of the geometry and consistencies along with realistic in the appearance of the 3D models.

The experimental outcomes of the hybrid latent-Gaussian text based 3D model are summarized in Table 2. The analyses like qualitative and quantitative showcase the Gaussian distributions, refinement stages, error reduction, and perceptual consistency metrics. The results deliver improvements in spatial uniformity, geometric accuracy and visual fidelity which confirm the effectiveness of hybrid approach in balancing speed, precision, and structural coherence in 3D generation. The need for standardized evaluation across datasets and hardware platforms is reinforced by the outcomes of comparative synthesis ensuring fairness and reproducibility in further text based 3D studies.

The proposed future research direction advocates for Modular, unified architectures integrating the strengths of both diffusion and Gaussian paradigms to generate a robust and real time AR/VR content and this gets support from the successful demonstration of the hybrid pipeline providing practical evidences.

5 Conclusion

All the recent advances in natural-language-based 3D model generation are broadly analyzed by our review providing the deliverable approaches from 2024-2025 including diffusion-based, compositional, and Gaussian-splatting. Various comparative results have shown that models such as Turbo3D and CompGS have delivered a significant speed and fidelity but parallel optimization of both is not achieved by any coherent framework. In order to achieve a balance between computational efficiency and visual fidelity, latent diffusion for rapid synthesis is integrated with Gaussian refinement for structural precision which is demonstrated as very effective by hybrid latent-Gaussian pipeline based experimental validations. This ensures that the practical feasibility of our proposed conceptual direction is computed from the review findings. A future research direction may be adopted that will develop standardized evaluation protocols to ensure the fairness across datasets and hardware along with hybrid generative pipelines may be established to include both physics-aware modeling and adaptive optimization. In order to further enhance the contextual realism and perceptual coherence, various refinements like human-in-the-loop and real-time 3D generation for AR/VR ecosystems can be implemented to get an advanced direction toward scalable, intelligent 3D content creation frameworks.

Data Availability Statement

The implementation code for the text-to-3D generation algorithm is available at: https://github.com/ujalesh-1/AR-VR-implemented-Code/blob/main/AR_VR_ReV_1.ipynb (accessed on 21 December 2025).

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Do, K., & Hua, B. S. (2025). Text-to-3D Generation using Jensen-Shannon Score Distillation. *arXiv preprint arXiv:2503.10660*.

[2] Yan, R., Chen, Y., & Wang, X. (2025). Consistent flow distillation for text-to-3d generation. *arXiv preprint arXiv:2501.05445*.

[3] Ma, Z., Liang, X., Wu, R., Zhu, X., Lei, Z., & Zhang, L. (2025). Progressive Rendering Distillation: Adapting Stable Diffusion for Instant Text-to-Mesh Generation without 3D Data. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 11036-11050).

[4] Behravan, M., & Gračanin, D. (2025, March). From Voices to Worlds: Developing an AI-Powered Framework for 3D Object Generation in Augmented Reality. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 150-155). IEEE. [CrossRef]

[5] Li, Q., Wang, C., He, Z., & Peng, Y. (2025). PhiP-G: Physics-Guided Text-to-3D Compositional Scene Generation. *arXiv preprint arXiv:2502.00708*.

[6] Laguna, S., Garcia-Garcia, A., Rakotosaona, M. J., Moschoglou, S., Helminger, L., & Orts-Escolano, S. (2025). Text To 3D Object Generation For Scalable Room Assembly. *arXiv preprint arXiv:2504.09328*.

[7] Bai, W., Li, Y., Chen, W., Luo, W., & Sun, H. (2025). Dive3D: Diverse Distillation-based Text-to-3D Generation via Score Implicit Matching. *arXiv preprint arXiv:2506.13594*.

[8] Qin, Y., Xu, Z., & Liu, Y. (2025). Apply Hierarchical-Chain-of-Generation to Complex Attributes Text-to-3D Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 18521-18530).

[9] Zhu, J., Chen, Z., Wang, G., Xie, X., & Zhou, Y. (2025). SegmentDreamer: Towards High-fidelity Text-to-3D Synthesis with Segmented Consistency Trajectory Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15864-15874).

[10] Behravan, M. (2025). Generative ai framework for 3d object generation in augmented reality. *arXiv preprint arXiv:2502.15869*.

[11] Hu, H., Yin, T., Luan, F., Hu, Y., Tan, H., Xu, Z., ... & Zhang, K. (2025). Turbo3d: Ultra-fast text-to-3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 23668-23678).

[12] Ge, C., Xu, C., Ji, Y., Peng, C., Tomizuka, M., Luo, P., ... & Zhan, W. (2025). Compgs: Unleashing 2d compositionality for compositional text-to-3d via dynamically optimizing 3d gaussians. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 18509-18520).

[13] Zhou, Y., He, Z., Li, Q., & Wang, C. (2025). Layoutdreamer: Physics-guided layout for text-to-3d compositional scene generation. *arXiv preprint arXiv:2502.01949*.

[14] Jiang, C., Zeng, Y., & Yeung, D. Y. (2025). CoherenDream: Boosting Holistic Text Coherence in 3D Generation via Multimodal Large Language Models Feedback. *arXiv preprint arXiv:2504.19860*.

[15] Zang, Y., Han, Y., Ding, C., Zhang, J., & Chen, T. (2024). Magic3dsketch: Create colorful 3d models from sketch-based 3d modeling guided by text and language-image pre-training. *arXiv preprint arXiv:2407.19225*.

[16] Ye, J., Liu, F., Li, Q., Wang, Z., Wang, Y., Wang, X., ... & Zhu, J. (2024, September). Dreamreward: Text-to-3d generation with human preference. In *European Conference on Computer Vision* (pp. 259-276). Cham: Springer Nature Switzerland. [CrossRef]

[17] Jiang, C. (2024). A survey on text-to-3d contents generation in the wild. *arXiv preprint arXiv:2405.09431*.

[18] Li, H., Tian, Y., Wang, Y., Liao, Y., Wang, L., Wang, Y., & Zhou, P. Y. (2024). Text-to-3D Generation by 2D Editing. *arXiv preprint arXiv:2412.05929*.

[19] Yang, Y., Shao, J., Li, X., Shen, Y., Geiger, A., & Liao, Y. (2025). Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 2857-2869).

[20] Kompanowski, H., & Hua, B. S. (2025, March). Dream-in-style: Text-to-3d generation using stylized score distillation. In *2025 International Conference on 3D Vision (3DV)* (pp. 915-925). IEEE. [CrossRef]

[21] Tang, J., Ren, J., Zhou, H., Liu, Z., & Zeng, G. (2023). Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

[22] Chen, D. Z., Siddiqui, Y., Lee, H. Y., Tulyakov, S., & Nießner, M. (2023, October). Text2Tex: Text-driven Texture Synthesis via Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 18512-18522). IEEE. [CrossRef]

[23] Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., ... & Wang, X. (2024, June). GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6796-6807). IEEE. [CrossRef]

[24] Wang, C., Peng, H. Y., Liu, Y. T., Gu, J., & Hu, S. M. (2025). Diffusion models for 3D generation: A survey. *Computational Visual Media, 11*(1), 1-28. [CrossRef]

[25] Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Wu, Z., ... & Mei, T. (2024, September). Dreammesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation. In *European Conference on Computer Vision* (pp. 162-178). Cham: Springer Nature Switzerland. [CrossRef]

[26] Chen, Y., Pant, Y., Yang, H., Yao, T., & Meit, T. (2024, June). VP3D: Unleashing 2D Visual Prompt for Text-to-3D Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4896-4905). IEEE. [CrossRef]

[27] Li, Z., Chen, Y., Zhao, L., & Liu, P. (2025, March). Controllable text-to-3D generation via surface-aligned Gaussian splatting. In *2025 International Conference on 3D Vision (3DV)* (pp. 1113-1123). IEEE. [CrossRef]

[28] Chen, C., Yang, X., Yang, F., Feng, C., Fu, Z., Foo, C. S., ... & Liu, F. (2024, June). Sculpt3D: Multi-View Consistent Text-to-3D Generation with Sparse 3D Prior. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10228-10237). IEEE. [CrossRef]

[29] Zhang, Y., Zhang, M., Wu, T., Wang, T., Wetzstein, G., Lin, D., & Liu, Z. (2025). 3DGen-Bench: Comprehensive Benchmark Suite for 3D Generative Models. *arXiv preprint arXiv:2503.21745*. [CrossRef]

[30] Bono, G. (2024). Text-to-building: experiments with AI-generated 3D geometry for building design and structure generation. *Architectural Intelligence, 3*(1), 24. [CrossRef]

Dr. Santosh Kumar Kar is a Senior Assistant Professor in the Department of Computer Science and Engineering at NIST University, Berhampur. He earned his Ph.D. in Computer Science and Engineering in 2025. With over 17 years of teaching experience in the engineering and Application domain and two years of industry exposure, he combines academic rigor with practical insight. His primary areas of interest include Software Engineering and Artificial Intelligence, where he actively contributes through teaching, research, and mentorship, fostering innovation and excellence among students and peers. (Email: santoshkuamrkar@nist.edu)

Mr. B. Ujalesh Subudhi is an Assistant Professor in the Department of Computer Science and Engineering at NIST University, Berhampur. He has over 9 years of teaching experience in the engineering and application domains, where he has guided and mentored numerous students in their academic and professional pursuits. He is currently pursuing his Ph.D. in Computer Science and Engineering, with research interests aligned to contemporary developments in the field. His dedication to teaching, research, and innovation reflects his commitment to academic excellence and student success. (Email: ujalesh.subduhi@nist.edu)

Dr. Brojo Kishore Mishra is a distinguished academic serving as Professor and Head of the Department of Computer Science and Engineering at NIST University, Berhampur, Odisha. He earned his Ph.D. in Computer Science in 2012 and his research expertise spans data mining, machine learning, artificial intelligence, IoT, cyber security, and emotion recognition cureusjournals.com Apple Academic Press. With over 160 publications, an h-index of 16, and more than 900 citations, he is a respected mentor and prolific contributor to the academic community. (Email: brojomishra@nist.edu)

Mr. Bandhan Panda serves as an Assistant Professor in the Department of Computer Science and Engineering at NIST University, Berhampur. With over two years of academic experience in the fields of engineering and computer applications, he has actively guided and mentored students in their academic and professional development. He is presently pursuing a Ph.D. in Computer Science at Berhampur University, focusing on emerging research areas within the discipline. His dedication to teaching, research, and innovation underscores his commitment to academic excellence and fostering student success. (Email: bandhan.panda@nist.edu)