BEYOND RGB: POLARIZATION-GUIDED SURFACE-AWARE 3D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-view 3D generation relies on strong RGB generative priors. However, 3D inference from RGB appearance alone remains geometrically ambiguous: a model may produce a visually plausible object while failing to recover local surface orientation and fine-scale geometry, particularly in weakly textured or highly reflective regions. Unlike RGB appearance, the angle of linear polarization (AoLP) is related to surface orientation and provides a complementary cue to local geometry. We propose polarization-guided surface-aware 3D generation, which retains the prior of an RGB-conditioned generator while using polarization to refine its modeling of local surfaces. A surface-aware feature fusion module jointly encodes semantic and appearance information from RGB and orientation cues from AoLP, yielding a normal-aware condition for the denoising process of a pretrained 3D generator. To support training and evaluation, we also construct OBJMV3D with 60,000 paired groups of RGB images, AoLP observations, surface normals, and high-quality 3D assets. Experiments on this benchmark show that polarization-derived surface cues improve fine-scale geometry while preserving overall generation quality, with more pronounced geometric benefits where RGB evidence is limited by reflectance or weak texture.

1 INTRODUCTION

Single-view 3D generation aims to recover a 3D asset with complete geometry and realistic appearance from one 2D observation. Recent advances in diffusion and large-scale generative models have substantially improved this task. By learning semantic, shape, and appearance priors from large RGB datasets, existing systems can generate objects with coherent global structure and high visual quality from a single image. The field has also progressed from optimization-based representations toward more efficient feed-forward generators (Hong et al., 2023; Tochilkin et al., 2024; Wang et al., 2024; Xiang et al., 2025; Li et al., 2025b). These results show the strength of RGB-based global 3D priors.

Global plausibility, however, does not guarantee accurate local surface geometry. An RGB image records the combined effects of object geometry, material, and illumination; different surface shapes or reflectance properties can produce similar appearance. The local surface orientation is therefore difficult to determine uniquely from RGB alone. Generators infer the missing geometry from statistical regularities in their training data. Although this can yield plausible objects, it can also smooth local structure, produce incorrect curvature, or hallucinate details. Weakly textured surfaces offer little appearance variation from which to infer geometry, while highly reflective surfaces are dominated by view-dependent highlights and environmental reflections. The metallic object in Figure 1 illustrates this difficulty: an RGB-based method recovers its approximate form but not its surface structure as reliably.

This motivates reconsidering the information available to a single-view generator. If RGB already supplies a strong generative prior, the missing input is a complementary observation of surface geometry rather than another source of semantic information. Polarization imaging offers such an observation. Reflection changes the polarization state of light according to surface orientation and reflectance; in particular, the angle of linear polarization (AoLP) has a physical relationship to surface azimuth (Ba et al., 2020). Instead of inferring geometry indirectly from texture, color, or shading, AoLP provides local orientation evidence. Previous shape-from-polarization and polarimetric reconstruction studies have demonstrated its value for recovering complex surfaces (Lei et al., 2022; Zhao et al., 2022).

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.

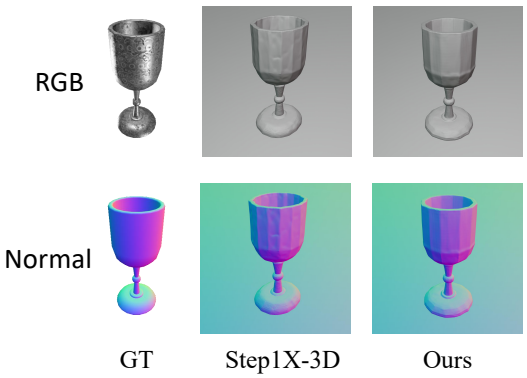


Figure 1: Representative synthetic comparison on a reflective object. RGB-based generation captures the approximate object shape, while the polarization-assisted result better preserves local surface structure in this example.

Our central premise is that polarization should complement the strong 3D prior learned from RGB, refining the surface geometry that RGB appearance cannot reliably determine. RGB supplies object category, global shape, and appearance; AoLP more directly reflects local orientation. We therefore build a surface-aware representation in which the RGB prior remains the foundation and polarization strengthens sensitivity to normals and fine-scale structure. This division allows the generator to retain the benefits of large-scale RGB training while applying polarization where it provides distinct geometric evidence.

We introduce Polarization-Guided Surface-Aware 3D Generation. Rather than treating RGB and AoLP as interchangeable modalities in a generic fusion block, we extract features according to their different roles and align them in a surface-aware fusion module. The resulting normal-aware condition is injected into a pretrained 3D generator so that polarization-derived surface cues influence the generated 3D latent.

Conditioning alone cannot ensure that the final explicit surface is accurate: the generator operates in a 3D latent space, whereas the desired improvement concerns mesh geometry. We therefore constrain both the latent and the renderable surface. A latent consistency term aligns the predicted representation with the ground-truth 3D latent. Separately, we render the generated object from multiple views and compare its normals with normals rendered from the target mesh. This passes local orientation supervision back into the 3D generation process.

Polarization-guided generation also lacks large paired datasets containing RGB, polarization, and high-quality 3D geometry. We construct OBJMV3D with 60,000 paired groups of RGB images, AoLP observations, surface normals, and corresponding 3D assets, providing data for training and evaluating surface-aware generation.

Our contributions are:

- We propose a polarization-guided surface-aware 3D generation framework that retains the global prior of an RGB generator while using AoLP orientation evidence to refine local geometry.
- We design surface-aware RGB–polarization fusion and geometric supervision. A normal-aware condition guides 3D generation, while latent consistency and multi-view rendered-normal losses act on the generated shape.
- We construct OBJMV3D with 60,000 paired RGB, AoLP, normal, and 3D asset groups. Experiments show improved fine-scale geometry, especially on reflective and weakly textured objects where RGB alone provides limited geometric evidence.

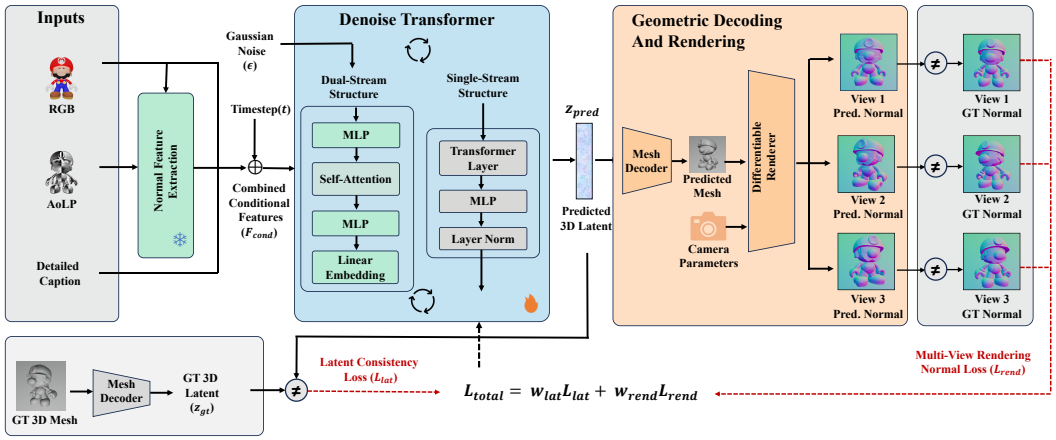


Figure 2: Polarization-guided surface-aware 3D generation. RGB–AoLP features form a normal-aware condition for a pretrained DiT. During training, the target mesh provides the ShapeVAE latent and rendered normals; latent consistency and multi-view rendered-normal losses supervise the generator.

2 Related Work

2.1 Single-view 3D generation

Single-view 3D generation has progressed from category-specific adversarial models (Smith & Meger, 2017; Zhu et al., 2018) to optimization-based systems that distill 2D diffusion priors (Qian et al., 2023; Tang et al., 2023; Wang et al., 2023) and feed-forward models that predict multi-view images or native 3D representations (Liu et al., 2023; Hong et al., 2023; Long et al., 2024; Wang et al., 2024). Recent systems, including TripoSR, Era3D, TRELLIS, Hash3D, and Step1X-3D, improve generation speed, resolution, and representation capacity (Tochilkin et al., 2024; Li et al., 2024b; Xiang et al., 2025; Yang et al., 2025; Li et al., 2025b). Their large-scale RGB training provides strong semantic and global-shape priors. However, RGB appearance depends jointly on geometry, material, and illumination, leaving local surface orientation ambiguous, particularly on reflective or weakly textured objects.

Our framework builds on pretrained Step1X-3D and introduces an aligned AoLP observation to provide complementary surface-orientation evidence. The experiments therefore compare systems operating under different sensing conditions. Comparisons with RGB-only generators characterize overall performance, while the component ablations evaluate the contributions of surface conditioning and rendered-normal supervision within the adapted framework.

2.2 Polarization and geometric inference

Shape from Polarization relates measured polarization to surface orientation through Fresnel reflection. Analytical and learned approaches combine polarization cues with contextual priors or multi-view constraints to address ambiguous normals and complex reflection behavior (Ba et al., 2020; Lei et al., 2022; Muglikar et al., 2023; Zhao et al., 2022). AoLP provides evidence about surface azimuth, but its periodicity and dependence on the reflection regime prevent unique recovery of surface normals from polarization alone. RGB context supplies complementary information for interpreting these measurements.

Polarization has also been incorporated into neural inverse rendering and Gaussian-splatting pipelines, including NeISF, NeISF++, and PolGS (Li et al., 2024a; 2025a; Han et al., 2025). These methods optimize a representation for an individual observed scene. Our framework learns a conditional generator across objects: a normal-supervised RGB–AoLP encoder produces surface-aware features that condition a pretrained 3D generation backbone.

2.3 LATENT 3D GENERATION AND GEOMETRIC SUPERVISION

Learned shape representations encode geometry in compact latent spaces and recover surfaces through a decoder (Park et al., 2019). Transformer-based diffusion architectures support conditional generation (Vaswani et al., 2017; Peebles & Xie, 2023), and recent 3D generators apply such architectures to structured or shape latents (Xiang et al., 2025; Li et al., 2025b). These representations allow a pretrained generator to retain broad shape priors while accepting additional conditioning information.

Our method introduces geometric constraints at both the conditioning and output stages. A normal-supervised surface representation supplies local orientation evidence to the generator. Latent consistency aligns the predicted clean latent with the target ShapeVAE encoding, while multi-view rendered-normal supervision constrains the decoded mesh. Together, these objectives connect the input surface cues to the geometry of the generated asset.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK

Given a single-view RGB image I_{rgb}, its pixel-aligned AoLP observation I_{pol}, and text c, we generate a mesh

$$\hat{\mathcal{M}} = \mathcal{G}_\Theta(I_{\mathrm{rgb}}, I_{\mathrm{pol}}, c). \tag{1}$$

RGB appearance depends jointly on geometry, material, and illumination, leaving local surface orientation ambiguous. Our framework retains a pretrained RGB-based 3D prior and adds a normal-aware surface condition. Training supervises both the shape latent and multi-view normals rendered from the generated mesh. Target meshes and their renderings are used only for training and evaluation; inference uses the three input conditions, with text derived from the observed RGB image.

3.2 POLARIZATION-GUIDED SURFACE-AWARE REPRESENTATION

Complementary orientation evidence. For linear-polarization measurements at $0°$, $45°$, $90°$, and $135°$, we compute

$$\begin{aligned} S_0 &= I_0 + I_{90}, \quad S_1 = I_0 - I_{90}, \quad S_2 = I_{45} - I_{135}, \\ \phi &= \tfrac{1}{2}\arctan2(S_2, S_1) \pmod{\pi}. \end{aligned} \tag{2}$$

Under ideal diffuse and specular reflection models, AoLP relates to normal azimuth φ_n as $\phi_{\mathrm{d}} = \varphi_n \pmod{\pi}$ and $\phi_{\mathrm{s}} = \varphi_n + \pi/2 \pmod{\pi}$, respectively (Ba et al., 2020; Lei et al., 2022). Its π-periodicity and the unknown reflection regime prevent unique normal inversion. RGB context helps resolve these ambiguities, while AoLP supplies complementary local orientation evidence.

Surface encoding and normal supervision. A frozen DINOv3 encoder (Siméoni et al., 2025) extracts multi-scale RGB features; a trainable U-Net branch encodes AoLP. Channel projections align the two streams, and a U-Net decoder aggregates their fused features across scales to produce $\mathbf{F}_{\mathrm{surf}}$. A normal head predicts $\tilde{\mathbf{N}} = D_{\mathrm{n}}(\mathbf{F}_{\mathrm{surf}})$, with $\hat{\mathbf{n}}(p) = \tilde{\mathbf{N}}(p)/(\|\tilde{\mathbf{N}}(p)\|_2 + \varepsilon)$ and a small stabilizer $\varepsilon > 0$. On valid foreground pixels Ω, the module is trained with

$$\mathcal{L}_{\mathrm{normal}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \big[1 - \langle \hat{\mathbf{n}}(p), \mathbf{n}_{\mathrm{gt}}(p) \rangle\big]. \tag{3}$$

For unit normals, this penalizes $1 - \cos\Delta\theta$ rather than vector magnitude. We then freeze the trained surface encoder and normal head. A projection constructs surface tokens $\mathbf{C}_{\mathrm{surf}} = P_{\mathrm{surf}}(\mathbf{F}_{\mathrm{surf}}, \hat{\mathbf{N}}) \in \mathbb{R}^{N_s \times d_c}$, where N_s is the token count and d_c the DiT condition dimension. Figure 3 summarizes this geometric bottleneck.

3.3 SURFACE-AWARE 3D LATENT GENERATION

The pretrained ShapeVAE encodes the target mesh as $\mathbf{z}_{\mathrm{gt}} = E_{\mathrm{3D}}(\mathcal{M}_{\mathrm{gt}})$. We retain the backbone's RGB and text encoders and concatenate their conditions with the surface tokens:

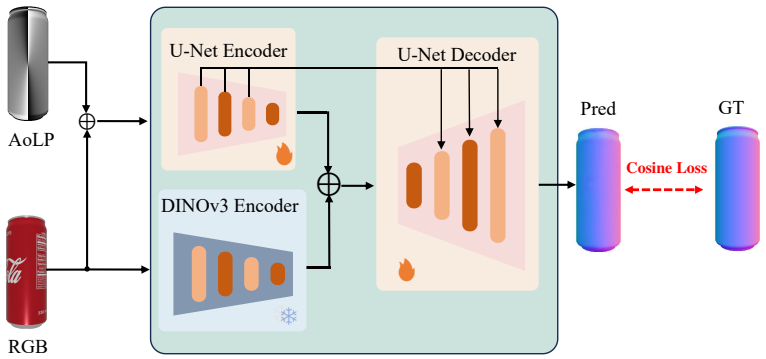


Figure 3: Surface encoder. A trainable AoLP branch and frozen DINOv3 RGB branch produce fused features and normals under cosine supervision. The trained representation supplies surface condition tokens.

$\mathbf{C} = \mathrm{Concat}(\mathbf{C}_{\mathrm{rgb}}, \mathbf{C}_{\mathrm{surf}}, \mathbf{C}_{\mathrm{text}})$. The original noise schedule and output parameterization are preserved:

$$\begin{aligned} \mathbf{z}_t &= \alpha(t)\mathbf{z}_{\mathrm{gt}} + \sigma(t)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{o}_t &= D_\theta(\mathbf{z}_t, t, \mathbf{C}), \qquad \hat{\mathbf{z}} = \Psi_t(\mathbf{z}_t, \mathbf{o}_t). \end{aligned} \tag{4}$$

Here Ψ_t is the base model's deterministic conversion from its output to a clean latent. Surface tokens extend the conditioning sequence while preserving the 3D latent state space. The mesh is decoded as $\hat{\mathcal{M}} = D_{\mathrm{mesh}}(\hat{\mathbf{z}})$.

3.4 Explicit Surface Constraint in Rendering Space

For cameras $\{\pi^{(v)}\}_{v=1}^{V}$, a differentiable renderer produces $\hat{\mathbf{N}}^{(v)} = \mathcal{R}_N(\hat{\mathcal{M}}, \pi^{(v)})$ and target normals $\mathbf{N}_{\mathrm{gt}}^{(v)} = \mathcal{R}_N(\mathcal{M}_{\mathrm{gt}}, \pi^{(v)})$. The target maps come directly from the ground-truth mesh. Both meshes use identical camera intrinsics, extrinsics, resolutions, and normal-coordinate conventions. With binary foreground-validity masks $M^{(v)}$, the loss is

$$\mathcal{L}_{\mathrm{rend}} = \frac{\sum_{v=1}^{V}\sum_{p} M^{(v)}(p)\|\hat{\mathbf{N}}^{(v)}(p) - \mathbf{N}_{\mathrm{gt}}^{(v)}(p)\|_2^2}{\sum_{v=1}^{V}\sum_{p} M^{(v)}(p) + \varepsilon}. \tag{5}$$

For unit normals, squared error equals $2(1 - \cos\Delta\theta)$. This objective aggregates orientation error over valid pixels across views, with gradients passing through the renderer and mesh decoder to the DiT. It constrains the final surface in addition to the input condition.

3.5 Optimization and Training

Training first learns the surface module with $\mathcal{L}_{\mathrm{normal}}$ and freezes its encoder and normal head. The second stage adapts pretrained Step1X-3D (Li et al., 2025b) through LoRA and the added conditioning modules while retaining the original backbone weights. Only the latent and rendered-normal objectives optimize 3D generation:

$$\begin{aligned} \mathcal{L}_{\mathrm{lat}} &= \frac{1}{N_z}\|\hat{\mathbf{z}} - \mathbf{z}_{\mathrm{gt}}\|_2^2, \\ \mathcal{L}_{\mathrm{total}} &= \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}} + \lambda_{\mathrm{rend}}\mathcal{L}_{\mathrm{rend}}. \end{aligned} \tag{6}$$

Here N_z is the number of latent elements. Latent consistency preserves global shape, while rendering supervision regularizes local surface orientation. The normal-head loss is used to learn the first-stage bottleneck and is not an additional second-stage generator loss.

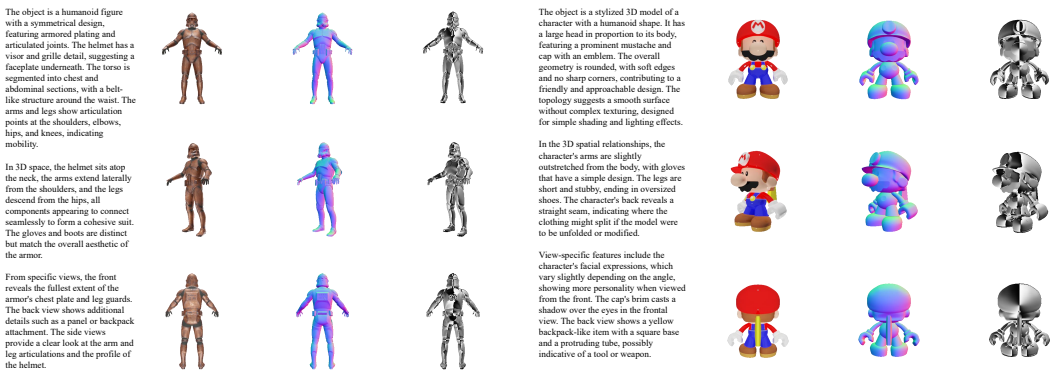


Figure 4: Representative OBJMV3D annotations. Two assets illustrate structured training text and camera-aligned RGB, mesh normals, and AoLP. The displayed views are subsets of the six canonical views stored for each object.

4 OBJMV3D: A Controlled Synthetic Benchmark

OBJMV3D comprises 60,000 curated Objaverse assets (Deitke et al., 2023), including everyday objects, industrial parts, and objects with complex surface structures. Each asset has a standardized mesh, text annotations, and six camera-aligned views at 512×512 resolution. The 360,000 aligned view tuples contain RGB, AoLP, mesh-rendered normals, foreground masks, and camera intrinsics and extrinsics. Figure 4 shows representative paired annotations.

We retain single, closed objects and remove scene compositions, low-poly assets with insufficient surface detail, and invalid meshes with conspicuous holes, non-manifold structures, or flipped normals. The construction pipeline includes geometric repair and standardization. Assets are centered at the world origin, normalized to a unit bounding box, and placed in a consistent canonical orientation. These steps reduce differences in scale and coordinates that would otherwise affect rendering and geometric supervision.

The pipeline uses BlenderProc and Mitsuba 3 for material and physical-property configuration, illumination setup, automated multi-view rendering, and post-processing. Six fixed cameras observe each asset from the front, front-left, front-right, left, right, and back. RGB, AoLP, normals, and masks share the same camera parameters within a view; outputs are indexed by object identity and view. Batch rendering supports parallel processing, providing repeatable cross-modal correspondence for surface encoding and multi-view mesh supervision.

For training annotations, Qwen-2.5-VL receives a 2×3 grid of the six RGB views and describes category, shape, color, material, part relationships, and view-specific features. Assets scoring below 6 on the VLM aesthetic assessment are discarded, followed by manual removal of descriptions with clear errors or semantic ambiguity. These multi-view descriptions are training annotations. At test time, the Spark VLM instead derives text from the single input RGB image, as specified in Section 5; unseen views and target geometry are not available to the generator. The 100 test objects are disjoint from the adaptation training objects.

5 Experiments

5.1 Experimental setup

We conduct all quantitative and qualitative evaluations on an independent set of 100 Objaverse objects (Deitke et al., 2023), disjoint at the object level from the samples used for polarization adaptation. For each object, we render a single RGB image and a pixel-aligned AoLP image using the training render protocol. At test time, the Spark vision-language model $\mathcal{V}$ generates text from the

input RGB image alone:

$$c_{\mathrm{rgb}} = \mathcal{V}(I_{\mathrm{rgb}}), \qquad \mathbf{x}_{\mathrm{test}} = (I_{\mathrm{rgb}}, I_{\mathrm{pol}}, c_{\mathrm{rgb}}). \tag{7}$$

RGB and AoLP come from the same view. At inference, the model uses only the aligned RGB–AoLP pair and text derived from the input RGB image.

5.2 Implementation details

We initialize the 3D generator from pretrained Step1X-3D (Li et al., 2025b) and adapt it with LoRA. Its original weights $\mathbf{W}_0$ are frozen; only LoRA parameters and the added surface-conditioning modules are optimized. This limits changes to Step1X-3D's RGB generative prior when learning from the smaller paired RGB–polarization–3D dataset. We use AdamW with a base learning rate of 1×10^{-4} on eight NVIDIA RTX 4090 GPUs. Full generation takes approximately 20 seconds per object.

As defined in Section 3, the 3D generation stage minimizes

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}} + \lambda_{\mathrm{rend}}\mathcal{L}_{\mathrm{rend}}, \qquad \mathcal{L}_{\mathrm{lat}} = \frac{1}{N_z}\|\hat{\mathbf{z}} - \mathbf{z}_{\mathrm{gt}}\|_2^2. \tag{8}$$

The latent term aligns the predicted clean latent with the ShapeVAE encoding of the target mesh. The rendering term compares multi-view normals rendered from the predicted and target meshes with identical cameras, directly supervising the final surface orientation.

5.3 Evaluation metrics

Chamfer Distance. For point sets P and Q sampled from the predicted and target mesh surfaces, respectively, we measure their overall spatial discrepancy by

$$\mathrm{CD}(P, Q) = \frac{1}{|P|}\sum_{\mathbf{p}\in P}\min_{\mathbf{q}\in Q}\|\mathbf{p}-\mathbf{q}\|_2^2 + \frac{1}{|Q|}\sum_{\mathbf{q}\in Q}\min_{\mathbf{p}\in P}\|\mathbf{q}-\mathbf{p}\|_2^2. \tag{9}$$

Lower CD indicates closer overall agreement with the target geometry.

CLIP-SIM. CLIP-SIM (Radford et al., 2021) measures visual semantic consistency with the input. We render each generated object i from V standard views. Let $\hat{I}_i^{(v)}$ denote view v and let E_{CLIP} be the CLIP image encoder. We compute

$$s_i = \frac{1}{V}\sum_{v=1}^{V}\cos\Big(E_{\mathrm{CLIP}}(I_i^{\mathrm{rgb}}), E_{\mathrm{CLIP}}(\hat{I}_i^{(v)})\Big), \tag{10}$$

$$\text{CLIP-SIM} = \frac{1}{100}\sum_{i=1}^{100} s_i. \tag{11}$$

5.4 Comparison with single-view 3D generators

We evaluate Era3D, Hash3D, TripoSR, Step1X-3D, TRELLIS, and our model on the same 100-object test set. As shown in Table 1, our model achieves a Chamfer Distance of 0.0089 and a CLIP-SIM of 0.9277. Step1X-3D obtains the lowest CD of 0.0064, whereas our model yields the highest CLIP-SIM among the evaluated methods. Table 2 further evaluates the contribution of the proposed surface-conditioning components within our framework.

Our CD (0.0089) is lower than Era3D, Hash3D, TripoSR, and TRELLIS, and remains in the same order as Step1X-3D (0.0064). LoRA adapts the base RGB prior with substantially fewer paired examples from a different distribution and an added modality, which may perturb its original generation distribution. TOur model achieves a CD of 0.0089 and a CLIP-SIM of 0.9277.

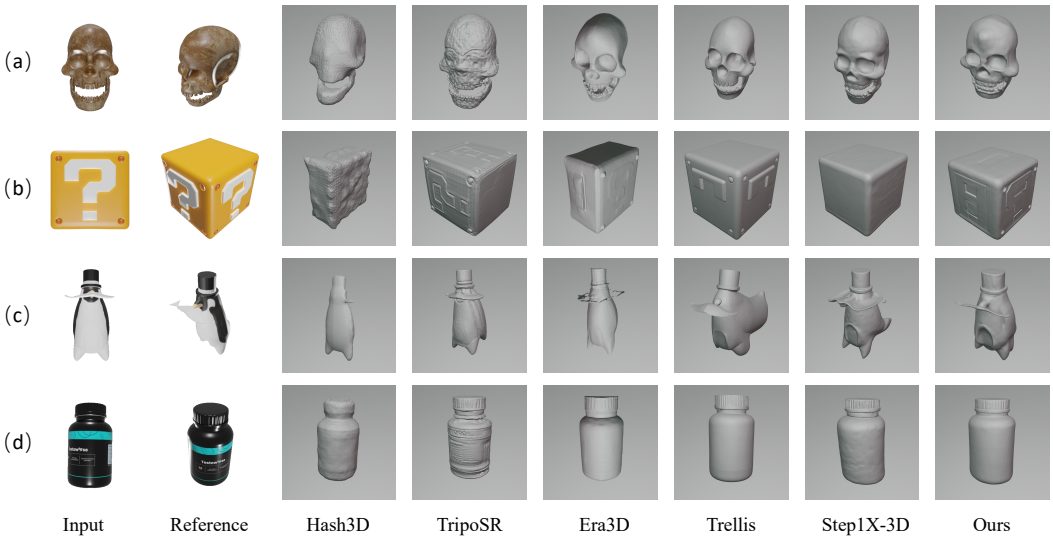


Figure 5: Qualitative comparison on synthetic objects. Columns show the input, source-asset reference, RGB-only baselines, and our RGB–AoLP model. Examples cover reflective surfaces, indentations, and thin structures.

5.5 Comparison with baseline and qualitative analysis

Our model retains the semantic and global-shape priors of Step1X-3D while adapting them to RGB–polarization inputs using limited paired data.Step1X-3D obtains a lower CD, while Figures 1 and 5 provide complementary qualitative comparisons of local structures such as indentations, sharp edges, and thin components. We therefore further examine the qualitative results in Figures 1 and 5, focusing on reflective surfaces, weakly textured indentations, sharp edges, and thin structures. These regions are particularly challenging for RGB-only reconstruction: view-dependent highlights can obscure the underlying geometry, weak texture provides limited shape evidence, and strong generative priors may smooth narrow grooves or slender components. The RGB-based methods generally recover the object category and overall silhouette, whereas the adapted model preserves more local orientation changes, sharper boundaries, clearer indentations, and more complete thin structures in these examples. These observations indicate that AoLP provides useful surface-orientation cues in regions where RGB evidence is ambiguous.

5.6 Ablation study

We keep the Step1X-3D backbone, LoRA configuration, training data, and hyperparameters fixed and remove one surface-aware component at a time. Table 2 reports both metrics.

Table 1: Comparison on 100 independent Objaverse test objects. Our model uses RGB and AoLP; the listed baselines use RGB only.

Method	CD ↓	CLIP-SIM ↑
Era3D	0.0281	0.6987
Hash3D	0.1750	0.8540
TripoSR	0.1693	0.9121
Step1X-3D	**0.0064**	0.8667
TRELLIS	0.0168	0.8926
Ours	0.0089	**0.9277**

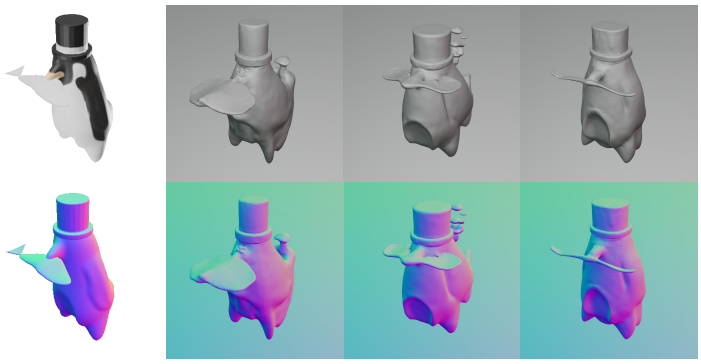


Figure 6: Qualitative ablation. Removing $\mathbf{C}_{\text{surf}}$ changes overall structure; removing $\mathcal{L}_{\text{rend}}$ loses some thin details. These examples complement the component ablation in Table 2.

Removing the surface condition leaves only the original RGB and text conditions. CD rises from 0.0089 to 0.0175; the full model reduces CD relative to this ablation by 49.1%. The RGB–polarization surface representation thus provides useful additional geometric conditioning under this protocol.

With the surface condition retained but $\mathcal{L}_{\text{rend}}$ removed, CD is 0.0121. Adding the multi-view normal loss reduces it to 0.0089, a relative decrease of 26.4%. Input conditioning alone does not guarantee that all surface evidence reaches the final geometry; comparing the predicted and target mesh normals gives direct output-side supervision. The two components play complementary roles: $\mathbf{C}_{\text{surf}}$ provides local surface evidence to the generator, while $\mathcal{L}_{\text{rend}}$ regularizes the generated surface. The full model attains the lowest CD in this ablation.

Table 2: Component ablation on the same 100 test objects.

Variant	CD ↓	CLIP-SIM ↑
w/o $\mathbf{C}_{\text{surf}}$	0.0175	0.9072
w/o $\mathcal{L}_{\text{rend}}$	0.0121	0.8896
Ours (full)	**0.0089**	**0.9277**

6 Conclusion

We presented a polarization-guided surface-aware framework for single-view 3D generation that augments a pretrained RGB-based generator with complementary surface-orientation cues from AoLP. The proposed method introduces a normal-aware RGB–AoLP representation for 3D latent generation and multi-view rendered-normal supervision to constrain the decoded surface. We further constructed OBJMV3D, a controlled multimodal benchmark containing aligned RGB, AoLP, surface-normal, camera, text, and 3D asset annotations for training and evaluation.

Experiments on 100 held-out Objaverse objects show that our model achieves a Chamfer Distance of 0.0089 and a CLIP-SIM of 0.9277. Qualitative comparisons further show improved preservation of local structures such as indentations, sharp boundaries, orientation changes, and thin components, while the ablation study demonstrates the contributions of both surface conditioning and rendered-normal supervision. These results support polarization as a complementary geometric cue for generative 3D modeling. Future work will extend the framework to more diverse conditions and real polarimetric captures.

AI USE STATEMENT

Generative AI was used to polish the title, abstract, introduction, method, dataset description, and experiments from the authors' drafts, and to check LaTeX formatting.

ETHICS STATEMENT

This work uses synthetic renderings of publicly available 3D assets and does not involve human participants or personal data. The principal practical consideration is the specialized sensing hardware required by the proposed input modality.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and objectives; Section 4 describes dataset construction; Section 5 reports the implementation, test protocol, and metrics. OBJMV3D includes camera parameters, masks, and aligned modalities to support subsequent reproduction studies.

REFERENCES

Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *European Conference on Computer Vision*, pp. 554–571. Springer, 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.

Yufei Han, Bowen Tie, Heng Guo, Youwei Lyu, Si Li, Boxin Shi, Yunpeng Jia, and Zhanyu Ma. PolGS: Polarimetric gaussian splatting for fast reflective surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28073–28082, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. *arXiv preprint arXiv:2311.04400*, 2023.

Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 12632–12641, 2022.

Chenhao Li, Taishi Ono, Takeshi Uemori, Hajime Mihara, Alexander Gatto, Hajime Nagahara, and Yusuke Moriuchi. NeISF: Neural incident stokes field for geometry and material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21434–21445, 2024a.

Chenhao Li, Taishi Ono, Takeshi Uemori, Sho Nitta, Hajime Mihara, Alexander Gatto, Hajime Nagahara, and Yusuke Moriuchi. NeISF++: Neural incident stokes field for polarized inverse rendering of conductors and dielectrics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26493–26503, 2025a.

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, 2024b.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1X-3D: Towards high-fidelity and controllable generation of textured 3D assets. *arXiv preprint arXiv:2505.07747*, 2025b.

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, 2023.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 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.

Manasi Muglikar, Leonard Bauersfeld, Diederik Paul Moeys, and Davide Scaramuzza. Event-based shape from polarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1547–1556, 2023.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Edward J Smith and David Meger. Improved adversarial systems for 3D object generation and reconstruction. In *Conference on Robot Learning*, pp. 87–96. PMLR, 2017.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 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*, pp. 22819–22829, 2023.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3D object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. CRM: Single image to 3D textured mesh with convolutional reconstruction model. In *European conference on computer vision*, pp. 57–74. Springer, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025.

Xingyi Yang, Songhua Liu, and Xinchao Wang. Hash3D: Training-free acceleration for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21481–21491, 2025.

Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8798–8812, 2022.

Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3D representations. *Advances in neural information processing systems*, 31, 2018.

A Extended Limitations

Synthetic-to-real gap. All reported experiments use rendered polarization with exact cross-modal alignment. Real sensors introduce calibration error, demosaicing artifacts, mixed reflection, and illumination variation. We do not claim direct generalization to such data.

Different sensing regime. AoLP requires a polarization-capable camera or a polarizer capture sequence. The proposed system therefore targets a different input setting from RGB-only generation; comparisons to RGB systems are not modality-matched.

Incomplete causal attribution. The current ablations establish that surface conditioning and rendered-normal supervision help the full pipeline, but they do not compare against normals predicted from RGB by an external estimator. Such a control is needed to separate the value of polarization-specific measurement from the general value of normal conditioning.

Evaluation breadth. The qualitative poses are not uniformly aligned across all external methods, and evaluation covers 100 synthetic objects using CD and CLIP-SIM. These results do not establish broad perceptual or semantic superiority across real-world categories and capture conditions.

Material assumptions. Transparent objects, strong subsurface transport, and multiple refraction can violate the surface-reflection assumptions underlying AoLP cues. Full Stokes measurements and material-aware polarization models are promising directions, but are outside the present evidence.

B Further Related Work

B.1 Single-view 3D generation

Single-view 3D generation has progressed from category-specific adversarial models (Smith & Meger, 2017; Zhu et al., 2018) to optimization-based systems that distill 2D diffusion priors (Qian et al., 2023; Tang et al., 2023; Wang et al., 2023) and feed-forward models that predict multi-view images or native 3D representations (Liu et al., 2023; Hong et al., 2023; Long et al., 2024; Wang et al., 2024). Recent systems such as TripoSR, Era3D, TRELLIS, Hash3D, and Step1X-3D improve speed, resolution, or representation capacity (Tochilkin et al., 2024; Li et al., 2024b; Xiang et al., 2025; Yang et al., 2025; Li et al., 2025b). These methods learn strong shape priors, but their geometric evidence is normally derived from RGB. Our study considers an augmented sensing regime in which the observed view also contains polarization. Comparisons to RGB-only systems are therefore contextual rather than input-matched evidence for a polarization-specific advantage.

B.2 Polarization and geometric inference

Shape from Polarization relates measured polarization to surface orientation through Fresnel reflection. Analytical and learned approaches use these cues together with priors or multi-view constraints to handle ambiguous normals and complex materials (Ba et al., 2020; Lei et al., 2022; Muglikar et al., 2023; Zhao et al., 2022). Polarization has also been incorporated into per-scene neural inverse rendering and Gaussian-splatting pipelines, including NeISF and PolGS (Li et al., 2024a; 2025a; Han et al., 2025). Those systems optimize a representation for a particular captured scene. In contrast, we study a feed-forward conditional generator trained across objects. Our use of a normal bottleneck is deliberately modest: AoLP contributes orientation evidence, but RGB context remains necessary because polarization alone is ambiguous.

B.3 Latent 3D generation and geometric supervision

Diffusion Transformers provide a backbone for conditional 3D denoising (Peebles & Xie, 2023; Vaswani et al., 2017). Learned shape latents and mesh decoders permit generation in a compact space before recovering an explicit surface (Park et al., 2019). Our focus is the surface representation used to condition this backbone and the geometric feedback applied after decoding.

C Additional qualitative results

Figure 7 provides additional paired-modality and normal-prediction examples, while Figure 8 gives further generation cases.

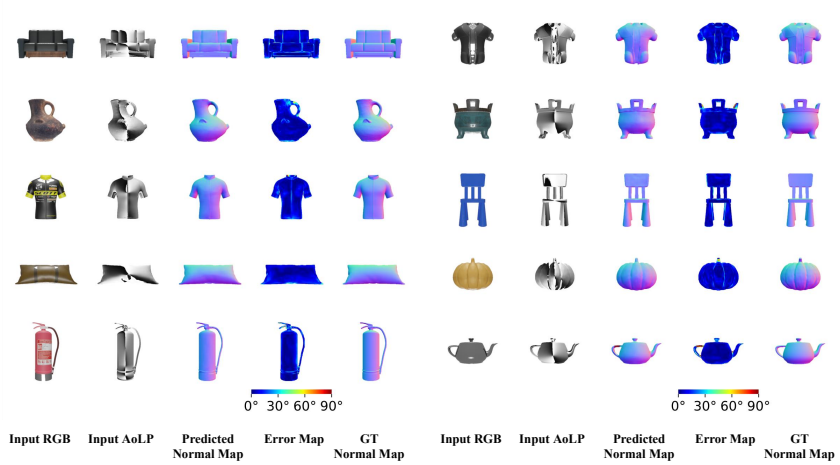


Figure 7: Additional synthetic examples of aligned RGB, AoLP, predicted normals, angular-error visualization, and mesh-rendered target normals.

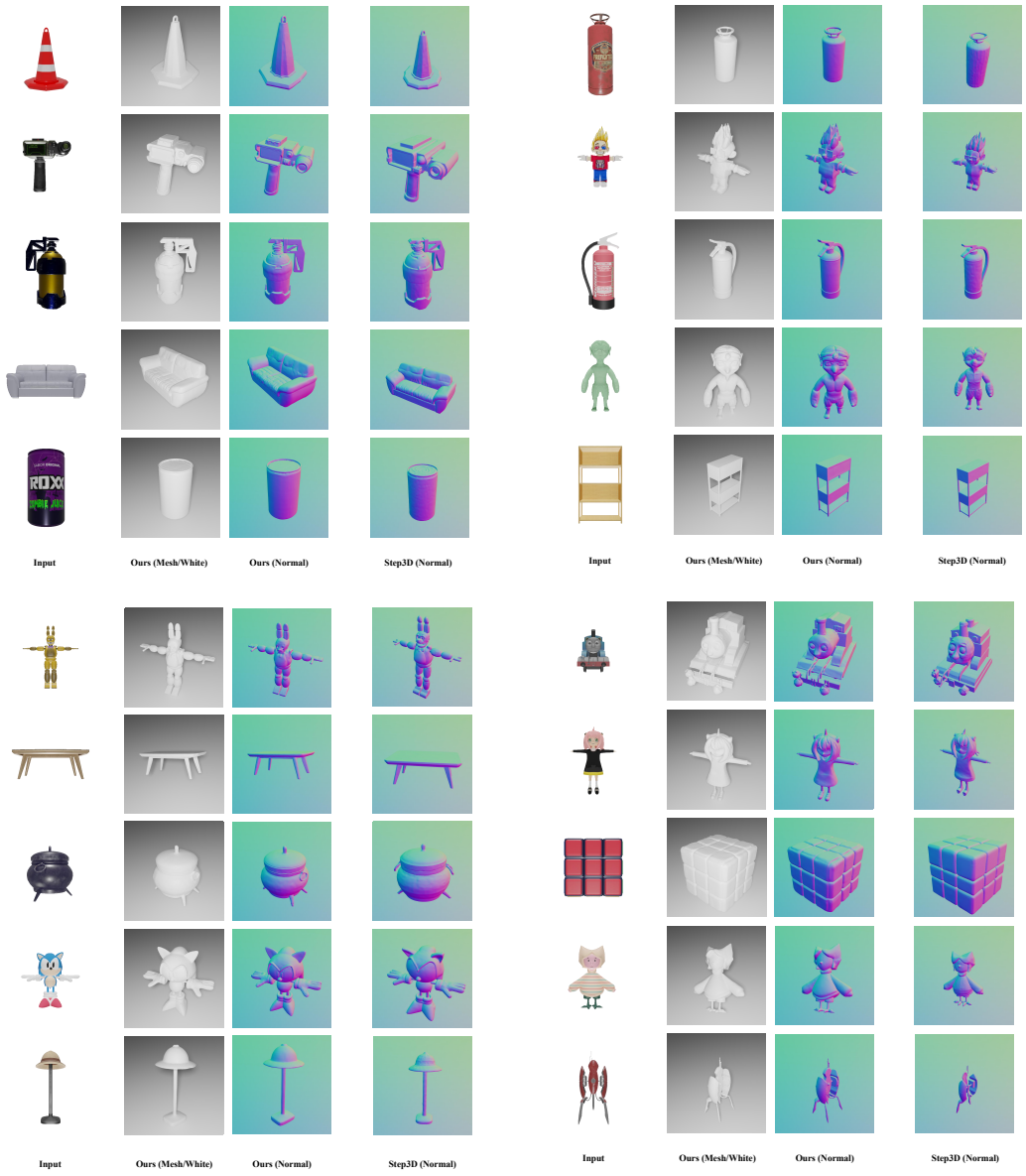


Figure 8: Additional qualitative generation cases from the original supplementary material. Each panel shows the input and corresponding mesh/normal outputs.