Twist and Compute: The Cost of Pose in 3D Generative Diffusion

Kyle Fogarty
University of Cambridge
ktf25@cam.ac.uk

Jack Foster
University of Cambridge
jwf40@cam.ac.uk

Jing Yang
University of Cambridge
jy496@cam.ac.uk

Cengiz Öztireli
University of Cambridge
aco41@cam.ac.uk

Abstract

Despite their impressive results, large-scale image-to-3D generative models remain opaque in their inductive biases. We identify a significant limitation in image-conditioned 3D generative models: a strong canonical view bias. Through controlled experiments using simple 2D rotations, we show that the state-of-the-art Hunyuan3D 2.0 model can struggle to generalize across viewpoints, with performance degrading under rotated inputs. We show that this failure can be mitigated by a lightweight CNN that detects and corrects input orientation, restoring model performance without modifying the generative backbone. Our findings raise an important open question: Is scale enough, or should we pursue modular, symmetry-aware designs?

1 Introduction

The expressivity of generative models, particularly diffusion architectures, has led to unprecedented success in 3D content creation from single images. These models promise to learn the complex distribution from a mixture of curated 3D training data and image foundation models. However, this success raises a foundational question: What inductive biases are actually learned during training, and how do they impact generalization? In the physical world, objects retain their identity despite changes in position and orientation. However, when these objects are projected into an observation space, such as an image, transformations result in structured variations in appearance. A model that genuinely understands 3D structure should therefore maintain object identity across these transformations, exhibiting equivariance between the 3D world space and its representation in the observation space.
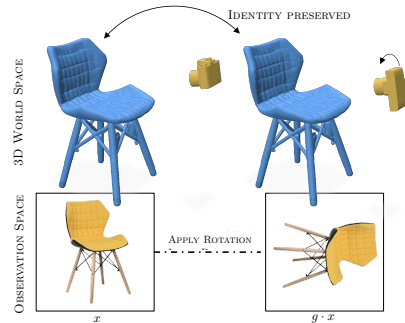

Figure 1: Equivariance between world and observation spaces preserves identity.

Established works argue that symmetry should be encoded rather than merely discovered: group-equivariant CNNs enforce rotation/reflection structure in 2D [2], and SE(3)-equivariant networks extend this to 3D representations [11, 3]. In contrast, popular image-to-3D pipelines often inherit dataset view biases (e.g., canonical front views), which can encourage shortcut solutions over true geometry [1, 6]. Recent efforts try to counter this bias through two main strategies: fine-tuning models on orientation-aligned data to directly produce canonical outputs [8], or normalizing the input
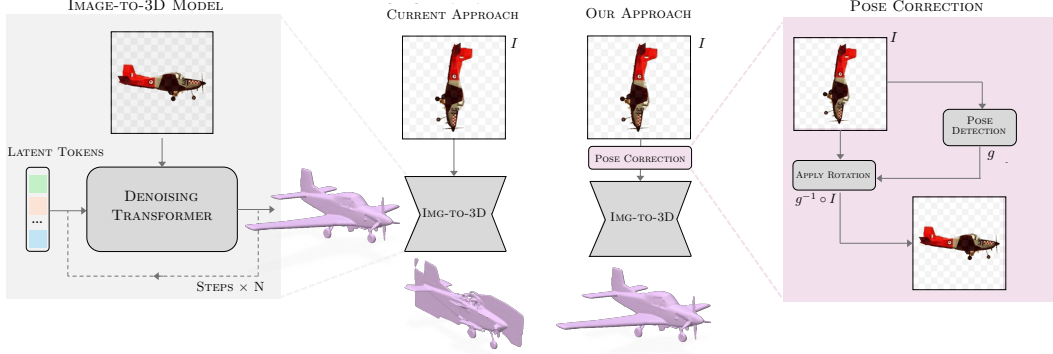
Figure 2: Rotating the input image breaks the image-to-3D pipeline (canonical-view bias), while a lightweight CNN predicts the rotation and applies its inverse to re-canonicalize the image before 3D generation, restoring performance. Note that the degraded 3D mesh is rotated to the canonical frame for better visualization.

pose before reconstruction. The viability of the latter is supported by foundational work showing that simple networks can effectively predict 2D image rotation [4].

We position our study within this landscape, using in-plane rotations of the input image as a controlled probe of whether the Hunyuan3D 2.0 image-to-3D pipeline [15], a current state-of-the-art model, exhibits genuine equivariance or instead relies upon on canonical views. A model that internalizes 3D structure should be equivariant to such transformations, maintaining object identity as the observation rotates. However, the prevalence of canonical orientations in large-scale image corpora may incentivize shortcut learning, yielding a preference for canonical views rather than robust 3D geometry. We refer to this as *canonical-view bias*, which limits generalization to arbitrarily oriented inputs.

Contributions This work investigates the impact of viewpoint biases in image-to-3D generation. (i) We empirically identify a strong canonical view bias in the Hunyuan3D generative model, particularly for object categories with well-defined orientations, such as airplanes, chairs, and cars. (ii) We demonstrate that this bias can significantly degrade 3D generation quality when input images are rotated away from their canonical poses. (iii) We show that a lightweight CNN-based pre-processing module, trained to detect and correct image orientation, can effectively restore generation quality, without requiring any changes to the generative model itself.

2 Methodology

2.1 Image-to-3D Generative Model

Our investigation focuses on Hunyuan3D [15], a state-of-the-art flow-matching architecture for single-image 3D generation. The model adopts a decoupled pipeline, separating 3D representation learning from the generative process. To obtain the 3D representation, high-level semantic features are extracted from the input image using a frozen DINOv2-based encoder [9]. These features are projected into a latent space as vector sets [14], which serve as implicit representations of complex 3D shapes. The generative component is a flow-matching diffusion transformer trained in this latent space to predict object token sequences from the input image. These token sequences are decoded into Signed Distance Functions (SDFs), which are subsequently converted into triangle meshes via iso-surfacing [7].

This architecture implies that the model's synthesis quality is bottlenecked by the feature fidelity of its 2D encoder. If the DINOv2 encoder exhibits a bias toward canonical object views, it may produce distorted feature representations when processing rotated inputs.

2.2 Probing for Canonical View Bias

Curated Dataset. To investigate the model's sensitivity to input orientation, we constructed a targeted evaluation dataset comprising three object categories with well-established canonical poses: airplanes, chairs, and cars. Images were sourced from publicly available repositories and manually curated to ensure the presence of a single, dominant object in each image, minimal occlusion, and sufficient diversity across object instances.

Rotational Transformations. To systematically evaluate robustness to viewpoint variation, we apply a set of in-plane 2D rotations to each image in the dataset. Specifically, each source image I is rotated by angles $\theta \in \{0°, 90°, 180°, 270°\}$, producing a set of transformed inputs $\{I'_0, I'_{90}, I'_{180}, I'_{270}\}$. Here, $0°$ denotes the original, canonical orientation.

Evaluation Metric: Cross-Modal Similarity (ULIP). To quantitatively assess the semantic fidelity of the generated 3D shapes, we employ the ULIP (Unified Language-Image Pre-training) score [13], which enables direct comparison between modalities by embedding 2D images and 3D shapes into a shared semantic space. In our setup, each generated mesh is converted into a point cloud of 8,192 points, and 3D features are extracted using a pretrained Point-BERT variant integrated within the ULIP framework. The ULIP score is computed as the cosine similarity between the image and point cloud embeddings, with higher scores indicating stronger semantic alignment between the input image and the generated 3D shape.

Evaluation Procedure. For each rotated image I'_θ, we use the Hunyuan3D model to generate a corresponding 3D shape M'_θ, and compute the associated ULIP score:

$$\mathcal{S}_\theta = \text{ULIP}(I'_\theta, M'_\theta).$$

By comparing scores across rotation angles, we assess the model's robustness to input orientation and identify any performance degradation when the input deviates from the canonical view.

2.3 A Lightweight CNN-based Orientation Corrector

To mitigate orientation bias, we introduce a lightweight pre-processing module: a compact orientation classifier based on EfficientNetV2, equipped with a 4-way softmax output. We provide more information on the orientator in Appendix A. Given an input image I, the classifier produces logits $\mathbf{z}(I) \in \mathbb{R}^4$, corresponding to discrete rotation angles $\{0°, 90°, 180°, 270°\}$. The predicted rotation $r^\star$ is obtained by

$$r^\star = \arg\max_k z_k(I),$$

after which the image is rotated back to a canonical orientation. We employ publicly available pretrained weights for the classifier without any additional fine-tuning. With only 20.3M parameters, negligible compared to the 2.8B parameters of the generative model, the classifier achieves 96.2% accuracy in our evaluations.

3 Results and Analysis

3.1 Canonical-View Bias in Hunyuan3D 2.0

We observe a consistent decline in ULIP similarity scores as the input image is rotated away from the canonical $0°$ orientation across all evaluated categories: *airplanes*, *chairs*, and *cars*. In each case, the $0°$ view yields the highest similarity, while rotated views at $90°$, $180°$, and $270°$ result in significantly lower scores, revealing a clear sensitivity to input orientation rather than robust 3D understanding (Fig. 3).

This trend is further supported by qualitative analysis. Inputs at non-canonical angles often lead to systematic geometric failures; for example, collapsed or sheared airplane wings, and misaligned or duplicated chair legs, whereas outputs from canonical views remain structurally coherent and faithful to the object's shape (Fig. 4).

Effect of Inference Steps. We evaluated whether increasing the number of diffusion inference steps mitigates the orientation bias. Across categories and view angles, we observed no consistent trend: additional steps sometimes yielded small gains for non-canonical inputs, but the improvements were unstable and insufficient to close the gap to the $0°$ view. This suggests the bias is embedded in the learned representation rather than an artifact of under-converged sampling.
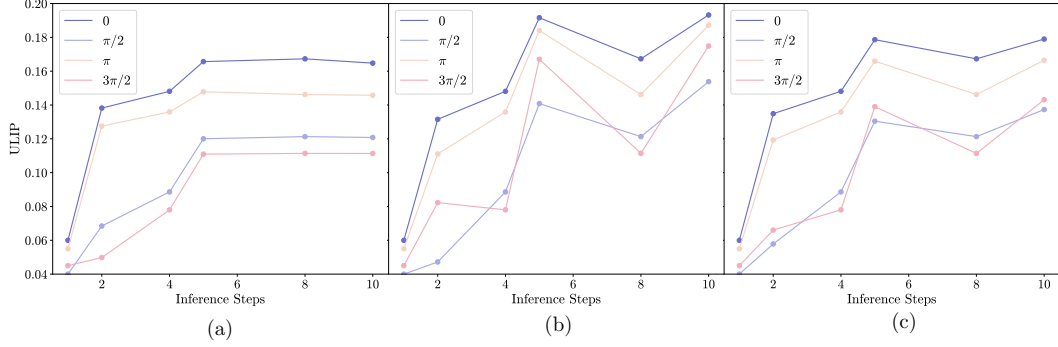
Figure 3: ULIP similarity (higher is better) versus diffusion inference steps for Hunyuan3D 2.0 under four in-plane input rotations $\{0°, 90°, 180°, 270°\}$, shown separately for (a) airplanes, (b) chairs, and (c) cars. Across all categories, the canonical $0°$ view consistently achieves the highest ULIP, while non-canonical rotations suffer substantial drops.
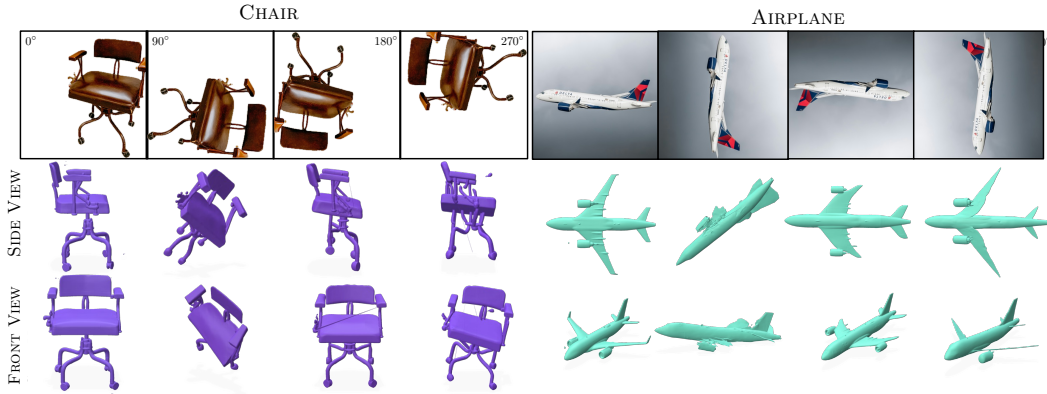

Figure 4: Qualitative effect of input rotation on Hunyuan3D 2.0. For each object, the input image is rotated by $\{0°, 90°, 180°, 270°\}$. To ensure comparability, all generated meshes are reoriented to a common camera before being rendered from canonical front and side views. Non-canonical inputs induce systematic geometric failures (e.g., collapsed airplane wings, misaligned/duplicated chair legs), whereas the $0°$ view remains stable, illustrating a strong canonical-view bias.

3.2 Lightweight Orientation Corrector

To address the model's sensitivity to input orientation, we introduce a compact CNN-based orientation corrector that re-canonicalizes input images prior to 3D generation. Importantly, this correction is applied as a pre-processing step and does not require any modification to the generative backbone. After correction, ULIP scores closely align with those obtained from images originally presented at the canonical $0°$ view (Table 1), demonstrating the effectiveness of this lightweight intervention.

Table 1: ULIP scores (mean $\pm$ std) by category and input condition. Our lightweight CNN corrector restores performance close to the canonical view across all categories.

Input Image Condition	Airplanes	Chairs	Cars
Canonical View ($0°$)	0.166 ± 0.047	0.184 ± 0.07	0.176 ± 0.07
Rotated View (avg. non-canonical)	0.126 ± 0.048	0.155 ± 0.06	0.132 ± 0.06
Rotated View + CNN Corrector	0.166 ± 0.047	0.181 ± 0.06	0.170 ± 0.07

4 Discussion and Conclusion

Discussion. Our findings reveal a clear canonical-view bias in a state-of-the-art image-to-3D pipeline. Specifically, rotating input images by $90°$, $180°$, or $270°$ consistently leads to lower ULIP similarity scores across object categories such as airplanes, chairs, and cars, with the $0°$ (canonical) view performing best in all cases (Fig. 3). Qualitative results further illustrate structured geometric artifacts in outputs from non-canonical views, for example, collapsed or sheared wings and duplicated chair legs, while canonical inputs yield stable and coherent reconstructions (Fig. 4).

Notably, increasing the number of diffusion inference steps fails to mitigate this discrepancy, suggesting that the degradation stems from limitations in the learned representation itself rather than from insufficient sampling or premature convergence. To address this, we demonstrate that a lightweight auxiliary network, trained to predict input rotation, can effectively recover performance to near-canonical levels, without any modification to the underlying generative model (Table 1).

Limitations and Future Work. We identify three primary limitations of our current study, each of which motivates future research directions:

1. **Discrete Rotation Group** (C_4)**:** Our analysis is limited to $90°$ in-plane rotations, corresponding to the discrete cyclic group C_4. Extending this evaluation to continuous in-plane rotations $(SO(2))$ remains an open direction. Additionally, it would be valuable to explore whether orientation-invariant or equivariant architectures, such as steerable CNNs that produce continuously equivariant features, can mitigate this bias more effectively than discrete classification approaches.

2. **Model Scope:** Our experiments focus exclusively on Hunyuan3D 2.0 [15]. However, other image-to-3D models such as TripoSR [12], OpenLRM [5], and others, are also promising candidates for assessing rotation-induced degradation and for testing the generality of our correction strategy.

3. **Category Coverage:** This study focuses on object categories with well-defined canonical views, such as airplanes, chairs, and cars. An open question is whether similar canonical-view biases persist in more diverse or less geometrically constrained categories.

Conclusion. These findings suggest that scaling alone is unlikely to resolve the observed failure modes. We demonstrate that a lightweight CNN, trained to detect and correct input orientation, can effectively mitigate these issues and restore model performance, all without modifying the generative backbone. While our study focuses on a constrained problem setting, the results highlight the importance of encoding natural symmetries into the representation pipeline, raising interesting questions for future research.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, Jun 2022.

[2] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, NY, USA, Jun 2016. PMLR.

[3] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023.

[6] 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)*, pages 9298–9309, Oct 2023.

[7] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

[8] Yichong Lu, Yuzhuo Tian, Zijin Jiang, Yikun Zhao, Yuanbo Yang, Hao Ouyang, Haoji Hu, Huimin Yu, Yujun Shen, and Yiyi Liao. Orientation matters: Making 3d generative models orientation-aligned. *arXiv preprint arXiv:2506.08640*, 2025.

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[10] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[11] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. In *Proceedings of the NeurIPS 2018 Workshop on Machine Learning for Molecules and Materials*, 2018.

[12] 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.

[13] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023.

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

[15] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.