VIEWROBUST-3D: BENCHMARKING AND IMPROVING VIEWPOINT ROBUSTNESS IN SINGLE-VIEW 3D GENERATION

Anonymous authors
Paper under double-blind review

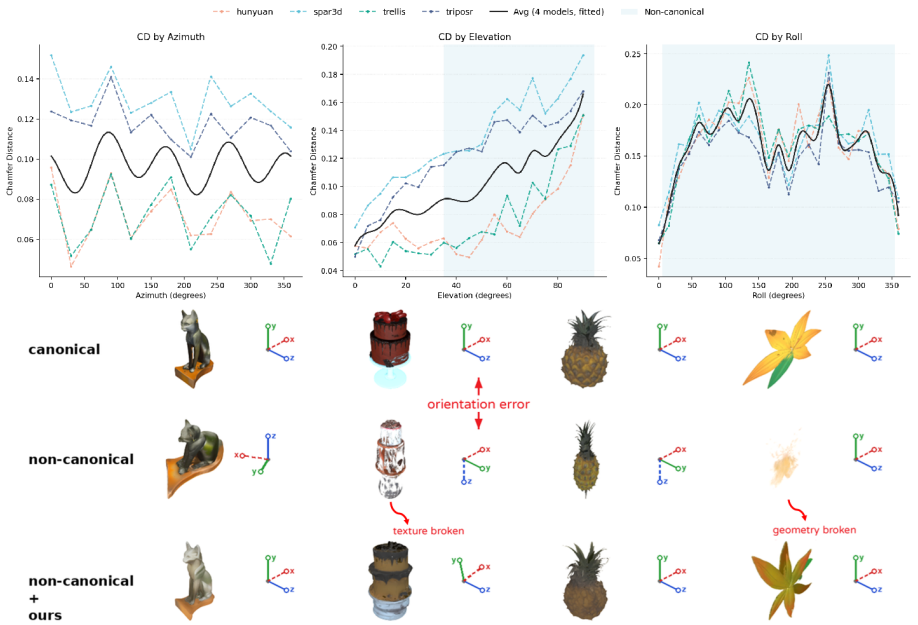


Figure 1: **Viewpoint-induced degradation in single-view 3D generation and qualitative result of our fix methods. Top:** Chamfer Distance (CD) of four representative models across azimuth, elevation, and roll, showing substantial degradation under non-canonical inputs (low elevation, non-zero roll viewpoints), especially at high elevation and non-zero roll. **Bottom:** From top to bottom: references generated from canonical views, failures under non-canonical views,Models are rotation-aligned for visual comparison, with the original coordinate-frame orientation shown alongside. The unrotated results are provided in Appendix B.The qualitative result show that our method alleviates both orientation and geometric errors.

ABSTRACT

In recent years, single-view 3D generation has advanced rapidly and shown great potential in many areas. However, existing models are primarily developed and evaluated under canonical views, leaving their robustness to broader view variations poorly understood. We observe that non-canonical inputs (e.g., high elevation or non-zero roll angles) can lead to reconstruction discrepancies. Motivated by this observation, we introduce **ViewRobust-3D**, a controlled benchmark to systematically analyze the impact of view variations on single-view 3D generation. By isolating view factors through controlled changes in azimuth, elevation, and roll, the benchmark reveals performance degradation and interactions between these factors across multiple models. Guided by these findings, we further propose **CanoView**, a lightweight view normalization strategy to normalize non-canonical inputs toward canonical views. Experiments validate **ViewRobust-3D** in exposing robustness gaps and demonstrate the effectiveness of **CanoView** in improving the performance of single-view 3D generation models under non-canonical views. Our benchmark and method will be publicly released upon publication.

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.

1 INTRODUCTION

Single-view 3D generation has emerged as a practical technology for rapidly creating 3D assets, with applications in gaming, film, AR/VR, e-commerce, robotics, and digital twins. Driven by advances in visual foundation models, large-scale 3D datasets, and efficient 3D representations, recent methods like LRM (Hong et al., 2024), TripoSR (Tochilkin et al., 2024), Hunyuan3D (Yang et al., 2024; Zhao et al., 2025), TRELLIS (Xiang et al., 2025; 2026), and SPAR3D (Huang et al., 2025) have substantially improved generation quality and generalization capabilities, transitioning this technology from experimental research to practical utility.

Despite these advancements, existing models are predominantly evaluated under canonical views (usually refer to the front, low-elevation, zero-roll images), leaving their robustness to non-canonical inputs such as high/low elevations and arbitrary roll angles poorly understood. However, in practical applications (e.g., robotics and user-captured photos), images often exhibit arbitrary viewpoints, requiring models to reliably generate 3D structures beyond commonly observed views. More importantly, evaluating models under non-canonical views provides a complementary perspective for assessing whether their 3D generation capability genuinely generalizes beyond familiar viewpoints. As illustrated in Figure 1, existing models can suffer from severe reconstruction inaccuracies when presented with non-canonical views, revealing substantial viewpoint sensitivity.

This significant viewpoint sensitivity underscores the importance of systematically evaluating single-view 3D generation under non-canonical viewpoints. Although recent studies such as MVG-Bench and Objaverse have revealed the viewpoint sensitivity of visual models (Deitke et al., 2023b; Xie et al., 2025), Cue3D (Li et al., 2025) has examined the role of different visual cues in single-view 3D generation, the systematic sensitivity and interaction effects of single-view 3D generation under varying elevation, azimuth, and roll remain unclear.

To address these evaluation gaps, we introduce **ViewRobust-3D**, a controlled benchmark that systematically evaluates single-view 3D generation under non-canonical viewpoints. Our initial observations on an Objaverse subset reveal that models are highly sensitive to elevation and roll, yet remain relatively robust to azimuth variations. Motivated by this finding, we construct a controlled evaluation set to deeply analyze the independent and interactive effects of elevation and roll. Specifically, we divide the benchmark into four distinct subsets, each revealing unique insights into model performance: (1) Canonical (low elevation, zero roll), which serves as the standard baseline where models generally perform optimally; (2) Elevation-Noncanonical (high elevation, zero roll), which shows that increased elevation alone leads to significant performance degradation across all models; (3) Roll-Noncanonical (low elevation, non-zero roll), where the generation performance drops rapidly as the roll deviates from 0 degrees, with a slight recovery at a completely inverted state of 180 degrees; and (4) Joint-Noncanonical (high elevation, non-zero roll), which reveals a surprising interaction effect: the performance drop at high elevation under non-zero roll is actually smaller, even surpassing the performance of low elevation under non-zero roll

To mitigate the performance degradation under these challenging viewpoints, we propose View Normalization, which transforms non-canonical inputs to a canonical view, thereby reducing model sensitivity. This approach only requires training a view transformation module, functioning as a generic, plug-and-play preprocessing step compatible with various downstream models. The processing pipeline includes: estimating the input elevation, converting it to a unified canonical view, performing local noise suppression, and finally feeding the result directly into the 3D generation model. Experiments demonstrate that our method successfully handles atypical viewpoint inputs with high elevation or roll angles, improving the model's performance on these inputs.

In summary, our main contributions are as follows:

- Systematically discovered and analyzed the sensitivity of single-view 3D generation models to various viewpoint parameters.
- Constructed a controlled non-canonical view evaluation set and conducted extensive cross-model quantitative and qualitative evaluations.
- Proposed a plug-and-play non-canonical view normalization method that bypasses downstream model retraining, with extensive empirical evaluations proving its capabilities.

2 Related Work

2.1 Single-View 3D Reconstruction and Generation

Beyond early neural implicit representations like NeRF (Mildenhall et al., 2020), LRM (Hong et al., 2024) and TripoSR (Tochilkin et al., 2024) pioneered **feed-forward 3D reconstruction** (Xu et al., 2024), while diffusion models advanced **generative 3D modeling** (Poole et al., 2023; Zhang et al., 2024; Shi et al., 2024; Long et al., 2024). Recent systems directly produce high-fidelity 3D assets, such as the point-cloud-based SPAR3D (Huang et al., 2025), Hunyuan3D 2.1 (Hunyuan3D Team et al., 2025), and TRELLIS (Xiang et al., 2025) and TRELLIS.2 (Xiang et al., 2026) with structured 3D latents. Following these two major routes—feed-forward reconstruction (TripoSR, SPAR3D[1]) and 3D latent generative modeling (TRELLIS.2, Hunyuan3D 2.1)—we select these four representative models to examine how viewpoint robustness varies across architectures.

2.2 3D Generation Evaluation and Benchmarks

Although Objaverse and MVGBench have revealed viewpoint sensitivity and multi-view consistency issues (Deitke et al., 2023b; Xie et al., 2025), Cue3D (Li et al., 2025) further quantifies the role of individual image cues, and recent evaluation suites (Wu et al., 2024) focus on text-to-3D alignment and aesthetics. However, existing evaluations still struggle to adequately characterize systematic sensitivity to non-canonical views: none of them systematically addresses the underlying patterns and interaction effects of final generation quality concerning viewpoint variations, which is precisely the gap ViewRobust-3D fills.

2.3 Novel View Synthesis

Diffusion-based novel view synthesis generates target views from a single image (Liu et al., 2023; Stability AI, 2023), with follow-ups (Deitke et al., 2023a; Kong et al., 2024). Since Stable Zero123 performs stably across varying elevations and rolls with a moderate parameter count, we adopt it as our view transformation module that maps non-canonical inputs to canonical views.

3 Benchmarking Non-Canonical Viewpoint Robustness

This section presents **ViewRobust-3D**, our controlled benchmark that systematically evaluates how robust existing single-view 3D generation methods are to non-canonical viewpoints. Specifically, it answers three questions: whether current methods generally degrade under non-canonical views or only specific individual models suffer from degradation (Section 3.2), how elevation, roll, and azimuth affect performance individually (Section 3.3), and whether these factors act interactively or independently (Section 3.4).

3.1 Controlled Benchmark Design

ViewRobust-3D isolates viewpoint factors by rendering the same objects under controlled conditions. We consider azimuth, elevation, and roll. As shown in Fig. 1, azimuth shows no clear global trend, while elevation and roll exhibit substantial and systematic effects. We therefore adopt an elevation $\times$ roll factorial design, fixing azimuth within each object group.

For each object-azimuth group, we construct four conditions: Canonical (C), low elevation ($[0°, 35°]$) with roll $= 0°$; Roll-Noncanonical (R), the same low elevation with nonzero roll; Elevation-Noncanonical (E), high elevation ($(35°, 90°]$) with roll $= 0°$; and Joint-Noncanonical (ER), the same high elevation with the same roll as R. Nonzero roll is uniformly sampled from $(0°, 360°)$ and shared by R and ER. The $35°$ boundary is chosen because Fig. 1 shows a stable CD regime below it followed by a sharp increase, while roll $= 0°$ is used as the canonical reference since small deviations already cause rapid CD degradation.

[1] Although SPAR3D contains point diffusion stage, it uses point diffusion only to generate 512 sparse points, followed by feed-forward mesh generation; we therefore group it with TripoSR.

Table 1: **Quantitative evaluation on the controlled evaluation set.** All models degrade under non-canonical viewpoints; roll causes larger average degradation than elevation, and their joint effect is not simply additive.

	Canonical		Elevation-Noncanonical		Roll-Noncanonical		Joint-Noncanonical	
Model	CD↓	PSNR↑	CD↓	PSNR↑	CD↓	PSNR↑	CD↓	PSNR↑
TripoSR	0.0917	19.41	0.1422 (+0.0505)	18.47 (-0.94)	**0.1486** (+0.0569)	18.44 (-0.97)	0.1512 (+0.0595)	18.26 (-1.15)
SPAR3D	0.1029	**19.94**	0.1552 (+0.0523)	**19.26** (-0.68)	0.1653 (+0.0624)	**19.36** (-0.58)	0.1724 (+0.0695)	**19.19** (-0.75)
TRELLIS	**0.0539**	18.38	0.0932 (+0.0393)	17.87 (-0.51)	0.1651 (+0.1112)	16.99 (-1.39)	0.1477 (+0.0938)	17.32 (-1.06)
Hunyuan3D	0.0617	17.71	**0.0815** (+0.0198)	17.27 (-0.44)	0.1629 (+0.1012)	16.23 (-1.48)	**0.1369** (+0.0752)	16.58 (-1.13)
Average	0.0776	18.86	0.1180 (+0.0405)	18.22 (-0.64)	0.1605 (+0.0829)	17.76 (-1.10)	0.1521 (+0.0745)	17.84 (-1.02)

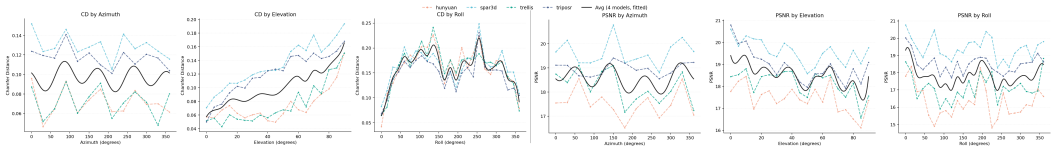


Figure 2: **Quantitative results across viewpoint factors.** CD and PSNR of four representative models versus azimuth, elevation, and roll. Azimuth mainly causes periodic fluctuations, whereas higher elevation and nonzero roll generally degrade generation quality.

ViewRobust-3D contains approximate 350 objects rendered from the aesthetically high-scoring subset of Objaverse (Deitke et al., 2023b), ranked by the aesthetic scores of GT23D-Bench (Cai et al., 2025). In total, this construction yields 588 object-azimuth groups, each containing the four conditions (C, E, R, and ER), for a total of 2352 test inputs. Before evaluation, prediction and GT are scale-normalized to their bounding-box diagonals and aligned only along azimuth using ICP (Besl & McKay, 1992) because elevation and roll have unambiguous reference orientation; aligning them would also mask spatial-orientation discrepancies in the generated models. We report Chamfer Distance (CD↓) for reconstruction fidelity (Fan et al., 2017) and Peak Signal-to-Noise Ratio (PSNR↑) for appearance consistency, computed as the average PSNR over the four front, back, left, and right views (Mildenhall et al., 2020).

3.2 How Robust Are Current Models to Non-Canonical Views?

We first investigate whether performance degradation under non-canonical viewpoints is specific to individual models or a general limitation of current image-to-3D systems. To this end, we evaluate four representative models—TripoSR (Tochilkin et al., 2024), SPAR3D (Huang et al., 2025), TRELLIS 2 (Xiang et al., 2026), and Hunyuan3D 2.1 (Hunyuan3D Team et al., 2025)—across the four controlled subsets. Results are summarized in Table 1.

A consistent robustness gap is observed across all four models. (i) **General degradation.** All models show higher CD and lower PSNR under non-canonical views, with average CD rising from 0.0776 (C) to 0.1180 (E) and 0.1605 (R); (ii) **Stronger roll sensitivity.** Roll causes larger CD and PSNR changes than elevation ($+0.0829$ vs. $+0.0405$; -1.10 vs. -0.64); (iii) **Model-dependent responses.** TRELLIS and Hunyuan3D are relatively robust to E ($+0.0393$, $+0.0198$), whereas TripoSR and SPAR3D are more robust to R ($+0.0569$, $+0.0624$ vs. E: $+0.0505$, $+0.0523$); (iv) **Non-additive interactions.** ER does not necessarily amplify R degradation, with average CD even lower than R (0.1521 vs. 0.1605). Overall, all models exhibit limited robustness to non-canonical views.

Having established the cross-model robustness gap, we next examine which viewpoint factors contribute to it and how performance degrades as the corresponding angles vary.

3.3 How Do Different Viewpoint Factors Affect 3D Generation?

We next examine how azimuth, elevation, and roll individually affect 3D generation. We first conduct quantitative experiments over the three viewpoint factors, as summarized in Fig. 2. Then we use the qualitative traversals in Fig. 3 to further interpret and validate these quantitative trends.

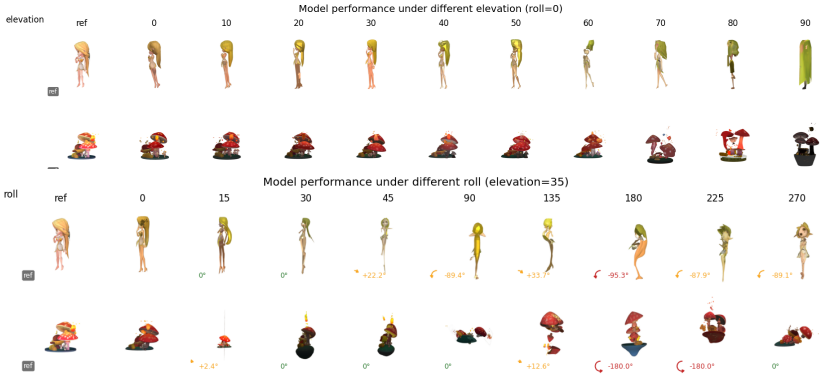


Figure 3: **Qualitative generations under varying viewpoints.** TRELLIS results with varying elevation at roll $= 0°$ (top) and varying roll at elevation $= 35°$ (bottom); Roll-traversal generations are manually rotated for visual comparison; arrows and degree labels indicate the correction amount.

Table 2: **Quantification of the elevation–roll interaction.** The difference-in-differences (DiD) contrast on log-CD; negative $I_{E\times R}$ indicates a smaller roll-induced relative CD increase at high elevation. All models exhibit negative interactions: higher elevation consistently attenuates the relative sensitivity to roll.

Model	$I_{E\times R}$	95% CI	d_z
TripoSR	-0.44	$[-0.50, -0.39]$	-0.85
SPAR3D	-0.39	$[-0.45, -0.34]$	-0.74
TRELLIS	-0.73	$[-0.85, -0.62]$	-0.69
Hunyuan3D	-0.55	$[-0.67, -0.43]$	-0.48
Mean	-0.53	—	—

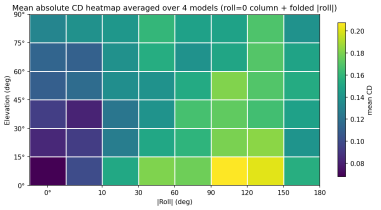


Figure 4: **Visualization of the elevation–roll interaction.** Mean CD of four models versus elevation and folded roll. Roll causes a sharper CD increase at low elevation, whereas high elevation flattens the roll response and yield lower CD than low elevation at moderate-to-large roll angles.

Effect of azimuth As shown in Fig. 2, azimuth mainly induces periodic fluctuations in performance. CD and PSNR vary in a broadly coordinated manner, yet neither exhibits a clear global trend across the full angular range.

Effect of elevation Elevation has a clear global effect on generation quality (2. Mean CD generally increases as elevation rises, while TRELLIS and Hunyuan3D remain relatively stable at lower elevations before degrading rapidly beyond a model-dependent threshold. These findings are further supported by the elevation traversal in Fig. 3. For the human, abnormal hair begins to appear around $40°$, body parts become increasingly distorted beyond $60°$, and the human shape is no longer recognizable at $90°$. For the mushroom, errors begin to emerge around $70°$, including the root detaching from the ground and incorrect relative positions. At $80°$, a nonexistent person appears in the generation, while at $90°$ the texture becomes nearly completely black.

Effect of roll The quantitative curves in Fig. 2 show CD rises sharply after departing from $0°$, followed by a more gradual increase over a broad plateau, and shows a slight recovery around $180°$. This trend is corroborated by the roll traversal in Fig. 3, where slight anomalies already appear at $15°$. Beyond $30°$, the human exhibits severe structural errors, with the body and hair becoming twisted, while the mushroom develops darker textures, missing parts, and spatial displacement. Both objects also exhibit severe spatial orientation errors, including upside-down generations.

3.4 Do elevation and roll interact?

Section 3.2 revealed a counterintuitive **non-additive interaction**: jointly introducing elevation and roll does not necessarily cause greater degradation than roll alone. To determine whether the two factors act independently or one modulates the other, we conduct a factorial experiment that evaluates elevation and roll jointly under matched object and azimuth conditions.

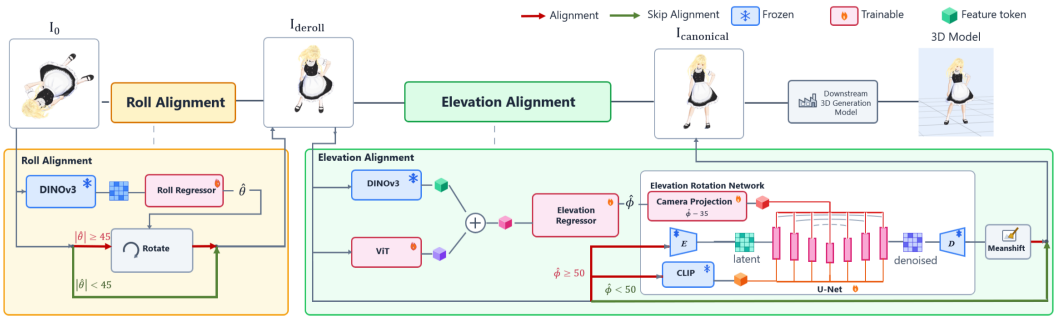


Figure 5: **Overview of our proposed method CanoView.** CanoView consists of Roll Alignment (RA) and Elevation Alignment (EA). Given an input image I_0, it is sequentially processed by RA and EA. Each module estimates the corresponding viewpoint angle and selectively applies a rotation when needed to bring the image toward the canonical range. The resulting image is then fed to downstream 3D generation models.

To quantify this interaction, we adopt the widely used difference-in-differences (DiD) framework (Bertrand et al., 2004; Angrist & Pischke, 2009). Since objects differ in CD scale, we conduct the analysis in log-CD space, where differences correspond to relative changes in CD. As the two interaction directions $E \times R$ and $R \times E$ are algebraically equivalent, their common interaction contrast is expressed as

$$I_{E\times R} = I_{R\times E} = \log \mathrm{CD}(ER) - \log \mathrm{CD}(E) - \log \mathrm{CD}(R) + \log \mathrm{CD}(C)$$

As shown in Table 2, all four models exhibit negative interaction effects (mean -0.5301, all 95% CIs excluding zero), indicating that higher elevation reduces the relative sensitivity to roll.

Figure 4 further visualizes this interaction across the continuous viewpoint space: since the roll response is highly symmetric around $180°$ (Fig. 2), we fold the roll angle by merging r and $360° - r$. The heatmap shows that CD increases rapidly as roll departs from $0°$ at low elevation, whereas the roll response is substantially flatter at high elevation; at moderate-to-large roll angles ($60°$–$150°$), high-elevation inputs can even achieve lower CD than low-elevation inputs.

These results demonstrate a systematic elevation–roll interaction: higher elevation attenuates the effect of roll, indicating that the two factors are neither independent nor simply additive.

4 Mitigating Non-Canonical Viewpoint Failures

4.1 Method Overview

As shown in Section 3, non-canonical roll and elevation can substantially degrade single-view 3D generation. We therefore propose **CanoView**, a plug-and-play view transformation framework that mitigates the performance degradation of single-view 3D models under non-canonical views. CanoView consists of two sequential components, **Roll Alignment (RA)** and **Elevation Alignment (EA)**. Given an input image I_0, RA first canonicalizes its in-plane rotation, after which Elevation Alignment further transforms its elevation toward the canonical range. The overall pipeline is illustrated in Figure 5. Training details for the angle estimators and ERN are provided in Appendix A.

4.2 Roll Alignment

To perform roll-level canonicalization, we first estimate the input image's in-plane rotation and then determine whether correction is needed.

Roll Estimator. Our roll estimator consists of a frozen DINOv3 (Siméoni et al., 2025) backbone and a trainable roll regressor. To learn the relation between visual features and in-plane rotation, we apply a randomly sampled 2D rotation to each rendered image and use the applied angle as the ground-truth roll. The rotated image is encoded by DINOv3 into an ordered 32×32 patch-token

grid. Rather than aggregating these features into a global token, the roll regressor preserves their spatial ordering to capture rotation-dependent patterns. It predicts $(\cos\hat{\theta}, \sin\hat{\theta})$, which is decoded into the signed roll angle $\hat{\theta}$ and supervised by the applied rotation. All training objects are excluded from ViewRobust-3D.

Roll Canonicalization. After estimating the roll $\hat{\theta}$, we determine whether correction is required. Specifically, when $|\hat{\theta}|$ exceeds a predefined threshold, we remove the background, rotate the foreground by $-\hat{\theta}$, and rescale it to remain within the image boundaries. When the estimated roll is below the threshold, the input is retained. We set the threshold to $45°$:

$$I_{\text{deroll}} = \text{RollAlignment}(I_0) = \begin{cases} I_0, & |\hat{\theta}| < 45°, \\ \text{Rotate}(I_0, -\hat{\theta}), & |\hat{\theta}| \geq 45°. \end{cases} \tag{1}$$

Here, $\hat{\theta}$ denotes the estimated roll angle, and $\text{Rotate}(\cdot)$ denotes 2D image rotation.

4.3 Elevation Alignment

After RA, we obtain an image with its in-plane rotation largely removed. We next adjust its elevation toward the canonical range.

Elevation Estimator. We first estimate the elevation of the derolled image I_{deroll}. Similar to the roll estimator, the elevation estimator combines a frozen DINOv3 (Siméoni et al., 2025) backbone with a trainable elevation regressor, but adopts a two-stream architecture. The first stream extracts the global CLS feature from DINOv3, while the second uses a trainable ViT (Dosovitskiy et al., 2021) to encode the input image and obtain its CLS token. The projected DINOv3 feature is added to the ViT CLS token as global semantic guidance, and the fused representation is then mapped by the elevation regressor to the elevation $\hat{\phi}$. The model is trained on multi-view images with elevation labels to learn this visual-to-elevation mapping. Training objects are likewise excluded from ViewRobust-3D.

Elevation Rotation Network (ERN). Once the elevation $\hat{\phi}$ is estimated, we further need a model that can transform the high-elevation input into the desired target view according to the required elevation change. We therefore introduce the **Elevation Rotation Network (ERN)**, a diffusion-based (Ho et al., 2020) view transformation network fine-tuned from Stable Zero123 (Stability AI, 2023). During fine-tuning, the U-Net denoising network and camera projection layer are optimized, while the condition image encoder (Radford et al., 2021) and VAE (Rombach et al., 2022) remain frozen. The fine-tuning set likewise excludes all benchmark objects.

Meanshift Postprocessing Because diffusion-based generation may introduce artifacts and high-frequency noise that degrade subsequent 3D generation, we apply MeanShift-based refinement (Comaniciu & Meer, 2002) to the generated image. Specifically, pyramid MeanShift filtering with spatial radius $s_p = 40$ and color radius $s_r = 13$ identifies locally consistent regions, after which each region is reconstructed using its mean color from the original generated image.

Elevation Canonicalization. With the estimated elevation $\hat{\phi}$, we determine whether correction is required. When $\hat{\phi}$ exceeds $50°$, ERN transforms the image toward a target elevation of $35°$; otherwise, the derolled image is retained:

$$I_{\text{canonical}} = \text{ElevationAlignment}(I_{\text{deroll}}) = \begin{cases} I_{\text{deroll}}, & \hat{\phi} < 50°, \\ \mathcal{M}\Big(\text{ERN}\left(I_{\text{deroll}}, \hat{\phi} - 35°\right)\Big), & \hat{\phi} \geq 50°. \end{cases} \tag{2}$$

Here, $\mathcal{M}$ denotes the MeanShift-based refinement, and $35°$ is the canonical target elevation.

After these two alignment steps, we obtain $I_{\text{canonical}}$, which can be directly used as input to downstream single-view 3D generation models.

Table 3: **Main results of CanoView.** CD and PSNR of direct generation and generation after view transformation on the E, R, ER subsets; CanoView improves both metrics across nearly all models and subsets, with larger gains on the R and ER subsets.

Subset	Model	Direct baseline		CanoView (ours)		Change	
		CD↓	PSNR↑	CD↓	PSNR↑	CD	PSNR
E	TRELLIS	0.0932	17.87	0.0890	19.08	-4.6%	+6.8%
	SPAR3D	0.1552	19.26	0.1434	19.68	-7.6%	+2.2%
	TripoSR	0.1422	18.47	0.1305	19.13	-8.2%	+3.6%
	Hunyuan3D	0.0815	17.27	0.0888	17.94	+9.0%	+3.9%
R	TRELLIS	0.1651	16.99	0.1225	18.76	-25.8%	+10.4%
	SPAR3D	0.1653	19.36	0.1449	19.84	-12.3%	+2.5%
	TripoSR	0.1486	18.44	0.1247	19.42	-16.1%	+5.3%
	Hunyuan3D	0.1629	16.23	0.1269	18.21	-22.1%	+12.2%
ER	TRELLIS	0.1477	17.32	0.0977	18.76	-33.9%	+8.3%
	SPAR3D	0.1724	19.19	0.1483	19.63	-14.0%	+2.3%
	TripoSR	0.1512	18.26	0.1327	19.11	-12.2%	+4.7%
	Hunyuan3D	0.1369	16.58	0.0972	17.73	-29.0%	+6.9%

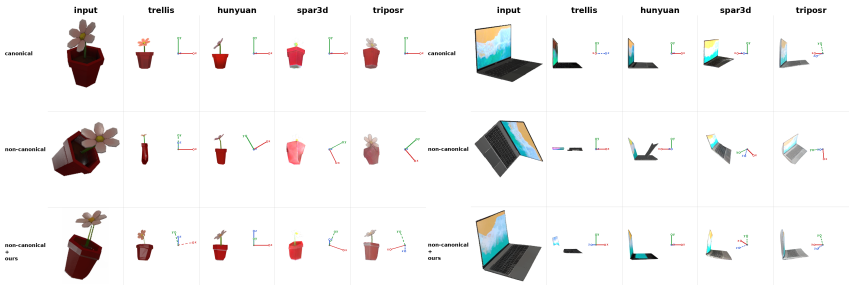


Figure 6: **Qualitative comparison on representative non-canonical cases.** Non-canonical inputs cause geometric distortion, incorrect spatial structures, and planarized collapse, while CanoView yields more complete and geometrically consistent generations. The unrotated version and more cases can be found in the appendix

4.4 Results and Ablation

Main Results. We evaluate the proposed framework on the controlled viewpoint subsets introduced in Section 3, comparing direct generation from the original input with generation after applying CanoView on the R, E, and ER subsets.

From Table 3 we obtain the following conclusions: ***i)*** The largest improvement is observed on the **ER subset**, demonstrating the benefit of our method under the most complex setting where roll and elevation deviations occur jointly. ***ii)*** A considerable improvement is also observed on the **R subset**, confirming that our method remains effective for non-canonical views caused solely by roll deviations. ***iii)*** The **E subset** also shows consistent gains across most models, with only Hunyuan exhibiting a slight CD degradation, possibly due to its limited robustness to the appearance changes introduced by the input transformation.

These quantitative improvements are further reflected in the qualitative comparisons in Figure 6, which demonstrate three main advantages of our method: **(i) Canonical and faithful processed inputs.** CanoView removes severe elevation and roll deviations while preserving the original visual information, producing canonical and faithful processed inputs. **(ii) Improved geometric accuracy.** It corrects geometric distortions, such as the elongated vase from TRELLIS and the wrinkled, inward-collapsed vase from SPAR3D, bringing them closer to the reference geometry. **(iii) Fewer structural failures.** It prevents severe generation failures, such as the erroneous middle panel in Hunyuan and the collapsed structures from SPAR3D and TripoSR, restoring them to geometries broadly consistent with the reference.

Table 4: **Ablation study on the Joint-Noncanonical TRELLIS subset.** Changes in parentheses are relative to Ours.

	CD↓	PSNR↑
Ours	0.0977	18.76
w/o roll and w/ elev	0.1255 (+28.5%)	18.24 (-2.8%)
w/ roll and w/ (elev w/o mean)	0.1016 (+4.0%)	18.55 (-1.1%)
w/ roll and w/o elev	0.1306 (+33.7%)	18.71 (-0.3%)
w/o roll and w/o elev	0.1477 (+51.2%)	17.32 (-7.7%)

Ablation Study. We further conduct an ablation study on the Joint-Noncanonical TRELLIS subset. Starting from the CanoView result reported in Table 3, we evaluate variants that remove RA, remove EA, remove both, or remove the MeanShift operation applied after EA.

Table 4 indicates that the three components play complementary roles: RA and EA perform the main viewpoint correction, while MeanShift further reduces residual transformation errors. Specifically, **(i) w/ roll and w/ (elev w/o mean).** moderately degrades both CD and PSNR, showing the role of MeanShift in reducing residual transformation errors after alignment. **(ii) w/ roll and w/o elev.** substantially degrades CD with only a minor PSNR change, indicating the primary contribution of Elevation Alignment to reconstruction accuracy. **(iii) w/o roll.** also degrades both metrics, confirming the importance of roll correction for restoring a canonical input orientation. The **w/o roll and w/o elev** variant produces the largest degradation, showing the combined importance of both alignment steps.

5 Insights and Discussions

We summarize the findings from our controlled benchmark (Section 3) and mitigation experiments (Section 4) as four insights on viewpoint robustness, evaluation, and model design.

The Canonical-View Mirage: canonical views overestimate 3D capability. Average CD nearly doubles from canonical to Roll-Noncanonical views (0.0776 → 0.1605), with all four models degrading (Table 1), highlighting the need to include non-canonical views in evaluation and training.

Viewpoint responses are anisotropic, nonlinear, and non-independent. Azimuth mainly causes periodic fluctuations, while non-canonical roll increases CD more than non-canonical elevation (+0.0829 vs. +0.0405) (Fig. 2). The increase is also non-monotonic with roll, and adding non-canonical elevation to roll can partially recover performance (Fig. 4), indicating an interaction between the two factors.

Architecture-dependent sensitivity: no universally robust model. According to Table 1,TRELLIS and Hunyuan3D show greater elevation robustness, while SPAR3D and TripoSR are more robust to roll. This suggests architectural differences may lead to distinct robustness patterns and a potential trade-off between spatial and texture consistency.

Input-space correction is feasible and effective. Input canonicalization without retraining reduces TRELLIS CD on the Joint-Noncanonical subset by 33.9% (0.1477 → 0.0977), nearly reaching its Elevation-Noncanonical level (0.0932) (Table 3), showing that input-side correction is a high-leverage mitigation.

6 Conclusion

This paper reveals that current image-to-3D models exhibit systematic performance degradation under non-canonical viewpoints, driven by elevation and roll variations that interact rather than act independently. By proposing CanoView, a plug-and-play view transformation module, we demonstrate that these robustness gaps can be effectively mitigated without retraining downstream models.

AI DISCLOSURE

In this work, we used generative AI tools for implement methods, assist with translation, interpret results. We have not used generative AI tools for propose or refine hypotheses,support qualitative, and thematic data analysis and generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for suggest experimental parameters, create or edit software code, draft parts of a research paper, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature. We have reviewed all AI-assisted work. AI-generated text, analyses, code, and other outputs were carefully reviewed and verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004. doi: 10.1162/003355304772839588.

Paul J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Xiao Cai, Sitong Su, Jingkuan Song, Pengpeng Zeng, Ji Zhang, Qinhong Du, Mengqi Li, Heng Tao Shen, and Lianli Gao. GT23D-Bench: A comprehensive general text-to-3D generation benchmark, 2025. URL https://arxiv.org/abs/2412.09997.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1554. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/70364304877b5e767de4e9a2a511be0c-Paper-Datasets_and_Benchmarks.pdf.

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)*, June 2023b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *International Conference on Learning Representations (ICLR)*, 2024.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M. Rehg, and Varun Jampani. SPAR3D: Stable point-aware reconstruction of 3D objects from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hunyuan3D Team, 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.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Escher-Net: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiang Li, Zirui Wang, Zixuan Huang, and James M. Rehg. Cue3D: Quantifying the role of image cues in single-image 3D generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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 (ICLR)*, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Stability AI. Stable Zero123. `https://huggingface.co/stabilityai/stable-zero123`, 2023.

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

Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, et al. GPT-4V(ision) is a human-aligned evaluator for text-to-3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Xianghui Xie, Chuhang Zou, Meher Gitika Karumuri, Jan Eric Lenssen, and Gerard Pons-Moll. MVGBench: A comprehensive benchmark for multi-view generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

Xianghui Yang et al. Hunyuan3D 1.0: A unified framework for text-to-3D and image-to-3D generation. *arXiv preprint arXiv:2411.02293*, 2024.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, et al. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *arXiv preprint arXiv:2406.13897*, 2024.

Zibo Zhao et al. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

A Training Details

A.1 Angle Estimators

The roll and elevation estimators were trained on multi-view rendered objects using a training split of 205,704 objects and a validation split of 2,000 objects. During training, one view was sampled per object and resized to 512×512 pixels; random horizontal flipping was applied as data augmentation. The DINOv3 backbone was kept frozen, while the task-specific regression modules were optimized. For elevation prediction, the target angle was clamped to the valid range and normalized by 90° before computing a mean-squared-error (MSE) regression loss. The roll estimator used the corresponding sine and cosine representation to avoid the discontinuity at the angle boundary. We optimized the trainable parameters with AdamW using a learning rate of 1×10^{-5} and weight decay 1×10^{-2}. Training was configured with mixed bfloat16 precision and gradient clipping at 1.0. Objects from ViewRobust-3D were excluded from the estimator training data.

A.2 Elevation Rotation Network

ERN was fine-tuned from Stable Zero123 on paired reference and target views constructed for a target elevation of 35°, with reference views filtered to elevations of at least 40°. Images were rendered and processed at 256×256 resolution. Fine-tuning used four GPUs with a batch size of 8 per GPU, corresponding to an effective batch size of 32, for up to 100,000 optimization steps. The U-Net denoiser and camera projection layer were updated, while the CLIP image encoder and VAE remained frozen. We used AdamW with learning rate 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1×10^{-2}, together with a cosine learning-rate schedule. The training objective was the standard epsilon-parameterized noise-prediction MSE; the variational lower-bound term was disabled.

B Original Unrotated Results

Figure 7 shows the original, unrotated model outputs corresponding to the rotation-aligned visualization in Figure 1. These results preserve the coordinate-frame orientation produced by each model.



Figure 7: **Original unrotated model results of Teaser Image.** The outputs are shown in their original coordinate-frame orientations, before the rotation alignment used in the main-paper visualization.

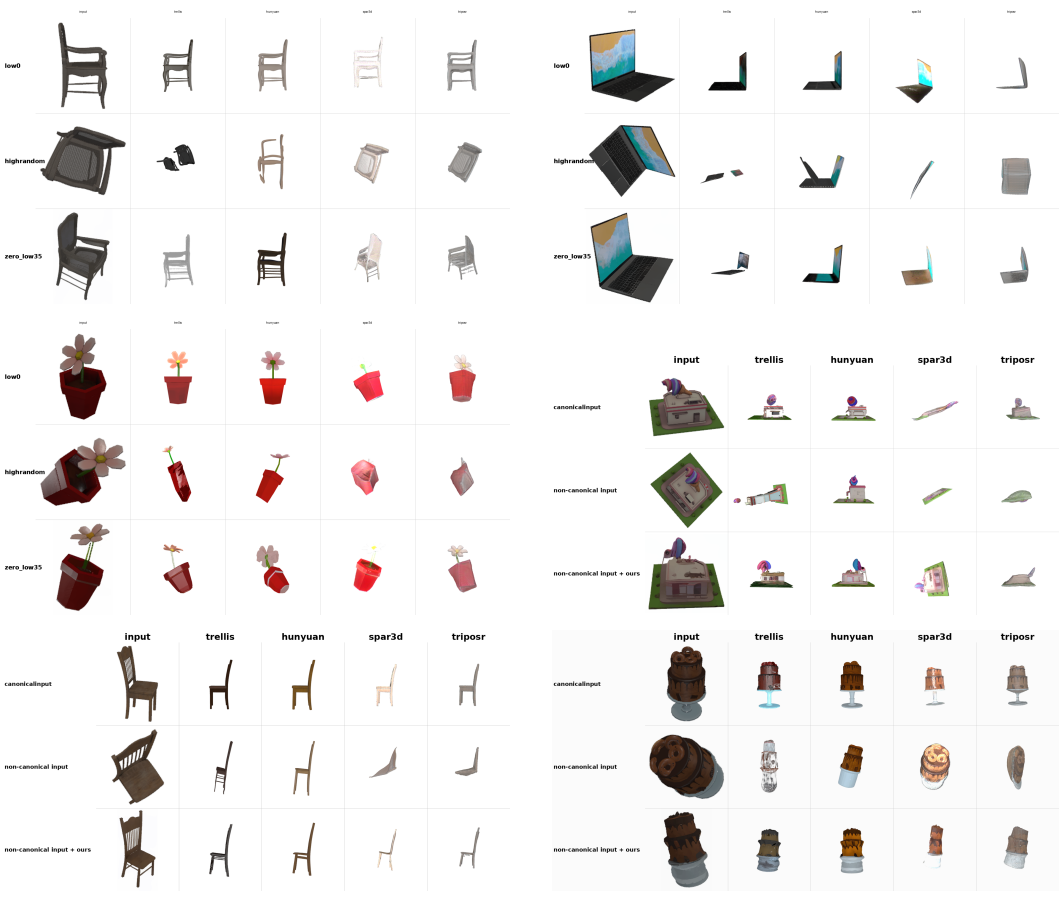


Figure 8: Unrotated results and additional qualitative cases. Each row contains two grid visualizations.