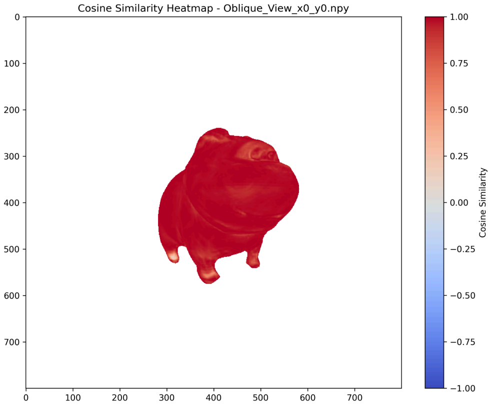
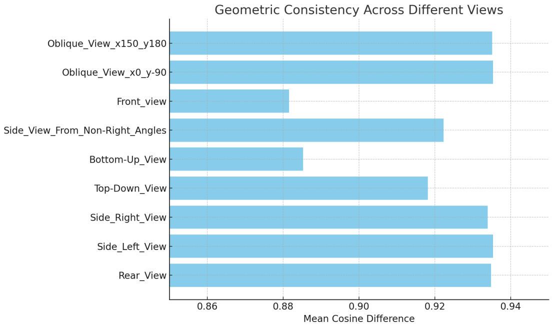
Evaluating the Geometric Consistency of Text-to-3D generated models using Surface Normal Analysis

The field of text-to-3D generative methods has seen remarkable progress in recent times, driven by a series of breakthroughs. Despite this progress, the existing evaluation metrics often focus on a single criterion, such as the alignment between the input text and the generated 3D models[1], but they do not comprehensively evaluate the quality of the generated 3D model itself. Traditional methods for evaluating 3D models typically measure the distance between generated and reference shape distributions. However, these methods are not readily applicable to text-conditioned generative tasks due to the difficulty in obtaining a comprehensive reference set, given the vast range of natural language inputs[1]. In this work, we propose a novel approach to evaluate the geometric consistency of generated 3D models using surface normal analysis. Surface normals provide crucial information about the geometry of a surface, describing aspects such as surface orientation, curvature, and shape.

To evaluate the geometric consistency of the generated 3D models, we use surface normal analysis as a key metric. First, we generate 3D models from text inputs using a state-of-the-art text-to-3D model, represented as triangular meshes. Surface normals are then computed directly from the mesh geometry, serving as ground truth for comparison. To analyze the models from different perspectives, we capture 2D images from both canonical (e.g., front, side) and non-canonical (e.g., oblique, tilted) viewing angles. For surface normal prediction, we utilize DSINE[2], a robust model designed to predict surface normals from images under complex lighting and geometric conditions. The predicted normals from DSINE are compared with the mesh-derived normals using cosine difference as the primary metric, which measures the angular discrepancy between the two sets of normals. To ensure that only valid regions of the model are evaluated, a masking procedure is applied to exclude irrelevant pixels from the background. This approach allows us to assess the geometric fidelity of the 3D models across multiple views and varying levels of complexity, providing insight into the performance of text-to-3D generative models.



Our results on evaluating 20 3D models generated by 5 generative models, including the most recent work Prolificdreamer [3], show that canonical views (Rear and Side_Left), demonstrated high geometric consistency, with the highest mean cosine difference reaching 0.93, indicating strong alignment between the predicted and ground truth surface normals. In contrast, non-canonical views (Side View from Non-Right Angles, Bottom-Up, and Top-Down), showed comparatively lower consistency, with the lowest mean cosine difference being 0.88. Although these non-canonical views also displayed relatively good consistency, these findings emphasize the importance of focusing on non-canonical views to enhance the overall geometric fidelity of text-to-3D generative models.



[1] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation, 2024.
[2] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
[3] 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. In Advances in Neural Information Processing Systems (NeurIPS), 2023.39