HYBRID3D-RL: HYBRID-SPACE REINFORCEMENT LEARNING FOR HIGH-FIDELITY 3D GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: We introduce **Hybrid3D-RL**, a hybrid-space reinforcement-learning framework for higher-quality 3D generation.

ABSTRACT

Reinforcement-learning post-training has become a standard approach for improving image and video generation, yet its extension to high-fidelity 3D generation remains challenging. Reward signals computed only from rendered images cannot fully characterize geometric quality, while direct supervision with explicit 3D metrics is ill-posed for single-image generation because unobserved regions have no uniquely determined geometry. We introduce **Hybrid3D-RL**, a hybrid-space post-training framework for VoxSet-based 3D generation. Its first component, the **Implicit Latent Reward Model (I-LRM)**, reuses the geometric prior of a pre-trained generator to assess mesh quality directly from latent representations. We train this model without human annotation using preference pairs constructed through multi-granularity VAE re-encoding. Its second component, **HRPO** (**H**ybrid-space **R**anking **P**reference **O**ptimization), separates online rollout from reward evaluation in a modular training pipeline. HRPO combines an explicit appearance reward measured from a camera-corrected rendering with the implicit latent reward, and optimizes stochastic flow trajectories using relative advantages and reference-policy regularization. Experiments demonstrate reliable latent quality discrimination and consistent improvements in image-to-3D generation across quantitative and qualitative evaluations.

1 INTRODUCTION

Recent advances in 3D content generation, driven by diffusion models and large-scale datasets (Deitke et al., 2023b;a), have enabled state-of-the-art methods (Zhang et al., 2023; 2024; Hunyuan3D et al., 2025; Xiang et al., 2025; Wu et al., 2025) to synthesize high-fidelity 3D assets from text or image inputs. Despite this progress, aligning generated 3D content with human preferences and quality standards remains an open problem. In 3D, such preferences often concern subtle

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.

geometry, topology, and image-mesh alignment failures that standard metrics miss. In 1D and 2D domains, reinforcement learning from human feedback (RLHF) has become a standard post-training technique (Ouyang et al., 2022; Grattafiori et al., 2024; Xu et al., 2023; Wu et al., 2023), where a reward model is trained on preference data and the policy is then optimized with a variety of algorithms, such as PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023), and GRPO (Shao et al., 2024). These techniques have substantially improved the output quality of language and image generation models, raising a natural question: *can we bring similar capabilities to 3D generation?*

Recent reinforcement-learning approaches have demonstrated the value of post-training for improving mesh quality (Zhao et al., 2025; Liu et al., 2025b; Zhou et al., 2026). Some methods construct preference pairs in advance using human judgments or geometric quality scores (Zhao et al., 2025; Liu et al., 2025b), while others employ rule-based rewards for surface integrity, topological regularity, and consistency with 3D observations (Zhou et al., 2026). These designs have advanced the generation of explicit, artist-style meshes. Extending this progress to high-resolution latent 3D generation introduces a complementary objective: recovering fine geometric details while remaining faithful to a single input image. In this setting, the visible surface provides direct evidence, whereas unobserved regions admit multiple plausible completions. Structural quality rewards and image-conditioned consistency therefore capture different aspects of the desired output. This motivates a post-training framework that combines feedback from the observed view with a learned prior over complete shape quality, and constructs geometric preferences without manual quality annotation.

We introduce **Hybrid3D-RL**, a post-training framework that combines rewards from explicit image space and implicit latent space. We first develop an **Implicit Latent Reward Model (I-LRM)** that attaches a lightweight reward head to intermediate features of a pre-trained VoxSet generator. These features encode geometric priors learned from large-scale 3D data and allow the model to score fine-grained mesh quality without decoding each candidate. We construct its preference data automatically through multi-granularity VAE re-encoding, where compression differences yield controlled quality variations ranging from coarse to subtle.

We then propose **HRPO**, short for **Hybrid-space Ranking Preference Optimization**. HRPO organizes online post-training into grouped trajectory sampling, hybrid reward evaluation, and flow-policy optimization. Its explicit appearance reward decodes a terminal latent, renders the mesh from the corrected conditioning camera, and evaluates the similarity between the rendered mesh and the input image. This view-aligned design supervises only observable evidence rather than imposing a full-shape 3D target. The frozen I-LRM scores the terminal latent directly, supplying a learned prior over geometric fidelity, including regions not explained by the visible view. HRPO compares candidates using the fused reward and updates the flow policy with group-relative advantages and a reference-policy constraint, following the stochastic flow optimization formulation of Flow-GRPO (Liu et al., 2025c).

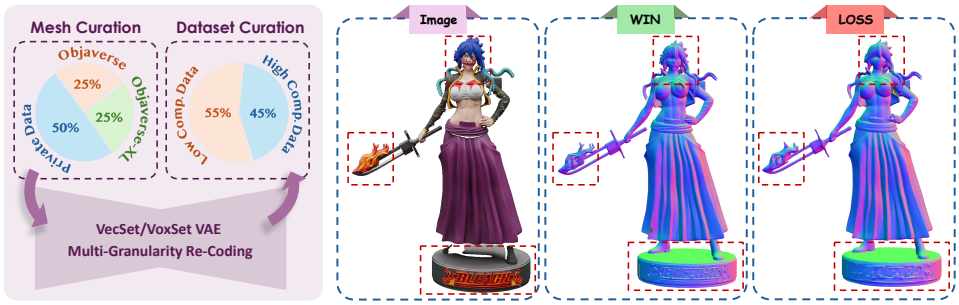
Experiments evaluate latent reward discrimination across noise levels and downstream image-to-3D generation using Uni3D, OpenShape, and ULIP-2. Hybrid3D-RL improves all three metrics over the pre-trained generator and produces cleaner, more detailed geometry in qualitative comparisons. Our contributions are:

- We introduce **Hybrid3D-RL**, a hybrid-space reinforcement-learning framework for post-training high-fidelity VoxSet-based 3D generators.
- We propose **I-LRM** that uses pre-trained geometric features and automatically curated multi-granularity preferences for decoding-free quality assessment.
- We develop **HRPO**, a hybrid-space ranking preference optimization method that combines camera-aligned appearance supervision with an implicit latent quality prior.
- Experiments on image-to-3D generation demonstrate improved Uni3D, OpenShape, and ULIP-2 scores over the pre-trained model and competitive methods.

2 Related Work

2.1 Native 3D Generation

Recent 3D generation has converged on two native latent-space paradigms. VecSet-based methods (Zhang et al., 2023; 2024; Zhao et al., 2023; Hunyuan3D et al., 2025; Lai et al., 2025; Li et al.,



(a) Data Curation Pipeline **(b) Representative Preference Pair**

Figure 2: Mesh quality comparison data. (a) Automated construction pipeline. (b) Representative preference pairs.

2025) compress shapes into fixed-length coordinate-free latent vectors and apply diffusion in this space, though scaling the VecSet size for higher fidelity incurs steep computational costs. Voxel-based methods (Xiang et al., 2025; Wu et al., 2025; He et al., 2025; Xiang et al., 2026; Ren et al., 2024) assign tokens to spatially localized regions, enabling finer local geometry at the expense of cubic memory growth with resolution; coarse-to-fine strategies (Chen et al., 2025; Lai et al., 2026; Jia et al., 2025) mitigate this through multi-stage generation. Despite rapid progress in native 3D generation quality, effective post-training alignment with human preferences remains an open challenge.

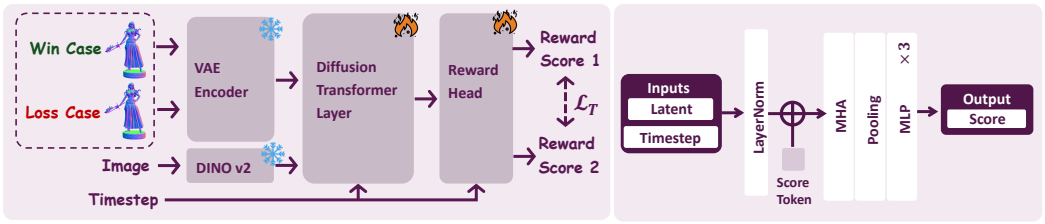
2.2 Reinforcement Learning in 3D Generation

Existing 3D post-training methods commonly rely on rendered observations or explicit mesh measurements. DreamReward (Ye et al., 2024; Liu et al., 2025a) scores rendered views, and DreamCS (Zou et al., 2025) introduces structural guidance for score-distillation-based generation. DeepMesh (Zhao et al., 2025) and Mesh-RFT (Liu et al., 2025b) apply preference optimization to explicit mesh sequences. Mesh-Pro (Zhou et al., 2026) uses rule-based geometric rewards for artist-style meshes. For single-image generation, however, full-shape explicit metrics can penalize plausible geometry in unobserved regions. Our method instead combines supervision from a corrected observed view with a learned latent prior over complete shape quality.

2.3 Generative Model Post-Training

Preference-based post-training has expanded from language models (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024) to diffusion and flow generators. Learned visual rewards (Xu et al., 2023; Wu et al., 2023) support both direct preference learning and online policy optimization. Diffusion-DPO (Wallace et al., 2024) adapts direct preference optimization to diffusion models through a denoising-based objective on preferred and dispreferred samples. Flow-DPO (Liu et al., 2025d) extends preference optimization to rectified-flow video generators using human feedback. These approaches provide a way to learn from pre-collected pairwise preferences without online reward evaluation during each policy update.

Online methods instead evaluate samples from the evolving policy. Flow-GRPO (Liu et al., 2025c) introduces stochastic sampling through an ODE-to-SDE conversion and applies group-relative policy optimization to flow matching models. Dense-GRPO (Deng et al., 2026) further estimates step-wise reward gains from intermediate clean predictions and adjusts exploration according to the reward signal, providing more localized feedback along the denoising trajectory. Building on flow-based online optimization, Hybrid3D-RL focuses on reward design for single-image 3D generation, combining similarity to the observed image with an implicit latent prior over geometric quality.



(a) I-LRM **(b) Reward Head**

Figure 3: **I-LRM framework.** **(a)** Preference pairs are encoded into VoxSet latents and scored using a diffusion transformer and a reward head. **(b)** Reward-head design with a learnable score token, LayerNorm, multi-head attention, pooling, and an MLP for scalar quality prediction.

3 Methodology

3.1 Overview

Hybrid3D-RL consists of an implicit latent-space quality model and a hybrid-space reinforcement-learning procedure. We first describe the automated preference data used to train the I-LRM in Sec. 3.2. Section 3.3 introduces its architecture and training objective. Section 3.4 then presents HRPO, which combines grouped online rollout with explicit appearance and implicit latent rewards.

3.2 Data Curation

Our source meshes comprise 25% Objaverse (Deitke et al., 2023b), 25% Objaverse-XL (Deitke et al., 2023a), and 50% licensed private data. As illustrated in Figure 2, we construct preference pairs automatically by encoding and decoding each mesh with a pre-trained 3D VAE at multiple latent resolutions. Resolution-dependent compression provides a controlled quality hierarchy: reconstructions with more latent tokens generally preserve finer geometry and are designated as winners over lower-resolution reconstructions.

We partition the Mesh Quality Comparison Data according to the latent-token gap within each pair. Pairs with a large gap form the High-Comparison Data and account for 45% of the dataset, while pairs with a smaller gap form the Low-Comparison Data and constitute the remaining 55%. The former captures pronounced quality differences, whereas the latter emphasizes subtle geometric distinctions. Representative pairs are shown in Figure 2(b).

3.3 I-LRM: Implicit Latent Reward Model

The I-LRM evaluates intrinsic mesh quality from the native representation of a pre-trained VoxSet generator. It reuses the generator's geometric prior and predicts a scalar reward from intermediate features, avoiding mesh decoding and rendering during latent reward evaluation.

Model Architecture. As illustrated in Figure 3(a), a frozen VAE encoder maps a mesh M to its native latent $\mathbf{z}_0 \in \mathbb{R}^{N\times d_z}$, while a frozen DINOv2 encoder (Oquab et al., 2023) extracts image tokens $\mathbf{c} \in \mathbb{R}^{L\times d_c}$ from the conditioning image I. To support quality assessment along the flow trajectory, we sample $\tau \sim \mathcal{U}(0,1)$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}. \tag{1}$$

The noisy latent, timestep, voxel coordinates, and image tokens are processed by the first eight blocks of the pre-trained DiT:

$$\mathbf{H} = \mathrm{DiT}_{1:8}(\mathbf{z}_\tau, \tau, \mathbf{c}) \in \mathbb{R}^{(N+1)\times d}. \tag{2}$$

As detailed in Figure 3(b), the reward head normalizes $\mathbf{H}$, aggregates it with learnable score queries, and maps the pooled representation to a scalar:

$$r_\phi = g_\phi(\mathrm{Pool}(\mathbf{Q}, \mathrm{LN}(\mathbf{H}))) \,. \tag{3}$$

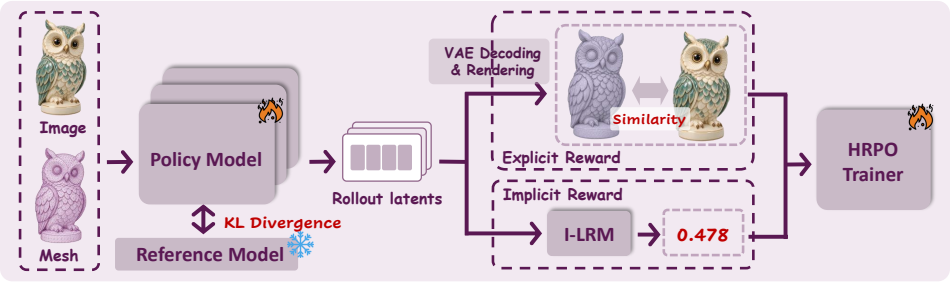


Figure 4: HRPO framework. Generated latents receive explicit rendered-view similarity and implicit latent quality rewards, which guide policy updates with frozen-reference regularization.

Training Objective. For a comparison $(I, M_a, M_b, y) \sim \mathcal{D}_q$, where $y = 1$ indicates that M_a is preferred, the model predicts r_a and r_b and is trained with a Bradley–Terry loss:

$$\mathcal{L}_{\mathrm{BT}} = -\mathbb{E}_{\mathcal{D}_q}[y \log \sigma(r_a - r_b) + (1 - y) \log \sigma(r_b - r_a)]\,. \tag{4}$$

We regularize the reward scale with $\mathcal{L}_{\mathrm{reg}} = \mathbb{E}_{\mathcal{D}_q}[r_a^2 + r_b^2]$. The complete objective is

$$\mathcal{L}_T = \lambda_{\mathrm{BT}} \mathcal{L}_{\mathrm{BT}} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}}. \tag{5}$$

3.4 HRPO: Hybrid-space Ranking Preference Optimization

HRPO post-trains the VoxSet flow policy using candidate rankings induced jointly by observable appearance and implicit 3D quality. Given an image condition, the policy generates a group of stochastic flow trajectories. Each terminal latent is evaluated by two complementary reward branches, and the fused reward is converted into a relative advantage for policy optimization. Figure 4 summarizes the framework; Appendix A details the sampling and optimization procedure.

3.4.1 Hybrid Reward Design

Explicit Appearance Reward. Direct supervision with a full-shape 3D metric is not well suited to single-image generation: only the visible surface is supported by the observation, while occluded geometry may admit multiple plausible completions. Penalizing all unseen regions against one target mesh imposes arbitrary supervision. We instead decode the terminal latent, transform the mesh into the conditioning-camera coordinate system, and render it from the corrected input view. Let $\widehat{I}_i$ be this rendering for candidate i. The explicit reward combines identity-sensitive and semantic visual similarity:

$$r_{\mathrm{app}}^{(i)} = 0.6\, s_{\mathrm{ID}}(I, \widehat{I}_i) + 0.4\, s_{\mathrm{DINO}}(I, \widehat{I}_i). \tag{6}$$

We compute s_{ID} with ID-Sim (Chae et al., 2026) using its default DINOv3 ViT-L/16 model and released pre-trained weights, without additional fine-tuning. The semantic score s_{DINO} uses a separate frozen DINOv3 (Siméoni et al., 2025) ViT-L/16 encoder with its released pre-trained weights. This camera-aligned 2D signal evaluates whether the generated visible geometry explains the evidence in the conditioning image without prescribing an unobserved back surface. Camera correction reduces viewpoint mismatch as a source of score variation, making the comparison more directly sensitive to the candidate's visible structure.

Implicit Latent Reward. The I-LRM reliably distinguishes reconstructions across VAE granularities and therefore supplies a learned prior over complete mesh quality. It scores the clean terminal latent directly, together with its voxel coordinates $\mathbf{v}_i$ and image context $\mathbf{c}$:

$$r_{\mathrm{lat}}^{(i)} = \mathcal{R}_\phi\left(\mathbf{z}_{\mathrm{term}}^{(i)}, \mathbf{v}_i, \mathbf{c}\right), \tag{7}$$

where $\mathcal{R}_\phi$ is frozen during HRPO. This branch requires neither VAE re-encoding nor rendering and complements the view-specific appearance signal with implicit geometric knowledge. The latent score supplies a learned quality preference for the complete shape while leaving its unobserved geometry unconstrained by any particular ground-truth completion.

Table 1: Pairwise ranking accuracy (%) of the I-LRM across timestep intervals on held-out preference pairs.

Timestep Interval (τ)					
$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0]$	**Overall**
96.80	96.60	92.50	81.30	73.40	88.12

Table 2: Quantitative comparison on the shared 90-image test suite.

Method	**Uni3D ↑**	**OpenShape ↑**	**ULIP-2 ↑**
Hunyuan3D-2.1 (Hunyuan3D et al., 2025)	0.3658	0.2283	0.2287
Direct3D-S2 (Wu et al., 2025)	0.3652	0.2274	0.2273
UltraShape-1.0 (Jia et al., 2025)	0.3678	0.2303	0.2295
TRELLIS.2 (Xiang et al., 2026)	0.3661	0.2301	0.2874
Pretrained Model	0.3657	0.2275	0.2865
Ours (Hybrid3D-RL)	**0.3730**	**0.2348**	**0.2920**

Hybrid Reward and Flow Optimization. For valid decoded and rendered candidates, the terminal reward is

$$R_i = \lambda_{\text{app}} r_{\text{app}}^{(i)} + \lambda_{\text{lat}} r_{\text{lat}}^{(i)}, \qquad \lambda_{\text{app}} = \lambda_{\text{lat}} = 0.5. \tag{8}$$

Failed decoding, rendering, or coverage checks trigger resampling and discard the trajectory. Only after collecting G valid candidates do we compute rewards and normalized, clipped group-centered advantages A_i. Update direction depends on advantage, not the raw I-LRM score's sign.

For a selected stochastic flow transition at step k, let $\rho_{i,k}(\theta) = \exp(\log p_\theta(\mathbf{z}_{k+1}^{(i)}|\mathbf{z}_k^{(i)}, I) - \log p_{\text{old}}(\mathbf{z}_{k+1}^{(i)}|\mathbf{z}_k^{(i)}, I))$. Following Liu et al. (2025c), HRPO minimizes

$$\mathcal{L}_{\text{HRPO}} = -\mathbb{E}_{i,k}[\min(\rho_{i,k}A_i, \text{clip}(\rho_{i,k}, 1-\epsilon, 1+\epsilon)A_i)] + \beta_{\text{ref}}\, \mathbb{E}_{i,k}[D_{\text{KL}}(p_\theta \| p_{\text{ref}})]\,, \tag{9}$$

where clipping limits policy changes and the frozen reference preserves the pre-trained generative prior.

4 Experiments

4.1 Experimental Setup

Implementation Details. For Stage 1, we directly adopt the default settings of Hunyuan3D-2.1 (Hunyuan3D et al., 2025), with a VecSet latent sequence of 4096 tokens. For Stage 2, following the architectural design of UltraShape (Jia et al., 2025), we build a 4B-parameter DiT and a 3B-parameter VAE. For HRPO post-training, we optimize the Stage 2 DiT using AdamW with a learning rate of 10^{-6}, while keeping the VAE, reward models, and reference policy frozen. Each conditioning image is used to collect a group of four valid candidates using 20 stochastic flow steps per rollout; eight transitions per accepted trajectory are randomly selected for optimization. We set the policy clip range to 0.001 and the reference KL coefficient to 0.04.

Training Data. For pretraining, our model is trained on 520K meshes from Objaverse-XL (Deitke et al., 2023a) and private licensed meshes. From this corpus, we randomly sample 10K meshes to construct the I-LRM training data using the multi-granularity VAE re-encoding pipeline in Sec. 3.2. We independently sample 4K image-conditioned examples from the same corpus for HRPO training and use them for online candidate generation and policy optimization.

4.2 I-LRM Evaluation

We evaluate the I-LRM on 1,000 held-out preference pairs from the Mesh Quality Comparison Data. For each pair, we sample a timestep τ and perturb the clean latents with Gaussian noise following Eq. (1). Pairwise ranking accuracy is the proportion of comparisons for which the I-LRM assigns

Table 3: Ablation of the ℓ_2 reward regularizer λ_{reg} on the I-LRM. Pairwise ranking accuracy (%) is reported across timestep intervals and overall.

	Timestep Interval (τ)					
Setting	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0]$	**Overall**
w/o λ_{reg}	90.80	89.50	84.90	78.30	62.40	81.18
w/ λ_{reg}	96.80	96.60	92.50	81.30	73.40	88.12

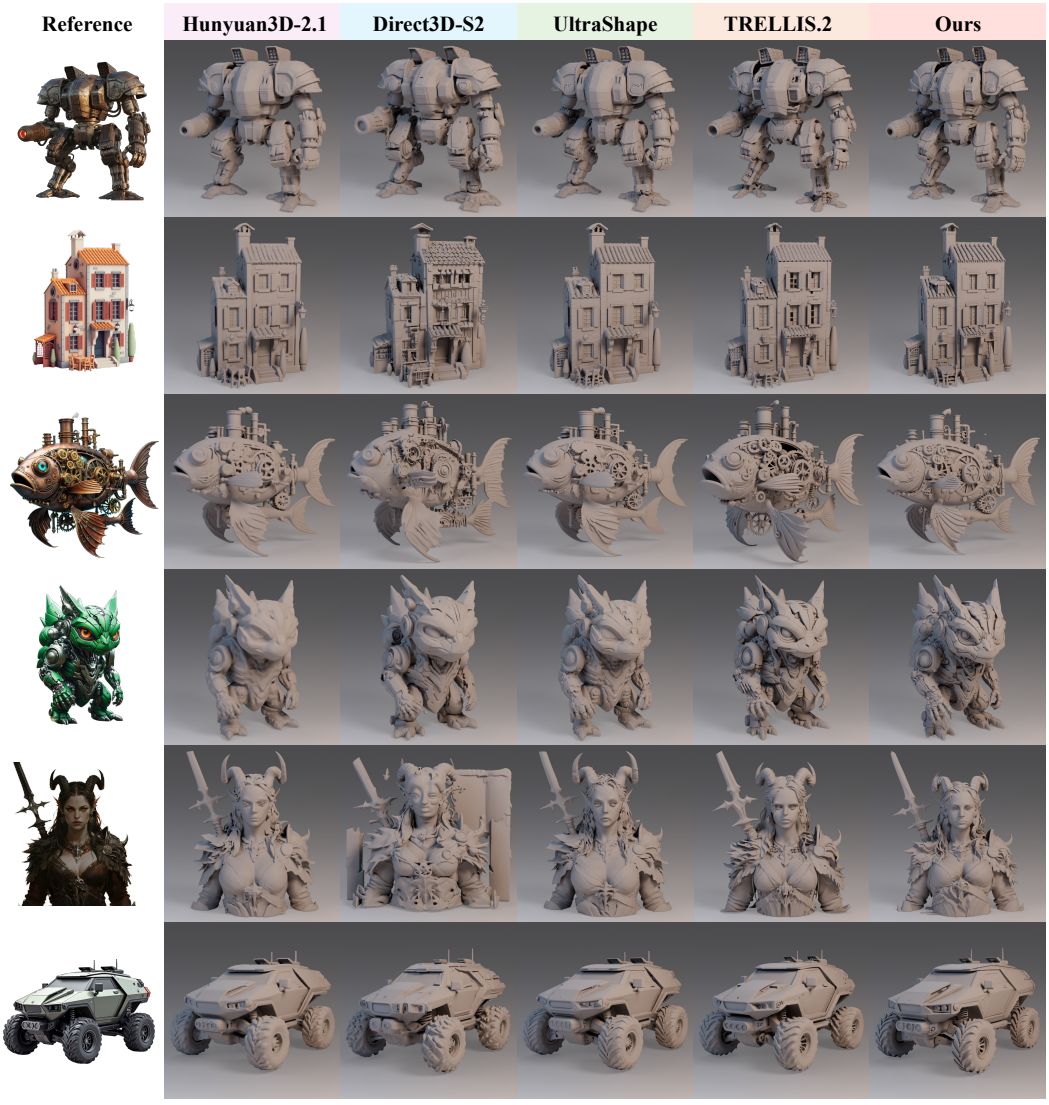


Figure 5: Qualitative comparison with recent image-to-3D methods. Hybrid3D-RL produces cleaner surfaces, more complete structures, and finer local geometry.

a higher score to the higher-quality reconstruction. To characterize reward reliability along the generation trajectory, we divide $\tau \in [0, 1]$ into five equal-width intervals and report both per-interval and overall accuracy in Table 1.

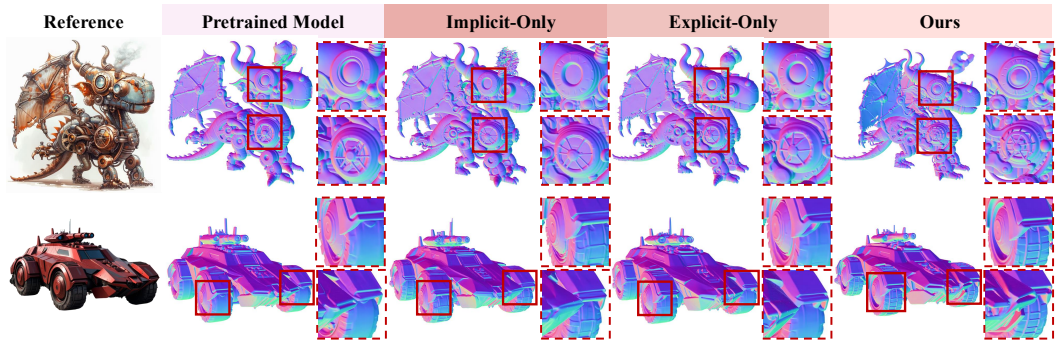


Figure 6: Qualitative ablation of reward components: the pre-trained model, implicit-only and explicit-only rewards, and our full method.

Table 4: Ablation of reward components.

Setting	Uni3D ↑	OpenShape ↑	ULIP-2 ↑
Pretrained Model	0.3657	0.2275	0.2865
Explicit Reward Only	0.3683	0.2269	0.2865
Implicit Reward Only	0.3674	0.2281	0.2872
Ours	**0.3730**	**0.2348**	**0.2920**

Table 5: Comparison of RL algorithms.

Algorithm	Uni3D ↑	OpenShape ↑	ULIP-2 ↑
Pre-trained Model	0.3657	0.2275	0.2865
Diffusion-DPO (Wallace et al., 2024)	0.3669	0.2286	0.2872
HRPO (Ours)	**0.3730**	**0.2348**	**0.2920**

4.3 Quantitative Evaluation of HRPO

We evaluate Hybrid3D-RL against Hunyuan3D-2.1 (Hunyuan3D et al., 2025), Direct3D-S2 (Wu et al., 2025), UltraShape (Jia et al., 2025), and TRELLIS.2 (Xiang et al., 2026) on the same 90 test images. Following Direct3D-S2, we use Uni3D (Zhou et al., 2024), OpenShape (Liu et al., 2023), and ULIP-2 (Xue et al., 2024) as image-referenced shape-alignment metrics. Each metric encodes the conditioning image and a point cloud sampled from the generated mesh into a shared feature space, then measures their cosine similarity. Scores are averaged over the test set. These metrics assess consistency with the input image without requiring a ground-truth mesh. As shown in Table 2, Hybrid3D-RL improves all three scores over the Pretrained Model and achieves the highest scores among the evaluated methods.

4.4 Subjective Evaluation

We conduct two user studies on image-conditioned 3D generation, assessing *Geometric Quality* and *Image–Geometry Consistency*. The first study collects 1–5 ratings for five methods. The second compares the pre-trained model with ours and records a preference or tie for each criterion. These results complement the qualitative comparisons below.

Figure 5 compares Hybrid3D-RL with recent image-to-3D methods. Hybrid3D-RL produces more complete structures, cleaner surfaces, and finer local geometry, particularly in thin parts and high-frequency surface regions. These visual differences complement the feature-based metrics by revealing geometric quality that is difficult to summarize with a single score.

Detailed results of our user studies are provided in Appendix B. Appendix C presents more detailed comparisons with the pre-trained model and additional qualitative results.

4.5 Ablation Studies

The Effect of Reward Regularization. We conduct this ablation on the VoxSet-based I-LRM. We ablate the ℓ_2 reward regularizer λ_{reg} in the I-LRM objective (Eq. (5)). We train the I-LRM with and without this term under the same data and optimization setting, then evaluate both variants using the timestep-bucketed ranking protocol described in Sec. 4.2. Table 3 reports the resulting accuracy across the generation trajectory. The regularized model consistently achieves higher ranking accu-

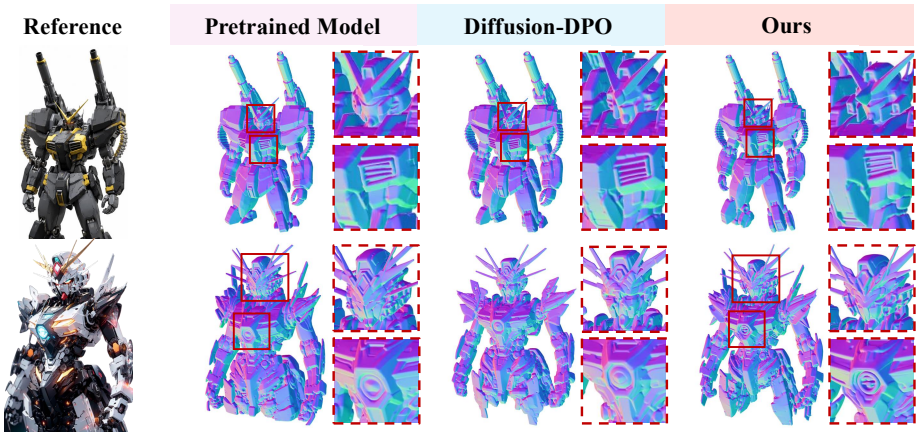


Figure 7: Qualitative comparison of post-training strategies: the pre-trained model, Diffusion-DPO, and HRPO (Ours).

racy, demonstrating that λ_{reg} stabilizes reward learning and preserves reliable discrimination under noisy latent inputs.

The Effect of Reward Components. We compare the pre-trained model with explicit-only, implicit-only, and combined rewards, holding the generator, data, and optimization settings fixed. Table 4 reports Uni3D, OpenShape, and ULIP-2 on the 90-image test set. Combining both rewards achieves the best scores across all three metrics, while single-reward variants exhibit metric-dependent trade-offs. In Figure 6, the combined reward better preserves local geometric details and their correspondence with the reference image, illustrating the complementary roles of latent quality assessment and appearance supervision.

Comparison of RL Algorithms. We compare Diffusion-DPO (Wallace et al., 2024) and our Flow-GRPO-based HRPO using the three metrics (Table 5). Diffusion-DPO uses the offline preference pairs from Sec. 3.2, without I-LRM or other reward functions. Figure 7 shows that HRPO recovers finer geometric detail and better correspondence with the input image than Diffusion-DPO and the pre-trained model. These visual improvements complement the quantitative gains and support combining online sampling with hybrid reward supervision. We also tested Dense-GRPO (Deng et al., 2026), whose step-wise scoring requires intermediate predictions. At early denoising steps, these predictions do not reliably decode into meshes suitable for explicit appearance scoring. Dense-GRPO is therefore unsuitable for our current decode-and-render reward pipeline and is excluded from the quantitative comparison.

5 Conclusion

We introduced Hybrid3D-RL, a post-training framework for high-fidelity image-conditioned 3D generation. Its I-LRM evaluates geometric quality at the latent level using pre-trained geometric features. Learned from preference pairs constructed through multi-granularity VAE re-encoding, it distinguishes coarse and subtle quality differences without mesh decoding or rendering. HRPO combines this implicit geometric reward with explicit appearance supervision in a hybrid-space formulation. The explicit reward evaluates camera-aligned renderings against the conditioning image, while I-LRM supplies a complementary geometric prior. Together, they promote fine details and image–geometry consistency without prescribing a unique completion of unseen regions. Experiments show that Hybrid3D-RL improves upon the pre-trained generator and outperforms the evaluated methods, producing more detailed meshes with better input-image alignment. These findings demonstrate that combining latent-level quality assessment with image-based supervision offers a practical path toward improving 3D generation beyond pre-training.

AI USE STATEMENT

Generative AI tools assisted with manuscript revision and language polishing, LaTeX formatting, literature lookup and reference checking, and figure preparation. For the user-study visualizations, AI-assisted scripts aggregated questionnaire ratings and preferences and plotted the resulting summaries. The authors are responsible for verifying AI-assisted text, references, numerical summaries, and figures against their underlying sources and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies image-conditioned 3D generation and includes human evaluations of generated outputs. We report aggregate ratings and preferences without disclosing participant identifiers or contact information. The use and redistribution of training assets, conditioning images, and pre-trained models must respect their respective licenses and permissions. Generated assets may reflect biases in the training data or be used to create misleading content, and improved geometric fidelity does not establish factual accuracy or suitability for safety-critical applications. Application-specific review and safeguards remain necessary.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the preference-data construction procedure, and Sections 3.3 and 3.4 specify the reward model and post-training objectives. Section 4.1 reports the model configurations, training data, and optimization settings, while Appendix A provides the HRPO algorithm, reward computation, and advantage normalization. The evaluation procedures and ablations are described in Section 4, with user-study results in Appendix B. Exact reproduction of training is limited by access to the licensed private portion of the data.

REFERENCES

Julia Chae, Nicholas Kolkin, Jui-Hsien Wang, Richard Zhang, Sara Beery, and Cusuh Ham. ID-Sim: An identity-focused similarity metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11250–11262, June 2026.

Yiwen Chen, Zhihao Li, Yikai Wang, Hu Zhang, Qin Li, Chi Zhang, and Guosheng Lin. Ultra3d: Efficient and high-fidelity 3d generation with part attention. *arXiv preprint arXiv:2507.17745*, 2025.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:35799–35813, 2023a.

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 (CVPR)*, pp. 13142–13153, 2023b.

Haoyou Deng, Keyu Yan, Chaojie Mao, Xiang Wang, Yu Liu, Changxin Gao, and Nong Sang. DenseGRPO: From sparse to dense reward for flow matching model alignment. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14822–14833, 2025.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

Tanghui Jia, Dongyu Yan, Dehao Hao, Yang Li, Kaiyi Zhang, Xianyi He, Lanjiong Li, Yuhan Wang, Jinnan Chen, Lutao Jiang, et al. Ultrashape 1.0: High-fidelity 3d shape generation via scalable geometric refinement. *arXiv preprint arXiv:2512.21185*, 2025.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Qingxiang Lin, Jingwei Huang, Chunchao Guo, and Xiangyu Yue. Lattice: Democratize high-fidelity 3d generation at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19982–19992, 2026.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

Fangfu Liu, Junliang Ye, Yikai Wang, Hanyang Wang, Zhengyi Wang, Jun Zhu, and Yueqi Duan. Dreamreward-x: Boosting high-quality 3d generation with human preference alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025a.

Jian Liu, Jing Xu, Song Guo, Jing Li, Jingfeng Guo, Jiaao Yu, Haohan Weng, Biwen Lei, Xianghui Yang, Zhuo Chen, Fangqi Zhu, et al. Mesh-rft: Enhancing mesh generation via fine-grained reinforcement fine-tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 38: 156033–156058, 2025b.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:40783–40818, 2025c.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:82155–82192, 2025d.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:44860–44879, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:53728–53741, 2023.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4219, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, 2024.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, Yao Yao, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 38: 170778–170804, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

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 (CVPR)*, pp. 21469–21480, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14419–14429, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:15903–15935, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27091–27101, 2024.

Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 259–276. Springer, 2024.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

Ruowen Zhao, Junliang Ye, Zhengyi Wang, Guangce Liu, Yiwen Chen, Yikai Wang, and Jun Zhu. Deepmesh: Auto-regressive artist-mesh creation with reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10612–10623, 2025.

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 73969–73982, 2023.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *The Twelfth International Conference on Learning Representations (ICLR)*, volume 2024, pp. 46766–46782, 2024.

Zhen Zhou, Jian Liu, Biwen Lei, Jing Xu, Haohan Weng, Yiling Zhu, Zhuo Chen, Junfeng Fan, Yunkai Ma, Dazhao Du, et al. Mesh-pro: Asynchronous advantage-guided ranking preference optimization for artist-style quadrilateral mesh generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34248–34258, 2026.

Xiandong Zou, Ruihao Xia, Hongsong Wang, and Pan Zhou. Dreamcs: Geometry-aware text-to-3d generation with unpaired 3d reward supervision. *arXiv preprint arXiv:2506.09814*, 2025.

APPENDIX A: COMPLETE HRPO PIPELINE

This appendix details the complete Hybrid-space Ranking Preference Optimization pipeline summarized in Sec. 3.4. Figure 4 presents the separation between rollout generation, hybrid reward evaluation, and policy training.

A.1 EXPLICIT APPEARANCE REWARD

For a terminal candidate $\mathbf{z}_{\text{term}}^{(i)}$, we first undo the generator's latent scaling and decode it with the frozen VAE while preserving its native voxel support $\mathbf{v}_i$:

$$\widehat{M}_i = \mathcal{D}_{\text{VAE}}\Big(\text{Unscale}(\mathbf{z}_{\text{term}}^{(i)}), \mathbf{v}_i\Big). \tag{10}$$

The conditioning camera is transferred into the normalized mesh coordinate system using the saved intrinsic and extrinsic parameters and the model-to-world transformation. We render $\widehat{M}_i$ from this corrected pose to obtain $\widehat{I}_i$. This coordinate conversion uses the recorded camera rather than optimizing a new pose. Unlike an arbitrary canonical-view reward, the rendered evidence corresponds to the actual observed view of the input image.

For s_{ID}, we use the released ID-Sim (Chae et al., 2026) DINOv3 ViT-L/16 backbone and adapter weights without fine-tuning. The configured scorer clamps the similarity score between referenced image and rendered image to $[0, 1]$. The semantic branch uses a separate frozen DINOv3 (Siméoni et al., 2025) ViT-L/16 encoder with released pre-trained weights. It computes the cosine similarity of normalized features and maps the result to $[0, 1]$ as $(1 + \cos)/2$. The ID branch emphasizes instance-specific structure, while DINOv3 measures broader semantic and structural consistency:

$$r_{\text{app}}^{(i)} = 0.6\, s_{\text{ID}}(I, \widehat{I}_i) + 0.4\, s_{\text{DINO}}(I, \widehat{I}_i). \tag{11}$$

A candidate is accepted only if VAE decoding and camera-aligned rendering succeed and foreground coverage lies in $[0.01, 0.98]$. Failed candidates and their stored transitions are discarded and resampled under the same conditioning input until G valid candidates are collected. They receive no reward and do not enter the optimization buffer.

A.2 IMPLICIT LATENT REWARD

The frozen I-LRM directly evaluates the already-scaled terminal VoxSet latent:

$$r_{\text{lat}}^{(i)} = \mathcal{R}_\phi\Big(\mathbf{z}_{\text{term}}^{(i)}, \mathbf{v}_i, \mathbf{c}_i; t_{\text{clean}}\Big). \tag{12}$$

The implementation evaluates the clean endpoint at native flow time $t = 1$ (equivalently $\tau = 0$ in Eq. (1)), without additional noise, VAE re-encoding, latent rescaling, or rendering. Candidate order, voxel coordinates, and image contexts are preserved exactly between rollout and reward evaluation. The reward model is trained using the multi-granularity comparisons from Sec. 3.2; its score therefore encodes a learned preference for geometric fidelity that complements the visible-view evidence of Eq. (11).

A.3 HYBRID REWARD AND RELATIVE ADVANTAGE

For a valid candidate, the two branches are combined as

$$R_i = 0.5\, r_{\text{app}}^{(i)} + 0.5\, r_{\text{lat}}^{(i)}. \tag{13}$$

The latent score is used in its raw form, without a sigmoid or score calibration. Only the G accepted candidates form the rollout group $\mathcal{G}$. After this group is complete, we compute their fused rewards, center them against the group mean, and normalize them by an exponential moving estimate s_{EMA} of the reward scale:

$$A_i = \text{clip}\left(\frac{R_i - \frac{1}{|\mathcal{G}|}\sum_{j\in\mathcal{G}} R_j}{\max(s_{\text{EMA}}, \epsilon_A)}, -A_{\max}, A_{\max}\right). \tag{14}$$

Here, ϵ_A prevents division by a small reward scale and $A_{\max}$ bounds the advantage magnitude. Only transitions from accepted candidates are retained for optimization.

Algorithm 1 HRPO: Grouped Rollout, Hybrid Scoring, and Policy Update

Require: Policy π_θ, frozen reference π_{ref}, reward models, group size G
Ensure: Post-trained policy π_θ

while training **do**
  Sample image I, camera C, and voxel support $\mathbf{v}$; encode context $\mathbf{c}$
  $\pi_{\text{old}} \leftarrow \pi_\theta$; $\mathcal{G} \leftarrow \varnothing$; $\mathcal{B} \leftarrow \varnothing$
  for $i = 1, \ldots, G$ **do**
    repeat
      $\mathcal{B}_i \leftarrow \varnothing$
      Roll out π_{old} with fresh noise; store selected transitions in $\mathcal{B}_i$
      Decode and render the terminal latent; check decoding, rendering, and coverage
    until the candidate is valid
    Add the accepted candidate and its temporary buffer $\mathcal{B}_i$ to $\mathcal{G}$
  end for
  for each accepted candidate $i \in \mathcal{G}$ **do**
    Compute $r_{\text{app}}^{(i)}$, $r_{\text{lat}}^{(i)}$, and fused reward R_i
    $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{B}_i$
  end for
  Compute group-centered, normalized advantages from $\{R_i\}_{i \in \mathcal{G}}$
  Update π_θ on $\mathcal{B}$ with the clipped objective and reference regularization
end while

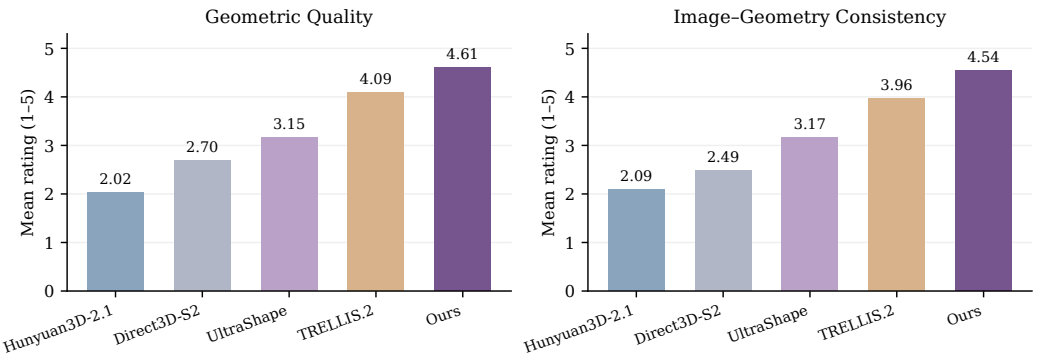


Figure 8: Mean user ratings for five methods on geometric quality and image–geometry consistency.

Signed Rewards. I-LRM may output negative scores, but policy updates depend on group-centered advantages rather than the sign of the raw reward. A negative score above the group mean can still yield a positive advantage.

A.4 Policy Update

Following Liu et al. (2025c), HRPO evaluates the likelihood ratios of the accepted transitions against their stored rollout-policy probabilities and optimizes Eq. (9). The clipped objective limits policy changes, while the frozen-reference KL term regularizes the update. Algorithm 1 summarizes the procedure.

Appendix B: User Study Results

We conduct a user study to evaluate *Geometric Quality* and *Image–Geometry Consistency*. Figure 8 compares our method with other methods, with participants rating each result from 1 to 5. Figure 9 compares our method with the pre-trained model through pairwise preferences. Both figures show that users generally favor our method over the competing methods and the pre-trained model on both criteria.

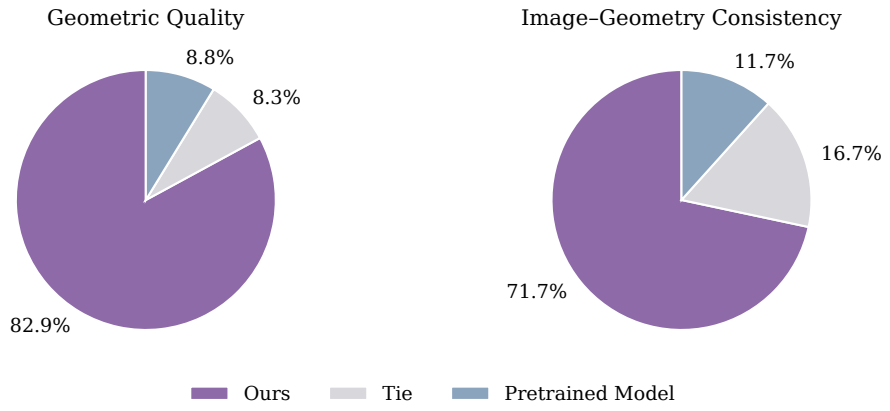


Figure 9: User preferences between our method and the pre-trained model, with ties included.

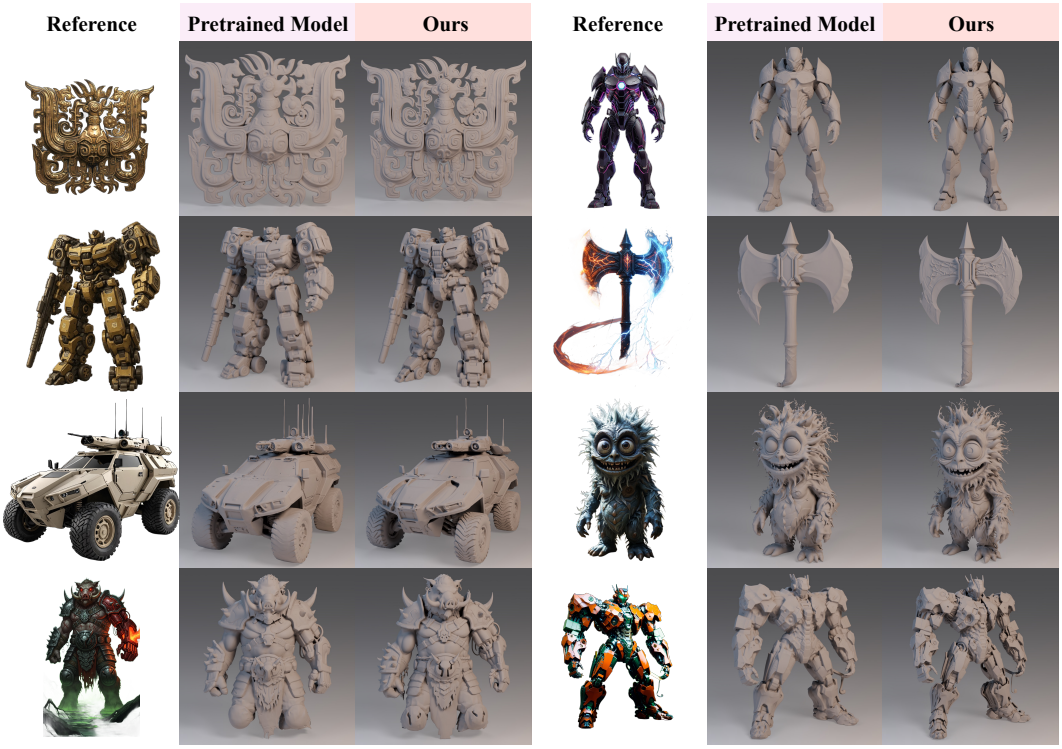


Figure 10: Additional comparisons with the pre-trained model. Each group shows the reference image, the pre-trained output, and our post-trained output.

APPENDIX C: ADDITIONAL QUALITATIVE RESULTS

Comparison with the Pre-trained Model. Figure 10 presents additional comparisons between the pre-trained generator and Hybrid3D-RL under the same image conditions. The side-by-side results highlight changes in local geometry and correspondence with the reference image after post-training, complementing the reward and RL-strategy ablations in the main paper.

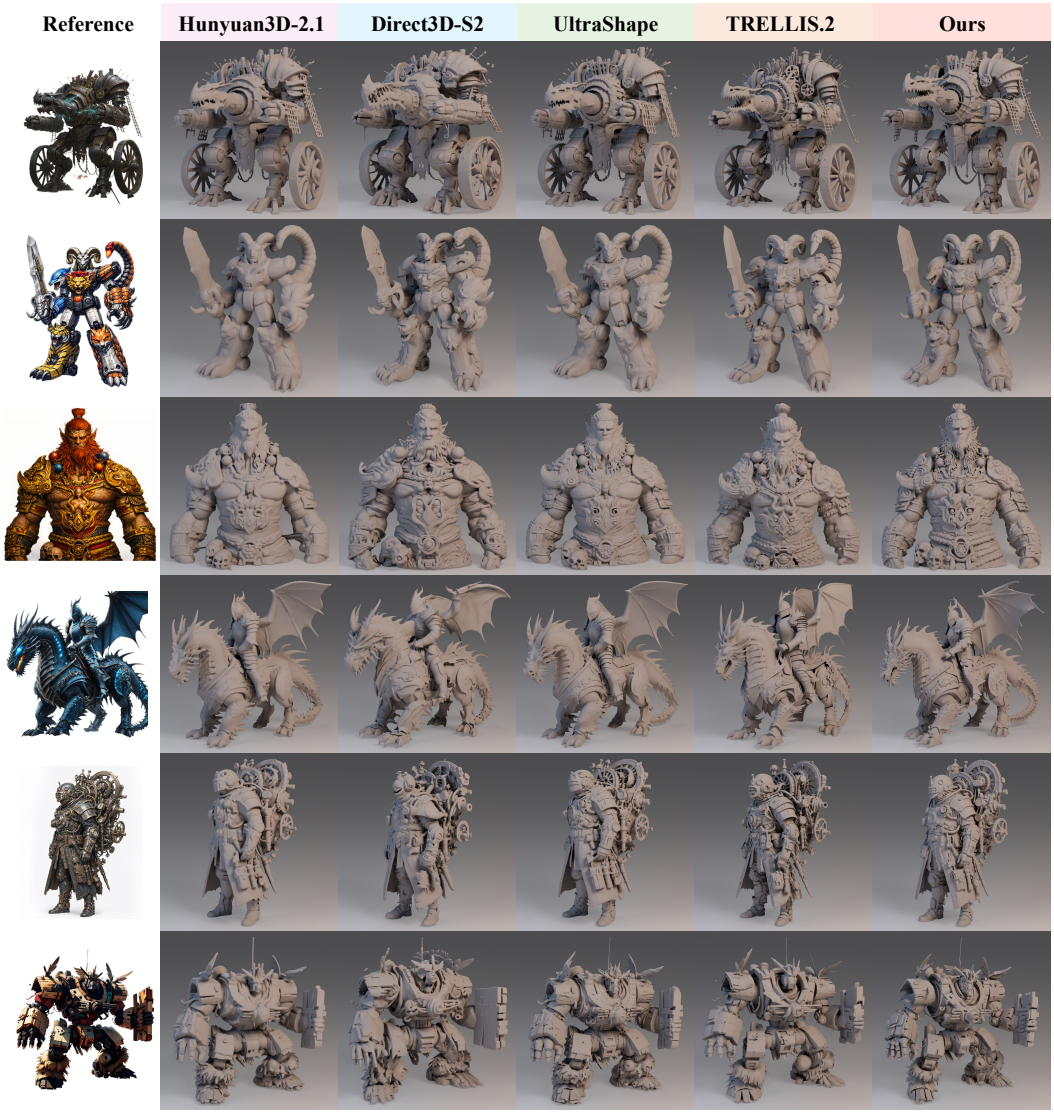


Figure 11: Additional qualitative comparison with recent image-to-3D methods. Hybrid3D-RL produces more complete structures, cleaner surfaces, and finer local geometry across diverse inputs.

Comparison with Other Methods. Figure 11 provides additional comparisons with recent image-to-3D methods. Across diverse object categories, Hybrid3D-RL recovers more complete structures and cleaner local geometry, particularly for thin components and detailed surface regions.

Additional Visual Results. Figure 12 presents additional image-conditioned meshes generated by Hybrid3D-RL, with each input image paired with its corresponding output. These examples cover characters, creatures, and mechanical objects, illustrating the generated geometry for thin structures, layered components, and detailed surface patterns. All generated meshes in this figure are outputs of our model, rather than comparisons with other methods.

APPENDIX D: LIMITATIONS AND FUTURE WORK

Our I-LRM learns a scalar preference from reconstructions produced at different VAE granularities. This supervision captures quality differences associated with reconstruction granularity, but does not

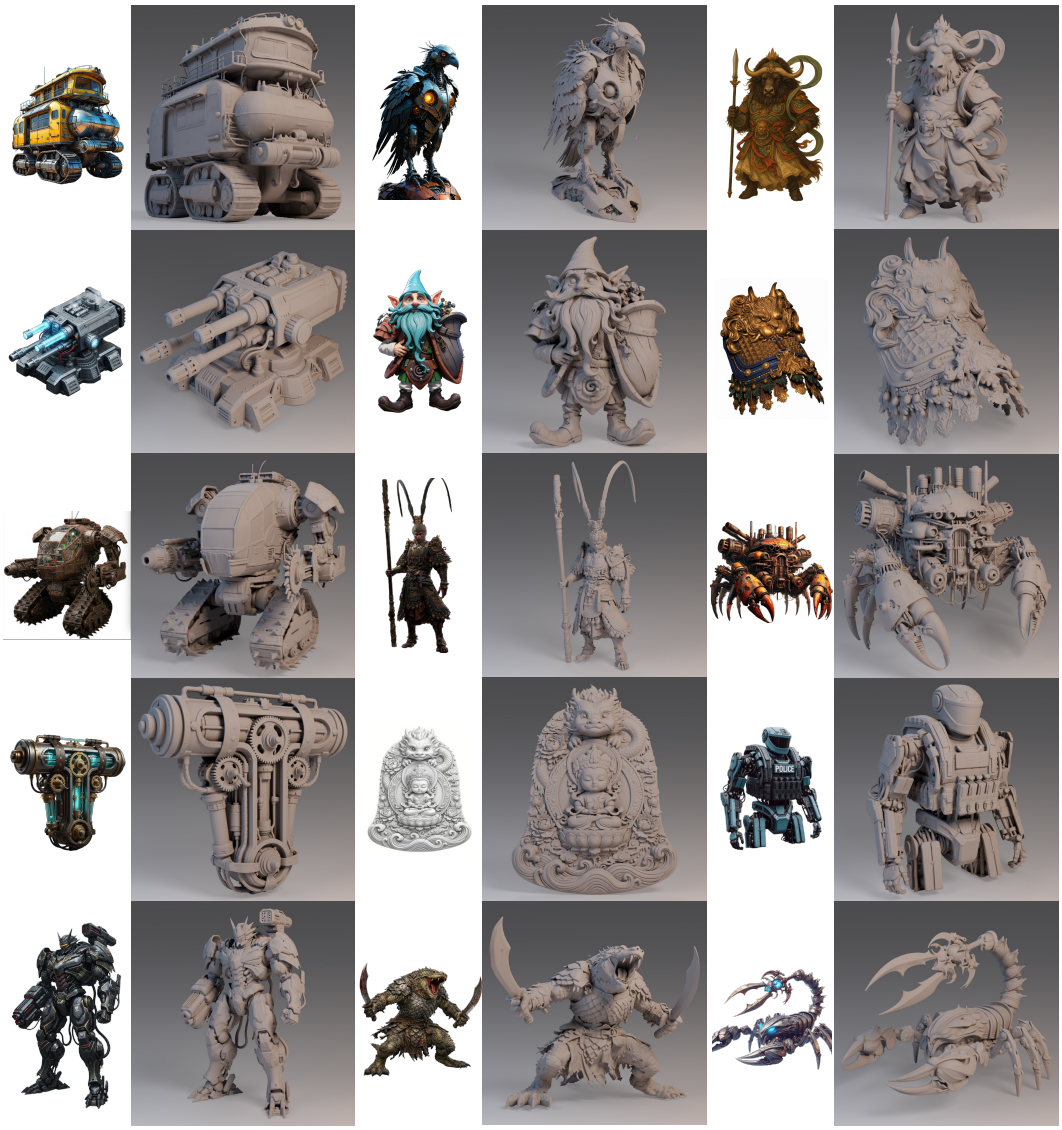


Figure 12: Additional visual results generated by Hybrid3D-RL. Each pair shows the input image (left) and our generated mesh (right).

disentangle geometric quality from image–geometry consistency. Changes in latent resolution can simultaneously affect the recovered geometry and its correspondence with the conditioning image. Consequently, the learned score does not provide independent control over these two objectives. Different applications may require different trade-offs. For example, asset creation may prioritize detailed, coherent geometry and plausible completion of unseen regions, whereas reference-driven reconstruction may place greater emphasis on fidelity to the observed image. A single entangled preference score cannot directly express these application-specific priorities. Exploring disentangled latent rewards and controllable post-training objectives that separately supervise geometric quality and image–geometry consistency is therefore an important direction for future work.