

# Shape from Semantics: 3D Shape Generation from Multi-View Semantics

LIANGCHEN LI, University of Science and Technology of China, China

CAOLIWEN WANG, University of Science and Technology of China, China

YUQI ZHOU, University of Science and Technology of China, China

BAILIN DENG, Cardiff University, United Kingdom

JUYONG ZHANG\*, University of Science and Technology of China, China

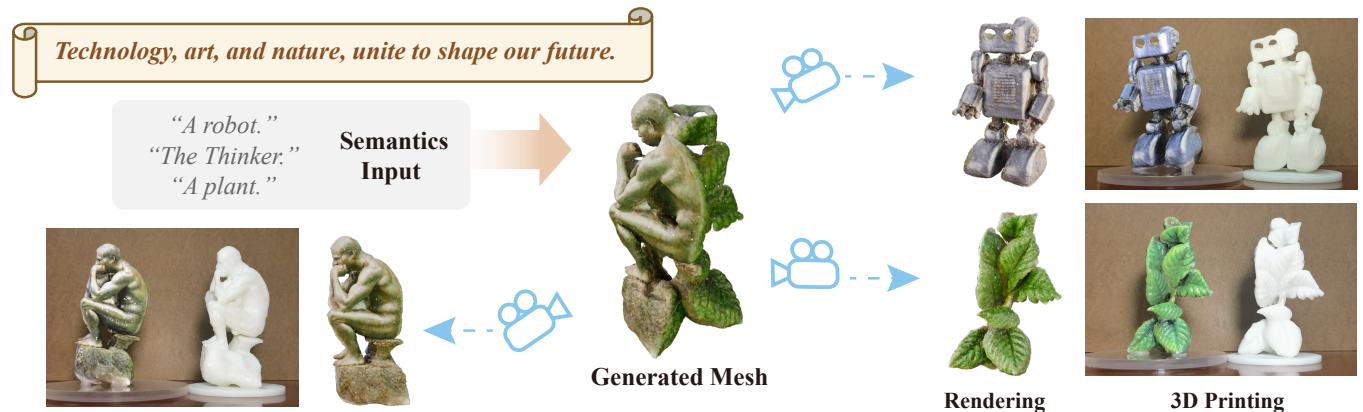


Fig. 1. We propose and address **Shape from Semantics**, a novel generative problem. Given a set of semantics and corresponding views as input, our method can produce high-quality shapes that exhibit geometry and appearance consistent with the semantics from each view and are feasible for real-world fabrication.

Existing 3D reconstruction methods utilize guidances such as 2D images, 3D point clouds, shape contours and single semantics to recover the 3D surface, which limits the creative exploration of 3D modeling. In this paper, we propose a novel 3D modeling task called “Shape from Semantics”, which aims to create 3D models whose geometry and appearance are consistent with the given text semantics when viewed from different views. The reconstructed 3D models incorporate more than one semantic elements and are easy for observers to distinguish. We adopt generative models as priors and disentangle the connection between geometry and appearance to solve this challenging problem. Specifically, we propose Local Geometry-Aware Distillation (LGAD), a strategy that employs multi-view normal-depth diffusion priors to complete partial geometries, ensuring realistic shape generation. We also integrate view-adaptive guidance scales to enable smooth semantic transitions across views. For appearance modeling, we adopt physically based rendering to generate high-quality material properties, which are subsequently baked into fabricable meshes. Extensive experimental results demonstrate that our method can generate meshes with well-structured, intricately detailed geometries, coherent textures, and smooth transitions, resulting in visually appealing 3D shape designs.

CCS Concepts: • Computing methodologies → Shape modeling; Rendering; Machine learning approaches.

Additional Key Words and Phrases: inverse modeling, generative priors

\*Corresponding author (juyong@ustc.edu.cn).

Authors' addresses: Liangchen Li, liangchen@mail.ustc.edu.cn, University of Science and Technology of China, China; Caoliwen Wang, wclw1021@mail.ustc.edu.cn, University of Science and Technology of China, China; Yuqi Zhou, sasuke18@mail.ustc.edu.cn, University of Science and Technology of China, China; Bailin Deng, DengB3@cardiff.ac.uk, Cardiff University, United Kingdom; Juyong Zhang, juyong@ustc.edu.cn, University of Science and Technology of China, China.

## 1 INTRODUCTION

Reconstructing 3D shapes from various inputs is a cornerstone of computer graphics and vision, vital for applications like cultural heritage, film, and architecture. Conventional methods reconstruct 3D geometry and renderings using specific visual data such as RGB images, point clouds, or surface normals. While these inputs facilitate accurate surface reconstruction, they impose strict constraints that can limit the generation of imaginative and novel 3D assets crucial for design, AR/VR, and art.

In this paper, we introduce a novel “Shape from Semantics” problem, which utilizes textual descriptions to guide 3D model generation. Here, semantics are high-level, human-interpretable concepts describing an object’s desired characteristics. This semantic-driven approach enables the creation of 3D models that convey intended visual properties from multiple viewpoints, offering a new level of flexibility and creativity in 3D content generation. The resulting 3D models provide a more intuitive and immersive experience through direct observation from different views, compared to 2D designs or projections. Furthermore, semantics-based operations are inherently user-friendly, empowering even non-professionals to produce detailed meshes and intricate textures from just a few text prompts, thereby significantly lowering the barrier to artistic creation.

This task is non-trivial as it requires matching geometry and appearance with input semantics from different viewpoints rather than specific input images; this makes existing multi-view reconstruction techniques inapplicable. Current text-to-3D generation models are also unsuitable, as they typically use a single prompt to describe a

single object, whereas our method employs multiple prompts to define different object appearances from multiple views. The research most similar to ours uses information like shadows or 2D contours as guidance for reconstruction and design. For instance, Shadow Art [Mitra and Pauly 2009] designs objects whose projections match given 2D shapes under specific lighting. Wire Art [Hsiao et al. 2018; Qu et al. 2024; Tojo et al. 2024] focuses on generating wireframe geometries that align with 2D line drawings or outline shapes consistent with semantic inputs. However, these methods primarily offer a two-dimensional visual experience; directly observing the 3D objects often makes it challenging to perceive the intended embedded semantic information. Additionally, such techniques frequently depend on specific setups (e.g., light sources, projection planes) and face fabrication challenges, limiting their practical use.

To address our challenging problem, we leverage the text understanding capabilities of generative models to create a 3D model that matches input semantics from different observation directions. Our approach disentangles the generation process into separate geometry and appearance stages. For geometry, our core insight is that required geometric parts are derived from complete geometries corresponding to the input semantics, which motivates the use of 3D-consistent priors. To this end, we propose Local Geometry-Aware Distillation (LGAD), a strategy employing a multi-view normal-depth diffusion model [Qiu et al. 2024] as a prior to construct high-quality geometry, represented using Tetrahedron Splatting [Gu et al. 2024]. We also introduce a view-adaptive guidance scale to promote smooth semantic transitions across views. For appearance, we employ a physically based rendering (PBR) pipeline, and utilize a Depth-conditioned Albedo diffusion model to generate and bake high-quality material properties into the fabricable meshes.

Extensive experiments demonstrate our method’s capacity for high creativity, generating models that surpass traditional spatial intuition or non-semantic inputs. The resulting 3D models feature well-structured, intricately detailed geometry, coherent textures, and smooth transitions, presenting fascinating and surprising creative designs. In summary, our contributions include:

- We introduce a novel “Shape from Semantics” problem for 3D generation from semantics of different views, which provides a powerful modeling tool for design and artistic creation.
- We propose Local Geometry-Aware Distillation for robust 3D geometry from limited per-semantic views by directly guiding local normal-depth features with a 3D prior; a view-adaptive guidance strategy for coherent multi-semantic integration; and a PBR-based appearance modeling approach utilizing an albedo diffusion prior to generate high-quality textures.
- Our method enables creating high-quality meshes with detailed textures and rich geometry from just a few prompts.

## 2 RELATED WORK

*Shape from X.* Traditional “Shape from X” methods focus on high-precision reconstruction of existing objects using known specific visual data, such as RGB images [Goesele et al. 2007; Moulou et al. 2013; Schönberger and Frahm 2016; Snavely et al. 2006; Wang et al. 2021], depth [Dai et al. 2017; Newcombe et al. 2011] and normals [Cao et al. 2022; Kadambi et al. 2015]. A related body of research

explores constructing single, fixed objects that offer multiple visual interpretations; these methods achieve diverse visual perceptions by leveraging factors such as viewing distance [Oliva et al. 2006], figure-ground organization [Kuo et al. 2017], illumination from different directions [Alexa and Matusik 2010; Baran et al. 2012; Bermano et al. 2012], light reflections [Sakurai et al. 2018; Wu et al. 2022], viewing angles [Hsiao et al. 2018; Qu et al. 2024; Sela and Elber 2007; Tojo et al. 2024; Zeng et al. 2021], and shadow casting on external planar surfaces [Mitra and Pauly 2009; Sadekar et al. 2022]. However, the goal of these works is typically to produce different 2D information perceptions from an object—whether as a contour [Hsiao et al. 2018; Qu et al. 2024; Tojo et al. 2024], a projection [Mitra and Pauly 2009; Sadekar et al. 2022], or a picture [Min et al. 2017; Schwartzburg et al. 2014; Wu et al. 2022]. Our work enables direct 3D perception, allowing the characteristics of 3D objects to be experienced firsthand. This is the first work to explore creating multiple 3D interpretations of a single object. In addition, we leverage semantics as a substitute for traditional inputs, similar to [Qu et al. 2024; Tojo et al. 2024], significantly expanding the creative space.

*3D Data Representations.* The representation of 3D data is a core topic in computer graphics and vision. Beyond traditional point cloud and mesh representations, many novel 3D representations have recently demonstrated significant advantages. Mildenhall et al. [2021] propose Neural Radiance Fields (NeRF), which represent a scene with a neural implicit function guided by neural rendering. NeRFs have been widely applied to multi-view reconstruction [Li et al. 2023b; Wang et al. 2021, 2023a], sparse reconstruction [Jain et al. 2021; Liu et al. 2023; Niemeyer et al. 2022; Wynn and Turmukhambetov 2023; Yu et al. 2021], and generation tasks [Chen et al. 2023; Jain et al. 2022; Lin et al. 2023; Poole et al. 2022; Tang et al. 2023b; Wang et al. 2023b], thanks to its capability in representing objects with rich details. However, its optimization can be time-consuming and computationally intensive. Recently, 3DGS [Kerbl et al. 2023] brings new possibilities for rendering [Lu et al. 2024; Yan et al. 2024; Yu et al. 2024] and reconstruction problems [Fu et al. 2024; Guédon and Lepetit 2024; Luiten et al. 2024; Zhu et al. 2023] thanks to its flexible model design and efficient differentiable rendering framework. Tang et al. [2023a] incorporate a generative model into 3DGS, enabling rapid generation of textured meshes. However, the geometry generated by 3DGS often suffers from significant detail loss, excessive surface undulations, and suboptimal mesh quality.

Other representations [Guo et al. 2024; Yariv et al. 2024] have also demonstrated advantages in reconstruction and generation tasks. DMTET [Shen et al. 2021] combines implicit and explicit representations by predicting surfaces on a deformable tetrahedral grid and extracting meshes via Marching Tetrahedra, enhancing accuracy and efficiency. Fantasia3D [Chen et al. 2023] successfully applies this representation to 3D generation tasks. Tetrahedron Splatting [Gu et al. 2024] combines precise mesh extraction enabled by tetrahedral grids with efficient optimization of volumetric rendering and demonstrates outstanding performance in generation tasks. In this work, we utilize this geometric representation to achieve high-fidelity geometry and detailed textures while reducing computational costs.

*3D Generation.* While generative models have recently gained widespread attention in computer vision and graphics, the task of 3D

generation continues to pose substantial challenges, primarily due to the limited availability of extensive, high-quality 3D datasets [Deitke et al. 2023, 2022; Koch et al. 2019; Wu et al. 2023]. Recently, many 3D generation methods [Chen et al. 2023; Jain et al. 2022; Lin et al. 2023; Poole et al. 2022; Tang et al. 2023b; Wang et al. 2023b] utilize 2D information as supervision to guide 3D generation, using various representations of 3D data. DreamFields [Jain et al. 2022] pioneers the use of diffusion models for semantic-based 3D generation. DreamFusion [Poole et al. 2022] introduces the score distillation sampling (SDS) loss, which leverages semantic information and 2D rendering results, and this approach has since been widely adopted. However, as these methods inherently rely on supervision from 2D rendering results, they often face challenges with multi-view inconsistency. While many existing works aim to mitigate such inconsistencies [Liu et al. 2023; Shi et al. 2023], our approach leverages such potential inconsistency to generate creative objects with multiple visual interpretations. Moreover, researchers incorporate various priors (normal, depth, etc.) into 3D generation tasks to enhance the realism of models. SweetDreamer [Li et al. 2023a] and RichDreamer [Qiu et al. 2024] integrate canonical coordinate maps and normal-depth priors into the loss function, respectively. Meanwhile, Wonder3D [Long et al. 2023] and CRM [Wang et al. 2024] directly utilize these priors to construct corresponding meshes.

Researchers also try to use 3D datasets directly for 3D generation tasks. PolyGen [Nash et al. 2020], MeshGPT [Siddiqui et al. 2024], and XCube [Ren et al. 2024] represent geometry natively using mesh vertices, mesh surface sequences, and voxels, respectively. SDFusion [Cheng et al. 2023] and 3DGen [Gupta et al. 2023] leverage 3D Variational Autoencoders (VAEs) to encode geometry, employing Signed Distance Fields (SDFs) and triplanes as geometric representations. Methods such as Shap-E [Jun and Nichol 2023] and 3DShape2VecSet [Zhang et al. 2023] adopt transformer-based architectures to encode geometry, while more recent methods such as TRELLIS [Xiang et al. 2024] and CLAY [Zhang et al. 2024] focus on constructing more compact and versatile latent spaces for decoding into diverse representations. DeepMesh [Zhao et al. 2025] and Oct-GPT [Wei et al. 2025] enhance pretraining efficiency and stability via autoregressive modeling. However, they are typically trained and evaluated on datasets such as ShapeNet [Chang et al. 2015] and Objaverse [Deitke et al. 2022], which constrains the diversity and complexity of the generated shapes. In contrast, our method seeks to unlock the creative potential of generative models to synthesize astonishing geometric forms that transcend common objects.

### 3 METHOD

We take as input  $n$  semantic labels  $\mathcal{Y} = \{y_i\}$ , each being a textual prompt, and their corresponding view directions  $\mathcal{V} = \{v_i \in SO(3)\}$ . We call  $\mathcal{V}$  the *observation views*, which can either be predefined or initialized randomly. We aim to generate a colored 3D shape  $S$  whose texture and geometry align with the associated semantic class  $C(y_i)$  when observed from any main review  $v_i$ .  $S$  should possess a simple, intuitive, and compact design suitable while retaining key geometric features that define its appearance. Meanwhile, the generated shape should be highly recognizable and visually elegant.

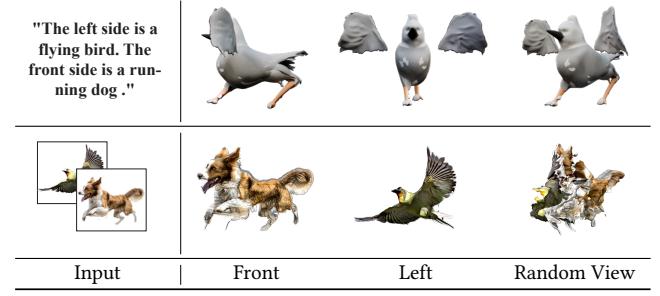


Fig. 2. **Failure of Naive Solutions.** We expect the shape to combine “a running dog” (front side) and “a flying bird” (left side). The top row shows the text-to-3D generation result from TRELLIS [Xiang et al. 2024], which mixes the semantics directly. The bottom row shows a baseline approach that first generates 2D images using Stable Diffusion [Esser et al. 2024], then performs multi-view reconstruction [Wang et al. 2021], whose result shows meaningless geometry. Our result of this case can be found in Fig. 13.

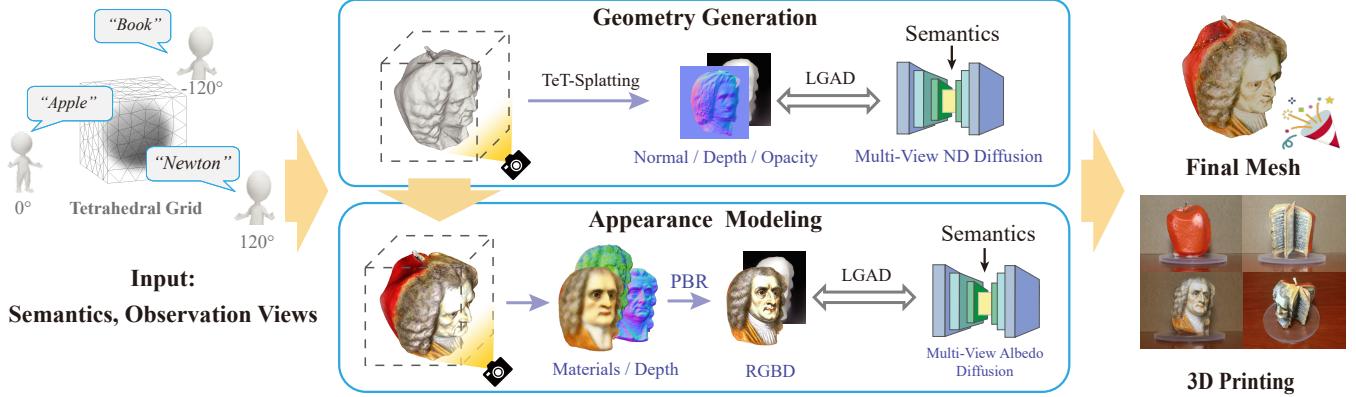
This task is inherently challenging. Despite recent advances in generative models, their direct application to our problem yields unsatisfactory results. For instance, state-of-the-art text-to-3D models struggle with our task (Fig. 2, top) due to their limited understanding of directional descriptions, leading to semantic blending across different views. An alternative approach involves generating an image for each semantic label using a text-to-image model (e.g., Stable Diffusion [Esser et al. 2024]) followed by 3D reconstruction from these multi-view images [Wang et al. 2021]. However, this often results in meaningless or distorted geometries (Fig. 2, bottom) because the generated images lack the necessary 3D geometric information for robust reconstruction, leading to flattened or deformed shapes.

To overcome these limitations, we propose a novel solution (see Fig. 3) that disentangles geometry and appearance generation into a two-stage process, ensuring geometrically plausible and semantically coherent results. Section 3.1 introduces our Local Geometry-Aware Distillation (LGAD) approach that leverages geometric priors from pre-trained diffusion models to achieve high-quality geometry under the limited view range for each semantic constraint. Then, Section 3.2 presents our 3D geometry representation and complementary strategies for effective geometry generation. Finally, Section 3.3 describes our PBR approach for appearance modeling, which produces high-quality, fabricable textures.

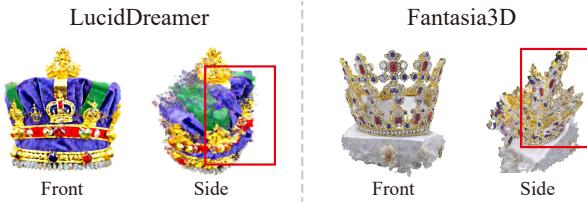
#### 3.1 Local Geometry-Aware Distillation

Given the absence of suitable datasets for our novel task, training a data-driven 3D generative model directly is infeasible. Therefore, we adopt the score distillation sampling (SDS) framework [Poole et al. 2022] to leverage powerful pre-trained diffusion models as priors for 3D shape generation. In a typical SDS iteration, a camera pose  $c$  and corresponding semantics  $y(c)$  are sampled. An image  $I = I(\theta, c)$  is rendered from the current 3D shape representation  $\theta$ . This image is then guided by a text-to-image diffusion model to match  $y(c)$ . The SDS gradient is commonly expressed as:

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}}(\theta) \triangleq \mathbb{E}_{t, \epsilon, c} \left[ \omega(t) (\epsilon_{\text{pre}}(I_t; t, y(c)) - \epsilon) \frac{\partial I}{\partial \theta} \right], \quad (1)$$



**Fig. 3. Shape from Semantics Pipeline.** We use TeT-Splatting [Gu et al. 2024] as the 3D representation, and disentangle geometry and appearance generation into a two-stage process. In the geometry generation stage, we render the normal and depth map through alpha blending, and optimize the geometry using the proposed LGAD method. In the appearance modeling stage, we use physically based rendering to obtain the RGBD map for diffusion and learn view-independent realistic textures. Finally, the colored mesh is extracted and can be crafted into visually appealing art pieces.



**Fig. 4. SDS under Limited Sampling Views.** To simulate the constraints we encounter, we reduced the SDS sampling azimuth range from  $[-180^\circ, 180^\circ]$  to  $[-45^\circ, 45^\circ]$ . The semantics is “an imperial state crown of England”. While LucidDreamer [Liang et al. 2024] with modified SDS achieves high-quality frontal rendering, significant shape collapse and fragmentation become evident upon rotation (marked by red box). Even with normal maps applied for geometric optimization [Chen et al. 2023], the shape still suffers from structural collapse/deformation.

where  $\mathcal{I}_t$  is the noised rendered image with sampled noise  $\epsilon$ ,  $\epsilon_{pre}$  is the noise predicted by the 2D diffusion prior, and  $\omega(t)$  is a weighting function. However, effective SDS relies on dense and varied view sampling. Our problem inherently restricts the view range for each semantic, as multiple distinct semantics must be expressed from specific viewpoints of a single object. This leads to weak geometric supervision, often resulting in a mismatch between the intended shape and the rendered appearance, or incorrect details (Fig. 4).

To address this challenge and achieve robust geometry under limited-view supervision, we introduce *Local Geometry-Aware Distillation* (LGAD). The core principle is that any plausible local geometric attribute (e.g., surface normals and depth observed from a viewpoint  $v_i$ ) corresponding to a semantic  $y_i$  must be consistent with some complete 3D shape  $\theta^*$  that fully embodies  $y_i$ . For a given semantic  $y(c)$  from a camera view  $c$ , let  $\mathbf{g} = P(\theta, c)$  be the currently rendered local geometric attributes (specifically, normal and depth maps, or ND maps) from our evolving shape  $\theta$ . LGAD aims to guide  $\theta$  such that  $\mathbf{g}$  aligns with the local attributes  $\mathbf{g}^* = P(\theta^*, c)$  that would be observed from such an ideal shape  $\theta^*$  along view  $c$ .

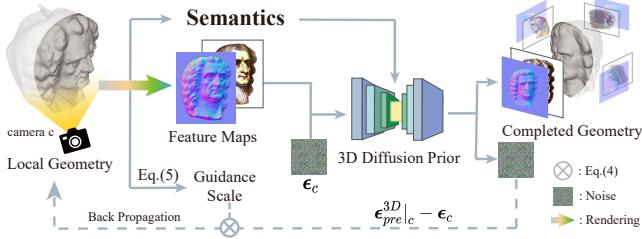
Rather than explicitly reconstructing  $\theta^*$ , LGAD uses a pre-trained 3D-aware diffusion model as a prior. The key is to shift the distillation target from 2D RGB images to ND maps. As our 3D-aware prior, we employ the multi-view normal-depth diffusion model from RichDreamer [Qiu et al. 2024], which is represented as a noise prediction network  $\epsilon_{pre}^{3D}$  conditioned on the semantic  $y(c)$  and a set of camera views  $C$  (which includes the observation view  $c$  and other views surrounding the object). We then formulate a loss that measures the deviation between the predicted and ground-truth noises for the view  $c$ , similar to the SDS loss in Eq. (1). To satisfy the RichDreamer prior’s requirement for multi-view ND inputs for all views in  $C$  while focusing on guidance from  $y(c)$ , we render a single noise-free ND map  $g_0(\theta, c)$  from the observation view  $c$ , and duplicate it for each view in  $C$  with individually added noise per view, creating a set of noised maps  $\{g_{t,c'}\}_{c' \in C}$  that share the same underlying geometry. This set is then input to  $\epsilon_{pre}^{3D}$  along with the semantics  $y(c)$  and the views  $C$ . We then use the noise prediction  $\epsilon_{pre}^{3D}(\{g_{t,c'}\}_{c' \in C}; t, y(c), C)|_c$  corresponding to the observation view  $c$  and compare it with the ground truth  $\epsilon_c$ . Our final LGAD loss gradient is:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{t, \epsilon, c} \left[ \omega(t) \left( \epsilon_{pre}^{3D}(\{g_{t,c'}\}_{c' \in C}; t, y(c), C)|_c - \epsilon_c \right) \frac{\partial g}{\partial \theta} \right]. \quad (2)$$

A detailed pseudo-code is shown in the supplementary materials.

### 3.2 Geometry Representation and Generation

*Tetrahedron Splatting.* In 3D generation tasks, implicit representations like NeRF [Mildenhall et al. 2021] can involve lengthy training, while explicit representations such as 3DGS [Kerbl et al. 2023; Tang et al. 2023a] may produce unstructured or low-quality geometry. Instead, we follow [Gu et al. 2024] and adopt tetrahedral splatting as our representation, which constructs a tetrahedral grid encoding a Signed Distance Field (SDF) in 3D and uses alpha blending for tetrahedron rendering. To enhance geometric quality during training,



**Fig. 5. Local Geometry-Aware Distillation.** In each iteration, we sample a camera and render feature maps of the local geometry. Semantics and guidance scale are obtained through Eq. (5). Afterwards, 3D diffusion prior is utilized to denoise the features to match the front view feature of a complete geometry, and finally back propagate to optimize the local geometry.

we incorporate an eikonal loss and a normal consistency loss:

$$\mathcal{L}_{\text{eik}} = \sum_{\delta} (\|\nabla f_{\delta}\|_2 - 1)^2, \quad \mathcal{L}_{\text{nc}} = \sum_e (1 - \cos(\mathbf{n}_{e_1}, \mathbf{n}_{e_2})), \quad (3)$$

where  $\nabla f_{\delta}$  is the SDF gradient of each tetrahedron  $\delta$ , and  $\mathbf{n}_{e_1}$  and  $\mathbf{n}_{e_2}$  are the surface normals at the vertices connected by edge grid  $e$ .

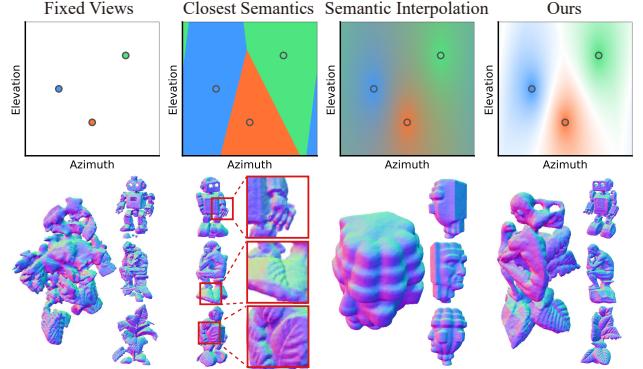
**View-Adaptive Guidance.** While our LGAD loss induces strong geometric supervision, simply applying it at the predefined observation views  $v_i \in \mathcal{V}$  often fails to produce satisfactory overall 3D geometry (see Fig. 6, left). For a coherent shape, it is beneficial to also apply LGAD guidance from camera views that are near the observation views. However, this introduces a challenge: the input semantics are explicitly defined only for the observation views; for any other views  $c \notin \mathcal{V}$ , the intended semantics  $y(c)$  are ambiguous.

A naive approach is to assign  $y(c)$  based on the semantics of the closest observation view. However, this leads to abrupt and potentially conflicting transitions in areas where the influence of two observation views meets (Fig. 6, second column). An alternative is to interpolate the embeddings of the surrounding observation view semantics  $\{y_i\}$  by weighting them based on proximity to  $c$ , e.g.,  $\text{Emb}[y(c)] = (\sum_i w_i \cdot \text{Emb}[y_i]) / (\sum_i w_i)$ , where  $w_i = 1/(1 - c \cdot v_i)$  are influence weights, and  $\text{Emb}(\cdot)$  is the text encoding function. As shown in Fig. 6 (third column), this semantic blending significantly decreases the geometric expressiveness and distinctiveness of the intended multiple interpretations. To address these issues, we propose a View-Adaptive Guidance strategy that utilizes Classifier-Free Guidance (CFG) [Ho and Salimans 2022]. CFG allows modulation of the semantic guidance strength via a scale parameter  $s$ :

$$\tilde{\epsilon}_{\text{pre}}(\mathcal{I}_t, t, y) := s\epsilon_{\text{pre}}(\mathcal{I}_t, t, y) + (1-s)\epsilon_{\text{pre}}(\mathcal{I}_t, t, \emptyset), \quad (4)$$

where  $\tilde{\epsilon}_{\text{pre}}$  is the guided noise prediction,  $\epsilon_{\text{pre}}$  is the model’s raw noise prediction (conditioned on semantics  $y$  or an unconditional prompt  $\emptyset$ ), and  $\mathcal{I}_t$  is the noised input. We can dynamically adjust  $s$  to enforce stronger guidance when the camera view  $c$  is closer to an observation view, and weaker guidance when it is in an ambiguous transition zone. Specifically, we sort the influence weights of each observation view on the current view  $c$  in descending order:  $\{w'_0 \geq \dots \geq w'_{n-1}\}$ , and compute the guidance scale as:

$$s = s_0(w'_0 - w'_1) / \sum_i w'_i, \quad (5)$$



**Fig. 6. View-Adaptive Guidance.** The top row shows the guidance scale variation across camera views, with circles marking observation views. Point colors denote semantic components, while transparency indicates the guidance scale. Four strategies are tested on the case in teaser: training on fixed observation views and their semantics, with randomly sampled views and their closest semantics, semantics interpolation, and our method. The bottom row shows normal maps of generated geometries, with the second case contains three middle views, and the other contain three observation views and a random view. Training with fixed views leads to fractured structures, semantics interpolation makes shape less expressive, and directly choosing closest semantics causes geometric feature blending at semantic boundaries, exemplified by generating a human hand for the robot, or plant leaves on the sculpture and the robot body.

where  $s_0$  is a hyperparameter. This ensures  $s$  is largest when  $c$  aligns with a single observation view and diminishes as  $c$  moves into regions where multiple observation views have comparable influence. Additionally, to avoid semantic blending in the prior conditioning, we provide the LGAD diffusion prior with the semantics corresponding to the observation view closest to  $c$ . Fig. 6 shows that when the dominant semantic influence transitions from one observation view to another,  $s$  naturally passes through or near zero. This creates continuous and smooth semantic supervision across views, leading to more coherent and expressive geometric results.

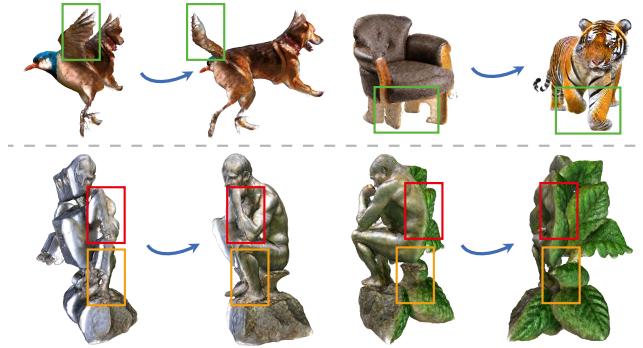
**Training Details.** Our geometry generation employs a structure-to-detail process. We initialize the tetrahedral SDF field as a sphere, then apply LGAD to obtain a coarse geometry. Afterwards, for detailed geometric refinement, we lower the timestep sampling range in the diffusion process. Throughout the training process, we also integrate vanilla Stable Diffusion [Rombach et al. 2022] as an additional guidance complementing our LGAD optimization:

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t, \epsilon, c} \left[ \omega(t) \left( \epsilon_{\phi}(\mathcal{I}; y(c), t) - \epsilon \right) \frac{\partial \mathcal{I}}{\partial \theta} \right], \quad (6)$$

where  $\mathcal{I}_t$  is the noised rendered normal maps from view  $c$ , and  $\epsilon_{\phi}(\mathcal{I}; y(c), t)$  is the noise estimated by the UNet  $\epsilon_{\phi}$  of the 2D prior.

### 3.3 Appearance Modeling

With the geometry established, this stage focuses on adding rich color and realistic surface appearance to the 3D model. To this end, the well-trained tetrahedral SDF field is first converted into a polygonal mesh using the Marching Tetrahedra algorithm [Shen et al.



**Fig. 7. Geometric Structures Details.** Our generated results achieve sophisticated visual effects by sharing geometric elements across different semantic components.

2021]. For the subsequent appearance optimization, we employ Physically Based Rendering (PBR) better disentangle material properties and achieve more realistic results. The material properties at any surface point  $p$  are determined by the diffuse color  $k_d \in \mathbb{R}^3$  (albedo), roughness  $k_r \in \mathbb{R}$ , metallic term  $k_m \in \mathbb{R}$ , and tangent-space normal variation  $k_n \in \mathbb{R}^3$ . These spatially varying attributes are encoded using a hash grid  $\Phi_\Theta$  with parameter  $\Theta$ :  $(k_d, k_r, k_m, k_n) = \Phi_\Theta(p)$ . To ensure the final baked appearance on the extracted mesh is high-fidelity and consistent with training renders, our PBR setup decouples materials from view-dependent lighting effects, making all physical attributes spatially invariant for faithful extraction.

We use a Depth-conditioned Albedo diffusion model [Qiu et al. 2024] as the appearance prior, capable of producing multi-view albedo maps conditioned on semantics and camera poses. The LGAD framework from Sec. 3.1 is adapted to optimize these PBR materials by using rendered RGBD images as the distillation target  $g$  in Eq. (2). Additionally, the view-adaptive guidance strategy and auxiliary SDS loss from Sec. 3.2 are utilized to further refine the appearance.

## 4 EXPERIMENTS

*Implementation Details.* During training, camera views are sampled with an azimuthal range of  $\pm 50$  degrees around each observation view and an elevation range of  $\pm 25$  degrees, further enhanced with adaptive scale adjustments in Eq. (5) with  $s_0 = 70$ . The geometry generation stage takes 3,000 iterations, which includes 1,000 iterations for initial coarse shape formation and 2,000 iterations for subsequent geometric refinement. Appearance modeling is then performed for an additional 2,000 iterations. The entire training procedure is performed on a single NVIDIA RTX 3090 GPU (24GB VRAM) and completes in approximately 1.5 hours. We maintain a consistent tetrahedral grid resolution of  $256^3$  for both the training and mesh extraction stages.

*Main Results.* We apply our method to generate various multi-semantic texture meshes, which are shown in Fig. 11. Our inputs cover diverse semantic and view inputs, demonstrating the rich creativity of our method. The results are highly consistent with the expected semantics in terms of geometry and appearance. We 3D-print some cases with results provided in the supplementary

**Table 1. CLIP Similarity (%) (higher is better) between the Semantics and the Observed Meshes in Fig. 11.** Each result is displayed as scores of with/without textures.

View/Metric	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
View1	38.84/37.36	37.44/35.37	35.98/36.74	40.63/37.65	38.39/34.17	36.41/36.36	34.68/36.94
View2	41.07/37.32	30.77/23.95	33.36/30.77	37.25/36.47	31.51/23.23	25.29/25.45	34.81/32.57
View3	36.24/31.81	22.27/22.32	40.76/39.95	30.39/22.11	31.74/33.32	41.57/37.37	39.45/38.31
Mean Score	38.72/35.49	30.16/27.21	36.70/35.82	36.09/32.08	33.88/30.24	34.43/33.06	36.31/35.94

**Table 2. Scores of the User Study for Results in Fig. 11.** The rating range is 0-10, with higher scores indicating better results.

Metric	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
w/o Texture	6.51	5.46	7.39	6.99	5.87	6.50	7.40
w/ Texture	9.35	7.47	9.38	9.47	8.94	8.54	9.44
Semantic Pref.	8.82	8.96	8.80	9.04	8.82	8.82	8.92
Overall Pref.	8.61	8.87	8.71	8.87	8.96	9.00	8.83

materials. The fabricated objects are highly consistent with the expected design and exhibit an aesthetic appeal.

A notable feature of our generated results is their sophisticated geometric structures. As demonstrated in Fig. 7, the LGAD technique enables the same geometric components to serve distinct semantic roles. For instance, a bird’s wings transform into a dog’s tail when rotated, while the stone bench simultaneously serves as a plant leaf. This kind of combination achieves geometric-semantic transitions during rotational observing, exhibiting rich playfulness.

*Quantitative Evaluation.* To evaluate the consistency between the generated results and the input semantics, we render the generated 3D models from each observation view and use the CLIP score [Radford et al. 2021] to measure their semantic similarity to the input. Tab. 1 presents the CLIP scores for textured and non-textured cases. For each observation view, we allow random variations within a 20-degree latitude and longitude range to render the images. The CLIP model then evaluates the captured results 1,000 times, and the average score is taken as the score for that observation view. The results indicate that the generated models effectively convey semantic information, regardless of whether textures are applied. Observers can also discern the semantic representation of these geometric shapes even with slight changes in perspective.

In addition, we conducted a user study to further validate our method. The participants were shown the rendering of our generated models one at a time, and asked to sequentially answer the following questions with a score from 0 to 10:

- Q1: How well does the textureless rendering match the semantics?
  - Q2: How well does the textured rendering match the semantics?
- The last two questions focus on participants’ preferences between the results of ours and [Tojo et al. 2024] under the same semantics. A score closer to 10 indicates a stronger preference for our results:
- Q3: Which result better aligns with semantics?
  - Q4: Which overall result do you prefer?

We randomly and fairly selected participants, collecting 83 samples. The average scores from each observation view of each case are presented in Tab. 2. The results indicate that our 3D model design receives considerable recognition.

In preference scoring, our overall performance in semantics and aesthetics is significantly higher compared to the method we benchmarked against. This suggests that the combination of our rendering

Representation	Generated 3D Shapes			
	Front	Left	Middle	Details
NeRF				
DMTET				
3DGS				
TeT-Splatting (Ours)				

**Fig. 8. Comparison of Different Representations.** The semantics are “flower” & “butterfly”. We compare with Dreamfusion [Tang 2022] using NeRF, Fantasia3D [Chen et al. 2023] using DMTET [Shen et al. 2021], and DreamGaussian [Tang et al. 2023a] using 3DGS. We implement the baselines by modifying SDS into multi-semantics version according to Eq (1). Results show that the TeT-Splatting geometry is smoother and more detailed.

and geometry maintains strong semantic expressiveness. Moreover, when compared to monochrome or outline results, our overall design proved to be highly appealing.

**Qualitative Comparison.** As far as we know, no previous research has been conducted with the same purpose as ours. Therefore, we make comparisons with Shadow Art [Mitra and Pauly 2009] and Wire Art [Tojo et al. 2024]. Similar to us, they aim to represent diverse semantics from different views. Initially, we provide our final rendered images at observation views to both of them and compare their results with ours. As shown in Fig. 15, both methods convey information solely through contours and silhouettes, lacking color representation and meaningful geometric structure. In contrast, our shape representation integrates rendering, enabling the depiction of more intricate and complex shapes.

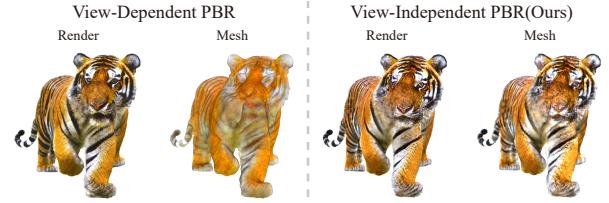
We also compare multi-semantic generation quality in Fig. 13. Given the same inputs, our generated shapes are more refined while maintaining better geometric-semantic consistency in local details. Additionally, as shown in the first row “chair” & “tiger” case, our design could convey distinct front/back semantics, which is challenging for contour-based representations.

Comparisons between different representations are presented in Fig. 8. The tetrahedral splatting presents superior geometric fidelity, significantly enhancing stability even compared to methods using similar structures like DMTet [Chen et al. 2023; Shen et al. 2021].

**Ablation Study.** To demonstrate the effectiveness of our LGAD strategy, we conduct an ablation study on different score distillation approaches, with results presented in Fig. 9. Results show that the absence of geometric supervision leads to degraded geometry. Using a single-view text-to-ND model, however, results in overfitting, distorted shapes, and the emergence of features from other semantics,



**Fig. 9. Ablation for LGAD.** Three methods use the same randomly sampled observation views, and semantics are [“Fragile Egg”, “Soaring Chicken”, “Fallen Feather”]. Here we present normal maps of three observation views and a random view.



**Fig. 10. Ablation for View-independent PBR.** We run appearance modeling based on the full PBR pipeline and our view-independent PBR on the same generated shape. The former mesh has the albedo extracted as texture.

such as the chicken claw extending out of the egg. In contrast, our method generates geometry that is rich in features while remaining clean, and faithfully conveying the semantics.

We also conduct an ablation study on different rendering methods, with results presented in Fig. 10. Our rendering achieves comparable texture details to the full PBR pipeline. However, as the mesh columns show, our extracted mesh faithfully preserves the rendered appearance, whereas the mesh generated by the full PBR approach (using albedo texture as the extraction basis) exhibits significant detail loss,

In Fig. 12 we explore how view distributions affect generation. While maintaining the same semantic inputs, we employ two distinct observation view sets for generation. Both groups achieve high-quality rendering and geometry, demonstrating our method’s robustness to view variations. Simultaneously, all three semantic instances exhibit distinct shapes, confirming the diversity of our generation approach. The geometry of each semantics adaptively composes spatially coherent structures, achieving geometric compatibility while showcasing rich creativity.

## 5 CONCLUSION & DISCUSSION

We introduced and addressed “Shape from Semantics,” a novel problem focused on generating 3D shapes from multi-view semantics. Our core approach leverages 3D diffusion priors for both shape and appearance optimization. Experiments show our method successfully produces impressive shapes that are aesthetically pleasing, semantically consistent with inputs, and readily manufacturable.

Our method still has some limitations. Complex semantics can introduce inherent multi-view conflicts that are difficult to fully

resolve. Additionally, while our strong geometric constraints effectively prevent oversimplified or flattened results, they can occasionally cause collapses or distortions. As shown in Fig. 14, selecting alternative input views can mitigate some of these conflicts. A promising avenue for future work is to treat observation views as optimizable parameters; this could improve semantic compatibility and allow for better integration of different shape characteristics.

## REFERENCES

- Marc Alexa and Wojciech Matusik. 2010. Reliefs as images. *ACM Trans. Graph.* 29, 4, Article 60 (July 2010), 7 pages. <https://doi.org/10.1145/1778765.1778797>
- Ilya Baran, Philipp Keller, Derek Bradley, Stelian Coros, Wojciech Jarosz, Derek Nowrouzezahrai, and Markus Gross. 2012. Manufacturing Layered Attenuators for Multiple Prescribed Shadow Images. *Comput. Graph. Forum* 31, 2pt3 (May 2012), 603–610. <https://doi.org/10.1111/j.1467-8659.2012.03039.x>
- Amit Bermano, Ilya Baran, Marc Alexa, and Wojciech Matusik. 2012. ShadowPix: Multiple Images from Self Shadowing. *Comput. Graph. Forum* 31, 2pt3 (May 2012), 593–602. <https://doi.org/10.1111/j.1467-8659.2012.03038.x>
- Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. 2022. Bilateral Normal Integration.
- Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR* abs/1512.03012 (2015). arXiv:1512.03012 <http://arxiv.org/abs/1512.03012>
- Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22246–22256.
- Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alex Schwing, and Liangyan Gui. 2023. SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. 2017. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. *ACM Trans. Graph.* 36, 3, Article 24 (May 2017), 18 pages. <https://doi.org/10.1145/3054739>
- Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).
- Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv preprint arXiv:2212.08051* (2022).
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling reflect flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.
- Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. 2024. COLMAP-Free 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20796–20805.
- Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. 2007. Multi-View Stereo for Community Photo Collections. In *2007 IEEE 11th International Conference on Computer Vision*. 1–8. <https://doi.org/10.1109/ICCV.2007.4408933>
- Chun Gu, Zeyu Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. 2024. Tetrahedron Splatting for 3D Generation. In *NeurIPS*.
- Antoine Guédon and Vincent Lepetit. 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. *CVPR* (2024).
- Minghao Guo, Bohan Wang, Kaiming He, and Wojciech Matusik. 2024. TetSphere Splatting: Representing High-Quality Geometry with Lagrangian Volumetric Meshes. *arXiv preprint arXiv:2405.20283* (2024).
- Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 2023. 3DGen: Triplane Latent Diffusion for Textured Mesh Generation. *arXiv:2303.05371* [cs.CV] <https://arxiv.org/abs/2303.05371>
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- Kai-Wen Hsiao, Jia-Bin Huang, and Hung-Kuo Chu. 2018. Multi-view Wire Art. *ACM Trans. Graph.* 37, 6, Article 242 (2018), 11 pages.
- Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. *CVPR* (2022).
- Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5885–5894.
- Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv:2305.02463* [cs.CV] <https://arxiv.org/abs/2305.02463>
- Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. 2015. Polarized 3D: High-Quality Depth Sensing with Polarization Cues. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3370–3378. <https://doi.org/10.1109/ICCV.2015.385>
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4, Article 139 (July 2023), 14 pages. <https://doi.org/10.1145/3592433>
- Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. 2019. ABC: A Big CAD Model Dataset For Geometric Deep Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ying-Miao Kuo, Hung-Kuo Chu, Ming-Te Chi, Ruen-Rone Lee, and Tong-Yee Lee. 2017. Generating Ambiguous Figure-Ground Images. *IEEE Transactions on Visualization and Computer Graphics* 23, 5 (2017), 1534–1545. <https://doi.org/10.1109/TVCG.2016.2535331>
- Weiyi Li, Rui Chen, Xuelin Chen, and Ping Tan. 2023a. SweetDreamer: Aligning Geometric Priors in 2D Diffusion for Consistent Text-to-3D. *arXiv:2310.02596* (2023).
- Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023b. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 2024. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6517–6526.
- Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. *arXiv:2303.11328* [cs.CV]
- Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuxin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2023. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv preprint arXiv:2310.15008* (2023).
- Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. 2024. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20654–20664.
- Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- Sehee Min, Jaedong Lee, Jungdam Won, and Jehee Lee. 2017. Soft Shadow Art. In *Computational Aesthetics*, Holger Winnemoeller and Lyn Bartram (Eds.). Association for Computing Machinery, Inc (ACM). <https://doi.org/10.1145/3092912.3092915>
- Niloy J. Mitra and Mark Pauly. 2009. Shadow art. In *ACM SIGGRAPH Asia 2009 Papers* (Yokohama, Japan) (*SIGGRAPH Asia ’09*). Association for Computing Machinery, New York, NY, USA, Article 156, 7 pages. <https://doi.org/10.1145/1661412.1618502>
- Pierre Moulon, Pascal Monasse, and Renaud Marlet. 2013. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In *2013 IEEE International Conference on Computer Vision*. 3248–3255. <https://doi.org/10.1109/ICCV.2013.403>
- Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. 2020. PolyGen: an autoregressive generative model of 3D meshes. In *Proceedings of the 37th International Conference on Machine Learning (ICML’20)*. JMLR.org, Article 669, 10 pages.
- Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. 127–136. <https://doi.org/10.1109/ISMAR.2011.6092378>
- Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Aude Oliva, Antonio Torralba, and Philippe G. Schyns. 2006. Hybrid images. *ACM Trans. Graph.* 25, 3 (July 2006), 527–532. <https://doi.org/10.1145/1141911.1141919>
- Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

- Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. 2024. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9914–9925.
- Zhiyu Qu, Lan Yang, Honggang Zhang, Tao Xiang, Kaiyue Pang, and Yi-Zhe Song. 2024. Wired Perspectives: Multi-View Wire Art Embraces Generative AI. In *CVPR*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. 2024. XCube: Large-Scale 3D Generative Modeling using Sparse Voxel Hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 10674–10685. <https://doi.org/10.1109/CVPR52688.2022.01042>
- Kaustubh Sadekar, Ashish Tiwari, and Shanmuganathan Raman. 2022. Shadow Art Revisited: A Differentiable Rendering Based Approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 29–37.
- Kaisei Sakurai, Yoshinori Dobashi, Kei Iwasaki, and Tomoyuki Nishita. 2018. Fabricating reflectors for displaying multiple images. *ACM Trans. Graph.* 37, 4, Article 158 (July 2018), 10 pages. <https://doi.org/10.1145/3197517.3201400>
- Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly. 2014. High-contrast computational caustic design. *ACM Trans. Graph.* 33, 4, Article 74 (July 2014), 11 pages. <https://doi.org/10.1145/2601097.2601200>
- Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. <https://doi.org/10.1109/CVPR.2016.445>
- Guy Sela and Gershon Elber. 2007. Generation of view dependent models using free form deformation. *Vis. Comput.* 23, 3 (Feb. 2007), 219–229. <https://doi.org/10.1007/s00371-006-0095-2>
- Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2023. MV-Dream: Multi-view Diffusion for 3D Generation. *arXiv:2308.16512* (2023).
- Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. 2024. MeshGPT: Generating Triangle Meshes with Decoder-Only Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.* 25, 3 (July 2006), 835–846. <https://doi.org/10.1145/1141911.1141964>
- Jiaxiang Tang. 2022. Stable-dreamfusion: Text-to-3D with Stable-diffusion. <https://github.com/ashawkey/stable-dreamfusion>.
- Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023a. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653* (2023).
- Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023b. Make-It-3D: High-fidelity 3D Creation from A Single Image with Diffusion Prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22819–22829.
- Kenji Tojo, Ariel Shamir, Bernd Bickel, and Nobuyuki Umetani. 2024. Fabricable 3D Wire Art. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) (*SIGGRAPH ’24*). Association for Computing Machinery, New York, NY, USA, Article 134, 11 pages. <https://doi.org/10.1145/3641519.3657453>
- Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).
- Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. 2023a. Neus2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213* (2023).
- Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Daqiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. 2024. CRM: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model. *arXiv preprint arXiv:2403.05034* (2024).
- Si-Tong Wei, Rui-Huan Wang, Chuan-Zhi Zhou, Baoquan Chen, and Wang Peng-Shuai. 2025. OctGPT: Octree-based Multiscale Autoregressive Models for 3D Shape Generation. In *SIGGRAPH*.
- Kang Wu, Renjie Chen, Xiao-Ming Fu, and Ligang Liu. 2022. Computational Mirror Cup and Saucer Art. *ACM Trans. Graph.* 41, 5, Article 174 (July 2022), 15 pages. <https://doi.org/10.1145/3517120>
- Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. 2023. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jamie Wynn and Daniyar Turmukhambetov. 2023. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*.
- Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. *arXiv preprint arXiv:2412.01506* (2024).
- Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. 2024. Multi-Scale 3D Gaussian Splatting for Anti-Aliased Rendering . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 20923–20931. <https://doi.org/10.1109/CVPR52733.2024.01977>
- Lior Yariv, Omri Puny, Oran Gafni, and Yaron Lipman. 2024. Mosaic-SDF for 3D Generative Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4630–4639. <https://doi.org/10.1109/CVPR52733.2024.00443>
- Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.
- Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2024. Mip-Splatting: Alias-free 3D Gaussian Splatting. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).
- Jiani Zeng, Honghao Deng, Yunyi Zhu, Michael Wessely, Axel Kilian, and Stefanie Mueller. 2021. Lenticular Objects: 3D Printed Objects with Lenticular Lens Surfaces That Can Change their Appearance Depending on the Viewpoint. In *The 34th Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST ’21)*. Association for Computing Machinery, New York, NY, USA, 1184–1196. <https://doi.org/10.1145/3472749.3474815>
- Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 2023. 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. *ACM Trans. Graph.* 42, 4, Article 92 (July 2023), 16 pages. <https://doi.org/10.1145/3592442>
- Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *arXiv preprint arXiv:2406.13897* (2024).
- Ruowen Zhao, Junliang Ye, Zhengyi Wang, Guangce Liu, Yiwen Chen, Yikai Wang, and Jun Zhu. 2025. DeepMesh: Auto-Regressive Artist-mesh Creation with Reinforcement Learning. *arXiv preprint arXiv:2503.15265* (2025).
- Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. 2023. FSGS: Real-Time Few-Shot View Synthesis using Gaussian Splatting. *arXiv:2312.00451 [cs.CV]*

## A PSEUDOCODE

---

**ALGORITHM 1:** Geometry Generation with LGAD

---

```

1: Input: Main views  $\mathcal{V} = \{v_i\}_{i=0}^{n-1}$  and semantics.  $\mathcal{Y} = \{y_i\}_{i=0}^{n-1}$ 
2: Choose: multi-view ND diffusion model with noise predictor  $\epsilon_{\text{pre}}^{3D}$ 
3: initialize Tetrahedron Splatting  $\theta$  as a sphere
4: while  $\theta$  is not converged do
5:   Sample: camera  $c$  in the neighbor of  $\mathcal{V}$ , horizontal surround
      cameras  $C = \{c_i\}$ 
6:    $\{w_i\} = \{1/(1 - c \cdot v_i)\}$ ,  $\{w'_i\} = \text{sort}(\{w_i\})$ 
7:    $y(c) = y_{\text{argmax}}\{w'_i\}$ 
8:    $\mathbf{g}_0 = \mathbf{n}_0$ ,  $\mathbf{d}_0 = P(\theta, c)$ .
9:   Sample:  $t, \epsilon_c$ .
10:  copy  $\mathbf{g}_t$  to obtain  $\{\mathbf{g}_{t,c'}\}_{c' \in C}$ 
11:   $\epsilon_{y,c} = \epsilon_{\text{pre}}^{3D} \left( \{\mathbf{g}_{t,c'}\}_{c' \in C}; t, y(c), C \right) |_c$ 
12:   $\epsilon_{\emptyset,c} = \epsilon_{\text{pre}}^{3D} \left( \{\mathbf{g}_{t,c'}\}_{c' \in C}; t, \emptyset, C \right) |_c$ 
13:   $s = s_0(w'_0 - w'_i)/\sum_i w'_i$ 
14:   $\hat{\epsilon}_c = s\hat{\epsilon}_{y,c} + (1-s)\epsilon_{\emptyset,c}$ 
15:   $\nabla_\theta \mathcal{L}_{\text{LGAD}} = \mathbb{E}_{t,\epsilon,c} \left[ \omega(t) (\hat{\epsilon}_c - \epsilon_c) \frac{\partial \mathbf{g}}{\partial \theta} \right]$ 
16:   $\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon,c} \left[ \omega(t) (\epsilon_\phi(\mathbf{n}_t; y(c), t) - \epsilon) \frac{\partial \mathbf{n}_0}{\partial \theta} \right]$ 
17:   $\mathcal{L}_{\text{eik}} = \sum_\delta (\|\nabla f_\delta\|_2 - 1)^2$ ,
18:   $\mathcal{L}_{\text{nc}} = \sum_e (1 - \cos(\mathbf{n}_{e1}, \mathbf{n}_{e2}))$ 
19:  update  $\theta$  with losses above
20: end while
21: return  $\theta$ 

```

---

## B EXPERIMENT DETAILS

*Selection of Semantics and observation Views.* We use ChatGPT to generate semantics, requiring certain correlations between each group of semantics to enhance aesthetic effects. When selecting observation views, some results adopt fixed patterns (e.g., horizontal surround, orthographic views), while others are randomly initialized before generation, with the constraint that the angle between any two observation views is no less than 120 degrees.

## C MORE EXPERIMENT RESULTS

To demonstrate our generation quality, we conduct comparisons with other prior-based 3D generation works under a single semantics as input. Since the guided semantics remain invariant to view changes in such a case, we sample 360-degree cameras and keep the guidance scale constant, and utilize four orthogonal ND maps as supervision. As shown in Fig. 16, our method achieves superior geometry and rendering quality compared to established works when given identical semantic inputs. And compared with the work also adopting tetrahedral splatting representations [Gu et al. 2024], our extracted mesh better approximates physically based rendering results with finer details.

We also conduct experiments with reduced distance between observation views, where shared geometry and rendering components significantly increased generation difficulty. As demonstrated in

Fig. 17, our method accomplishes the task even when the separation between two observation views is narrowed to 60 degrees.

## D 3D PRINTING RESULTS

We 3D-print several examples using both colored and white materials: white prints make it easy to observe geometric details, while colored prints can verify the final appearance.

	Input Semantics	Views	Mesh	View1		View2		View3	
				Render	Normal	Render	Normal	Render	Normal
1	“Apple” & “Isaac Newton” & “Open Book”	[0,0] [0,120] [0,-120]							
2	“Robot” & “Barren Planet” & “Lost Civilization”	[0,0] [0,120] [0,-120]							
3	“Skull” & “Compass” & “Treasure Chest”	[0,0] [0,120] [0,-120]							
4	“Cactus” & “Purple Succulent” & “Pond Dotted with Plants”	[0,0] [0,90] [90,0]							
5	“Grapes” & “Green Vine” & “Fox”	Random Views							
6	“Poison Apple” & “Crystal Coffin” & “Dwarf Cottage”	Random Views							
7	“Sydney Opera House” & “Great Wall of China” & “Taj Mahal”	[0,0] [0,120] [0,-120]							
8	“Blue Rose” & “Glass Cup” & “Pumpkin Carriage” & “Ball Gown”	[0,0] [0,90] [0,180] [0,-90]							
			Mesh	View1		View3		View4	

Fig. 11. **Gallery of Shape from Semantics.** We show the inputs, the generated colored mesh, the rendering and normal maps of each semantics. The textured meshes are rendered with Blender. The normal maps show that our generated shapes have meaningful geometries aligned with the renderings. Our method can complete generation with inputs of up to four semantics.

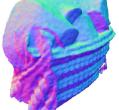
Mesh	View1 “Grandma wearing a red headscarf”	View2 “Basket with Goodies”	View3 “Grandma’s Cottage”
			
			
	[0,0]	[0,120]	[0,-120]
			
	[0,0]	[45,-127]	[-55,149]

Fig. 12. **Ablation Study on Different Observation Views.** Under fixed semantics, we present RGB and normal maps under an orbiting pattern (top row) and randomly sampled (bottom row). Our method adaptively generates matched shape compositions conditioned on view variations. It can be observed that with orbiting viewpoints, the relative independence of perspectives leads to slightly larger volumetric shapes for each semantic component.

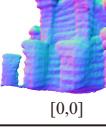
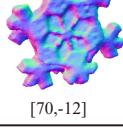
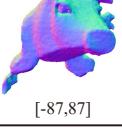
	View1 “Ice Castle”	View2 “Snowflake”	View3 “Reindeer”
Failure Case			
			
	[0,0]	[0,90]	[90,0]
Changed Views			
			
	[0,0]	[70,-12]	[-87,87]

Fig. 14. **Limitations of Our Results.** We present the renderings and normal maps of a failure case. Under certain semantics and observation views as inputs, the geometry we generate might collapse or distort. However, after changing the observation views, the generated results could become much more compatible.

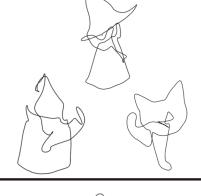
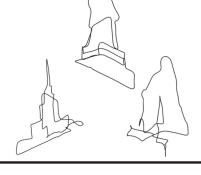
Input	Wire Art	Ours
“Chair” & “Tiger” [0,0] [0,180]		
“Running Dog” & “Flying Bird” [0,0] [0,-90]		
“Young Witch” & “Magic Cauldron” & “Black Cat” [0,0] [0,120] [0,-120]		
“Statue of Liberty” & “Empire State Building” & “Statue of Jesus” [0,0] [0,120] [0,-120]		

Fig. 13. **Comparison with Wire Art [Tojo et al. 2024].** We use the same semantics for comparison. The top-row result highlights the limitations of Wire Art, which arise from its dependence on projections to convey information, therefore, it is difficult to complement back/front design. All results demonstrate that our model can capture perceptual 3D characteristics while delivering high levels of creativity, visual appeal.

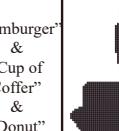
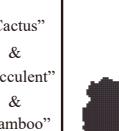
Input	Shadow Art	Wire Art	Ours
“Hamburger” & “Cup of Coffe” & “Donut”			
“Cactus” & “Succulent” & “Bamboo”			

Fig. 15. **Comparison with Similar Works.** We compare with Shadow Art [Mitra and Pauly 2009], Wire Art [Tojo et al. 2024]. Considering that Shadow Art only accepts binary images as input, the inputs for Wire Art during comparison are RGB images we rendered, while the inputs for Shadow Art are their masks. The input views are [0,0], [0,90], [90,0]. The results illustrate that our models effectively integrate multiple semantic elements, presenting the information in a manner that is more readily perceivable to observers.

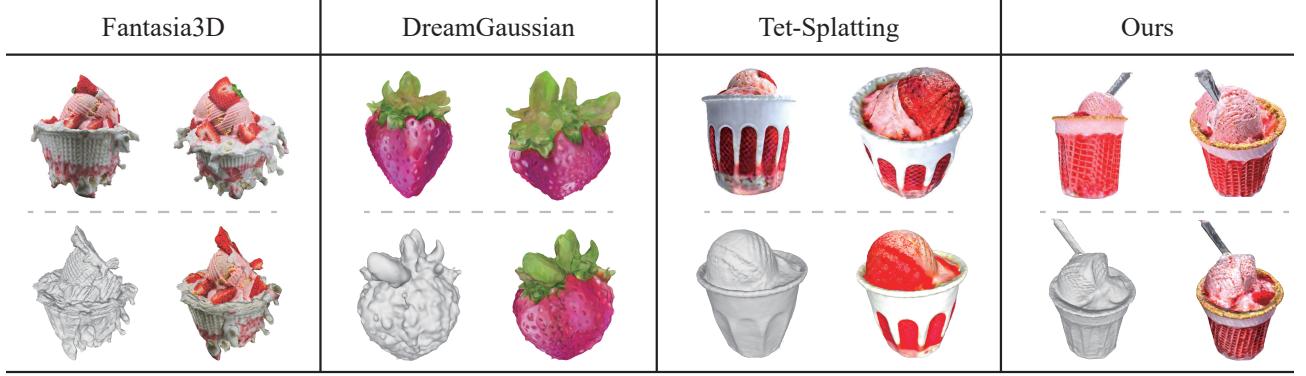


Fig. 16. **Qualitative Comparison of Text-to-3D Methods on a Single Semantic Concept.** The input semantics is “strawberry ice cream sundae in the cup”. The top row is the rendering and the bottom row is the extracted mesh with/without textures. The result of Fantasia3D [Chen et al. 2023] contains much noise, and the result of DreamGaussian [Tang et al. 2023a] deviates from semantics. Compared with Tet-Splatting [Gu et al. 2024], the mesh we extracted can better maintain the details in PBR.

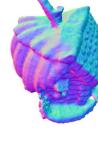
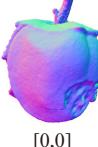
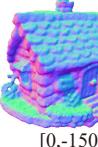
Mesh	View1 “Poison Apple”	View2 “Pumpkin Carriage”	View3 “Dwarf Cottage”
			
			
	[0,0]	[0,60]	[0,-150]

Fig. 17. **Experiments with Reducing Viewing Distance.** This shape achieves generation by associating the decayed portion of an apple with a wheel’s structure, which facilitates extensive geometry and rendering resource sharing.



Fig. 18. **3D Printing Results.** Results show that the manufactured outcomes are nearly identical to the simulations, delivering eye-catching visual effects.