GRPQ: STRUCTURE-PRESERVING ALIGNMENT OF 3D DIFFUSION MODELS WITH RGB-ONLY REWARDS

Anonymous authors
Paper under double-blind review

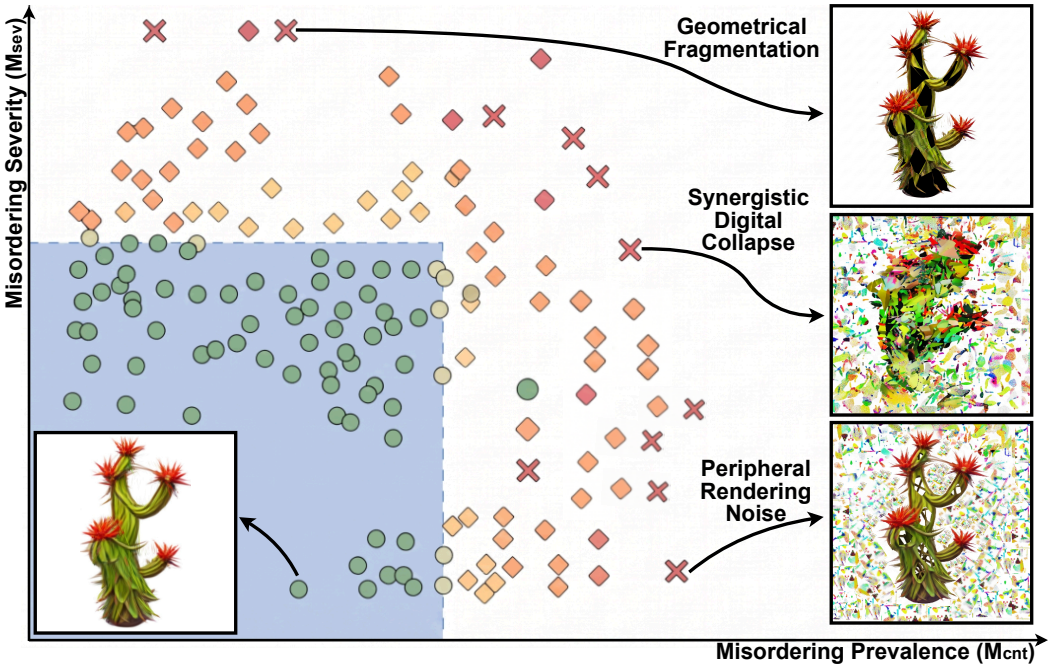

Figure 1: During fine-tuning of 3D generative models with 2D RGB rewards, information from views with **mismatched structural evidence richness and scores damages the 3D structural prior**, where widespread mismatches induce peripheral rendering noise, and severe mismatches cause geometrical fragmentation that degrades the generated structure towards 2D.

ABSTRACT

Aligning 3D diffusion models with specialized styles and human preferences is essential for producing usable assets, yet directly specifying such objectives in 3D remains difficult and costly. Generic 2D reward models provide abundant supervision, but optimizing their scores over rendered views can rapidly erode the structural prior of a pretrained 3D model. Existing safeguards often rely on structure-aware multi-view models or auxiliary depth and normal estimators, thereby limiting the range of applicable 2D rewards and potentially introducing estimator-specific biases. Our analysis reveals that pronounced disagreement between the cross-view rankings induced by 2D rewards and structural evidence often results in unreliable supervision and structural degradation; we refer to this phenomenon as *reward–structural-evidence mismatch*. Based on this finding, we introduce **GRPQ**, a prior-preserving alignment framework that uses the degree of mismatch as evidence for estimating the reliability of each view-level reward and assigns optimization weights accordingly. By suppressing harmful signals rather than introducing additional structural supervision, **GRPQ** enables generic RGB-only rewards to improve preference alignment while preserving the pretrained 3D structural prior. Extensive experiments across three alignment methods, multiple 3D diffusion models, and diverse reward models show that **GRPQ** improves

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.

structural integrity and prior retention while largely preserving reward gains and training efficiency. Process-level visualizations further show that **GRPQ** prevents optimization from favoring views that conceal geometric defects over those that reveal them, providing interpretable evidence for the proposed mechanism.

1 INTRODUCTION

Advances in mesh generation, implicit representations, structured 3D latents, and high-fidelity text-conditioned synthesis have enabled 3D diffusion models to generate plausible assets for games, film production, and robotics simulation Liu et al. (2024); Siddiqui et al. (2024); Wang et al. (2023); Xiang et al. (2025); Zhang et al. (2024b). However, these outputs are rarely ready for professional use Liu et al. (2026). Practical applications impose task-specific requirements on geometry, topology, materials, editability, and physical plausibility, together with stylistic preferences that are often difficult to specify through a single conditioning signal. Satisfying these requirements therefore relies heavily on costly expert post-processing. Preference-oriented post-training could internalize such corrections, but encounters a fundamental supervision bottleneck: professional requirements and subjective preferences are difficult to express as 3D objectives, while collecting large-scale 3D preference annotations is prohibitively expensive Wang et al. (2025); Liu et al. (2026).

2D models offer task-specific evaluators for aesthetics, semantics, and human preferences Schuhmann (2022); Wu et al. (2023); Ye et al. (2024a); Zhang et al. (2026a). Recent work therefore uses rendered-view rewards to align 3D diffusion models Ye et al. (2024b); Wang et al. (2025); Liu et al. (2026). Yet they score projections, not the underlying 3D object. Direct optimization can erode the pretrained 3D structural prior, causing structural collapse, multi-view inconsistency, and reward-hacking artifacts, as shown in Fig. 1 Hong et al. (2023); Liu et al. (2026); Leng et al. (2026). Existing methods use geometry-aware or multi-view 2D models Shi et al. (2024); Wang et al. (2025); Xie et al. (2024), or monocular normal/depth supervision Ye et al. (2024a); Zhang et al. (2026a). However, reliable 3D-aware 2D models remain scarce, while monocular geometry estimators may hallucinate or misestimate structure Liu et al. (2026). This constraint narrows the reward space, leaving many specialized RGB evaluators for diverse functional requirements and preferences difficult to use safely in 3D alignment. This raises a central question: *how can a 3D diffusion model be aligned with a generic 2D RGB reward without sacrificing its structural prior?*

To understand how projection-level rewards erode an object-level prior, we examine their cross-view effects. Structural degradation is view-dependent: misleading updates concentrate on particular viewpoints and progressively disrupt the cross-view balance of generation capacity, as shown in Fig. 1. Different views exhibit varying structural-evidence richness, i.e., the amount of evidence each projection reveals about the underlying 3D structure. A well-formed 3D prior accommodates this asymmetry while still preserving valid projections across all viewpoints. View-wise reward optimization may ignore these differences and follow local visual preferences, disturbing the pretrained prior's balance. We therefore first test whether structural-evidence richness directly predicts each view's generation capacity. Measurements on structurally intact assets show that this relationship is too weak for a reliable training constraint; results appear in the Appendix A.

Nevertheless, nearly all observed cases of structural collapses exhibit disagreement among sampled views between rankings by structural-evidence richness and reward scores, as shown in Fig. 1. We characterize its prevalence and severity by the number of misordered views and maximum rank displacement. A mismatched ranking can overreward views with limited structural evidence while penalizing those exposing stronger evidence. The resulting updates favor projections concealing geometric defects and suppress those revealing structural complexity, progressively disturbing the pretrained 3D prior's generation-capacity allocation.

This observation shows that view-level rewards are not uniformly reliable. We introduce **GRPQ** (**G**roup **R**ank-based 3D Structural **P**rior **Q**uality Preservation), a plug-and-play framework estimating view-wise reward confidence from structural-evidence rank mismatch and attenuating gradients from unreliable views. We evaluate **GRPQ** in two ways. Across every tested combination of three alignment methods, two generic 2D RGB rewards, and two base 3D diffusion models, **GRPQ** consistently improves structural-prior preservation with negligible loss in reward improvement or convergence efficiency. Paired denoising visualizations show that it suppresses misleading reward

influence and prevents error accumulation. Because confidence estimation does not require spatially aware reward models, **GRPQ** generalizes across alignment methods, rewards, and base models.

We make three contributions. **(1)** We identify reward–structure-evidence misordering as a failure pattern in 3D diffusion alignment and use it to suppress structurally unreliable updates while preserving reward gains and efficiency. **(2)** Experiments spanning three alignment methods, two rewards, and two base models demonstrate GRPQ's plug-and-play applicability across diverse models and rewards. **(3)** We provide a reference-free geometric self-consistency diagnostic that complements reward-based and image-space evaluations.

2 Related Work

2.1 Optimization Paradigms for 3D Alignment

Per-instance methods such as A3D Ignatyev et al. (2025) and DreamReward Ye et al. (2024b) optimize each 3D asset from rendered-view rewards or preferences, requiring instance-specific optimization. Post-training methods update generator parameters: DDPO Black et al. (2024) applies policy-gradient optimization, while ReFL Xu et al. (2023) and DRaFT Clark et al. (2024) backpropagate differentiable rewards through truncated or full computation graphs. Direct optimization against imperfect reward signals, however, can weaken pretrained priors. Distribution-level methods optimize reward-weighted generation distributions through stochastic-control or adjoint objectives Domingo i Enrich et al. (2025), soft-Q formulations Haarnoja et al. (2017), or GFlowNets Liu et al. (2025), whereas sampling-time guidance Liu et al. (2026) avoids retraining at the cost of repeated reward evaluation. MVReward Wang et al. (2025) and Carve3D Xie et al. (2024) further incorporate multi-view preference modeling and render–reconstruction consistency, respectively. DPO Rafailov et al. (2023) and Circular-DPO Li et al. (2026) avoid explicit scalar rewards but depend on high-quality preference pairs, while advantage estimation Zhang et al. (2026b) and trajectory trimming Yin & Liu (2026) improve credit assignment and rollout efficiency. Regardless of the optimization paradigm, reliable alignment ultimately depends on high-quality underlying reward or preference supervision.

2.2 2D Reward Modeling for 3D Alignment

Most 2D rewards operate directly on RGB renderings without explicitly encoding spatial geometry. Representative examples include Dream Fields Jain et al. (2022), which optimizes CLIP scores over rendered views, DreamFusion Poole et al. (2022), which introduces diffusion-based SDS, and preference-guided approaches Ye et al. (2024b) based on pretrained aesthetic Schuhmann (2022) or preference evaluators Wu et al. (2023). Other methods infer auxiliary spatial cues from the same RGB observations through normal estimation Ye et al. (2024a) or Depth-Normal Consistency (DNC) Zhang et al. (2026a). Such estimation adds computation and may propagate geometric errors into optimization, potentially reducing the effective signal-to-noise ratio for limited gain. Signal reliability therefore becomes critical for stable optimization, yet remains largely unexamined. DreamReward Ye et al. (2024b) and DreamDPO Zhou et al. (2025) construct human-preference supervision without estimating view-specific reward reliability; RewardSDS Chachy et al. (2026) weights samples by reward magnitude, which does not imply reliability; and GSD Mu et al. (2024) and MV-SDI Lupascu et al. (2026) improve gradient consistency or variance without validating the reward signal itself. Thus, existing methods improve supervision or optimization statistics without explicitly assessing per-view reward reliability. A further challenge is how to fuse multiple partial 2D observations into a faithful global 3D objective. DreamCS Zou et al. (2025) identifies inherent 2D reward bias and trains a geometry-aware 3D reward on more than 30,000 human-refined 3D preference samples, bypassing 2D reward fusion at the cost of dedicated 3D supervision. Methods retaining 2D rewards adopt diverse fusion strategies: Nabla-R2D3 Liu et al. (2026) averages rewards across camera poses; DreamDPO Zhou et al. (2025) and PSD Leng et al. (2026) optimize sampled views; MVReward Wang et al. (2025), HyperScore Zhang et al. (2025b), and 3DGen-Score Zhang et al. (2025a) learn joint multi-view assessments; Carve3D Xie et al. (2024) uses reconstruction-mediated consistency; and MVC-ZigAL Zhang et al. (2026b) balances single-view and joint-view rewards through constrained optimization. These varied formulations indicate that accurate global 3D reward approximation requires both reliable per-view signals and principled multi-view fusion.

2.3 Prior Degradation during 3D Alignment

Prior degradation causes cross-view inconsistency and geometric, semantic, or appearance failures. The Janus problem exemplifies this: individually plausible views correspond to a globally inconsistent object Hong et al. (2023). Existing methods mitigate such failures through score or prompt debiasing and multi-view diffusion priors Hong et al. (2023); Shi et al. (2024), pose-aware regularization Zheng et al. (2025), geometric regularization and priors Li et al. (2024); Wang et al. (2024); Qiu et al. (2024), and coarse-to-fine generation Li et al. (2025). Progressive and object-aware methods address missing objects and incorrect relations Cheng et al. (2024); Ge et al. (2025), while factorized representations, PBR priors, and synchronized texture generation reduce geometry–material entanglement and cross-view texture inconsistency Chen et al. (2023); Zhang et al. (2024c); Cheng et al. (2025). Other approaches address timestep or scale mismatch Lin et al. (2023); Wang et al. (2023) and mode collapse Ma et al. (2024); Chen et al. (2025). These methods compensate for degradation through additional priors or specialized designs. We instead target two upstream causes in reward-based alignment: unreliable 2D signal assessment and improper fusion into a 3D reward.

3 Preliminaries

3.1 Diffusion Models and RL-based Finetuning

Diffusion models generate data by reversing a Gaussian corruption process that progressively destroys the structure of clean samples. Given $x_0 \sim \mathcal{D}$, the forward process produces $x_t = \alpha_t x_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and α_t and σ_t denote timestep-dependent signal and noise scales. As t increases, the signal is gradually attenuated while Gaussian noise dominates, yielding a tractable terminal distribution. Generation starts from $x_T \sim \mathcal{N}(0, I)$ and applies the learned reverse transitions $p_\theta(x_{t-1} \mid x_t)$ to recover x_0 through iterative denoising. The denoising network predicts the injected noise using the weighted score-matching objective, thereby learning the direction required to progressively remove corruption at each timestep. Here, $w(t)$ is a timestep-dependent weighting function that balances contributions across noise levels, and $\epsilon_\theta(x_t, t)$ predicts the noise in x_t.

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{t\sim\text{Uniform}\{1,\dots,T\},\, \epsilon\sim\mathcal{N}(0,I),\, x_0\sim\mathcal{D}} \quad \left[w(t) \left\| \epsilon_\theta(\alpha_t x_0 + \sigma_t \epsilon, t) - \epsilon \right\|_2^2 \right]. \tag{1}$$

The stochastic reverse process defines an MDP with state $s_t = (x_t, t)$, action x_{t-1}, and policy $\pi_\theta(x_{t-1} \mid x_t) = p_\theta(x_{t-1} \mid x_t)$, thereby casting sequential denoising as a decision-making process. A denoising trajectory $\tau = (x_T, \dots, x_0)$ starts from Gaussian noise and receives terminal reward $R(x_0)$ over the full horizon. RL-based diffusion alignment maximizes $J(\theta) = \mathbb{E}\tau \sim \pi\theta[R(x_0)]$, adapting the model to downstream objectives not captured by the denoising loss Black et al. (2024).

3.2 Diffusion Alignment via Gradient-Informed RL

Reward finetuning adapts a pretrained diffusion model p_{base} to a desired objective while preserving its prior. The target is the reward-tilted distribution $p^\star(x_0) = Z^{-1} p_{\text{base}}(x_0) R(x_0)^\beta$, where Z is the normalizer and β balances reward optimization and prior preservation. This formulation increases high-reward sample probability while retaining the pretrained distribution as an explicit reference, thereby constraining alignment from drifting arbitrarily far from the original generative prior.

Policy-gradient methods treat diffusion trajectories as RL rollouts, optimizing terminal rewards without first-order information from differentiable reward models and assigning credit primarily from scalar feedback. In contrast, gradient-informed methods align transition scores with reward gradients, providing local directions to improve intermediate samples throughout denoising. Define the transition log-density ratio as $\rho_\theta(x_{t-1}, x_t) = \log p_\theta(x_{t-1} \mid x_t) - \log p_{\text{base}}(x_{t-1} \mid x_t)$ and the reward gradient as $g_t = \nabla_{x_{t-1}} \log R!\left(\hat{x}\theta(xt-1)\right)$, with $\hat{x}_\theta$ the clean-sample prediction and γ_t a timestep-dependent scale. These quantities guide prior-preserving transition updates along the reward model's local direction; substitution yields the gradient-informed training objective.

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{p_\theta}\left[\left\| \nabla_{x_{t-1}} \log \frac{p_\theta(x_{t-1} \mid x_t)}{p_{\text{base}}(x_{t-1} \mid x_t)} \quad -\gamma_t \beta \nabla_{x_{t-1}} \log R(\hat{x}_\theta(x_{t-1})) \right\|_2^2 \right]. \tag{2}$$

4 METHOD

4.1 VIEW-WISE REWARD–EVIDENCE MISORDERING

We study prior-preserving alignment of a 3D generator using a generic pretrained 2D RGB reward model. Given a generated asset x, a conditioning signal y, and $N \geq 2$ sampled camera poses $\{c_i\}_{i=1}^N$, let $I_i = \mathcal{T}(x, c_i)$ denote the rendered views. The reward model evaluates each view to produce a view-wise reward vector $\mathbf{r}$. To characterize the structural-evidence richness of each view, we use edge density as the sole proxy, where $\mathcal{E}_\tau(I_i) \in \{0,1\}^{H\times W}$ denotes the binary edge map from a fixed edge detector with threshold τ. The fraction of edge pixels defines a view-wise evidence vector $\mathbf{e}$, where a larger e_i indicates denser visible image-space edge evidence in view i.

$$\mathbf{r} = [\mathcal{R}_\psi(\mathcal{T}(x, c_i), y)]_{i=1}^N \in \mathbb{R}^N \qquad \mathbf{e} = \left[\frac{\|\mathcal{E}_\tau(\mathcal{T}(x, c_i))\|_1}{HW}\right]_{i=1}^N \in \mathbb{R}^N. \tag{3}$$

Because reward scores and edge densities lie on incomparable scales, we compare their induced view orderings by ranking views in descending order of reward and structural evidence, thereby obtaining scale-invariant comparisons. This avoids arbitrary normalization choices across the two signal domains while retaining only the relative cross-view preference encoded by each signal.

$$\pi_i^r = \mathrm{rank}^\downarrow\big(r_i; \{r_j\}_{j=1}^N\big), \quad \pi_i^e = \mathrm{rank}^\downarrow\big(e_i; \{e_j\}_{j=1}^N\big). \tag{4}$$

We characterize reward–evidence disagreement using the normalized signed rank displacement $\Delta_i = (\pi_i^e - \pi_i^r)/(N-1)$. Δ_i^+ indicates that view i receives a higher reward rank than evidence rank, whereas Δ_i^- indicates the reverse. Taking the absolute displacement yields the view-wise misordering, which directly measures the magnitude of ordering disagreement across views:

$$\mathbf{m} = [|\Delta_i|]_{i=1}^N = \left[\frac{|\pi_i^e - \pi_i^r|}{N-1}\right]_{i=1}^N \in [0,1]^N. \tag{5}$$

Spearman's footrule $\sum_i |\pi_i^e - \pi_i^r|$ has normalized view contribution m_i, giving scale-independent disagreement with standard permutation-statistical interpretation Diaconis & Graham (1977).

4.2 MISMATCH-AWARE REWARD CONFIDENCE

Positive displacement Δ_i^+ means reward outranks evidence, risking over-optimization of defective geometries. Negative displacement Δ_i^-, where $[a]_+ = \max(a, 0)$, implies evidence outranks reward, risking suppression of valid complexity. Both interpret misordering direction, not separate metrics. Asset-level intensity u weights mismatch prevalence and worst-case severity. Treating m_i as reliability residual, u attenuates disagreeing views, resembling robust estimation Huber (1992).

$$u = \frac{\lambda_{\mathrm{cnt}}}{N}\sum_{i=1}^N \mathbf{1}(m_i > \tau) + \lambda_{\mathrm{sev}} \max_{1\leq i\leq N} m_i, \quad \lambda_{\mathrm{cnt}}, \lambda_{\mathrm{sev}} \geq 0. \tag{6}$$

Asset-level intensity calibrates global confidence; m_i localizes disagreement by view. Their product is mapped to positive weights using $w_{\min} \in (0,1)$ and scaling $\kappa > 0$, then normalized view-wise:

$$w_i = w_{\min} + (1 - w_{\min})e^{-\kappa u m_i}, \quad \bar{w}i = \frac{w_i}{\sum_{j=1}^N w_j}. \tag{7}$$

To retain the total coefficient mass of standard uniform reward aggregation, we define view i confidence by rescaling normalized weight, guaranteeing unit-mean positive values:

$$\bar{w}_i = \frac{w_i}{\sum_{j=1}^N w_j}, \qquad q_i := N\bar{w}_i = \frac{N w_i}{\sum_{j=1}^N w_j}, \qquad \frac{1}{N}\sum_{i=1}^N q_i = 1.$$

Algorithm 1: GRPQ for 3D Diffusion Alignment

Input: Generator G_θ, reward model $\mathcal{R}_\psi$, base objective $\mathcal{O}_{\text{base}}$, camera distribution $\mathcal{C}$, edge detector $\mathcal{E}_{\tau_e}$; hyperparameters $\{N, \tau_m, w_{\min}, \kappa, \lambda_{\text{cnt}}, \lambda_{\text{sev}}\}$
Output: Aligned parameters θ

while not converged **do**
  $y \sim p_y, \quad x \leftarrow \text{BaseSample}(G_\theta, y; \mathcal{O}_{\text{base}}), \quad \{c_i\}_{i=1}^N \sim \mathcal{C}$
  for $i = 1, \ldots, N$ **do**
    $I_i \leftarrow \mathcal{T}(x, c_i), \quad r_i \leftarrow \mathcal{R}_\psi(I_i, y), \quad e_i \leftarrow \|\mathcal{E}_{\tau_e}(I_i)\|_1/(HW)$
  end for
  $\boldsymbol{\pi}^r \leftarrow \text{rank}^\downarrow(\mathbf{r}), \quad \boldsymbol{\pi}^e \leftarrow \text{rank}^\downarrow(\mathbf{e}), \quad m_i \leftarrow \frac{|\pi_i^e - \pi_i^r|}{N-1},\ i = 1, \ldots, N$
  $M_{\text{cnt}} \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{1}(m_i > \tau_m), \quad M_{\text{sev}} \leftarrow \max_i m_i, \quad u \leftarrow \lambda_{\text{cnt}} M_{\text{cnt}} + \lambda_{\text{sev}} M_{\text{sev}}$
  $w_i \leftarrow w_{\min} + (1 - w_{\min})\exp(-\kappa u m_i), \quad q_i \leftarrow \frac{N w_i}{\sum_{j=1}^N w_j}$
  $\mathcal{O}_{\text{base}}^{\text{GRPQ}} \leftarrow \mathcal{O}_{\text{base}}\left[\frac{1}{N}\sum_i (\cdot)_i \mapsto \frac{1}{N}\sum_i \text{sg}[q_i](\cdot)_i\right]$
  Update θ using $\mathcal{O}_{\text{base}}^{\text{GRPQ}}$
end while
return θ

The confidence-weighted reward attenuates structurally unreliable signals while preserving aggregate weight, and $\text{sg}[\cdot]$ blocks gradients through the confidence estimator. If reward and structural-evidence rankings coincide for all views, then $m_i = M_{\text{cnt}} = M_{\text{sev}} = u = 0$, yielding $w_i = 1$ and $q_i = 1$. The objective thus exactly recovers standard uniform multi-view reward aggregation:

$$R_{\text{conf}}(x) = \frac{1}{N}\sum_{i=1}^N \text{sg}[q_i]\mathcal{R}_\psi\left(\mathcal{T}(x, c_i), y\right). \tag{8}$$

4.3 Plug-and-Play Integration with Alignment

Confidence factors $\{q_i\}_{i=1}^N$ replace uniform camera averages with a mean-preserving weighted average $\frac{1}{N}\sum_{i=1}^N \text{sg}[q_i](\cdot)$, thereby down-weighting less reliable views while preserving the aggregate scale across cameras. Importantly, the original alignment objectives remain otherwise unmodified.

DDPO. DDPO fundamentally weights each denoising trajectory's score-function gradient by its corresponding terminal reward, thereby coupling updates to terminal outcomes. To enhance robustness, we replace uniform multi-view aggregation with a confidence-weighted reward:

$$\nabla_\theta J_{\text{DDPO}}^{\text{GRPQ}} = \mathbb{E}_{\tau \sim p_\theta(\tau|y)}\left[\sum_{t=1}^T \nabla_\theta \log p_\theta(x_{t-1} \mid x_t, y) \times \left(\frac{1}{N}\sum_{i=1}^N \text{sg}[q_i]\,\mathcal{R}_\psi(\mathcal{T}(x_0, c_i), y)\right)\right]. \tag{9}$$

ReFL. ReFL applies a reward-to-loss mapping to the one-step clean prediction from a random late denoising step. We only modify the reward supplied to this mapping, retaining the original pretraining loss without altering its optimization structure. $g_\theta(y)$ denotes this prediction.

$$\mathcal{L}_{\text{ReFL}}^{\text{GRPQ}} = \lambda\,\mathbb{E}_{y \sim \mathcal{Y}}\left[\phi\left(\frac{1}{N}\sum_{i=1}^N \text{sg}[q_i] \times \mathcal{R}_\psi(\mathcal{T}(g_\theta(y), c_i), y)\right)\right]. \tag{10}$$

∇-R2D3. ∇-R2D3's approximate gradient-matching objective averages per-view losses over sampled camera poses. We replace this uniform camera expectation with a confidence-weighted average of these per-view losses while preserving the original per-view loss structure:

$$\begin{aligned}\mathcal{L}_{\nabla\text{-}R2D3}^{\text{GRPQ}}(z_{t-1:t}) = \frac{1}{N}\sum_{i=1}^N \text{sg}[q_i]\Big\|&\nabla_{z_{t-1}} \log \widetilde{p}_\theta(z_{t-1} \mid z_t) \\ &- \gamma_t \beta\ \text{sg}\big[\nabla_{z_{t-1}} \log \mathcal{R}_\psi\big(\mathcal{T}(\hat{z}_\theta(z_{t-1}), c_i), y\big)\big]\Big\|_2^2.\end{aligned} \tag{11}$$

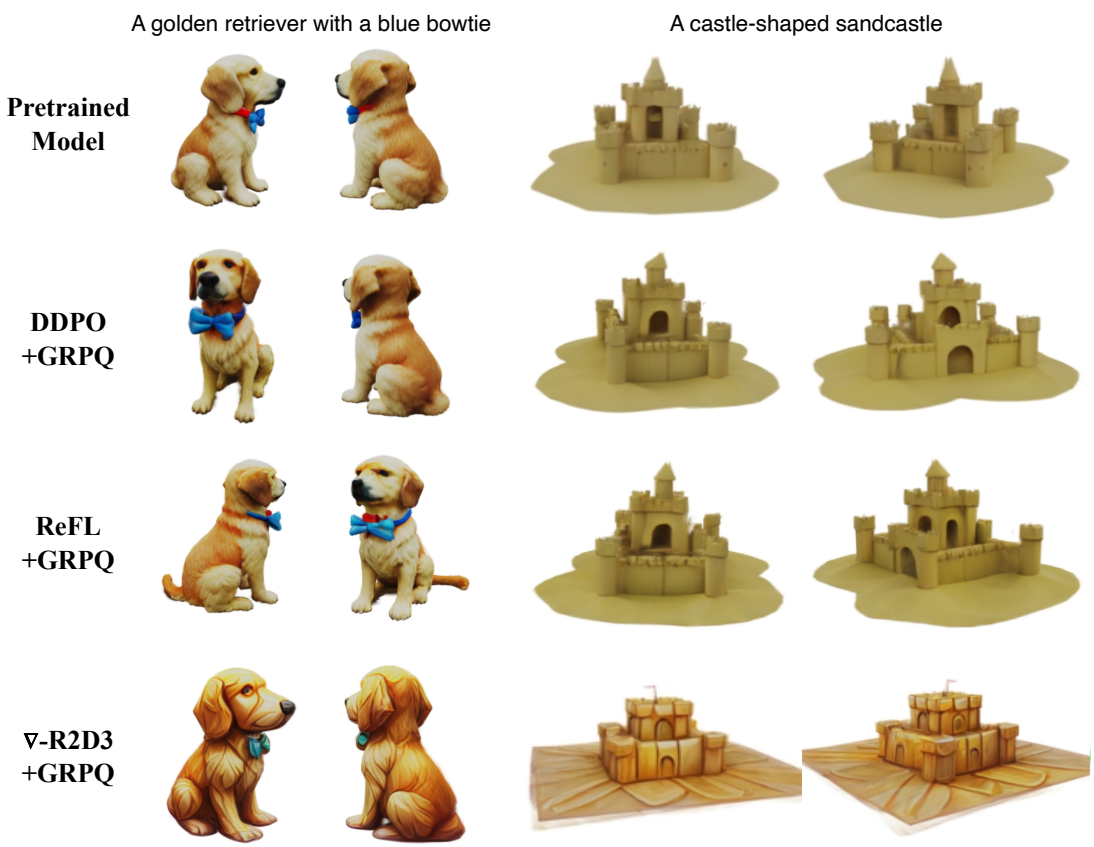

Figure 2: Qualitative examples generated by models trained with different alignment methods augmented with GRPQ reweighting for improved structural consistency across views.

5 Evaluation

5.1 Experimental Setup

Experimental Protocol We instantiate GRPQ with representative diffusion-alignment objectives, including DDPO Black et al. (2024), ReFL Xu et al. (2023), and Nabla-R2D3 Liu et al. (2026). We evaluate all methods using pretrained DiffSplat Lin et al. (2025) and GaussianCube Zhang et al. (2024a) checkpoints on the 100 prompts from T^3Bench He et al. (2023). All experiments are conducted on eight NVIDIA GeForce RTX 4080 SUPER GPUs under identical initialization, reward models, prompt and camera distributions, random seeds, numerical precision, batch construction, view sampling, rendering and sampling configurations, optimizer settings, training budgets, reward evaluation procedures, checkpoint-selection criteria, and evaluation protocols across methods.

5.1.1 Evaluation Metrics

We evaluate preference alignment, prior preservation, text–object correspondence, cross-view perceptual consistency, and geometric integrity using five metrics. Preference alignment is measured using the reward model employed during optimization, including Aesthetic Score Schuhmann (2022) and HPSv2 Wu et al. (2023). We further report MV-FID Gao et al. (2022); Heusel et al. (2017), MV-CLIP Radford et al. (2021), and A-LPIPS Zhang et al. (2018) for prior preservation, text–object correspondence, and perceptual consistency, respectively. All multi-view metrics are averaged across views, assets, and prompts. We introduce the Multi-view Geometric Integrity Score (MGIS), which combines three aspects: silhouette consistency, cross-view depth consistency, and local surface regularity. Its complete definition is provided in the Appendix B.

Table 1: Quantitative comparison of RGB-reward-based alignment on DiffSplat and GaussianCube. Each entry reports the original result, followed by the change obtained with GRPQ in parentheses, e.g., 20(+2). Higher values indicate better performance for Reward, MV-CLIP, and MGIS, whereas lower values indicate better performance for MV-FID and A-LPIPS.

Method	Aesthetic Score Alignment (+GRPQ)					HPSv2 Alignment (+GRPQ)				
	Reward ↑	**MV-FID ↓**	**MV-CLIP ↑**	**A-LPIPS ↓**	**MGIS ↑**	**Reward ↑**	**MV-FID ↓**	**MV-CLIP ↑**	**A-LPIPS ↓**	**MGIS ↑**
DiffSplat-SD1.5	4.539	–	23.234	0.036	0.718	0.189	–	23.171	0.036	0.834
DDPO	4.554 *(+0.034)*	52.113 *(-0.924)*	25.520 *(+0.018)*	0.034 *(-0.018)*	0.523 *(+0.032)*	0.203 *(+0.020)*	43.210 *(+1.941)*	25.776 *(+0.035)*	0.033 *(-0.021)*	0.417 *(+0.023)*
ReFL	4.601 *(+0.048)*	62.580 *(-0.462)*	25.877 *(+0.217)*	0.034 *(-0.009)*	0.326 *(+0.022)*	0.201 *(+0.017)*	70.765 *(+0.173)*	25.668 *(+0.114)*	0.035 *(-0.075)*	0.318 *(+0.012)*
Nabla-R2D3	5.468 *(+0.007)*	185.422 *(+1.523)*	25.296 *(+0.133)*	0.030 *(-0.007)*	–	0.213 *(+0.001)*	96.297 *(+5.941)*	26.148 *(+0.070)*	0.031 *(-0.013)*	0.341 *(+0.043)*
GaussianCube	4.353	–	0.184	0.089	0.689	0.134	–	0.196	0.084	0.732
DDPO	4.453 *(+0.017)*	2.925 *(+0.001)*	20.470 *(+0.025)*	0.083 *(-0.002)*	0.573 *(+0.132)*	0.149 *(+0.003)*	2.108 *(-0.369)*	20.533 *(+0.025)*	0.068 *(-0.002)*	0.526 *(+0.232)*
ReFL	4.421 *(+0.031)*	4.566 *(-0.026)*	21.215 *(+0.028)*	0.084 *(+0.000)*	0.471 *(+0.212)*	0.150 *(+0.008)*	5.042 *(+0.730)*	21.410 *(+0.079)*	0.073 *(-0.001)*	0.372 *(+0.212)*
Nabla-R2D3	4.467 *(+0.051)*	9.005 *(-2.275)*	22.267 *(+0.024)*	0.073 *(-0.001)*	0.431 *(+0.142)*	0.153 *(+0.002)*	3.509 *(+0.056)*	21.597 *(+0.054)*	0.072 *(-0.002)*	0.417 *(+0.142)*

$$\mathrm{MGIS}(\mathcal{I}) = \left(S_{\mathrm{occ}}^{\alpha_{\mathrm{occ}}} S_{\mathrm{dep}}^{\alpha_{\mathrm{dep}}} S_{\mathrm{surf}}^{\alpha_{\mathrm{surf}}}\right)^{\frac{1}{\alpha_{\mathrm{occ}}+\alpha_{\mathrm{dep}}+\alpha_{\mathrm{surf}}}}, \quad \alpha_{\mathrm{occ}} > 0, \quad \alpha_{\mathrm{dep}} > 0, \quad \alpha_{\mathrm{surf}} > 0. \tag{12}$$

5.2 Main Results and Training Efficiency

Main results. We compare the original alignment objectives with their GRPQ-augmented variants on DiffSplat and GaussianCube under Aesthetic Score and HPSv2 alignment. As summarized in Table 1, GRPQ improves reward scores and MV-CLIP across all evaluated configurations, with consistent MGIS gains in all reported comparisons. These improvements span both model architectures and reward settings, supporting the applicability of GRPQ beyond a single backbone or preference signal. Changes in MV-FID are mixed, however, indicating that improved geometric self-consistency does not uniformly translate into stronger prior preservation as measured by this metric. Taken together, the results suggest that GRPQ enhances geometric self-consistency and text–object correspondence without compromising reward optimization, offering a practical means of improving structural quality during RGB-reward-based alignment.

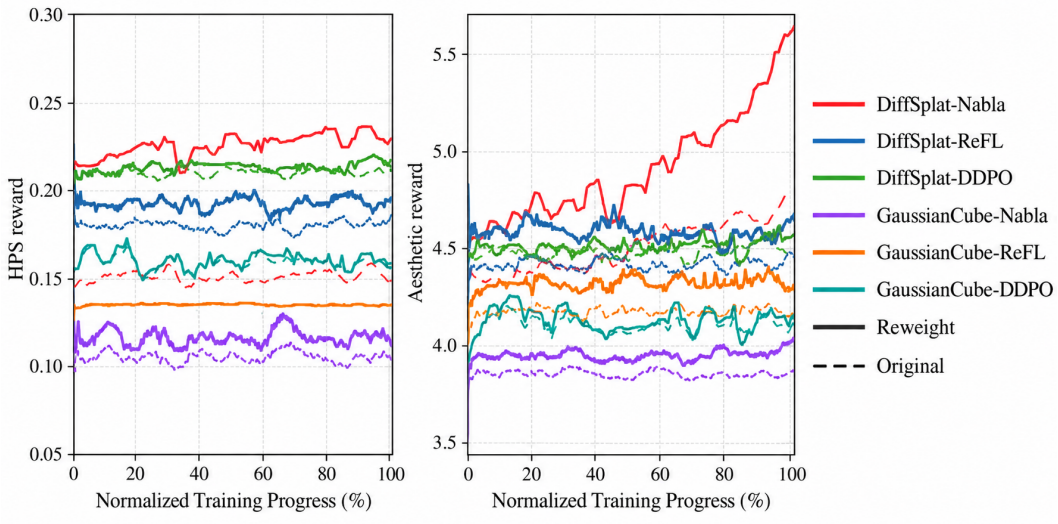

Figure 3: Training curves under different reward models.

Training efficiency. Figure 3 compare the optimization dynamics under the two reward models. With GRPQ, reward optimization efficiency remains nearly unchanged, while structure-related metrics improve consistently throughout training. This behavior suggests that the additional structural constraints do not materially slow preference optimization or alter its overall convergence tendency. Furthermore, the persistent gains in structural quality indicate that GRPQ provides preservation

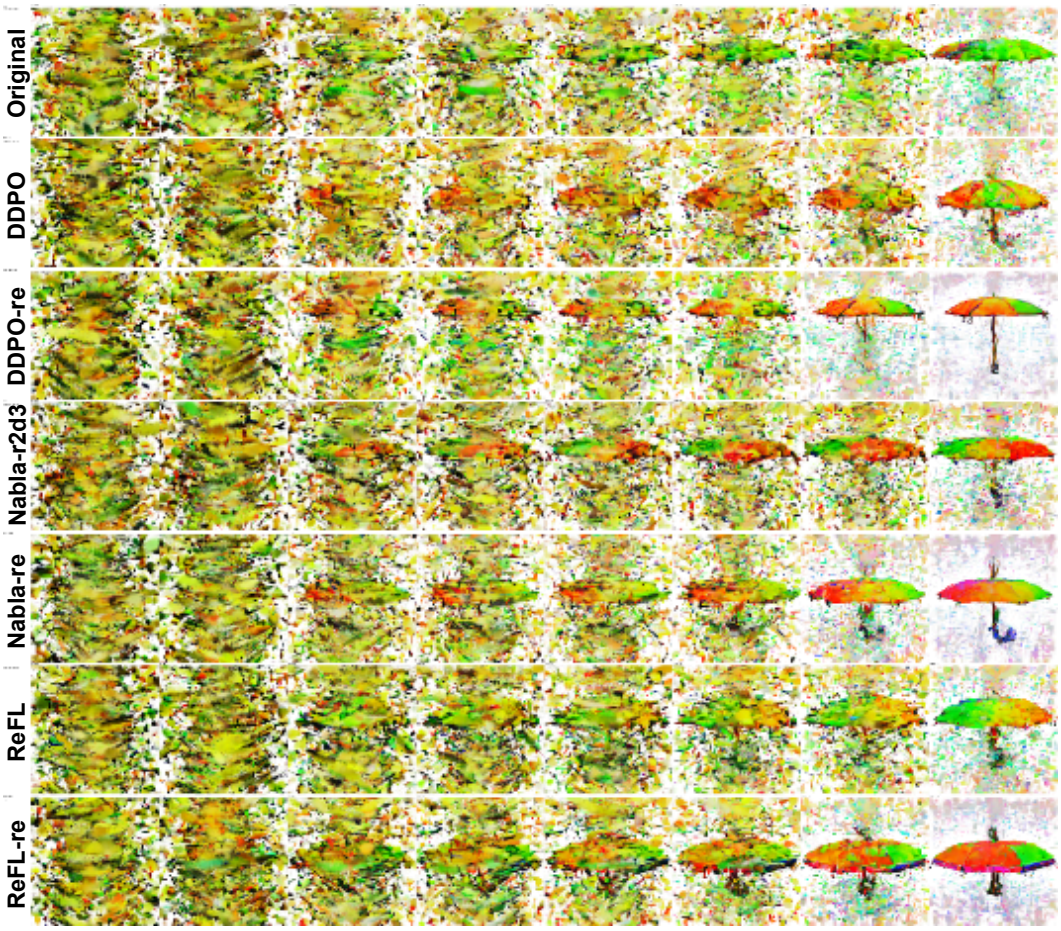

Figure 4: Denoising Trajectories of RGB Reward Alignment Methods With and Without GRPQ

benefits during the optimization process rather than only at the final checkpoint. Overall, GRPQ introduces limited optimization overhead while delivering more reliable structural preservation.

5.3 Qualitative Analysis of Denoising Trajectories

We analyze the mechanism by which GRPQ preserves the structural priors of pretrained 3D generators during RGB-reward-based alignment. Figure 4 visualizes the intermediate denoising trajectories produced by different alignment strategies. Throughout the denoising process, GRPQ identifies and suppresses low-confidence reward signals that may introduce unstable or structurally inconsistent updates. By limiting the influence of these potentially harmful signals, GRPQ prevents local appearance optimization from progressively distorting the underlying geometry. Consequently, the model improves RGB-level preference alignment while retaining coherent object structures and stable multi-view generation under diverse prompts and viewing conditions.

6 Conclusion

We present GRPQ, a plug-and-play framework for preserving the structural prior of pretrained 3D diffusion models during alignment with generic 2D RGB rewards. By identifying reward–structural-evidence misordering and down-weighting unreliable view-level supervision, GRPQ mitigates structural degradation during reward-driven adaptation without introducing additional depth, normal, or multi-view reward models. Experiments across multiple alignment methods, 3D generators, and reward functions show that GRPQ consistently improves structural integrity, prior preservation, and cross-view consistency while largely maintaining reward gains and training efficiency.

AI USE STATEMENT

Generative AI tools were used in this work to assist with methodological discussion and refinement of mathematical formulations, literature search and summarization, manuscript drafting and language editing, and the preparation and refinement of figures, captions, tables, and code. AI tools were also used to provide feedback on experimental design and the presentation and interpretation of results. All AI-assisted content, including equations, methodological suggestions, code, references, and written text, was carefully reviewed and verified by the authors. Experimental results were obtained from our own implementations and evaluation pipelines, and all scientific claims and conclusions were independently checked by the authors. We take full responsibility for the final content of this work, including any text, claims, code, or other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies the alignment of 3D generative models using pretrained 2D reward models and does not involve human-subject experiments or the collection of personally identifiable information. Our experiments use existing pretrained models, public benchmarks, and automatically generated 3D assets. As with other generative-model research, the behavior of the resulting systems may inherit biases from the underlying pretrained models, reward models, and evaluation data, and generated content may potentially be used in unintended ways. Our work focuses on improving the structural reliability of 3D generation rather than expanding capabilities toward harmful applications. We follow the ICLR Code of Ethics and take responsibility for the research practices, experimental results, and claims presented in this paper.

REPRODUCIBILITY STATEMENT

We provide the mathematical formulation and complete algorithm of GRPQ in Sec. 4, including the definition of reward–evidence misordering, confidence estimation, and its integration with different diffusion-alignment objectives. Sec. 5.1 specifies the experimental protocol, including the evaluated base models, reward models, prompts, hardware, camera sampling, optimization settings, random seeds, and evaluation procedure. The Appendix provides additional analyses and the complete definition of the proposed Multi-view Geometric Integrity Score (MGIS). Together, these descriptions specify the main methodological, experimental, and evaluation components required to reproduce the reported results.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Itay Chachy, Guy Yariv, and Sagie Benaim. Rewardsds: Aligning score distillation via reward-weighted sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4916–4926, 2026.

Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5881–5892, 2025.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22246–22256, 2023.

Wei Cheng, Juncheng Mu, Xianfang Zeng, Xin Chen, Anqi Pang, Chi Zhang, Zhibin Wang, Bin Fu, Gang Yu, Ziwei Liu, et al. Mvpaint: Synchronized multi-view diffusion for painting anything 3d. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 585–594, 2025.

Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Yuan Li. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. In *International Conference on Learning Representations*, volume 2024, pp. 53067–53087, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, volume 2024, pp. 4793–4822, 2024.

Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(2):262–268, 1977.

Carles Domingo i Enrich, Michal Drozdzal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, volume 2025, pp. 53791–53846, 2025.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems*, 35:31841–31854, 2022.

Chongjian Ge, Chenfeng Xu, Yuanfeng Ji, Chensheng Peng, Masayoshi Tomizuka, Ping Luo, Mingyu Ding, Varun Jampani, and Wei Zhan. Compgs: Unleashing 2d compositionality for compositional text-to-3d via dynamically optimizing 3d gaussians. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18509–18520, 2025.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3 bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation. *Advances in Neural Information Processing Systems*, 36:11970–11987, 2023.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Savva Ignatyev, Nina Konovalova, Daniil Selikhanovych, Oleg Voinov, Nikolay Patakin, Ilya Olkov, Dmitry Senushkin, Alexey Artemov, Anton Konushin, Alexander Filippov, et al. A3d: Does diffusion dream about 3d alignment? In *International Conference on Learning Representations*, volume 2025, pp. 47406–47430, 2025.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 867–876, 2022.

Jiaqi Leng, Shuyuan Tu, Haidong Cao, Sicheng Xie, Daoguo Dong, Zuxuan Wu, and Yu-Gang Jiang. Preference score distillation: Leveraging 2d rewards to align text-to-3d generation with human preference. *arXiv preprint arXiv:2603.01594*, 2026.

Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. In *International conference on learning representations*, volume 2024, pp. 21567–21589, 2024.

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native diffusion and interactive geometry refiner. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5307–5317, 2025.

Zejian Li, Jiarui Ma, Han Xu, Weiting Zheng, Yangrui Zhu, Chenye Meng, Pei Chen, Ling Yang, Zhiyuan Yang, Changyuan Yang, et al. Circular-dpo: Aligning multi-stage 3d generative models via preference feedback loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32592–32601, 2026.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable 3d gaussian splat generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Qingming Liu, Zhen Liu, Dinghuai Zhang, and Kui Jia. Nabla-r2d3: Effective and efficient 3d diffusion alignment with 2d rewards. *Advances in Neural Information Processing Systems*, 38: 47855–47884, 2026.

Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J Black, and David Ha. Ghost on the shell: An expressive representation of general 3d shapes. In *International Conference on Learning Representations*, volume 2024, pp. 52056–52085, 2024.

Zhen Liu, Tim Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed gflownets. In *International Conference on Learning Representations*, volume 2025, pp. 94806–94841, 2025.

Marian Lupascu, Mihai Sorin Stupariu, and Ionut Mironica. Variance reduction on the camera axis: Multi-view score distillation for 3d. *arXiv preprint arXiv:2606.29964*, 2026.

Zhiyuan Ma, Yuxiang Wei, Yabin Zhang, Xiangyu Zhu, Zhen Lei, and Lei Zhang. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Yuxuan Mu, Xinxin Zuo, Chuan Guo, Yilin Wang, Juwei Lu, Xiaofeng Wu, Songcen Xu, Peng Dai, Youliang Yan, and Li Cheng. Gsd: View-guided gaussian splatting diffusion for 3d reconstruction. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9914–9925, 2024.

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.

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*, 36:53728–53741, 2023.

Christoph Schuhmann. LAION-Aesthetics, 2022. Available at: https://laion.ai/blog/laion-aesthetics/. Accessed: 2024-09-27.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International conference on learning representations*, volume 2024, pp. 39838–39859, 2024.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19615–19625, 2024.

Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9037–9047, 2024.

Weitao Wang, Haoran Xu, Yuxiao Yang, Zhifang Liu, Jun Meng, and Haoqian Wang. Mvreward: Better aligning and evaluating multi-view diffusion models with human preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7898–7906, 2025.

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.

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*, pp. 21469–21480, 2025.

Desai Xie, Jiahao Li, Hao Tan, Xin Sun, Zhixin Shu, Yi Zhou, Sai Bi, Sören Pirk, and Arie E Kaufman. Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6369–6379, 2024.

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*, 36:15903–15935, 2023.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (ToG)*, 43(6):1–18, 2024a.

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 *European Conference on Computer Vision*, pp. 259–276. Springer, 2024b.

Zeyuan Yin and Xiaoming Liu. Trim: Scalable 3d gaussian diffusion inference with temporal and spatial trimming. *Advances in Neural Information Processing Systems*, 38:50356–50380, 2026.

Baowen Zhang, Chuan Fang, Rakesh Shrestha, Yixun Liang, Xiao-Xiao Long, and Ping Tan. Radegs: Rasterizing depth in gaussian splatting. *ACM Transactions on Graphics*, 45(2):1–14, 2026a.

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv preprint arXiv:2403.19655*, 2(3):6, 2024a.

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, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yuhan Zhang, Mengchen Zhang, Tong Wu, Tengfei Wang, Gordon Wetzstein, Dahua Lin, and Ziwei Liu. 3dgen-bench: Comprehensive benchmark suite for 3d generative models. *arXiv preprint arXiv:2503.21745*, 2025a.

Yujie Zhang, Bingyang Cui, Qi Yang, Zhu Li, and Yiling Xu. Benchmarking and learning multi-dimensional quality evaluator for text-to-3d generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18563–18574. IEEE, 2025b.

Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18, 2024c.

Ziyi Zhang, Li Shen, Deheng Ye, Yong Luo, Huangxuan Zhao, Meng Liu, Wei Yu, and Lefei Zhang. Refining few-step text-to-multiview diffusion via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2401–2411, 2026b.

Chenxi Zheng, Yihong Lin, Bangzhen Liu, Xuemiao Xu, Yongwei Nie, and Shengfeng He. Recdreamer: Consistent text-to-3d generation via uniform score distillation. *ICLR*, 1:3, 2025.

Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. *arXiv preprint arXiv:2502.04370*, 2025.

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.

A Correlation Between Richness and Rewards

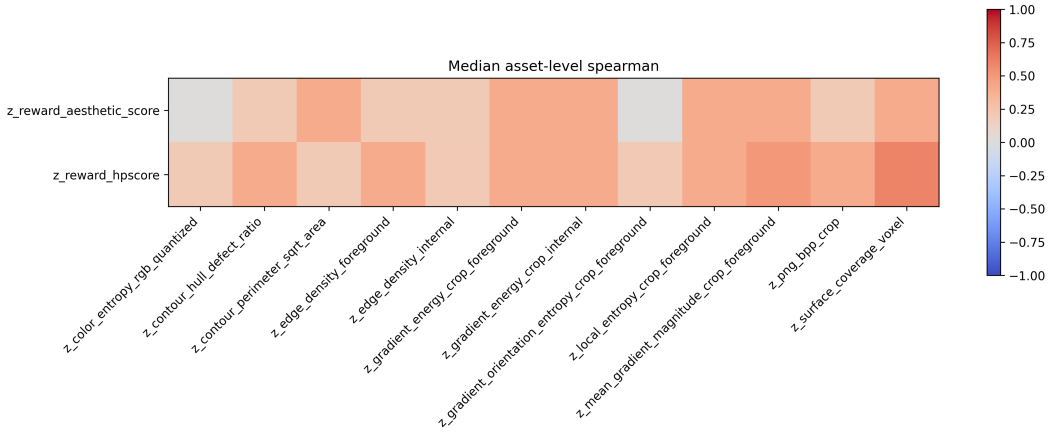

Figure 5: Reward-richness heatmap

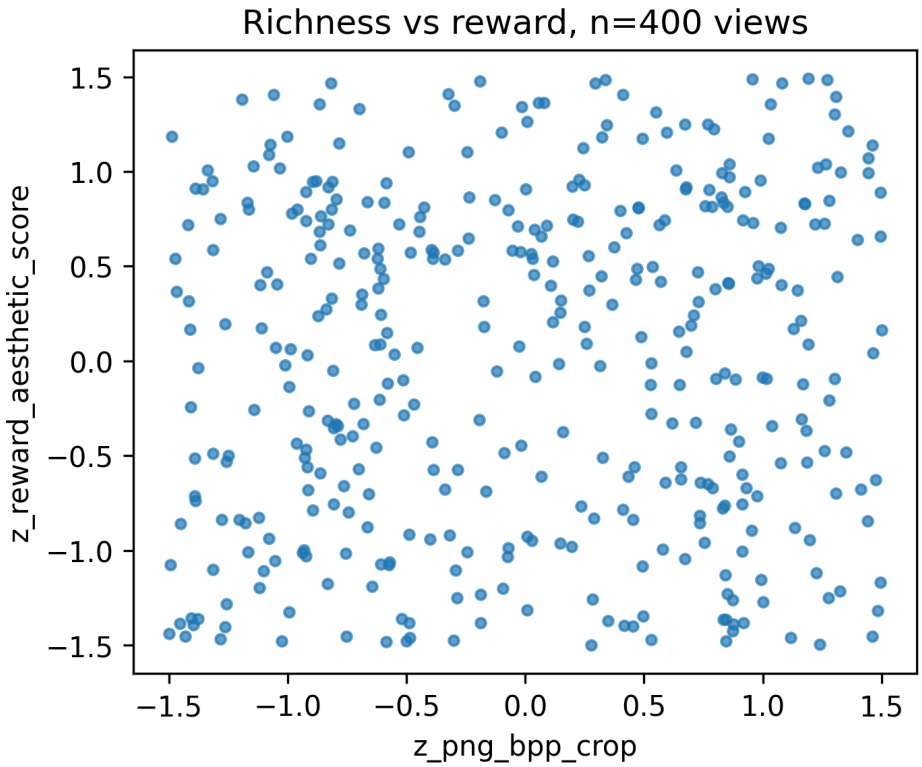

Figure 6: Reward-richness scatter

B Multi-view Geometric Integrity Score

We provide the complete definition of the Multi-view Geometric Integrity Score (MGIS), which evaluates the geometric self-consistency of a generated 3D asset across rendered views. Unlike reference-based 3D metrics, MGIS requires neither ground-truth geometry nor an additional learned depth or normal estimator. Instead, it operates directly on depth and opacity maps rendered from the generated asset using the fixed camera set adopted for evaluation.

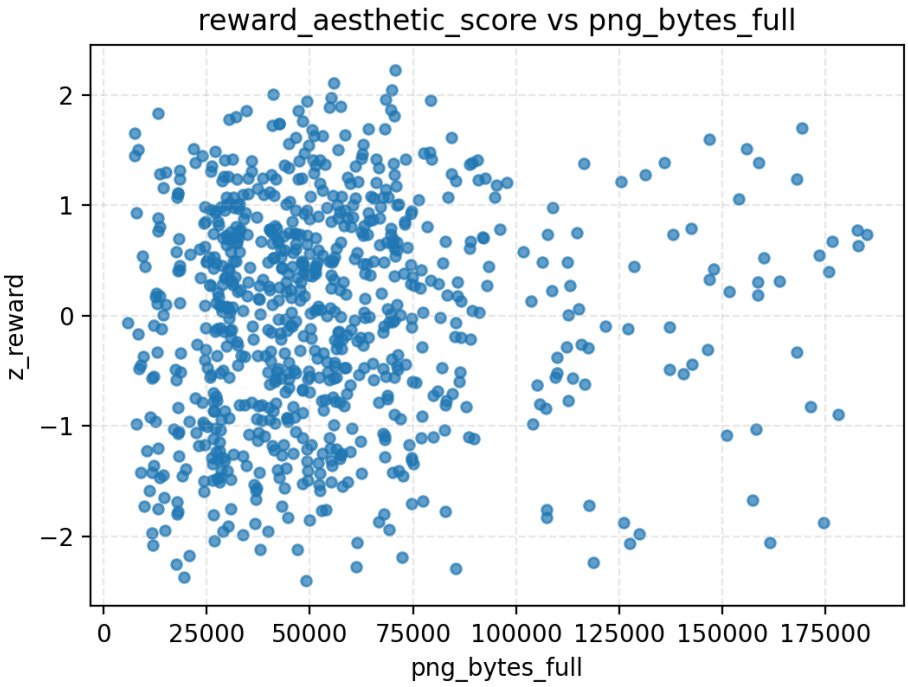

Figure 7: Aesthetic score and png bytes

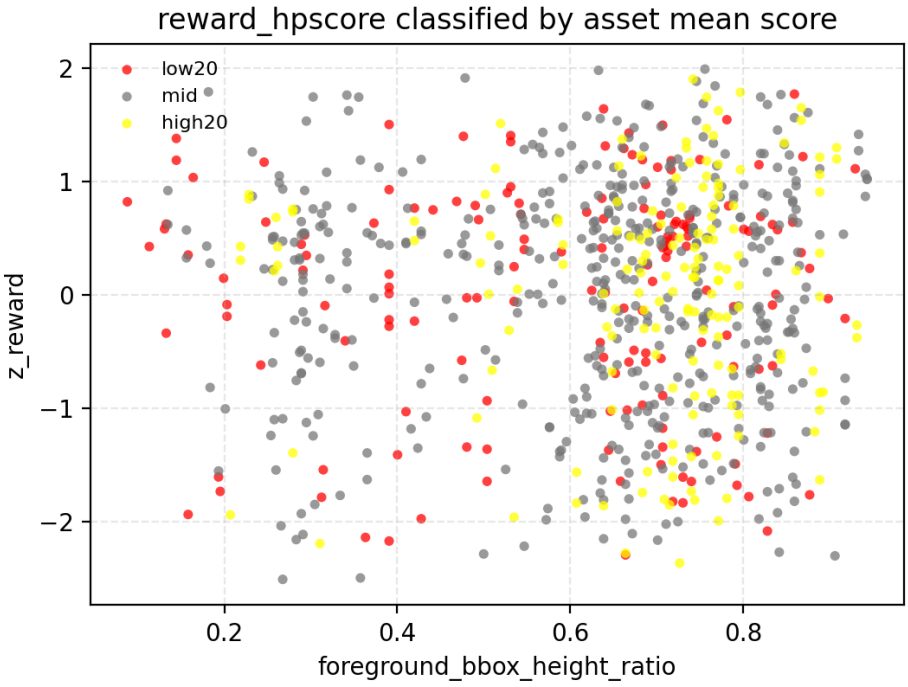

Figure 8: Hpscore and foreground bbox

Setup. Let $\mathcal{I} = \{I_i\}_{i=1}^{N}$ denote N rendered views with known camera projections $\{\Pi_i\}_{i=1}^{N}$. For each view i, we render a soft opacity map $A_i \in [0,1]^{H\times W}$ and a depth map $D_i \in \mathbb{R}_{+}^{H\times W}$. A valid

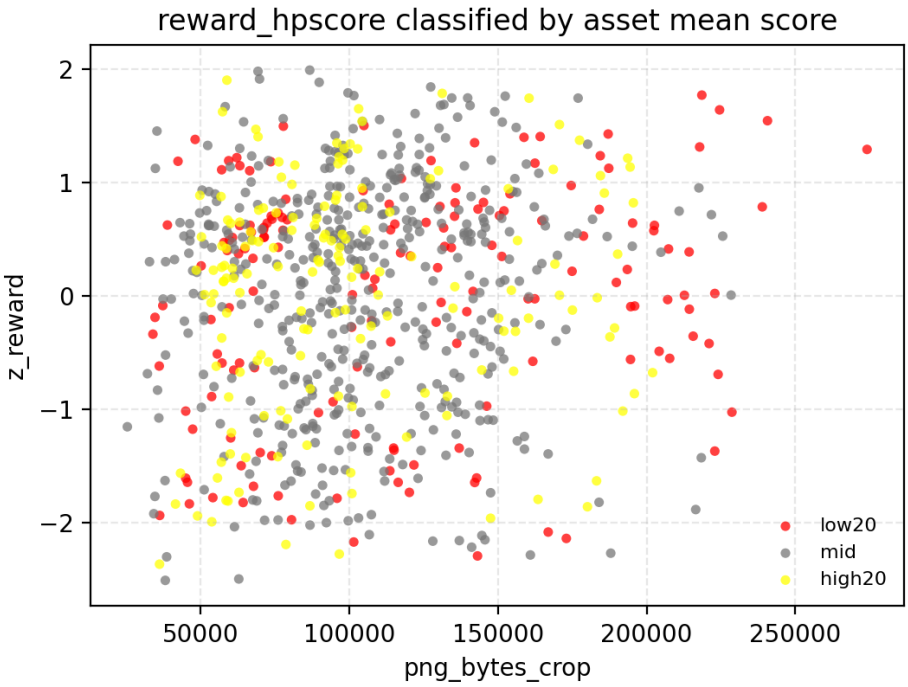

Figure 9: Hpscore and png bytes

pixel $\mathbf{p}$ in view i is back-projected to 3D and reprojected into view j as

$$\mathbf{X}_i(\mathbf{p}) = \Pi_i^{-1}(\mathbf{p}, D_i(\mathbf{p}))\,, \qquad \mathbf{p}_{i\to j} = \Pi_j(\mathbf{X}_i(\mathbf{p}))\,. \tag{13}$$

We forward-warp the source opacity and depth into view j, producing $\widetilde{A}_{i\to j}$ and $\widetilde{D}_{i\to j}$. Z-buffer visibility removes occluded correspondences. Let Ω_{ij} denote the resulting set of valid mutually visible target pixels, and let $\mathcal{P}$ denote the set of ordered camera pairs with overlapping fields of view. Throughout, $\varepsilon > 0$ is a small constant for numerical stability.

Silhouette consistency. Structural defects such as floating geometry, missing parts, or excessive opacity can produce inconsistent projected occupancy across viewpoints. We quantify this discrepancy using the soft intersection-over-union between the reprojected source opacity and target opacity:

$$S_{\text{occ}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{\sum_{\mathbf{p}\in\Omega_{ij}} \min\Big\{\widetilde{A}_{i\to j}(\mathbf{p}), A_j(\mathbf{p})\Big\}}{\sum_{\mathbf{p}\in\Omega_{ij}} \max\Big\{\widetilde{A}_{i\to j}(\mathbf{p}), A_j(\mathbf{p})\Big\} + \varepsilon}. \tag{14}$$

Higher S_{occ} indicates stronger agreement of projected object support across views.

Cross-view depth consistency. Silhouette agreement alone does not constrain the placement of visible surfaces in 3D. We therefore measure whether a surface point reconstructed from one view appears at a consistent depth when observed from another. For each valid correspondence, we define

the symmetric relative depth discrepancy and its aggregated consistency score as

$$d_{ij}(\mathbf{p}) = \frac{\left|D_j(\mathbf{p}) - \widetilde{D}_{i\to j}(\mathbf{p})\right|}{D_j(\mathbf{p}) + \widetilde{D}_{i\to j}(\mathbf{p}) + \varepsilon},$$
$$S_{\mathrm{dep}} = 1 - \frac{\sum_{(i,j)\in\mathcal{P}} \sum_{\mathbf{p}\in\Omega_{ij}} d_{ij}(\mathbf{p})}{\sum_{(i,j)\in\mathcal{P}} |\Omega_{ij}|}. \tag{15}$$

The relative normalization reduces sensitivity to the absolute depth scale. Thus, S_{dep} approaches one when mutually visible surfaces agree geometrically across viewpoints.

Local surface regularity. The preceding terms capture cross-view agreement but may not fully reflect high-frequency geometric artifacts within individual views. We therefore estimate local surface normals directly from the rendered depth maps. For valid central-difference stencils, define

$$\begin{aligned}
\mathbf{t}_i^x(\mathbf{p}) &= \mathbf{X}_i(\mathbf{p}+\mathbf{e}_x) - \mathbf{X}_i(\mathbf{p}-\mathbf{e}_x),\\
\mathbf{t}_i^y(\mathbf{p}) &= \mathbf{X}_i(\mathbf{p}+\mathbf{e}_y) - \mathbf{X}_i(\mathbf{p}-\mathbf{e}_y),\\
\mathbf{n}_i(\mathbf{p}) &= \frac{\mathbf{t}_i^x(\mathbf{p}) \times \mathbf{t}_i^y(\mathbf{p})}{\|\mathbf{t}_i^x(\mathbf{p}) \times \mathbf{t}_i^y(\mathbf{p})\|_2 + \varepsilon},
\end{aligned} \tag{16}$$

where $\mathbf{e}_x$ and $\mathbf{e}_y$ are one-pixel horizontal and vertical offsets. Let $\mathcal{E}_i$ contain valid 4-connected neighboring pixel pairs $(\mathbf{p}, \mathbf{q})$ in view i. We weight each pair by its joint opacity,

$$\omega_i(\mathbf{p}, \mathbf{q}) = A_i(\mathbf{p}) A_i(\mathbf{q}), \tag{17}$$

which suppresses background and weakly supported regions. Local surface regularity is then

$$S_{\mathrm{surf}} = \frac{\sum_{i=1}^{N} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{E}_i} \omega_i(\mathbf{p},\mathbf{q}) \frac{1 + \mathbf{n}_i(\mathbf{p})^\top \mathbf{n}_i(\mathbf{q})}{2}}{\sum_{i=1}^{N} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{E}_i} \omega_i(\mathbf{p},\mathbf{q}) + \varepsilon}. \tag{18}$$

This term measures local normal coherence and decreases for irregular, noisy, or fragmented surface geometry.

Overall MGIS. The three components capture complementary aspects of geometric integrity: S_{occ} measures projected occupancy consistency, S_{dep} measures cross-view surface agreement, and S_{surf} measures local surface regularity. We combine them using a weighted geometric mean:

$$\mathrm{MGIS}(\mathcal{I}) = \left(S_{\mathrm{occ}}^{\alpha_{\mathrm{occ}}} S_{\mathrm{dep}}^{\alpha_{\mathrm{dep}}} S_{\mathrm{surf}}^{\alpha_{\mathrm{surf}}}\right)^{\frac{1}{\alpha_{\mathrm{occ}}+\alpha_{\mathrm{dep}}+\alpha_{\mathrm{surf}}}}, \qquad \alpha_{\mathrm{occ}}, \alpha_{\mathrm{dep}}, \alpha_{\mathrm{surf}} > 0. \tag{19}$$

The multiplicative aggregation prevents strong performance in one component from fully compensating for severe degradation in another. The same weighting is used across all evaluated models and reward settings. Under equal weighting, $\alpha_{\mathrm{occ}} = \alpha_{\mathrm{dep}} = \alpha_{\mathrm{surf}} = 1$, MGIS simplifies to

$$\mathrm{MGIS}(\mathcal{I}) = \left(S_{\mathrm{occ}} S_{\mathrm{dep}} S_{\mathrm{surf}}\right)^{1/3}. \tag{20}$$

Higher MGIS indicates stronger overall geometric integrity. We compute MGIS for each generated asset over the fixed evaluation camera set and report the average across assets and prompts.