A Survey of Large Language Models and 3D Representation: Bridging Text and Geometry

www.surveyx.cn

Abstract

In the interdisciplinary domain of artificial intelligence and computer graphics, the integration of large language models (LLMs) with 3D representation is revolutionizing the generation, interpretation, and manipulation of three-dimensional content from textual descriptions. This survey explores the transformative potential of LLMs in bridging the semantic chasm between high-level linguistic inputs and detailed geometric representations, highlighting their role in enhancing multimodal information fusion, text-to-3D conversion, and inverse graphics. Key advancements include the application of diffusion models for coherent 3D generation, the integration of neural textures for realistic rendering, and the use of adaptive algorithms for optimizing computational models. Despite these advancements, challenges persist in aligning diverse data modalities, ensuring model efficiency, and achieving fine-grained control over digital objects. Future research directions emphasize the need for innovative solutions to improve model robustness, expand multilingual capabilities, and refine training datasets to encompass a broader range of concepts. By addressing these challenges, the field aims to develop sophisticated AI systems capable of seamless multimodal interactions, thereby advancing the fidelity and interactivity of digital environments. This synthesis underscores the expansive potential of combining LLMs with 3D representation, fostering advancements in both theoretical frameworks and practical applications.

1 Introduction

1.1 Interdisciplinary Domain Overview

The convergence of large language models (LLMs) and 3D representation is transforming artificial intelligence by integrating linguistics with spatial modeling to enhance the creation and interpretation of complex three-dimensional environments. This interdisciplinary field is propelled by advancements in LLMs that have significantly improved natural language processing capabilities and the increasing demand for automated technologies to analyze and comprehend video content, intersecting with video understanding [1]. The scaling of neural network capabilities further emphasizes the necessity of merging LLMs with 3D representation to tackle increasingly complex tasks across various disciplines [2].

In materials science, the absence of a unified representation for chemical structures in deep learning underscores the importance of integrating 3D representation with LLMs for a better understanding of molecular configurations [3]. This integration is essential for developing models that accurately simulate and predict material behavior at the molecular level.

The field also faces the challenge of efficiently processing large-scale 3D data, especially in edge computing scenarios where resource constraints are common. Efficient models capable of operating under such limitations are vital for advancing AI systems designed for real-time 3D understanding and reasoning [4].

Generated by SurveyX.cn.

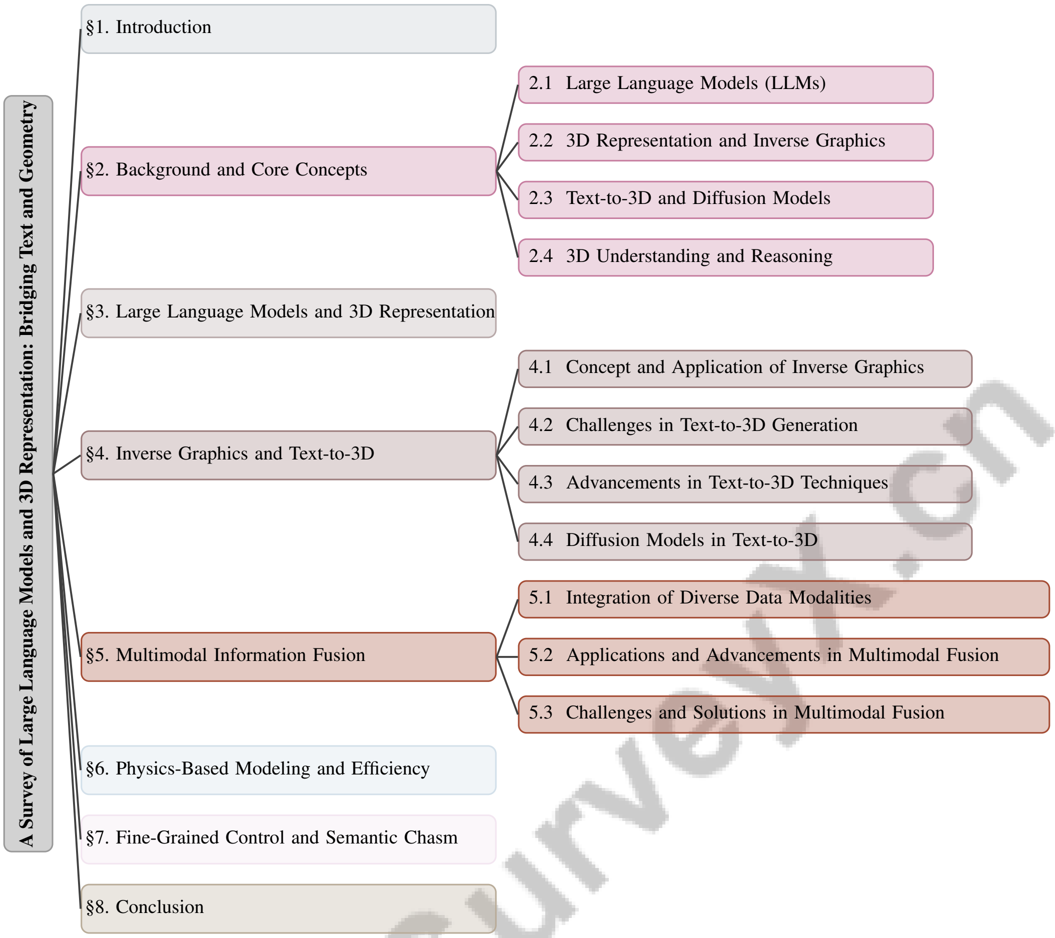

A Survey of Large Language Models and 3D Representation: Bridging Text and Geometry
§1. Introduction
§2. Background and Core Concepts
2.1 Large Language Models (LLMs)
2.2 3D Representation and Inverse Graphics
2.3 Text-to-3D and Diffusion Models
2.4 3D Understanding and Reasoning
§3. Large Language Models and 3D Representation
§4. Inverse Graphics and Text-to-3D
4.1 Concept and Application of Inverse Graphics
4.2 Challenges in Text-to-3D Generation
4.3 Advancements in Text-to-3D Techniques
4.4 Diffusion Models in Text-to-3D
§5. Multimodal Information Fusion
5.1 Integration of Diverse Data Modalities
5.2 Applications and Advancements in Multimodal Fusion
5.3 Challenges and Solutions in Multimodal Fusion
§6. Physics-Based Modeling and Efficiency
§7. Fine-Grained Control and Semantic Chasm
§8. Conclusion

Figure 1: chapter structure

These intersections highlight the vast potential of combining LLMs with 3D representation, fostering advancements in both theoretical frameworks and practical applications. This synthesis not only improves the generation and interpretation of 3D content but also paves the way for sophisticated AI systems that facilitate seamless multimodal interactions, thereby expanding the horizons of artificial intelligence and computer graphics.

1.2 Significance of Neural Network Integration

Integrating neural networks, particularly LLMs, into the synthesis of text and 3D geometry is crucial for enhancing artificial intelligence's capabilities in generating and interpreting detailed three-dimensional environments. LLMs have revolutionized natural language processing, enabling the conversion of textual descriptions into intricate 3D representations and significantly improving 3D content generation across various domains. The increasing size and complexity of these models underscore the pivotal role of neural networks in this context [2].

Specialized applications, such as MarineGPT, demonstrate the effectiveness of LLMs in recognizing marine objects and providing domain-specific responses, showcasing their potential to enhance 3D content generation [5]. The challenges of high-quality 3D data generation, including inefficiencies in existing representations like point clouds and neural fields, necessitate innovative neural network architectures [6]. Furthermore, the integration of ParticleGrid as a robust 3D representation highlights the significance of neural networks in modeling complex molecular structures, enhancing the understanding of spatial configurations [3].

Architectural disunity and a lack of labeled 3D data complicate 3D representation learning, necessitating robust neural network solutions to address these issues. Evaluation frameworks like PCA-Bench

AI-generated, for reference only.

aim to improve the assessment of Multimodal Large Language Models (MLLMs) by introducing error localization techniques that are essential for enhancing the integration of text and 3D geometry [7]. Additionally, synthesizing diverse research contributions related to LLMs is vital for enhancing their ability to perform complex language tasks with near-human proficiency [4].

Neural networks also play a critical role in addressing the complexities of 3D models and the ambiguities of text descriptions, which present significant challenges in tasks such as texture editing. Despite advancements, the absence of a 3D representation in current image generation methods hinders effective manipulation and understanding of objects from multiple viewpoints. Creating benchmarks to evaluate hierarchical understanding in visual language models (VLMs) is crucial for enhancing their performance, as exemplified by the HierarCaps dataset, which facilitates exploration of emergent visual-semantic hierarchies within these models. This initiative leverages the latent knowledge of existing foundation models while addressing the limitations of current multimodal hierarchical representation methods that often require extensive retraining. By optimizing hierarchical reasoning through targeted benchmarks, researchers can significantly enhance VLMs' capabilities in interpreting complex image-text relationships [8, 9, 10, 11]. The SPACE benchmark, for instance, evaluates spatial cognition capabilities, underscoring the necessity for models to navigate and comprehend physical environments effectively.

Neural networks are essential for the effective integration of text and 3D geometry, significantly enhancing both the generation and interpretation of 3D content. For example, models like MT3D utilize depth maps from high-fidelity 3D objects to mitigate viewpoint bias, ensuring geometric consistency in generated 3D shapes. SceneWiz3D employs a hybrid representation to synthesize complex 3D scenes from text, optimizing object placement through advanced algorithms. Additionally, HyperSD-Fusion leverages hierarchical structures in both language and geometry to improve the accuracy of 3D shape generation from textual descriptions. Collectively, these advancements illustrate the critical role of neural networks in creating diverse, detailed, and geometrically coherent 3D representations [12, 13, 14]. Their application in multimodal frameworks and vision-language tasks is vital for the ongoing evolution of AI technologies, despite persistent challenges that necessitate innovative solutions.

1.3 Structure of the Survey

This survey is structured to provide a comprehensive exploration of the intersection between large language models (LLMs) and 3D representation, emphasizing the synthesis of textual and geometric data. We begin with an **Introduction** that delineates the interdisciplinary nature of this domain, highlighting the transformative role of neural networks in bridging language and spatial modeling. Following this, the **Background and Core Concepts** section delves into fundamental theories, including LLMs, 3D representation, inverse graphics, and text-to-3D conversion, while discussing the relevance of diffusion models and multimodal information fusion in enhancing AI and computer graphics.

The survey examines the , investigating the pivotal role LLMs play in enhancing the generation and interpretation of 3D content. It highlights how these models facilitate the integration of textual descriptions with geometric data, addressing challenges such as information degradation and insufficient synergy between 3D representations and multimodal inputs. By leveraging advanced techniques like the Structured Multimodal Organizer (SMO) and Joint Multi-modal Alignment (JMA), the survey illustrates how LLMs can optimize 3D model performance and improve understanding in applications across computer vision, robotics, and autonomous driving [15, 16, 17, 18, 4]. This is followed by a detailed examination of **Inverse Graphics and Text-to-3D**, analyzing the challenges and advancements in converting text descriptions into 3D models, particularly focusing on the application of diffusion models in this process.

Next, we discuss **Multimodal Information Fusion**, highlighting the integration of diverse data modalities such as text, images, and 3D models to enhance 3D understanding and reasoning. The subsequent section on **Physics-Based Modeling and Efficiency** addresses the role of physics-based modeling in adding realism to 3D representations and optimizing computational models for performance and user engagement.

Finally, the survey explores **Fine-Grained Control and the Semantic Chasm**, examining techniques for achieving precise control over digital objects and strategies for bridging the gap between high-level semantic understanding and detailed geometric representation. The encapsulates essential findings regarding the integration of LLMs with 3D representation technologies, highlighting significant challenges such as information degradation from insufficient multi-view data, the lack of synergistic optimization between 3D models and multimodal features, and the underutilization of detailed learned representations. It also emphasizes future research directions aimed at overcoming these obstacles, particularly through innovative approaches like JM3D, which enhances 3D understanding by integrating point clouds, text, and images, and the novel paradigm for compositional 3D-aware video generation that utilizes LLMs for precise control over individual video concepts [18, 16].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) represent a transformative step in AI, excelling in understanding, generating, and processing human-like text from vast datasets. Notable models like GPT-3 and BERT have set benchmarks in tasks such as translation and summarization [19]. The progression from Statistical and Neural Language Models to Pre-trained and Large Language Models underscores significant advancements in language understanding. LLMs are increasingly integrated into multimodal frameworks, combining textual, visual, and spatial data, as demonstrated by the Valley model, which merges video, image, and language understanding [1]. Methods like VLLaVO further enhance real-world applications by integrating Vision Language Models with LLMs [20].

In visual reasoning, LLMs improve AI cognitive functions by generating and executing neural modules with minimal examples [4]. Their integration with 3D frameworks shows promise for learning intuitive physics models from visual inputs [2]. However, challenges remain in generating realistic 3D portraits from text, often relying heavily on geometric priors [19]. LLMs face inefficiencies in resource allocation within dynamic networks and challenges in multimodal response generation, particularly in visual-textual understanding [20]. Executing transformer networks on traditional platforms poses performance and energy efficiency issues.

LLMs drive AI innovation in natural language processing and multimodal integration. Evaluations using datasets like MULTI show certain multimodal LLMs surpass human experts on complex tasks, highlighting their potential for breakthroughs in AI [9, 4]. Their capacity to handle complex tasks with minimal supervision positions them as vital tools for advancing AI across various sectors.

2.2 3D Representation and Inverse Graphics

3D representation and inverse graphics are essential for linking textual descriptions to detailed 3D models, crucial for virtual reality and digital content creation. The evolution from image-based rendering to neural implicit functions enables realistic 3D model creation [21]. Inverse graphics facilitates text-to-3D conversion by inferring scene attributes like shape and lighting, supporting explicit surface geometry generation through multi-view outputs [22]. This enhances model flexibility and control, particularly in dynamic environments [23].

Challenges persist in segmenting 3D point clouds due to a lack of labeled datasets, crucial for applications like autonomous vehicles [24]. Existing methods struggle with non-watertight surfaces common in 3D objects [25]. Gaussian reconstruction models can cause inconsistencies and blurred textures [26]. Solutions like ParticleGrid, which provides a physics-inspired grid representation, are vital for effective text-to-3D conversion in molecular prediction [3]. VLLaVO addresses domain shift by converting images into detailed textual descriptions, enhancing 3D model robustness [20].

3D representation and inverse graphics form a robust framework for converting text to 3D models, advancing 3D content creation and manipulation. These principles foster AI and computer graphics innovations while addressing challenges like scarce large-scale 3D-text paired data and open vocabulary 3D scene understanding [27].

2.3 Text-to-3D and Diffusion Models

Text-to-3D conversion is a frontier in AI, enabling 3D model creation from textual descriptions through advanced computational methods. Diffusion models are pivotal, transforming text inputs into high-quality 3D representations by iteratively refining noisy data. The BrightDreamer framework exemplifies efficient 3D Gaussian generation from text prompts in a single stage [28]. Variational Distribution Mapping (VDM) enhances 3D generation fidelity by treating rendered images as degraded diffusion model outputs [29]. G-SHELL introduces joint parameterization of watertight and non-watertight meshes for efficient extraction and rendering [25].

Challenges like high computational requirements and training costs hinder advanced 3D content creation democratization [2]. Addressing these is crucial for robust text-to-3D systems. The Multimodal Relation Distillation (MRD) framework improves 3D shape understanding by distilling knowledge from Vision-Language Models [30]. Synthetic data bridges gaps in labeled datasets, training networks to learn geometric and texture cues. TextField3D introduces Noisy Text Fields (NTFs) for dynamic latent space adjustment, allowing flexible 3D generation [19].

Diffusion models advance text-to-3D conversion by generating coherent, detailed 3D representations. They address challenges in ensuring geometric consistency and enhancing multimodal integration. Models like MT3D use depth maps to mitigate viewpoint bias, improving geometric accuracy. Text-centric alignment techniques allow adaptation to varying modalities, enhancing generalizability across data combinations. These innovations contribute to sophisticated AI systems capable of seamless text-to-3D transformations [31, 14].

2.4 3D Understanding and Reasoning

3D understanding and reasoning are foundational for AI systems interpreting complex geometric data, enabling machine interaction with the world akin to human cognition. Recent research emphasizes fine-grained interactive edits, necessitating sophisticated 3D reasoning [32]. Disentangling 3D representations involves identifying and manipulating explanatory factors, facilitating spatial environment comprehension [33].

Approaches like PCA-Bench enhance 3D reasoning through multimodal decision-making frameworks, improving AI prediction accuracy [7]. Models like Valley integrate linguistic and visual data for nuanced video content understanding [1]. Challenges remain in optimizing model scaling and aligning LLMs with human values, crucial for enhancing 3D reasoning. Neural representation complexities and individual brain activity variability complicate cognitive task understanding [33]. Single-view image reliance and generic text descriptions often lead to ineffective 3D model representation.

Text-to-video (T2V) models highlight the need for benchmarks evaluating intuitive physics understanding, crucial for universal world simulators. Geometry Guided Self-Distillation (GGSD) leverages 2D and 3D data strengths, enhancing 3D representations. Innovations like the ULIP framework learn unified representations of language, images, and 3D point clouds, addressing small dataset limitations. The X-Dreamer approach uses pretrained 2D diffusion models for high-quality 3D content, bridging 2D and 3D synthesis gaps. These efforts lay the groundwork for sophisticated AI technologies in 3D content generation [15, 34, 35].

In recent years, the intersection of large language models (LLMs) and 3D representation techniques has garnered significant attention within the field of computer graphics and artificial intelligence. This paper aims to explore the advancements in these domains, particularly focusing on the role of LLMs in 3D content generation. To illustrate this complex relationship, Figure 2 provides a comprehensive overview of the hierarchical structure of advancements in these technologies. This figure details how LLMs facilitate the integration of textual descriptions with geometric data, thereby enhancing the generation of 3D models. Furthermore, it highlights key frameworks, methodologies, and challenges associated with these advancements, showcasing the transformative impact of these technologies in creating complex and high-quality 3D models. By analyzing these interactions, we can better understand the implications of LLMs in the realm of 3D representation and the future directions of this research area.

3 Large Language Models and 3D Representation

3.1 Role of Large Language Models in 3D Content Generation

Large Language Models (LLMs) have revolutionized 3D content generation by effectively bridging the gap between textual descriptions and complex geometric structures. The TextField3D framework

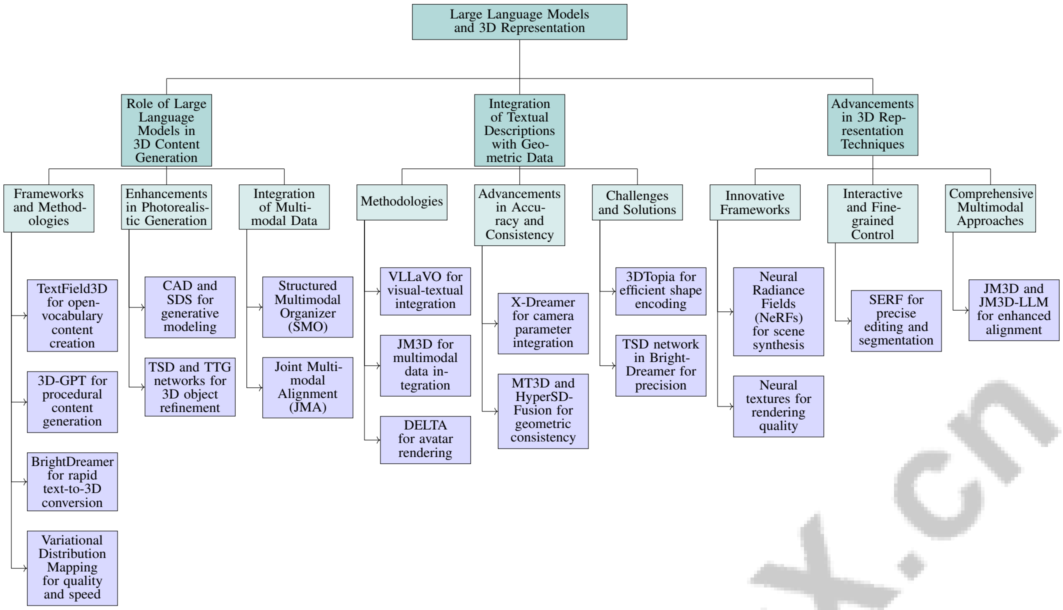

Large Language Models and 3D Representation
Role of Large Language Models in 3D Content Generation
Integration of Textual Descriptions with Geometric Data
Advancements in 3D Representation Techniques
Frameworks and Methodologies
Enhancements in Photorealistic Generation
Integration of Multimodal Data
Methodologies
Advancements in Accuracy and Consistency
Challenges and Solutions
Innovative Frameworks
Interactive and Fine-grained Control
Comprehensive Multimodal Approaches
TextField3D for open-vocabulary content creation
3D-GPT for procedural content generation
BrightDreamer for rapid text-to-3D conversion
Variational Distribution Mapping for quality and speed
CAD and SDS for generative modeling
TSD and TTG networks for 3D object refinement
Structured Multimodal Organizer (SMO)
Joint Multi-modal Alignment (JMA)
VLLaVO for visual-textual integration
JM3D for multimodal data integration
DELTA for avatar rendering
X-Dreamer for camera parameter integration
MT3D and HyperSD-Fusion for geometric consistency
3DTopia for efficient shape encoding
TSD network in Bright-Dreamer for precision
Neural Radiance Fields (NeRFs) for scene synthesis
Neural textures for rendering quality
SERF for precise editing and segmentation
JM3D and JM3D-LLM for enhanced alignment

Figure 2: This figure illustrates the hierarchical structure of advancements in large language models and 3D representation techniques, detailing the role of LLMs in 3D content generation, integration of textual descriptions with geometric data, and recent advancements in 3D representation techniques. It highlights key frameworks, methodologies, and challenges, showcasing the transformative impact of these technologies in creating complex and high-quality 3D models.

exemplifies this by mapping limited 3D data to dynamic textual fields, facilitating open-vocabulary content creation and enhancing model diversity [19]. Frameworks such as 3D-GPT advance procedural 3D content generation by interpreting user instructions and automating script generation for 3D software, significantly reducing manual intervention and streamlining modeling tasks through specialized agents [36, 16].

Figure 3 illustrates the role of Large Language Models in 3D content generation, highlighting key framework innovations, methodological enhancements, and multimodal integration approaches that contribute to advancements in the field. LLMs also enhance photorealistic 3D object generation through methodologies like Computer-Aided Design (CAD) and Score Distillation Sampling (SDS), addressing generative modeling challenges by decomposing complex queries and employing adversarial learning [29, 34, 37]. The BrightDreamer framework, for instance, converts text prompts into 3D Gaussian representations rapidly, utilizing networks like Text-guided Shape Deformation (TSD) and Text-guided Triplane Generator (TTG) to predict and refine 3D objects efficiently [38, 28, 39]. The Variational Distribution Mapping (VDM) method further enhances 3D asset creation by improving quality and speed while minimizing computational overhead.

The integration of multimodal data, including point clouds, text, and images, through innovations like the Structured Multimodal Organizer (SMO) and Joint Multi-modal Alignment (JMA), positions JM3D and JM3D-LLM as essential tools in AI-driven 3D modeling, addressing challenges such as information degradation and enhancing complex spatial information representation [15, 17, 18, 40].

3.2 Integration of Textual Descriptions with Geometric Data

Integrating textual descriptions with geometric data is essential for enhancing 3D representation. This integration leverages LLMs and computational techniques to unify symbolic and geometric information. The VLLaVO methodology illustrates this by generating textual descriptions for images using Vision Language Models (VLMs) and fine-tuning LLMs for image classification, effectively bridging visual and textual data [20].

Frameworks like JM3D demonstrate the power of multimodal integration by combining point clouds, text, and images, addressing issues of information degradation and ensuring 3D representations

AI-generated, for reference only.

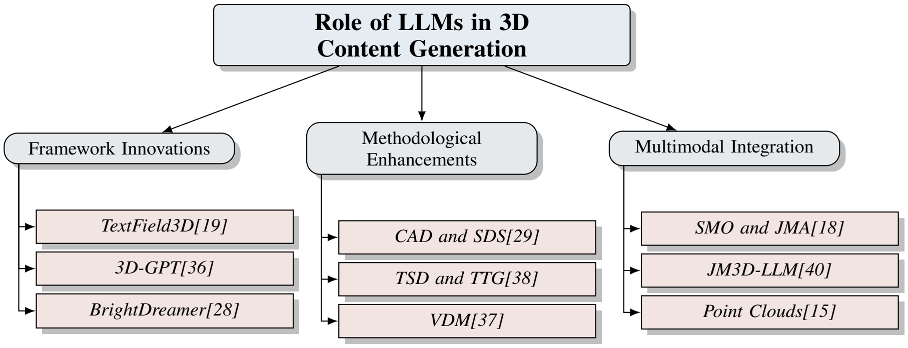


Figure 3: This figure illustrates the role of Large Language Models in 3D content generation, highlighting key framework innovations, methodological enhancements, and multimodal integration approaches that contribute to advancements in the field.

are informed by rich, multi-view data [41, 18, 40]. The DELTA framework showcases multimodal integration's versatility in avatar rendering and manipulation.

Incorporating camera parameters, as seen in the X-Dreamer framework, enhances 3D representation accuracy. Advancements in text-to-3D generation techniques, such as MT3D and HyperSDFusion, utilize depth maps and hierarchical representations to ensure geometric consistency and reduce viewpoint bias, leading to more controllable 3D content creation [12, 13, 14]. The MRD framework captures intra-modal and cross-modal relations, producing coherent and discriminative 3D shape representations.

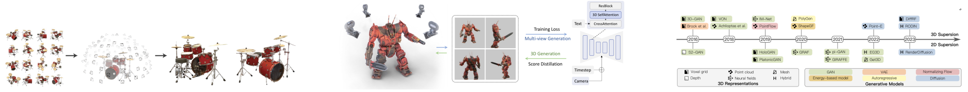
Challenges remain, such as ineffective explicit supervision leading to artifacts in shape synthesis. The 3DTopia method addresses these issues by encoding and rendering 3D shapes efficiently. The TSD network in BrightDreamer enhances 3D model generation by integrating textual descriptions with geometric data, improving flexibility and precision in text-to-3D synthesis [14, 28].

The integration of textual descriptions with geometric data is crucial for advancing 3D representation, with continuous advancements improving 3D model fidelity and coherence. Innovations like X-Dreamer and SceneWiz3D leverage neural representations and generative models to bridge gaps between 2D and 3D content creation, enhancing scene composition and texture editing for high-quality 3D assets [42, 13, 34, 35].

3.3 Advancements in 3D Representation Techniques

Recent advancements in 3D representation techniques are significantly driven by LLMs and innovative computational frameworks. Neural Radiance Fields (NeRFs) have emerged as transformative geometric primitives, enabling high-quality 3D scene synthesis from sparse input data [21]. Neural textures provide learned feature maps that enhance rendering quality and control, allowing nuanced manipulation of 3D objects [43].

The SERF framework enables precise 3D editing and segmentation without extensive training, enhancing interactivity and fine-grained control over 3D models [32]. The integration of LLMs and state-of-the-art computational methods, as seen in JM3D and JM3D-LLM, addresses challenges in 3D understanding by employing a comprehensive multimodal approach, enhancing alignment between point clouds, text, and images [17, 18, 35, 16]. These advancements push the boundaries of 3D modeling and visualization, paving the way for more complex and immersive digital experiences.



(a) Drum Kit Reconstruction and Visualization[21]

(b) Multi-view Generation and Score Distillation for 3D Robot Generation[44]

(c) Timeline of Generative Models in 3D and 2D Supervision[6]

Figure 4: Examples of Advancements in 3D Representation Techniques

As shown in Figure 4, significant advancements in large language models and 3D representation techniques illustrate the transformative impact of these technologies in creating and visualizing complex 3D models. Drum kit reconstruction showcases the ability to convert multiple 2D views into cohesive 3D models, highlighting the precision achievable with modern techniques. Similarly, the multi-view generation and score distillation approach for 3D robot generation enhances realism and detail. The timeline of generative models in 3D and 2D supervision provides an overview of milestones in this domain, underscoring continuous innovation in generative modeling. Collectively, these examples highlight the impact of LLMs and advanced 3D representation techniques across various applications, paving the way for more sophisticated digital representations [21, 44, 6].

4 Inverse Graphics and Text-to-3D

Category	Feature	Method
Concept and Application of Inverse Graphics	Structural Integrity	ViT3D[33], SERF[32], PG[3], HSF[12], PN++[24]
Challenges in Text-to-3D Generation	Cross-Domain Features	VLLaVO[20], VAL[1]
Advancements in Text-to-3D Techniques	Diversity and Richness Enhancement	TF3D[19], CAD[37], JM3D[18]
	Efficiency and Speed Enhancement	VDM[29], BD[28]
	Data Integration and Alignment	JM3D[17]
	Realism and Fidelity Improvement	G-SHELL[25], DELTA[45]
Diffusion Models in Text-to-3D	Temporal Sampling Strategies	TP-SDS[46]
	Image and View Integration	MVD[44], LGM[39], DE3DG[22]
	Sequential Attribute Generation	SALAD[47]

Table 1: This table provides a comprehensive summary of various methods and their applications in the field of inverse graphics and text-to-3D generation. It categorizes these methods into four main areas: the concept and application of inverse graphics, challenges in text-to-3D generation, advancements in text-to-3D techniques, and diffusion models in text-to-3D. Each category highlights specific features and methods, showcasing the diverse approaches and innovations driving progress in this domain.

Category	Feature	Method
Concept and Application of Inverse Graphics	Structural Integrity	ViT3D[33], SERF[32], PG[3], HSF[12], PN++[24]
Challenges in Text-to-3D Generation	Cross-Domain Features	VLLaVO[20], VAL[1]
Advancements in Text-to-3D Techniques	Diversity and Richness Enhancement	TF3D[19], CAD[37], JM3D[18]
	Efficiency and Speed Enhancement	VDM[29], BD[28]
	Data Integration and Alignment	JM3D[17]
	Realism and Fidelity Improvement	G-SHELL[25], DELTA[45]
Diffusion Models in Text-to-3D	Temporal Sampling Strategies	TP-SDS[46]
	Image and View Integration	MVD[44], LGM[39], DE3DG[22]
	Sequential Attribute Generation	SALAD[47]

Table 2: This table provides a comprehensive summary of various methods and their applications in the field of inverse graphics and text-to-3D generation. It categorizes these methods into four main areas: the concept and application of inverse graphics, challenges in text-to-3D generation, advancements in text-to-3D techniques, and diffusion models in text-to-3D. Each category highlights specific features and methods, showcasing the diverse approaches and innovations driving progress in this domain.

The exploration of inverse graphics as a foundational element in the text-to-3D domain reveals a multifaceted approach to understanding and reconstructing three-dimensional representations from various input modalities. This section delves into the concept and applications of inverse graphics, highlighting its significance in facilitating the generation of detailed 3D models from textual descriptions. Table 2 presents an organized overview of the methods and features pertinent to inverse graphics and text-to-3D generation, delineating the key advancements and challenges in this rapidly evolving field. By examining the methodologies employed and the practical implications of inverse graphics, we can better appreciate its role in enhancing the efficiency and accuracy of text-to-3D generation processes. Additionally, Table 4 presents another organized overview of the methods and features pertinent to inverse graphics and text-to-3D generation, further delineating the key advancements and challenges in this rapidly evolving field.

4.1 Concept and Application of Inverse Graphics

Inverse graphics represents a transformative paradigm in the realm of computer graphics and artificial intelligence, focusing on the reconstruction and interpretation of three-dimensional scenes from two-dimensional images or textual descriptions. This approach involves decomposing visual data into fundamental components such as shape, material, and lighting, thereby enabling the creation of detailed and manipulable 3D models. The generation of high-quality 3D shapes from text prompts necessitates an understanding of the hierarchical structure inherent in both the text and the shapes, as highlighted in recent studies [12].

Practical applications of inverse graphics are diverse and impactful, particularly in the domain of text-to-3D generation. For instance, interactive 3D segmentation is crucial for converting user inputs into detailed 3D models, allowing for fine-grained control over the resulting representations [32]. Additionally, the process of decoding visual stimuli from brain signals can be viewed as a form of inverse graphics, where complex neural signals are translated into 3D representations, thereby enhancing our understanding of neural processing and its applications in AI [33].

In the context of semantic segmentation, the generation of labeled synthetic datasets using simulators like CARLA plays a pivotal role in training models on real-world point clouds, exemplifying the utility of inverse graphics in creating accurate 3D models from multimodal inputs [24]. The inefficiency of traditional text-to-3D generation methods, which often require extensive optimization for each text prompt, underscores the importance of developing more efficient inverse graphics techniques [28].

Furthermore, inverse graphics is applied in the ParticleGrid framework, where molecular representations are converted into usable 3D models, demonstrating practical applications in scientific visualization and materials science [3]. The VLLaVO method further illustrates the potential of inverse graphics by leveraging large language models to generalize across domains, focusing on domain-invariant features extracted from textual descriptions [20].

Overall, inverse graphics serves as a pivotal concept in advancing the text-to-3D domain, providing robust methodologies for reconstructing and understanding complex 3D environments from minimal input. The applications of advanced AI and computer graphics technologies are extensive, encompassing diverse fields such as digital storytelling, where tools like Story3D-Agent enable the transformation of narratives into dynamic 3D visualizations; navigation systems that benefit from improved 3D model generation through cutting-edge techniques; virtual reality experiences enhanced by high-quality 3D content creation methods; and scientific visualization, which utilizes sophisticated 3D representations to illustrate complex data. This broad applicability highlights the critical role these technologies play in the ongoing advancement of both AI and computer graphics. [34, 48, 35]

4.2 Challenges in Text-to-3D Generation

The generation of 3D models from textual descriptions presents several formidable challenges, primarily due to the intricate task of aligning linguistic data with geometric representations. One significant obstacle is the inefficiency of per-prompt optimization, which necessitates numerous iterations to produce a single 3D object, as highlighted by the BrightDreamer framework [28]. This inefficiency is compounded by the limitations of Score Distillation Sampling (SDS), which often results in inaccurate and low-fidelity 3D assets from text inputs [29].

Figure 5 illustrates the primary challenges in text-to-3D generation, categorizing them into inefficiency issues, model limitations, and cross-domain challenges. It highlights inefficiencies in per-prompt optimization and Score Distillation Sampling, model limitations related to non-watertight meshes and efficiency trade-offs, and cross-domain challenges involving domain-invariant features and stereotype outputs.

Moreover, existing methods frequently rely on unsigned distance fields (UDF) for non-watertight meshes, complicating the extraction process and introducing modeling errors [25]. The communication overhead in large-scale distributed training further exacerbates these challenges, leading to diminishing returns in performance [2]. Additionally, the trade-offs between model size and efficiency often go unaddressed, posing significant computational resource requirements [4].

The exclusive focus on image modalities in existing methods limits their ability to learn domain-invariant features, resulting in sub-optimal performance in cross-domain tasks [20]. This is particularly problematic in video understanding, where models are often task-specific and fail to integrate

AI-generated, for reference only.

video and language understanding effectively [1]. Furthermore, the tendency of current methods to produce stereotype outputs when trained on small datasets restricts their ability to achieve open-vocabulary 3D content understanding [19].

These challenges highlight the urgent need for innovative solutions that can significantly improve the fidelity, efficiency, and scalability of text-to-3D generation processes. Recent advancements, such as the Variational Distribution Mapping (VDM) and Distribution Coefficient Annealing (DCA) techniques, have shown promise in enhancing the quality of generated 3D assets by addressing issues like geometric inconsistencies and viewpoint bias. Additionally, methods like X-Dreamer are bridging the domain gap between 2D and 3D representations, ensuring that generated content aligns better with textual descriptions and camera perspectives. Collectively, these developments underscore the necessity for ongoing research and refinement in the field to overcome existing limitations and leverage the full potential of text-to-3D generation technologies [29, 34, 35, 14]. Addressing these issues is crucial for advancing the field and enabling the development of robust AI systems capable of producing high-quality 3D models from textual descriptions. Continued research and innovation are essential to overcome these obstacles and drive progress in text-to-3D conversion.

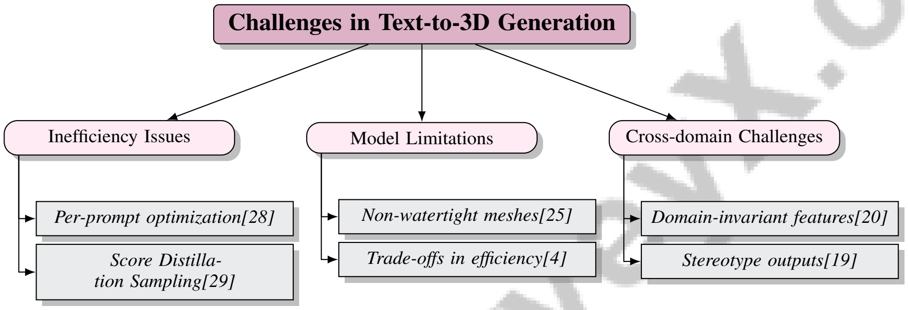

Challenges in Text-to-3D Generation
Inefficiency Issues
Model Limitations
Cross-domain Challenges
Per-prompt optimization[28]
Score Distillation Sampling[29]
Non-watertight meshes[25]
Trade-offs in efficiency[4]
Domain-invariant features[20]
Stereotype outputs[19]

Figure 5: This figure illustrates the primary challenges in text-to-3D generation, categorized into inefficiency issues, model limitations, and cross-domain challenges. It highlights inefficiencies in per-prompt optimization and Score Distillation Sampling, model limitations related to non-watertight meshes and efficiency trade-offs, and cross-domain challenges involving domain-invariant features and stereotype outputs.

4.3 Advancements in Text-to-3D Techniques

Method Name	Methodological Innovations	Optimization Efficiency	Application Domains
CAD[37]	Adversarial Learning	Fast Generation	Photorealistic Object Generation
BD[28]	Single-stage Framework	77 MS	Complex Semantic Prompts
VDM[29]	Trainable Degradation Process	Enhancing Optimization Efficiency	High-fidelity 3D
G-SHELL[25]	Jointly Parameterize Meshes	Efficient Mesh Extraction	Complex Geometries
TF3D[19]	Noisy Text Fields	Lower Latency	Open-vocabulary Generation
DELTA[45]	Hybrid Explicit-implicit	Differentiable Renderer Optimization	Virtual Environments
JM3D[18]	Dual-module Approach	Contrastive Learning	Zero-shot Classification
JM3D[17]	Structured Multimodal Organizer	Joint Multi-modal Alignment	3D Classification Tasks

Table 3: Comparison of recent advancements in text-to-3D generation methods, highlighting methodological innovations, optimization efficiency, and application domains. The table includes methods such as CAD, BrightDreamer, and VDM, each offering unique contributions to the field, from adversarial learning to hybrid explicit-implicit approaches. These methods address key challenges in photorealistic object generation, complex semantic prompts, and high-fidelity 3D modeling.

Recent advancements in text-to-3D generation have been driven by innovative methodologies that enhance the quality, fidelity, and efficiency of transforming textual descriptions into three-dimensional models. The CAD method significantly improves rendering quality and diversity compared to traditional Score Distillation Sampling (SDS) pipelines, effectively addressing limitations in previous methods related to oversaturation and lack of variety in generated outputs [37]. This advancement is crucial for applications requiring photorealistic 3D object generation.

The introduction of BrightDreamer marks a significant leap in text-to-3D techniques by offering a single-stage framework capable of generating 3D Gaussians rapidly and effectively, thereby addressing inefficiencies in traditional multi-step processes [28]. This innovation highlights the importance of streamlined approaches in enhancing the speed and quality of 3D content creation.

Furthermore, the Variational Distribution Mapping (VDM) method introduces a trainable degradation process that eliminates the need for complex Jacobian calculations in diffusion models, greatly enhancing optimization efficiency and the quality of 3D generation [29]. This approach not only simplifies the computational process but also improves the fidelity of the generated models.

G-SHELL demonstrates superior performance in reconstructing and generating non-watertight meshes, effectively handling complex geometries that are challenging for traditional methods [25]. This capability is particularly beneficial for applications requiring detailed and accurate 3D models.

TextField3D showcases significant advancements in open-vocabulary 3D generation by effectively mapping limited data to a broader range of textual concepts through the use of Noisy Text Fields, thereby enhancing the diversity and richness of generated 3D content [19]. This method underscores the potential of integrating linguistic flexibility into 3D modeling processes.

In addition to these advancements, future research directions aim to improve segmentation techniques, enhance the modeling of dynamics for clothing and hair, and explore the integration of lighting and material properties to further refine the realism and applicability of text-to-3D generation [45]. These efforts are essential for overcoming current limitations and expanding the scope of text-to-3D applications across diverse domains.

Overall, these advancements reflect a concerted effort to refine text-to-3D generation techniques, addressing key challenges such as geometric consistency, fidelity, and multimodal integration. Ongoing research and development in the field of text-to-3D generation is advancing the capabilities of artificial intelligence, with innovative approaches such as HyperSDFusion, X-Dreamer, SceneWiz3D, and DreamMapping. These methods leverage hierarchical structures, bridge the domain gap between 2D and 3D content, enhance scene composition, and refine distribution modeling techniques. As a result, they are paving the way for the creation of sophisticated AI systems that can seamlessly transform textual descriptions into high-quality, detailed 3D representations, significantly enhancing the realism and fidelity of generated assets [12, 34, 13, 29]. Table 3 provides a comprehensive overview of these recent advancements in text-to-3D generation techniques, detailing the methodological innovations, optimization efficiencies, and application domains of various methods.

4.4 Diffusion Models in Text-to-3D

Diffusion models have emerged as a central component in the text-to-3D conversion process, offering robust frameworks for generating detailed and consistent three-dimensional representations from textual descriptions. The application of diffusion models in this domain is exemplified by methods such as the Large Multi-view Gaussian (LGM) approach, which predicts and fuses 3D Gaussians from multi-view images to create comprehensive 3D models [39]. This technique leverages the strengths of multi-view data to enhance the geometric accuracy and detail of the resulting 3D representations.

The SALAD framework operates through a two-phase diffusion process, initially generating extrinsic parameters followed by intrinsic attributes, thereby facilitating detailed shape representation [47]. This structured approach ensures that both the form and finer details of 3D objects are accurately captured, contributing to the overall fidelity of the generated models.

Incorporating pre-trained 2D diffusion models, the method outlined by Wu et al. predicts multi-view depth maps alongside RGB and Gaussian features, effectively bridging the gap between 2D and 3D data [22]. This integration allows for the direct generation of explicit 3D content, enhancing the coherence and visual quality of the outputs.

MVDream illustrates the effectiveness of learning from both 2D and 3D data, maintaining generalizability while producing consistent multi-view outputs [44]. This capability is crucial for ensuring that the generated 3D models retain their integrity across various perspectives, which is essential for applications requiring high levels of detail and accuracy.

The TP-SDS method optimizes 3D content generation by prioritizing the sampling of timesteps based on the stage of optimization, thereby improving the alignment between the 3D model and the diffusion process [46]. This targeted approach enhances the efficiency of the generation process, reducing computational overhead while maintaining high-quality outputs.

The diffusion model-based techniques collectively underscore the revolutionary influence of sophisticated computational methods in text-to-3D conversion, effectively tackling significant challenges such as ensuring geometric consistency and integrating multimodal data. For instance, MT3D employs depth maps from high-fidelity 3D models to mitigate viewpoint bias and enhance geometric understanding, while DreamMapping introduces Variational Distribution Mapping (VDM) and Distribution Coefficient Annealing (DCA) to refine the generation process, resulting in high-quality, realistic 3D assets and improved optimization efficiency. [29, 14]. The continued refinement of these models is pivotal for advancing the capabilities of AI systems in generating sophisticated and realistic 3D content from textual inputs.

Feature	Inverse Graphics	BrightDreamer	Variational Distribution Mapping (VDM)
Optimization Approach	Decomposition-based	Single-stage	Trainable Degradation
Model Focus	3D Reconstruction	3D Gaussians	Diffusion Models
Key Advantage	Detailed Models	Rapid Generation	Improved Fidelity

Table 4: This table provides a comparative analysis of three distinct methodologies in the text-to-3D domain: Inverse Graphics, BrightDreamer, and Variational Distribution Mapping (VDM). It highlights the optimization approaches, model focuses, and key advantages of each method, offering insights into their contributions to the generation of three-dimensional representations from textual descriptions. Such a comparison is crucial for understanding the diverse strategies employed in enhancing the efficiency and fidelity of text-to-3D conversion processes.

5 Multimodal Information Fusion

The integration of diverse data modalities is crucial for enhancing 3D understanding and reasoning, forming the basis for creating comprehensive models that leverage varied inputs to improve the fidelity and coherence of 3D representations. This section explores methodologies and frameworks facilitating this integration, underscoring their significance in advancing AI systems' capabilities.

5.1 Integration of Diverse Data Modalities

Integrating diverse data modalities is essential for advancing 3D understanding and reasoning. The CAD method exemplifies this by utilizing pose pruning to stabilize GAN training, enhancing 3D content fidelity [37]. The PCA-Bench dataset supports this integration by providing diverse multimodal instances, such as images and action candidates, crucial for enhancing 3D reasoning [7]. Similarly, the WanJuan benchmark offers a high-quality dataset for multimodal tasks, facilitating model training and evaluation [49].

Practical applications, like the Valley model, demonstrate effective integration of video, images, and language, enhancing understanding and reasoning in complex scenes [1]. Additionally, synthetic data use in multimodal semantic segmentation reduces manual labeling needs, improving model training efficiency and 3D representation robustness [24].

Overall, integrating diverse data modalities is vital for advancing 3D reasoning, requiring innovative methodologies to combine and interpret varied inputs. Recent innovations, such as X-Dreamer and SceneWiz3D, leverage techniques like Camera-Guided Low-Rank Adaptation to bridge 2D and 3D synthesis gaps, improving generated models' quality [13, 34, 35, 40].

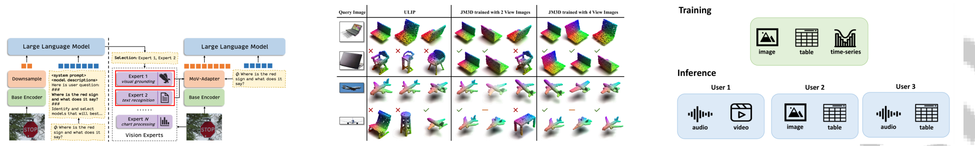
5.2 Applications and Advancements in Multimodal Fusion

Advancements in multimodal fusion, driven by integrating text, images, and 3D models, have expanded AI systems' capabilities in understanding and reasoning across domains. Language models' exploration for multimodal tasks highlights their potential in enhancing reasoning capabilities, as demonstrated by benchmarks showcasing effective data integration [50].

A key application is zero-shot learning, where multimodal feature integration enhances model generalization without specific training. The JM3D framework exemplifies this by combining point cloud, text, and image modalities, boosting zero-shot learning performance [40]. Bridging modality gaps to enhance generalizability and performance is also a focus, with methods like Text-Centric Alignment for Multimodality Learning (TAMML) improving data type integration [31].

Comprehensive datasets, like one containing 73,000 images with hierarchical captions, advance multimodal fusion by assessing lexical and textual entailment, providing insights into hierarchical modality relationships [8]. Research into Multimodal Large Language Models (MLLMs) explores vulnerabilities and defense strategies to enhance security and robustness, ensuring AI systems' reliability [51].

Innovative methodologies and extensive datasets in multimodal fusion enhance AI systems' ability to integrate and interpret diverse data modalities. This progress is evident in MLLMs, surpassing human baselines in complex tasks involving image-text comprehension and reasoning, highlighting their potential in applications like visual storytelling and geospatial analysis [9, 31, 52, 53].



(a) Large Language Model-Assisted Vision Expert System[11]

(b) Comparison of 3D Reconstruction Results Using ULIP and JM3D with Different View Images[40]

(c) Training and Inference of Machine Learning Models[31]

Figure 6: Examples of Applications and Advancements in Multimodal Fusion

Figure 6 illustrates significant advancements in multimodal fusion. The first example shows a Large Language Model-Assisted Vision Expert System, integrating language models with vision experts for efficient image processing and query response. The second example compares ULIP and JM3D 3D reconstruction techniques, highlighting multimodal inputs' importance in enhancing modeling accuracy. The third example depicts machine learning model training and inference, emphasizing their versatility in handling varied data types like images, tables, and time-series during training and extending applicability to diverse inputs like audio and video during inference. These examples underscore multimodal fusion's transformative potential in enhancing machine learning systems' capabilities [11, 40, 31].

5.3 Challenges and Solutions in Multimodal Fusion

Multimodal fusion presents challenges in integrating and interpreting diverse data modalities, necessitating innovative solutions to enhance AI systems' efficacy and robustness. A significant challenge is detecting and mitigating malicious information in images, which can manipulate model outputs and compromise AI systems' integrity [51]. Theoretical constructs of generative models may not fully capture real-world complexities, leading to discrepancies between predictions and outcomes [54].

Another challenge is models' limited robustness to rare categories, hindering performance in cross-modal applications where some data types may be underrepresented in training [15]. Addressing this requires developing sophisticated models capable of generalizing across underrepresented categories and integrating diverse data sources.

Solutions include implementing advanced defense mechanisms to detect and counteract malicious inputs, safeguarding multimodal systems' integrity. Future research should focus on enhancing model robustness by exploring novel architectures and training methodologies to improve AI systems' generalization capabilities, particularly in handling rare categories and cross-modal tasks [15].

Advancing multimodal fusion requires addressing challenges associated with modality mismatch and diverse data forms integration, like text, images, audio, and video. Ongoing research and innovation are crucial for developing robust AI systems that can seamlessly interpret and adapt to varying modalities, especially in real-world applications where data types may be dynamic and unpredictable. Recent advancements, including the use of Large Language Models and multimodal pretrained models, show promise in enhancing systems' generalizability and performance, paving the way for solutions transcending traditional fixed-modality frameworks [53, 31].

AI-generated, for reference only.

6 Physics-Based Modeling and Efficiency

The demand for realistic and interactive 3D representations emphasizes the importance of physics-based modeling, which enhances visual fidelity and user experiences by simulating real-world physics. The following subsections explore how this modeling achieves heightened realism and its implications for 3D technologies and applications.

6.1 Realism through Physics-Based Modeling

Physics-based modeling is crucial for achieving realism in 3D representations by simulating physical properties and interactions within virtual environments, essential for virtual reality, gaming, and simulation applications. For example, the ControlRoom3D framework uses semantic proxies to enhance 3D room realism, creating detailed and contextually rich environments [55]. Neural textures in deferred neural rendering techniques address geometric imperfections, providing flexibility across applications [43]. This flexibility ensures high-quality visualizations despite imperfect geometries, with efficient end-to-end training processes generating high-quality outputs with reduced computational overhead.

The JM3D framework exemplifies the integration of multi-modal information to enhance 3D models' cohesiveness and detail, leveraging diverse data sources for realistic representations that enrich user immersion [17]. The MVGamba method ensures multi-view information integrity, enabling efficient generation of high-quality 3D content with reduced model sizes, crucial for consistent representations across perspectives [26]. Adaptable systems like ARAA maintain optimal performance under varying conditions, contributing to realism and efficiency by dynamically adjusting to environmental factors [56].

Advancements in neural rendering techniques, such as Neural Radiance Fields (NeRFs), enable accurate view synthesis by representing scenes as continuous volumes. The integration of multimodal cues from 2D vision-language models addresses challenges related to data scarcity and information degradation. Systems like 3D-IntPhys enhance visual intuitive physics modeling in complex scenes, improving manipulation predictions without dense point tracking, significantly elevating virtual environments' fidelity [21, 57, 40]. These developments highlight the necessity of sophisticated computational techniques for accurate real-world physics simulation, advancing AI-driven 3D modeling and visualization capabilities.

6.2 Optimizing Computational Models

Optimizing computational models is essential for enhancing performance and efficiency in AI and 3D representation. The Large Autoregressive Model (LARM) exemplifies this optimization, efficiently predicting actions with high accuracy, significantly outperforming previous methods in task completion and speed [58]. Such improvements are critical for real-time processing and decision-making applications.

The Text-Centric Alignment for Multimodality Learning (TAMML) method enhances optimization by using text as a modality-invariant representation, reducing semantic gaps and improving generalizability across modalities [31]. By aligning diverse data types through a unified textual framework, TAMML facilitates seamless integration and processing, essential for optimizing multimodal tasks.

Adaptive algorithms, as seen in the ANDA method, dynamically adjust to data structures, enhancing performance and efficiency [59]. This adaptability is particularly beneficial in scenarios with data variability, allowing models to maintain accuracy and speed without excessive computational resources.

The optimization of computational models involves advanced methodologies that enhance performance and efficiency. By employing adaptive algorithms, utilizing modality-invariant representations, and improving predictive capabilities, these strategies empower AI systems to address intricate tasks and accommodate diverse data inputs. This comprehensive approach advances 3D representation technologies and addresses challenges such as modality mismatch and information degradation in multimodal learning, as demonstrated by innovations in models like AdaptVision and TAMM, which optimize processing for improved scene understanding and 3D shape comprehension [41, 31, 10, 40].

6.3 Efficiency and Interactivity in 3D Modeling

Efficiency and interactivity are crucial in 3D modeling, where swift and accurate manipulation of complex geometric data is essential for applications like virtual reality, gaming, and digital content creation. Advanced computational techniques and neural architectures significantly enhance the efficiency of 3D modeling processes. For instance, neural textures and deferred neural rendering techniques efficiently handle imperfect geometry, providing flexibility and high-quality visual outputs across various applications [43]. This optimization not only improves rendering processes but also enhances interactivity by reducing computational overhead and supporting real-time performance.

Frameworks like JM3D, which integrate multi-modal information, further emphasize the importance of efficiency in 3D modeling. By leveraging diverse data sources, these frameworks ensure that 3D models are both realistic and contextually accurate, thereby enhancing user interaction and engagement [17]. Additionally, methods like MVGamba maintain multi-view information integrity, enabling efficient generation of high-quality 3D content with reduced model sizes, vital for consistent representations across multiple perspectives [26].

Interactivity is further enhanced through adaptive algorithms and dynamic systems that adjust to varying environmental factors and data structures. Systems like ARAA exemplify this adaptability by maintaining optimal performance under fluctuating conditions, ensuring that 3D models can realistically respond to changes in lighting, texture, and other variables [56]. This adaptability is essential for creating interactive 3D environments that dynamically respond to user inputs and environmental changes, improving the overall user experience.

Optimizing computational models, particularly through techniques like LARM and TAMML, significantly enhances the efficiency and interactivity of 3D modeling processes. LARM generates coherent and contextually relevant outputs, while TAMML addresses modality mismatches by leveraging large language models to improve adaptability and generalizability in multimodal systems. This dual approach mitigates challenges such as information degradation and insufficient synergy in 3D shape understanding, enriching 3D data representation by integrating insights from diverse modalities, ultimately leading to more effective and nuanced 3D modeling capabilities [41, 31, 18].

The ongoing advancement of innovative methodologies and adaptive systems is crucial for enhancing the efficiency and interactivity of 3D modeling processes, especially given the rapid evolution of AI-driven technologies. Recent developments, including advanced neural representations and generative models, are improving the quality and diversity of 3D content generation. For example, the ITEM3D model shows how integrating diffusion models and differentiable rendering can effectively bridge the gap between text instructions and 3D representation, facilitating more precise texture editing. Such innovations streamline workflows and enable sophisticated applications in 3D modeling, ensuring that emerging technologies can meet the demands of contemporary digital landscapes [42, 35].

6.4 Enhancing User Engagement through Optimization

Enhancing user engagement in 3D modeling systems requires optimizing various aspects of the modeling process to create interactive and immersive experiences. A key strategy involves implementing adaptive algorithms that dynamically respond to user inputs and environmental changes, improving the interactivity and responsiveness of 3D environments. Systems like ARAA exemplify this adaptability, maintaining optimal performance under fluctuating conditions to ensure realistic responses to changes in lighting, texture, and other variables [56].

The integration of neural textures and deferred neural rendering techniques significantly enhances user engagement by efficiently handling imperfect geometry and providing flexibility across applications [43]. These techniques optimize rendering processes, reduce computational overhead, and improve real-time performance, which is vital for maintaining user interest and interaction.

Frameworks like JM3D, which integrate multi-modal information, ensure that 3D models are both realistic and contextually accurate, thereby enhancing user interaction and engagement [17]. By leveraging diverse data sources, these frameworks create immersive and engaging 3D environments that capture user attention.

Optimizing computational models through methods like LARM and TAMML further enhances user engagement by improving predictive capabilities and aligning diverse data types through unified frameworks. These strategies empower AI systems to manage intricate tasks and a wide array of data inputs, facilitating significant progress in 3D representation technologies. The introduction of advanced neural representations and generative models has accelerated high-quality 3D content generation, as evidenced by frameworks like X-Dreamer, which bridge the gap between text-to-2D and text-to-3D synthesis. Innovative approaches such as Point Prompt Training (PPT) leverage multiple datasets to improve 3D representation learning, while models like ULIP integrate multimodal information to bolster 3D understanding, showcasing the transformative impact of these methodologies across various applications in the field [15, 60, 34, 35].

To effectively enhance user engagement in 3D modeling systems, a comprehensive strategy is essential. This should integrate adaptive algorithms that leverage advanced neural representations, such as Neural Radiance Fields (NeRFs), to optimize scene representation and rendering. Employing efficient rendering techniques, including differentiable rendering and hybrid 3D representations, will facilitate the seamless integration of diverse data modalities, ensuring high fidelity and user-friendly interaction in creating and manipulating complex 3D environments [42, 21, 13, 35]. These efforts are vital for creating interactive and immersive 3D environments that captivate and retain user interest, ultimately advancing the capabilities of AI-driven 3D modeling and visualization.

7 Fine-Grained Control and Semantic Chasm

7.1 Techniques for Fine-Grained Control

Achieving fine-grained control over 3D digital objects enhances modeling precision and interactivity, crucial for applications that demand detailed manipulation of complex structures. The SERF framework exemplifies this by enabling precise 3D editing and segmentation without extensive training, facilitating interactive modifications that improve accuracy and responsiveness in 3D environments [32]. Interactive 3D segmentation techniques decompose objects into manageable components, allowing users to adjust attributes like texture, shape, and material properties for high customization. ITEM3D, for instance, employs diffusion models and differentiable rendering to optimize texture editing based on text instructions, addressing the complexities of 3D model manipulation [42, 35].

Adaptive algorithms and dynamic systems enhance real-time adjustments in 3D models, ensuring precise updates. The OneTo3D framework uses a Gaussian Splatting model to generate editable 3D models from single images, incorporating automatic generation and self-adaptive binding for object armatures. Similarly, the AdaptVision model dynamically adjusts visual token inputs based on image resolution and content, enhancing scene understanding and real-time modifications [61, 10]. Such adaptability is vital for applications requiring responsive control over digital objects, such as virtual reality and digital content creation.

Advancing techniques for fine-grained control is integral to AI-driven 3D modeling, enabling more detailed and interactive user experiences. Recent advancements highlight the role of sophisticated computational techniques in accurately manipulating intricate geometric data. Methods like Score Distillation Sampling (SDS) and the MT3D model illustrate how depth maps and deep geometric moments can address viewpoint bias and geometric inconsistencies, enhancing the quality and diversity of generated 3D models [35, 14].

7.2 Bridging the Semantic Chasm

Bridging the semantic chasm between high-level semantic understanding and detailed geometric representation is a critical challenge in 3D modeling. This involves reconciling abstract semantic information with precise geometric data. The SERF framework offers a promising approach by integrating multi-view reconstruction with 2D features to create a neural mesh representation, effectively bridging the gap and resulting in more accurate 3D models [32]. Multi-view reconstruction techniques capture intricate details of 3D objects, synthesizing diverse perspectives into cohesive models that enhance visual quality and rendering efficiency. The MVGamba framework, for instance, leverages multi-view Gaussian features and RNN-like State Space Models to improve information propagation, addressing challenges like multi-view inconsistency [15, 21, 39, 40, 26].

Neural mesh representations contribute to bridging the semantic chasm by providing flexible frameworks for modeling complex shapes. Advanced 3D representations encode complex geometric details while maintaining strong links to high-level semantic attributes. Innovations like the MT3D model and HyperSDFusion enhance shape generation, improving quality and usability across fields such as gaming, architecture, and simulation [62, 13, 12, 14, 6].

Effective strategies for bridging the semantic chasm are essential for enhancing AI-driven 3D modeling systems. Advances like HyperSDFusion and JM3D highlight the importance of sophisticated semantic alignment in evolving 3D modeling technologies. Initiatives like Story3D-Agent demonstrate the potential of large language models in creating complex 3D visual narratives, underscoring the significance of integrating semantic and geometric data to enhance the fidelity and coherence of 3D models [12, 18, 48].

7.3 Challenges in Multimodal Integration

Integrating multiple modalities in AI systems presents challenges that complicate the seamless fusion of diverse data types, necessitating innovative solutions to enhance system robustness and performance. Aligning and synchronizing data from different modalities, which possess distinct temporal and spatial characteristics, is a primary challenge. Misalignment can lead to discrepancies in data interpretation and hinder multimodal system performance [51]. Variability in data quality and representation across modalities further complicates integration, as visual data may suffer from noise while textual data might contain ambiguities [20]. Additionally, processing large volumes of multimodal data poses computational challenges, particularly in real-time applications [4].

Modality-specific biases further complicate integration, as models may inadvertently prioritize certain modalities, leading to skewed interpretations [31]. Solutions include developing advanced alignment techniques to synchronize data, enhancing data quality through robust preprocessing, and leveraging adaptive algorithms and scalable architectures to address computational challenges, ensuring efficient processing in real-time scenarios [59].

Addressing challenges in multimodal integration is critical for advancing AI system capabilities. By integrating cutting-edge computational techniques, researchers can improve multimodal system robustness, enabling advanced applications such as visual question answering and geo-localization, while addressing key challenges like modality mismatch and effective instruction tuning [9, 53, 31, 52].

7.4 Innovative Model Architectures

Innovative model architectures are crucial for addressing the semantic chasm and enhancing control over 3D digital environments. Key advancements include developing adaptive frameworks that integrate high-level semantic understanding with detailed geometric representation. The SERF framework exemplifies this integration by leveraging multi-view reconstruction and 2D features to create coherent 3D models, bridging the gap between semantic and geometric data [32]. Neural radiance fields (NeRFs) have emerged as transformative geometric primitives, offering flexible representations for synthesizing high-quality 3D scenes from sparse input data, enhancing realism and detail in visualizations [21].

The integration of neural textures and deferred neural rendering techniques enhances model architectures' capability to handle imperfect geometry and maintain high-quality visual outputs, ensuring precise manipulation of 3D models [43]. Frameworks like JM3D underscore the importance of leveraging diverse data sources to improve the fidelity of 3D representations, ensuring models are both semantically meaningful and geometrically precise [17].

Innovative model architectures are essential for advancing AI-driven 3D modeling systems. By integrating semantic information with geometric data, these architectures improve the accuracy and consistency of 3D models, addressing challenges like information degradation and viewpoint bias. This integration facilitates the creation of sophisticated digital experiences in fields like computer vision, robotics, and virtual environments [15, 13, 40, 14, 18].

AI-generated, for reference only.

8 Conclusion

8.1 Challenges and Future Directions

The fusion of large language models with 3D representations presents multifaceted challenges that necessitate ongoing exploration and innovation. Key among these is the need to optimize the efficiency and stability of 3D generative models to support a wide array of applications. Future research should focus on advancing generative modeling techniques and exploring more efficient architectures to enhance 3D modeling capabilities. Strengthening the robustness of models against domain-specific interruptions and expanding their applicability across various learning tasks is essential.

In materials science, the challenge of accurately modeling molecular properties persists, particularly in refining grid representations to better predict material behaviors. Expanding the integration of audio inputs and enhancing multilingual capabilities in models can significantly improve performance across diverse scenarios, addressing current multimodal integration limitations.

Further exploration of parallelization techniques may bolster the integration of LLMs with 3D representations, countering the diminishing returns seen in hardware scaling trends. Improving model efficiency, investigating novel training paradigms, and addressing ethical considerations related to LLM usage remain critical avenues for research.

Expanding training datasets to include a broader spectrum of concepts, particularly those involving actions or interactions, is crucial for enhancing the generative capabilities of models. Such expansion will enable the generation of more complex and varied 3D representations.

Addressing these challenges and pursuing these future research directions will be instrumental in developing advanced AI systems that seamlessly integrate linguistic and geometric data, ultimately enhancing the fidelity and interactivity of digital environments.

AI-generated, for reference only.

References

[1] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

[2] Jared Fernandez, Luca Wehrstedt, Leonid Shamis, Mostafa Elhoushi, Kalyan Saladi, Yonatan Bisk, Emma Strubell, and Jacob Kahn. Hardware scaling trends and diminishing returns in large-scale distributed training, 2024.

[3] Shehtab Zaman, Ethan Ferguson, Cecile Pereira, Denis Akhiyarov, Mauricio Araya-Polo, and Kenneth Chiu. Particlegrid: Enabling deep learning using 3d representation of materials, 2022.

[4] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[5] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public, 2023.

[6] Zifan Shi, Sida Peng, Yinghao Xu, Andreas Geiger, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey, 2023.

[7] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi Meng, Tianyu Liu, and Baobao Chang. Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain, 2024.

[8] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations, 2024.

[9] Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, Yingzi Ma, Situo Zhang, Zihan Zhao, Liangtai Sun, and Kai Yu. Multi: Multimodal understanding leaderboard with text and images, 2025.

[10] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[11] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context, 2024.

[12] Zhiying Leng, Tolga Birdal, Xiaohui Liang, and Federico Tombari. Hypersdfusion: Bridging hierarchical structures in language and geometry for enhanced 3d text2shape generation, 2024.

[13] Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Scenewiz3d: Towards text-guided 3d scene composition, 2023.

[14] Utkarsh Nath, Rajeev Goel, Eun Som Jeon, Changhoon Kim, Kyle Min, Yezhou Yang, Yingzhen Yang, and Pavan Turaga. Deep geometric moments promote shape consistency in text-to-3d generation, 2025.

[15] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding, 2023.

[16] Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. Compositional 3d-aware video generation with llm director, 2024.

[17] Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. Jm3d & jm3d-llm: Elevating 3d representation with joint multi-modal cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[18] Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. Jm3d jm3d-llm: Elevating 3d understanding with joint multi-modal cues, 2024.

[19] Tianyu Huang, Yihan Zeng, Bowen Dong, Hang Xu, Songcen Xu, Rynson W. H. Lau, and Wangmeng Zuo. Textfield3d: Towards enhancing open-vocabulary 3d generation with noisy text fields, 2024.

[20] Shuhao Chen, Yulong Zhang, Weisen Jiang, Jiangang Lu, and Yu Zhang. Vllavo: Mitigating visual gap through llms, 2024.

[21] Ravi Ramamoorthi. Nerfs: The search for the best 3d representation, 2023.

[22] Haoyu Wu, Meher Gitika Karumuri, Chuhang Zou, Seungbae Bang, Yuelong Li, Dimitris Samaras, and Sunil Hadap. Direct and explicit 3d generation from a single image, 2024.

[23] Baao Xie, Bohan Li, Zequn Zhang, Junting Dong, Xin Jin, Jingyu Yang, and Wenjun Zeng. Navinerf: Nerf-based 3d representation disentanglement by latent semantic navigation, 2024.

[24] Kartik Srivastava, Akash Kumar Singh, and Guruprasad M. Hegde. Multi modal semantic segmentation using synthetic data, 2019.

[25] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. Ghost on the shell: An expressive representation of general 3d shapes, 2024.

[26] Xuanyu Yi, Zike Wu, Qiuhong Shen, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Shuicheng Yan, Xinchao Wang, and Hanwang Zhang. Mvgamba: Unify 3d content generation as state space sequence modeling, 2024.

[27] Pengfei Wang, Yuxi Wang, Shuai Li, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Open vocabulary 3d scene understanding via geometry guided self-distillation, 2024.

[28] Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Brightdreamer: Generic 3d gaussian generative framework for fast text-to-3d synthesis, 2024.

[29] Zeyu Cai, Duotun Wang, Yixun Liang, Zhijing Shao, Ying-Cong Chen, Xiaohang Zhan, and Zeyu Wang. Dreammapping: High-fidelity text-to-3d generation via variational distribution mapping, 2024.

[30] Huiqun Wang, Yiping Bao, Panwang Pan, Zeming Li, Xiao Liu, Ruijie Yang, and Di Huang. Multi-modal relation distillation for unified 3d representation learning, 2024.

[31] Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. Text-centric alignment for multi-modality learning, 2024.

[32] Kaichen Zhou, Lanqing Hong, Enze Xie, Yongxin Yang, Zhenguo Li, and Wei Zhang. Serf: Fine-grained interactive 3d segmentation and editing with radiance fields, 2024.

[33] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction, 2024.

[34] Yiwei Ma, Yijun Fan, Jiayi Ji, Haowei Wang, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. X-dreamer: Creating high-quality 3d content by bridging the domain gap between text-to-2d and text-to-3d generation, 2024.

[35] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey, 2024.

[36] Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: Procedural 3d modeling with large language models, 2024.

[37] Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, and Leonidas Guibas. Cad: Photorealistic 3d generation via adversarial distillation, 2023.

[38] Xuening Yuan, Hongyu Yang, Yueming Zhao, and Di Huang. Dreamscape: 3d scene creation via gaussian splatting joint correlation modeling, 2024.

[39] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation, 2024.

[40] Haowei Wang, Jiji Tang, Jiayi Ji, Xiaoshuai Sun, Rongsheng Zhang, Yiwei Ma, Minda Zhao, Lincheng Li, zeng zhao, Tangjie Lv, and Rongrong Ji. Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation, 2024.

[41] Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding, 2024.

[42] Shengqi Liu, Zhuo Chen, Jingnan Gao, Yichao Yan, Wenhan Zhu, Jiangjing Lyu, and Xiaokang Yang. Directional texture editing for 3d models, 2024.

[43] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures, 2019.

[44] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[45] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023.

[46] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation, 2024.

[47] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. Salad: Part-level latent diffusion for 3d shape generation and manipulation, 2024.

[48] Yuzhou Huang, Yiran Qin, Shunlin Lu, Xintao Wang, Rui Huang, Ying Shan, and Ruimao Zhang. Story3d-agent: Exploring 3d storytelling visualization with large language models, 2024.

[49] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023.

[50] Sherzod Hakimov and David Schlangen. Images in language space: Exploring the suitability of large language models for vision language tasks, 2023.

[51] Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security, 2024.

[52] Chenjiao Tan, Qian Cao, Yiwei Li, Jielu Zhang, Xiao Yang, Huaqin Zhao, Zihao Wu, Zhengliang Liu, Hao Yang, Nemin Wu, Tao Tang, Xinyue Ye, Lilong Chai, Ninghao Liu, Changying Li, Lan Mu, Tianming Liu, and Gengchen Mai. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications, 2023.

[53] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

[54] Matteo Marchi, Stefano Soatto, Pratik Chaudhari, and Paulo Tabuada. Heat death of generative models in closed-loop learning, 2024.

[55] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, Peizhao Zhang, Bastian Leibe, Peter Vajda, and Ji Hou. Controlroom3d: Room generation using semantic proxy rooms, 2023.

[56] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models, 2024.

[57] Haotian Xue, Antonio Torralba, Joshua B. Tenenbaum, Daniel LK Yamins, Yunzhu Li, and Hsiao-Yu Tung. 3d-intphys: Towards more generalized 3d-grounded visual intuitive physics under challenging scenes, 2023.

[58] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, SerNam Lim, and Hengshuang Zhao. Larm: Large auto-regressive model for long-horizon embodied intelligence, 2025.

[59] Ziang Cao, Fangzhou Hong, Tong Wu, Liang Pan, and Ziwei Liu. Large-vocabulary 3d diffusion model with transformer, 2023.

[60] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training, 2024.

[61] Jinwei Lin. Oneto3d: One image to re-editable dynamic 3d model and video generation, 2024.

[62] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.