DREAMSURFELS: SINGLE-IMAGE 3D GENERATION WITH GEOMETRY-CONDITIONED VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating a detailed 3D mesh from a single image requires both plausible completion of unseen regions and consistent 3D geometry across the generated views. Existing methods that reconstruct 3D from synthesized views can propagate inconsistencies across viewpoints into the final mesh. Video diffusion provides a powerful prior for coherent view sequences, yet visual coherence does not guarantee that the views share a consistent 3D surface. To this end, we introduce ***DreamSurfels***, a geometry-conditioned video diffusion framework for single-image 3D mesh generation. To provide explicit geometric guidance to video diffusion, we lift the visible surface from the input view into Gaussian surfels and render the partial geometry at target viewpoints. Residual conditioning branches inject these renderings into the diffusion backbone, anchoring the generated views to the observed surface as the model completes unseen regions. Beyond consistent view synthesis, accurate mesh reconstruction requires surface orientation cues that RGB appearance alone cannot reliably provide. We therefore develop a joint RGB-normal video diffusion model that uses intermediate RGB features to guide surface normal predictions, aligning geometry with the generated views. With the resulting geometrically consistent RGB views and surface normals, our framework then reconstructs detailed textured meshes. Experiments show that ***DreamSurfels*** produces consistent multi-view geometry and high-quality mesh reconstructions.

1 INTRODUCTION

Generating high-fidelity 3D meshes from a single image is important for applications including virtual reality, digital content creation, robotic simulation, and manipulation, where geometrically faithful 3D assets are essential (Zitkovich et al., 2023; Liu et al., 2023a; Dasari et al., 2020). Despite its broad utility, this task remains challenging because a single image provides only partial observations of the underlying 3D geometry. Single-image 3D mesh generation requires completing surfaces that are invisible in the input while preserving its visible structure. Methods that reconstruct a mesh from synthesized views face an additional requirement: these views must provide compatible evidence for the same surface. Disagreement across viewpoints can lead to distorted geometry, even when individual images appear plausible.

Diffusion priors offer a way to complete unseen regions through synthesized views (Rombach et al., 2022; Liu et al., 2023b; 2024). In particular, video diffusion models trained on large-scale video data provide a strong prior for temporal consistency, making them well suited for synthesizing view sequences along a camera trajectory (Voleti et al., 2024; Team Wan et al., 2025). However, temporal consistency does not guarantee that the generated content is strictly aligned in 3D across views, and even small geometric disagreements can introduce artifacts into the reconstructed mesh. Moreover, RGB appearance alone leaves surface orientation ambiguous, particularly in unseen regions, where the same appearance may correspond to multiple plausible geometries.

To address these challenges, we present ***DreamSurfels***, a geometry-conditioned video diffusion framework for single-image 3D mesh generation that leverages explicit geometry derived from the input view to anchor the generated views to a shared 3D structure. Specifically, we represent the visible surface with oriented Gaussian surfels (Dai et al., 2024) and render it from the target viewpoints to obtain pixel-aligned geometric guidance. Unlike unoriented point clouds (Yu et al., 2025) or standard 3D Gaussians (Tang et al., 2024) used in prior methods, Gaussian surfels explicitly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

encode surface orientation, allowing us to distinguish front-facing and back-facing regions and prevent back-face information from leaking into the rendered guidance. We inject these renderings into the video diffusion model through residual conditioning branches, preserving the input-derived geometry while guiding the generation of unseen regions.

While the visible surface is anchored by geometric guidance from the input view, the geometry of unseen regions remains ambiguous: even with generated RGB views, the same appearance can correspond to multiple plausible surface orientations. To resolve this ambiguity, surface normals are needed in addition to RGB views for mesh reconstruction. These normals should remain consistent across views while agreeing with the produced RGB completion. We therefore develop RGB-guided normal denoising to generate surface normals that satisfy both requirements. Specifically, we use RGB-to-normal adapters to inject intermediate features from the RGB stream into the corresponding blocks of the normal stream, coupling normal prediction to the generated appearance. The normal stream additionally receives partial normal renderings as geometric guidance, anchoring its predictions to the observed surface.

With geometrically consistent and mutually aligned RGB views and surface normals, we then use a coarse-to-fine reconstruction pipeline to recover a detailed textured mesh. We first fit Instant-NSR to the generated observations to obtain a stable coarse surface. We initialize Gaussian surfels from this surface and refine them to capture finer geometric structures. Finally, we apply screened Poisson reconstruction to extract the textured mesh. Experiments show that our method consistently outperforms existing baselines in both quantitative metrics and qualitative comparisons.

In summary, our main contributions are as follows:

- We propose ***DreamSurfels***, a geometry-conditioned video diffusion framework for single-image 3D mesh generation. It represents the visible surface with oriented Gaussian surfels and injects their rendered geometric guidance through residual conditioning branches, anchoring the generated views to the observed geometry while completing unseen regions.
- We introduce RGB-guided normal denoising for video diffusion to infer the surface orientation of unseen regions. RGB-to-normal adapters inject intermediate RGB features into the normal stream, producing surface normals that remain consistent across views and aligned with the generated RGB completion.
- We develop a coarse-to-fine reconstruction pipeline to generate high-quality textured meshes from the synthesized RGB views and surface normals. It uses Instant-NSR to recover a stable coarse shape and Gaussian surfels to capture finer geometric details.

2 Related Works

3D Generation with 2D Priors. Pretrained image diffusion models support 3D generation through score distillation (Poole et al., 2023; Wang et al., 2023) or image-conditioned novel-view synthesis (Liu et al., 2023b). Multi-view models generate target views jointly (Liu et al., 2024; Shi et al., 2023); Wonder3D (Long et al., 2024) additionally generates color and normals with cross-view and cross-modal attention, while Unique3D (Wu et al., 2024) and Era3D (Li et al., 2024) improve resolution and viewpoint modeling. Video diffusion provides a prior over camera trajectories (Voleti et al., 2024; Team Wan et al., 2025). ViewCrafter (Yu et al., 2025) guides video generation with point-cloud renderings, and MVGenMaster (Cao et al., 2025) uses depth-warped RGB and canonical-coordinate priors for multi-view generation. Our geometric condition instead renders the oriented input-view surface as partial RGB and normal maps, together with its projected support. For cross-modal generation, Wonder3D exchanges information across color and normal predictions; our video model transfers intermediate features from a frozen RGB stream to a normal stream so that normal denoising follows the current RGB sample. We use the residual context-adapter mechanism of VACE (Jiang et al., 2025) to inject the partial renderings; the surfel-derived conditions and directed RGB-to-normal adapters define the task-specific design.

Geometry-Aware Surface Reconstruction. Accurate mesh extraction from generated multi-view observations requires a representation that preserves surface structure. NeuS (Wang et al., 2021) represents surfaces as the zero-level set of a signed distance function and learns them through volume rendering; efficient implementations such as Instant-NSR (Guo, 2022) provide a practical

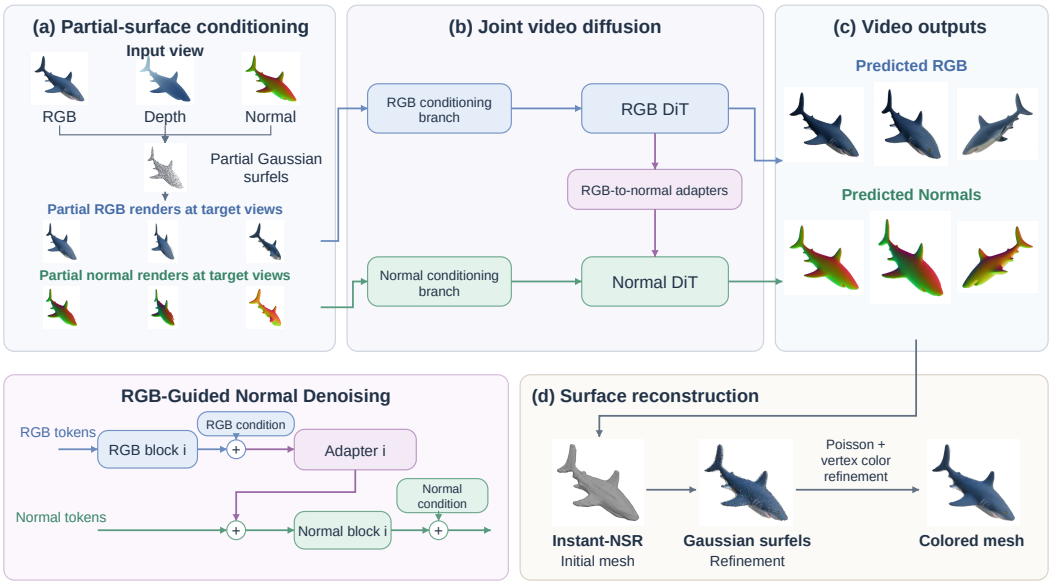

Figure 1: Overview of the ***DreamSurfels*** pipeline. (a) We lift the visible input-view surface into partial Gaussian surfels and render partial RGB and normal maps at target views as geometric guidance. (b) Residual conditioning branches guide the RGB and normal video diffusion streams, while RGB-to-normal adapters transfer intermediate RGB features to the normal stream, as illustrated in the lower panel. (c) The model predicts RGB views and aligned camera-space normals. (d) The generated views and normals support coarse-to-fine reconstruction: Instant-NSR provides an initial mesh, Gaussian surfels refine the surface, and screened Poisson reconstruction with color transfer yields the final mesh.

initialization for explicit surface refinement. Gaussian-based methods offer an alternative through efficient differentiable rasterization. While 3D Gaussian Splatting (Kerbl et al., 2023) uses volumetric primitives, SuGaR (Guédon & Lepetit, 2024) encourages surface alignment, and 2D Gaussian Splatting (Huang et al., 2024) and Gaussian Surfels (Dai et al., 2024) use flattened surface-oriented elements. GS2Mesh (Wolf et al., 2024) further studies mesh recovery from Gaussian representations using stereo observations. Our framework uses Gaussian surfels both to construct partial geometric conditions before generation and to refine a surface initialized by neural reconstruction, connecting the generative and mesh reconstruction stages.

3 Methods

3.1 Background

3.1.1 Video Diffusion Models

Video diffusion models (Blattmann et al., 2023; Team Wan et al., 2025) generate frame sequences in a shared video latent space. Our Wan backbone uses flow matching (Team Wan et al., 2025): given clean video latents $\mathbf{X}_0 \in \mathbb{R}^{F\times C\times H\times W}$ and Gaussian noise $\boldsymbol{\epsilon}$, the scheduler forms $\mathbf{X}_\sigma = (1-\sigma)\mathbf{X}_0 + \sigma\boldsymbol{\epsilon}$. The video model predicts the corresponding velocity $\mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{X}_0$ under conditioning $\mathbf{c}$, such as a reference image:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{X}_0,\boldsymbol{\epsilon},\sigma}\left[w(\sigma)\,\|\mathbf{v}_\theta(\mathbf{X}_\sigma,\sigma,\mathbf{c}) - \mathbf{v}^\star\|_2^2\right], \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $w(\sigma)$ is the scheduler's training weight. Previous work has explored how spatial observations can condition pretrained diffusion models without disrupting their generative priors. ControlNet augments a frozen image diffusion backbone with a trainable copy of its encoding blocks and connects the two branches through zero-initialized residual projections (Zhang et al., 2023). VACE extends this residual-conditioning design to video diffusion, using parallel context-adapter blocks to encode spatiotemporal observations and inject their outputs into the corresponding blocks

of the video diffusion transformer (Jiang et al., 2025). We follow this formulation to condition the video model on our partial geometric observations.

3.1.2 3D GAUSSIAN SURFELS

3D Gaussian Splatting (Kerbl et al., 2023) (3DGS) represents a scene as a collection of unstructured 3D Gaussians, each defined by a center position $\mu \in \mathbb{R}^3$, a covariance matrix Σ, an opacity value α, and spherical harmonics for view-dependent color c. To synthesize novel views, these Gaussians are projected onto the 2D image plane and blended along a camera ray $\mathbf{r}$ using point-based alpha compositing:

$$C(\mathbf{r}) = \sum_{i=1}^{N} c_i \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j'), \tag{2}$$

where α_i' denotes the 2D projected opacity of the i-th sorted Gaussian.

While this unconstrained volumetric formulation achieves real-time, high-fidelity rendering, the Gaussians often form chaotic, intersecting ellipsoids. This lack of explicit geometric constraints fails to accurately model underlying surfaces, resulting in noisy and discontinuous meshes that hinder geometry-aware downstream tasks. To bridge this gap, Gaussian Surfels (Dai et al., 2024) regularize the representation by constraining the 3D Gaussians to align with local surface manifolds. Specifically, the covariance matrix $\Sigma = \mathbf{RSS}^T\mathbf{R}^T$ is modified by flattening the scale matrix $\mathbf{S} = \text{diag}(s_x, s_y, s_z)$ along the local z-axis such that $s_z \to 0$. Concurrently, the rotation matrix $\mathbf{R}$ explicitly aligns this collapsed axis with the predicted surface normal $\mathbf{n}$. By structurally flattening the ellipsoids into 2D surface elements, this approach prevents the chaotic accumulation of volumetric density, encouraging the formation of continuous, physically plausible geometry that is ideal for incorporating generative diffusion priors.

3.2 OVERVIEW

Given a single reference RGB image I_{ref}, our goal is to generate a high-fidelity textured 3D mesh $\mathcal{M}$. Video diffusion provides a strong generative prior for synthesizing views along a camera trajectory $\mathcal{P} = \{\mathbf{p}_f\}_{f=1}^{F}$. However, its predictions may not remain geometrically consistent across views, causing artifacts when reconstructed into a 3D mesh. To address this challenge, ***DreamSurfels*** guides the generative process with geometry estimated from the input image. An overview of our framework is illustrated in Figure 1.

To encourage cross-view structural coherence, we represent the input-view geometry explicitly with a set of Gaussian surfels $\mathcal{S}_{\text{part}}$. Rendering them from each target-view pose $\mathbf{p}_f$ produces geometric guidance $\mathcal{G} = \{G_f\}_{f=1}^{F}$, where G_f denotes the partial rendering from that pose. These renderings guide the video diffusion model with projections of the same estimated surface, as detailed in Section 3.3.

Because RGB appearance alone cannot fully determine the surface orientations needed for mesh reconstruction, we introduce a joint RGB-normal video diffusion model. Geometric guidance $\mathcal{G}$ conditions RGB view synthesis, while intermediate features from the RGB stream guide a separate normal stream to predict surface normals aligned with those views, as described in Section 3.4. Finally, we use these outputs to reconstruct a detailed textured mesh, as described in Section 3.5.

3.3 GEOMETRY-AWARE VIDEO CONDITIONING

Video diffusion provides a strong prior for synthesizing a sequence of novel views, but its frame predictions are not inherently constrained to be projections of the same 3D surface. Even small cross-view geometric disagreements can produce artifacts in the reconstructed mesh. We therefore provide the denoiser with explicit geometric guidance $\mathcal{G}$ derived from geometry estimated from I_{ref}. Effective guidance should be pixel-aligned for dense spatial conditioning, rendered from a shared 3D representation across target poses, and efficient to generate. To this end, we adopt Gaussian surfels (Dai et al., 2024) as our explicit geometric representation. Surfels provide a compact surface model that can be efficiently rasterized into target views. Furthermore, unlike unoriented point clouds or standard 3D Gaussians, surfels explicitly encode local surface normals. This orientation awareness

supports visibility-aware rendering and helps prevent back-face appearance from leaking into the partial RGB and normal conditions.

Partial geometric guidance. Let Q_{ref} and N_{ref} denote a 3D point map and normal map inferred from I_{ref} and expressed in a common coordinate frame, and let Ω_{ref} be the set of valid object pixels. For each $\mathbf{u} \in \Omega_{\text{ref}}$, we initialize a Gaussian surfel with position, orientation, and color given by

$$\mathbf{x}_{\mathbf{u}} = Q_{\text{ref}}(\mathbf{u}), \qquad \mathbf{n}_{\mathbf{u}} = N_{\text{ref}}(\mathbf{u}), \qquad \mathbf{c}_{\mathbf{u}} = I_{\text{ref}}(\mathbf{u}). \tag{3}$$

These attributes define the partial Gaussian surfel set $\mathcal{S}_{\text{part}}$. We then render this set from each target-view pose $\mathbf{p}_f$ to obtain partial RGB and camera-space normal maps:

$$(\tilde{I}_f, \tilde{N}_f) = \mathcal{R}(\mathcal{S}_{\text{part}}, \mathbf{p}_f), \qquad f = 1, \dots, F. \tag{4}$$

Figure 2 highlights the effect of the surface-aware representation: while standard 3D Gaussians render observations without modeling surface orientation, Gaussian surfels suppress back-facing contributions when producing the partial conditions. As soft conditioning signals, these renderings encourage the generated views to follow the estimated visible geometry; they do not constrain unobserved surfaces or guarantee exact cross-view agreement.

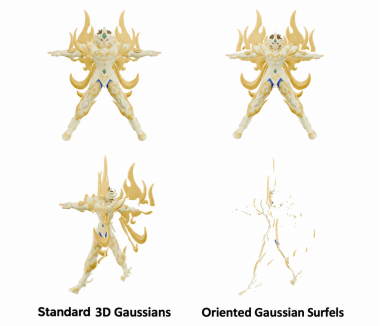

Figure 2: Differences between conditions rendered using standard 3D Gaussians and oriented Gaussian surfels.

Conditioning the video model. As illustrated in Figure 1(b), our joint video diffusion model has an RGB stream for appearance generation and a normal stream for surface orientation prediction. We use the partial RGB renderings to condition the RGB stream and the partial normal renderings to condition the normal stream. Following the context-adapter formulation (Jiang et al., 2025), we inject the geometric guidance into the video denoiser through additive residuals:

$$\mathbf{h}_{\ell}^{\text{RGB}} \leftarrow \mathbf{h}_{\ell}^{\text{RGB}} + \Delta_{\ell}^{\text{RGB}}(\tilde{\mathbf{I}}), \qquad \mathbf{h}_{\ell}^{\text{N}} \leftarrow \mathbf{h}_{\ell}^{\text{N}} + \Delta_{\ell}^{\text{N}}(\tilde{\mathbf{N}}). \tag{5}$$

Here $\mathbf{h}_{\ell}^{\text{RGB}}$ and $\mathbf{h}_{\ell}^{\text{N}}$ are the RGB and normal features at denoiser block ℓ; $\tilde{\mathbf{I}} = \{\tilde{I}_f\}_{f=1}^{F}$ and $\tilde{\mathbf{N}} = \{\tilde{N}_f\}_{f=1}^{F}$ denote the partial RGB and normal sequences. In practice, the residual output projections are initialized to zero, so these connections initially leave the pretrained denoiser unchanged at the beginning of training.

3.4 Joint RGB-Normal Video Diffusion

The geometric guidance $\mathcal{G}$ encourages the generated views to follow the input-view surface estimate, but the orientation of unseen surfaces must still be inferred. Since one input can admit several plausible completions, the normal maps should describe the particular completion produced by the RGB stream. Estimating normals from synthesized RGB views inherits the ambiguity of RGB appearance, while generating normals independently from the input view may disagree with those RGB views. We therefore couple RGB and normal video generation so that RGB features guide normal prediction toward agreement with the generated appearance.

Specifically, our joint model comprises two identically structured DiTs, one for RGB views and one for surface normals. We first train the RGB DiT for appearance synthesis with the geometric guidance described above. We then initialize the normal DiT from the trained RGB DiT and freeze the RGB stream while adapting the normal stream. The two DiTs process aligned video latent sequences along $\mathcal{P}$ at the same denoising timestep, with independent noise for the RGB and normal latents. Feature transfer is one-way: the normal stream receives features from the frozen RGB stream, while normal predictions do not feed back into RGB generation.

RGB-guided normal denoising. At each pair of corresponding transformer blocks, the RGB stream first updates its tokens under the partial RGB condition. Let $\mathbf{h}_{\ell,t}^{\text{RGB}}$ denote the resulting RGB tokens at block ℓ, and let $\mathbf{h}_{\ell,t}^{\text{N,in}}$ denote the input tokens of the corresponding normal block. An RGB-to-normal adapter A_{ℓ} transfers the aligned RGB features through an additive residual:

$$\bar{\mathbf{h}}_{\ell,t}^{\text{N,in}} = \mathbf{h}_{\ell,t}^{\text{N,in}} + A_{\ell}\big(\mathbf{h}_{\ell,t}^{\text{RGB}}\big), \tag{6}$$

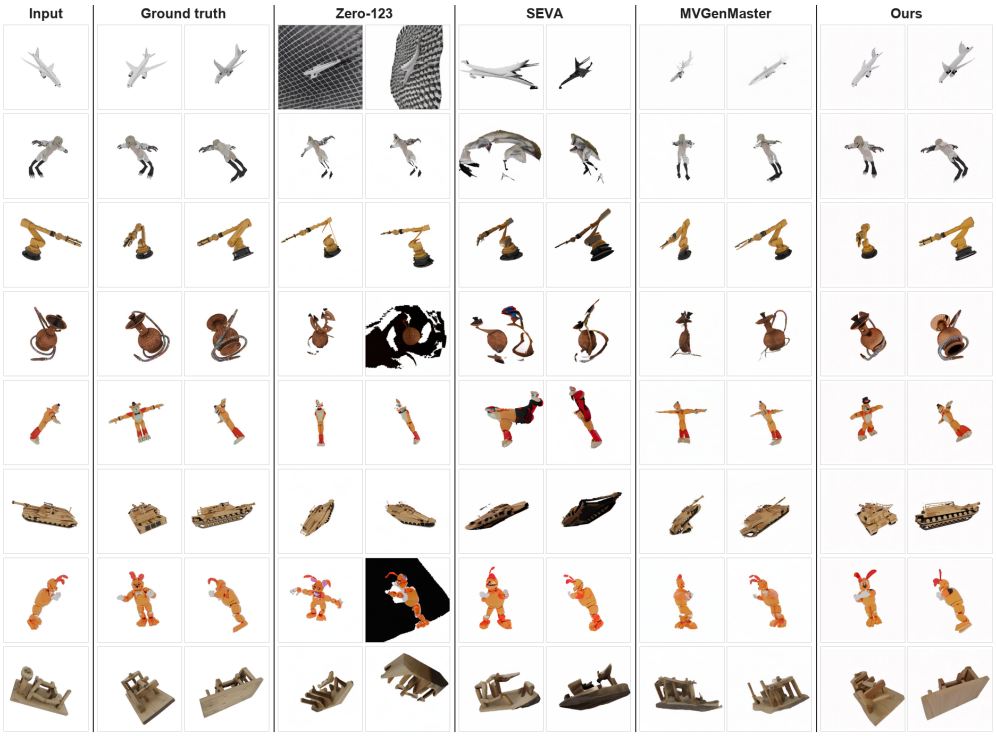

Figure 3: **Qualitative comparison of multi-view RGB generation on the Objaverse test set.** Each row shows a randomly rotated input and two matched target views for every method. Compared with the baselines, ***DreamSurfels*** better preserves object structure and appearance, with fewer missing parts and background artifacts.

The normal block then processes $\bar{\mathbf{h}}_{\ell,t}^{\mathrm{N,in}}$ under the partial normal condition, as illustrated in Figure 1(b). Each adapter A_ℓ is an MLP whose output projection is initialized to zero.

3.5 Surface Reconstruction and Mesh Extraction

Given the generated RGB views and surface normals, we reconstruct a textured 3D mesh. As illustrated in Figure 1(d), we first use Instant-NSR to recover a stable coarse shape and then refine it with Gaussian surfels to capture finer geometric structure.

Let $\hat{\mathcal{I}} = \{\hat{I}_f\}_{f=1}^F$ and $\hat{\mathcal{N}} = \{\hat{N}_f\}_{f=1}^F$ denote the generated RGB views and camera-space normal maps. We obtain foreground masks $\{\hat{M}_f\}_{f=1}^F$ for these views and fit Instant-NSR (Guo, 2022) using the generated observations, foreground masks, and known camera poses $\mathcal{P}$. Its implicit surface representation provides a stable estimate of the object's coarse shape, which we extract as an initial mesh $\mathcal{M}_0$.

To recover finer structure, we sample the surface of $\mathcal{M}_0$ to initialize a set of Gaussian surfels $\mathcal{S}_{\text{full}}$. We refine their geometry and appearance by matching rendered RGB images, normals, and silhouettes to $\hat{\mathcal{I}}$, $\hat{\mathcal{N}}$, and the foreground masks. Finally, we discard low-opacity surfels and apply screened Poisson reconstruction (Kazhdan & Hoppe, 2013) to obtain a continuous mesh. We remove small disconnected components and transfer the refined surfel colors to mesh vertices, yielding the final textured mesh $\mathcal{M}$.

4 Experiments

In this section, we conduct experiments to evaluate the performance of our proposed framework. Specifically, our evaluation aims to address three key questions: (a) whether Gaussian surfel guidance improves the visual quality and cross-view consistency of generated RGB views (Section 4.2); (b) whether RGB-guided normal denoising produces accurate surface normals aligned with the generated

Table 1: Quantitative results of multi-view RGB generation on 100 randomly rotated test objects. Metrics are computed over the 48 target views of each object, excluding the input view.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-123 (Liu et al., 2023b)	15.81	0.818	0.271
SEVA (Zhou et al., 2025)	14.05	0.867	0.232
MVGenMaster (Cao et al., 2025)	16.60	0.873	0.197
DreamSurfels	**18.42**	**0.909**	**0.132**

RGB views (Section 4.3); and (c) whether our coarse-to-fine reconstruction pipeline generates meshes with higher geometric fidelity than existing single-image methods (Section 4.4). Beyond these main evaluations, we also conduct ablation studies to validate our design choices. We present the ablation results in the supplementary material.

4.1 Implementation Details

Datasets: We use a subset of 103,679 objects from Objaverse (Deitke et al., 2023). We randomly select 100 objects for testing and use the remainder as our training pool. For training, we randomly rotate each object and render RGB images and camera-space normal maps along a fixed 49-frame camera trajectory at a resolution of 512×512 pixels.

Training Details: We use the pretrained Wan2.1-VACE-14B model (Team Wan et al., 2025; Jiang et al., 2025) as our video diffusion backbone. As described in Section 3.4, training proceeds in two stages. We first train the RGB stream, then initialize the normal stream from the trained RGB model and optimize it with RGB frozen. Both stages use rank-32 LoRA to adapt the DiT and conditioning branches. We train the RGB stream for approximately 30 k steps with a learning rate of 2×10^{-5}, taking about 5 days on 3 NVIDIA H100 80GB GPUs. We train the normal stream for approximately 10 k steps at 10^{-5}, taking about 3.5 days on 4 NVIDIA H100 80GB GPUs.

To initialize the partial Gaussian surfels from the input view, we use MoGe-2 (Wang et al., 2025) to estimate depth and camera-space normals from I_{ref}. We fine-tune MoGe-2 on the Objaverse training split using rendered RGB images, depth maps, and normal maps at 512×512 resolution. We freeze its image backbone and optimize other parts for 4 epochs with a learning rate of 10^{-5}.

4.2 Multi-view RGB Generation

We evaluate the quality and consistency of multi-view RGB generation from a single input image. For each of the 100 test objects, we randomly rotate the object as in training and render a 49-frame sequence along the fixed camera trajectory. The first frame corresponds to the input-view camera pose and is excluded from metric computation; we evaluate the remaining 48 target frames. We compare ***DreamSurfels*** with Zero-123 (Liu et al., 2023b), SEVA (Zhou et al., 2025), and MVGenMaster (Cao et al., 2025) using the same source view and target camera poses. ViewCrafter (Yu et al., 2025) also conditions video diffusion on point-cloud renderings, but its reported task-specific training uses scene videos from DL3DV and RealEstate10K rather than Objaverse object renders; we therefore exclude it from this object-centric benchmark.

Quantitative Results: We measure PSNR, SSIM, and LPIPS-Alex (Zhang et al., 2018) between the generated and ground-truth target frames; the results are reported in Table 1. ***DreamSurfels*** achieves a PSNR of 18.42, an SSIM of 0.909, and an LPIPS of 0.132. It improves PSNR over MVGenMaster by 1.82 dB and achieves the best SSIM and LPIPS among the compared methods. These results indicate that our method generates target views with greater fidelity to the ground truth.

Qualitative Results: As illustrated in Figure 3, baseline predictions can lose object parts, distort shapes, or introduce spurious backgrounds in novel views. MVGenMaster already uses depth-warped RGB and canonical-coordinate priors for geometric guidance, yet its results can still exhibit structural artifacts. In contrast, ***DreamSurfels*** preserves recognizable structures and appearance across the target viewpoints.

Table 2: Quantitative comparison of multi-view normal prediction on our Objaverse test set. All methods use the same canonical front-view RGB input and are evaluated at the four canonical orthographic views used by Unique3D.

Method	Cosine similarity ↑	Avg. angle diff. ↓
Wonder3D (Long et al., 2024)	0.4996	54.67°
Era3D (Li et al., 2024)	0.1923	81.71°
Unique3D (Wu et al., 2024)	0.4419	60.29°
DreamSurfels	**0.6418**	**43.74°**

Table 3: Quantitative results of 3D generation on our test set. All methods use the same front-view RGB input for each object.

Method	CD-L1 ↓	CD-L2^2 ↓	Vol-IoU ↑
Wonder3D (Long et al., 2024)	0.07446	0.00844	0.39195
Era3D (Li et al., 2024)	0.09431	0.01217	0.30732
Unique3D (Wu et al., 2024)	0.06806	0.00758	0.39869
DreamGaussian (Tang et al., 2024)	0.07599	0.00799	0.39798
InstantMesh (Xu et al., 2024)	0.08632	0.00892	0.33216
DreamSurfels	**0.05504**	**0.00587**	**0.44043**

4.3 Multi-view Normal Generation

We evaluate multi-view normal prediction on the 100 objects in our Objaverse test set. Since the compared methods use canonical orthographic cameras, we render the canonical front view of each object and use it as the common RGB input for all methods. We compare ***DreamSurfels*** with Wonder3D (Long et al., 2024), Era3D (Li et al., 2024), and Unique3D (Wu et al., 2024). Following the output camera configuration of Unique3D, we evaluate the predicted normal maps at four canonical orthographic views: front, right, back, and left. We measure cosine similarity and average angular error against the ground-truth normals on foreground pixels.

As shown in Table 2, ***DreamSurfels*** achieves the highest cosine similarity and the lowest average angular error, demonstrating more accurate multi-view normal prediction than the compared baselines.

4.4 3D Generation

We evaluate the final mesh reconstructions on our Objaverse test set. Era3D (Li et al., 2024) and Unique3D (Wu et al., 2024) exhibit reduced robustness to randomly rotated input views in our evaluation. We therefore use the rendered front view of each object as the common RGB input for all methods. We compare ***DreamSurfels*** with Wonder3D (Long et al., 2024), Era3D (Li et al., 2024), Unique3D (Wu et al., 2024), DreamGaussian (Tang et al., 2024), and InstantMesh (Xu et al., 2024).

Quantitative Results: We measure Chamfer distance (CD-L1 and CD-L2^2) and volumetric intersection over union (Vol-IoU) between the reconstructed and ground-truth meshes; the results are reported in Table 3. We normalize distances by each ground-truth mesh's bounding-box diagonal before computing the CD metrics. ***DreamSurfels*** achieves the best performance across all three metrics, with a CD-L1 of 0.05504, a CD-L2^2 of 0.00587, and a Vol-IoU of 0.44043. Compared with the strongest baseline, Unique3D, our method reduces CD-L1 and CD-L2^2 by 19.1% and 22.6%, respectively, while improving Vol-IoU by 4.17 percentage points. These consistent gains in surface distance and volumetric overlap demonstrate that our coarse-to-fine reconstruction produces meshes that more faithfully recover the ground-truth geometry.

Qualitative Results: In addition to the Objaverse test set, we evaluate three synthetic object images generated with OpenAI's image generation tool. Figure 4 compares ***DreamSurfels*** with DreamGaussian, Wonder3D, Era3D, Unique3D, and InstantMesh. The baselines can lose fine structures or introduce distorted surfaces and appearance in views unseen from the input. In contrast, ***DreamSurfels*** preserves the characteristic structures of the input objects and yields coherent geometry and appearance across the displayed views.

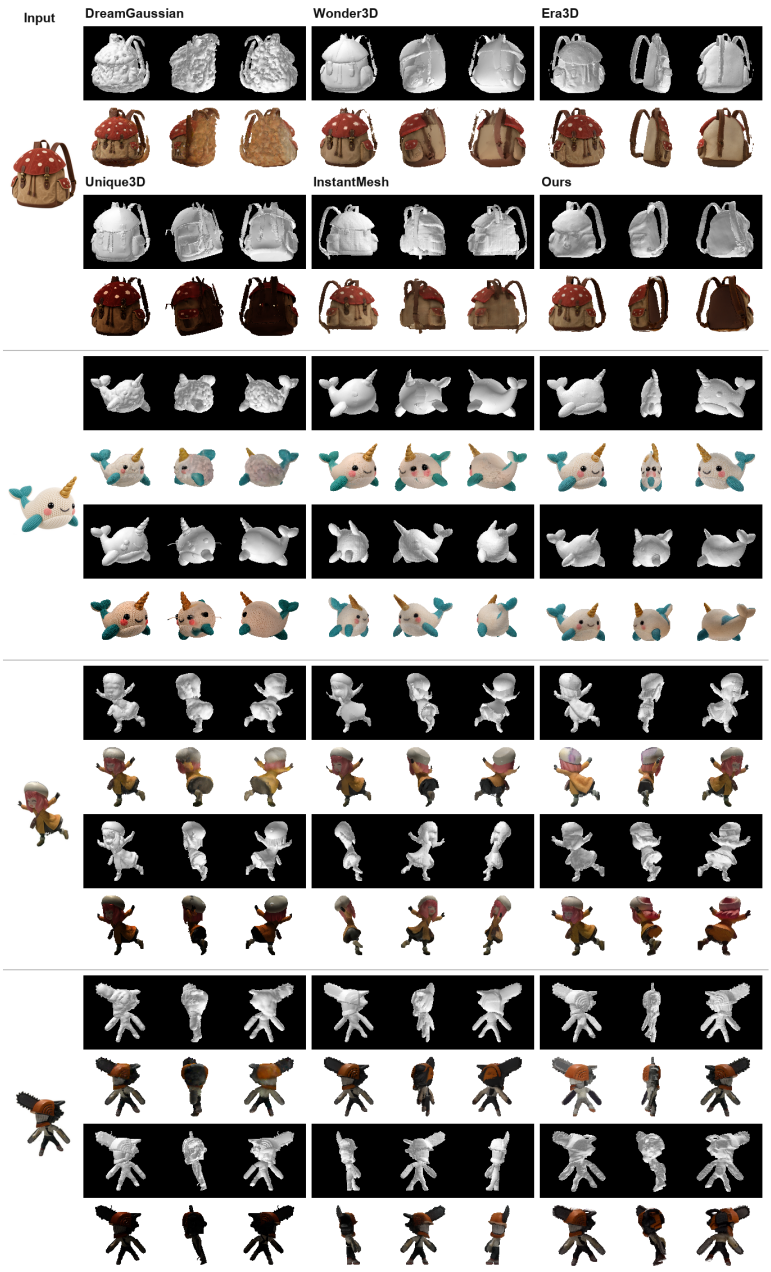

Figure 4: **Qualitative 3D reconstruction from image-generated inputs.** Each method uses the same input image. Within each method, the three columns show input-facing, side, and rear views; the upper and lower rows show geometry (white on black) and the colored mesh, respectively. Each mesh is centered and scaled by its bounding box for visualization.

5 Conclusion

We presented ***DreamSurfels***, a framework for high-fidelity single-image 3D mesh generation that uses partial Gaussian surfel renderings as geometric guidance for video diffusion. By coupling RGB views and surface normals through RGB-guided normal denoising, our model produces aligned predictions for coarse-to-fine mesh reconstruction. Experiments on the Objaverse dataset demonstrate that our approach generates geometrically consistent views and high-quality textured meshes faithful to the reference input.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6045–6056, 2025.

Pinxuan Dai, Jiamin Xu, Wenxiang Xie, Xinguo Liu, Huamin Wang, and Weiwei Xu. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657441.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. RoboNet: Large-scale multi-robot learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 885–897, 2020.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5354–5363, 2024.

Yuan-Chen Guo. Instant neural surface reconstruction. `https://github.com/bennyguo/instant-nsr-pl`, 2022.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657428.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3), 2013. doi: 10.1145/2487228.2487237.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. Datasets and Benchmarks Track.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12619–12629, 2023.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yaniv Wolf, Amit Bracha, and Ron Kimmel. Gs2mesh: Surface reconstruction from gaussian splatting via novel stereo views. In *European Conference on Computer Vision*, pp. 207–224. Springer, 2024.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3613256.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12405–12414, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023.

Table 4: Ablation of RGB-guided normal diffusion on the Objaverse test set. Both variants use the same partial-normal conditioning and differ only in whether intermediate RGB features are injected into the normal stream. RGB-normal foreground IoU measures the spatial alignment between the generated RGB views and predicted normal maps.

Method	Cosine similarity ↑	Avg. angle diff. ↓	RGB-normal fg. IoU ↑
w/o RGB guidance	0.4073	61.05°	0.4276
w/ RGB guidance	**0.6418**	**43.74°**	**0.9197**

A Reproducibility Statement

We describe the model architecture, geometric conditioning, RGB-guided normal diffusion, and surface reconstruction procedure in Sections 3.3 to 3.5. We report the dataset split, rendering protocol, training schedule, compute, and evaluation settings in Sections 4.1 to 4.4. The appendix provides inference-time and ablation details, and the supplementary archive contains the complete reconstructed meshes for our method and all evaluated baselines. We plan to release the training and inference code, model weights, and evaluation scripts upon publication.

B AI Use Statement

We used OpenAI Codex, an LLM-based assistant, to help implement our method and write and revise code for training, inference, and experimental evaluation. We also used it to revise and edit the manuscript, assist in preparing Figure 1, and arrange experimental comparison figures using output images from the respective methods. The authors reviewed the AI-assisted code, figures, and text and take responsibility for the final content of this paper.

C Inference Time

On a single NVIDIA A100-SXM4-80GB GPU, our pipeline takes approximately 15 minutes to generate a textured mesh from one input image, including input-view geometry estimation, geometric guidance, 49-view RGB and normal generation, and surface reconstruction.

D Ablation of RGB-Guided Normal Diffusion

We ablate the effect of RGB guidance in normal diffusion. We compare two variants with the same normal diffusion backbone and partial-normal conditioning. The first predicts complete multi-view normals directly from the partial normal renderings without using RGB information. The second uses our RGB-guided design, which injects intermediate features from the RGB stream into the normal stream through RGB-to-normal adapters. We evaluate both variants using the same canonical orthographic test protocol as in the main paper. We measure cosine similarity and average angular error against the ground-truth normals on foreground pixels. We additionally report the foreground IoU between the generated RGB views and predicted normal maps to evaluate their spatial alignment. This comparison isolates whether conditioning normal denoising on the generated RGB views improves both normal accuracy and RGB-normal consistency.

As shown in Table 4, RGB guidance substantially improves all three metrics, increasing cosine similarity from 0.4073 to 0.6418, reducing the average angular error by 17.31°, and improving RGB-normal foreground IoU from 0.4276 to 0.9197. These gains demonstrate that RGB guidance produces normals that are both more geometrically accurate and substantially better aligned with the generated RGB views.

E Additional Results

Additional results are provided in the supplementary ZIP archive. The archive contains the complete reconstructed meshes produced by ***DreamSurfels*** and all compared baseline methods for the evaluation examples, enabling inspection of their geometry and appearance from arbitrary viewpoints beyond the rendered views shown in the paper.

F LIMITATION AND BROADER IMPACT

Limitations: Our coarse-to-fine reconstruction requires per-object optimization, resulting in higher computational costs and longer inference times compared to direct feed-forward models. Additionally, the framework may struggle with extremely intricate topologies or severe self-occlusions where the single reference image provides insufficient geometric cues for accurate reconstruction.

Broader Impact: While benefiting digital content creation, this technology could be misused for intellectual property infringement. Furthermore, the significant compute required for training and optimization entails a non-negligible environmental footprint. We advocate for the development of 3D watermarking and more energy-efficient training paradigms to mitigate these risks.