Benchmarking Multi-View Image Generation for 3D Consistency Without Ground Truth

Julia Schneider Philipp Bauer
University of Tübingen, Tübingen AI Center
{julia5.schneider, philipp.bauer}@student.uni-tuebingen.de

Abstract

Evaluating the 3D consistency of multi-view image generation systems remains a significant challenge, especially without access to ground truth 3D data. Traditional metrics often fall short in generative settings where multiple plausible outputs can exist. In this project, we propose a novel benchmarking methodology that separately assesses geometric and texture consistency across synthesized views, without relying on ground truth. Building upon recent advances such as MEt3R [1] and leveraging self-supervised features (e.g., DINO [8]), our pipeline employs feature-based comparisons and view-alignment techniques to robustly quantify multi-view coherence in both geometry and appearance. We validate the method across several generative models, demonstrating its effectiveness in identifying perceptual and structural inconsistencies. This approach offers a scalable, interpretable alternative for evaluating 3D-aware image generation and paves the way for standardized benchmarking in this field.

1. Introduction

Multi-view image generation systems aim to synthesize novel views of a scene from limited inputs, enabling applications in virtual reality, 3D reconstruction, and content creation. A critical requirement for these systems is maintaining 3D consistency, ensuring that generated views correspond to a coherent underlying scene geometry and exhibit consistent appearance across viewpoints. However, evaluating 3D consistency remains a fundamental challenge, particularly in generative settings where ground truth 3D data or exact reference images are unavailable. Traditional evaluation metrics, such as pixel-wise errors or single-view perceptual scores, often fail to capture the nuanced multi-view coherence essential for realistic 3D-aware synthesis. Moreover, existing benchmarks typically conflate different sources of inconsistency, lacking the ability to disentangle geometric errors from texture or appearance variations.

Our work builds upon the recent MEt3R pipeline [1], which leverages self-supervised features (e.g., DINO [8]) to evaluate multi-view consistency without ground truth supervision. To evaluate the effectiveness of using only DINO features for multi-view comparison, we constructed a dedicated small dataset designed to isolate geometric and texture variations. Our experiments confirm that relying solely on DINO features primarily captures geometric consistency while being less sensitive to texture discrepancies. Motivated by this insight, we extend the benchmarking pipeline by incorporating an additional texture-based comparison, enabling a separate and interpretable evaluation of both geometry and texture consistency across synthesized views.

We validate our approach on several state-of-the-art generative models, demonstrating its ability to identify both structural and perceptual inconsistencies that existing metrics often overlook. By distinguishing geometry from texture coherence, our method provides a comprehensive and scalable framework for benchmarking 3D-aware image generation, contributing toward standardized evaluation protocols in this emerging field.

2. Related Work

Limitations of Geometry-Based Novel View Synthesis: Traditional approaches to novel view synthesis (NVS) typically reconstruct new perspectives by leveraging multiple images of the same scene, captured from densely sampled viewpoints. These methods often rely on explicit 3D geometry estimation, such as multi-view stereo reconstruction or structure-from-motion pipelines, to model the underlying scene structure. Once a consistent geometric representation is recovered, novel views are generated through reprojection and image-based rendering techniques, enabling accurate interpolation between observed viewpoints (e.g., [4, 6, 7, 9]). However, the effectiveness of such methods is

fundamentally constrained by the availability and coverage of the input views. When observations are sparse or limited to a narrow baseline, the reconstructed geometry becomes incomplete or unreliable, leading to noticeable artifacts in the synthesized images.

Neural Approaches to Novel View Generation from Sparse Observations: In contrast to traditional geometry-based pipelines, recent advances in generative neural networks have enabled novel view synthesis from sparse or even single input images by learning powerful priors over natural scenes. Methods such as MV-LDM [1], PhotoNVS [13] and GenWarp [10] can hallucinate plausible novel views that resemble the input observations without relying on explicit dense geometry reconstruction. While these approaches often produce visually coherent results and maintain a degree of appearance consistency across viewpoints—for example, enforcing approximate epipolar constraints, there is no guarantee that the synthesized views adhere to accurate 3D structure. Consequently, the generated images may contain inconsistencies or artifacts when evaluated under strict geometric criteria. Despite these limitations, neural rendering models have significantly expanded the applicability of novel view synthesis to scenarios with limited or narrow-baseline observations, where classical techniques typically fail.

3D Consistent Image Generation Evaluation: In the context of multi-view image generation, ground-truth novel views are inherently unavailable, since the model synthesizes previously unobserved content. This fundamental limitation necessitates alternative strategies for assessing multi-view consistency. Recent approaches address this by reconstructing an explicit 3D representation of the generated outputs, typically leveraging NeRF-based [12] or Gaussian Splatting pipelines. These reconstructions enable rendering from the original input camera perspectives, facilitating quantitative comparison between the reprojected images and the known input views to assess consistency and coherence.

For example, Feat2GS [5] introduces a framework that infers 3D Gaussian attributes—capturing both geometry and texture—directly from features extracted by visual foundation models (VFMs) applied to unposed images. This approach enables probing a model's 3D awareness through novel view synthesis without requiring any ground-truth 3D supervision, and further disentangles geometry and appearance consistency for more granular evaluation. While such methods yield valuable insights into a model's capacity to generate structurally coherent scenes, they rely on multiple synthesized views to optimize a faithful radiance field or Gaussian representation. Additionally, reconstructing NeRFs or 3D Gaussian splats remains computationally intensive, which can limit their scalability in large-scale benchmarks.

To mitigate these constraints, we adopt MEt3R [1], which leverages self-supervised DINO features to efficiently quantify multi-view consistency without requiring explicit 3D reconstruction. In our experiments, MEt3R provides a substantially faster evaluation pipeline and supports assessment even in scenarios with only a single generated view. However, we observed that DINO-based metrics are predominantly sensitive to geometric misalignments, exhibiting limited responsiveness to fine-grained texture or color inconsistencies across views.

Intrinsic Image Decomposition: Intrinsic image decomposition is a fundamental problem in computer vision that involves separating an image into its reflectance (albedo) and shading components. This decomposition enables the recovery of surface properties independent of illumination effects, which is crucial for applications such as material recognition, relighting, and texture analysis. Recent approaches have made significant progress in addressing the challenges of intrinsic decomposition in complex, real-world environments [3, 2]. By disentangling reflectance from shading, intrinsic image decomposition provides a robust foundation for consistent analysis of image content under varying lighting conditions. The specific techniques employed in this work are described in detail in the methodology section.

3. Methodology

Our methodology is grounded in the hypothesis that self-supervised DINO features predominantly encode geometric and structural information, while exhibiting limited sensitivity to variations in texture and color. To empirically test this assumption, we designed a controlled experiment to quantify the effect of global hue shifts on the resulting DINO feature representations.

We construct a small dataset of natural images denoted by

$$\mathcal{I}_{\text{orig}} = \{I_{\text{orig},1}, \ldots, I_{\text{orig},N}\}. \tag{1}$$

For each image $I_{\text{orig},i}$, we applied a uniformly sampled global hue shift, resulting in the transformed image $I_{\text{shift},i}$. Formally, this hue shift mapping is defined as

$$\mathcal{H}_\delta(I)(x,y) = \text{HSV}^{-1}\Big[\big(H(x,y) + \delta \bmod 1\big), \\ S(x,y),\, V(x,y)\Big], \tag{2}$$

where $\delta \sim \mathcal{U}(0,1)$, HSV^{-1} denotes conversion back to RGB, and H, S, V represent the hue, saturation, and value channels respectively. This transformation yields the set

$$\mathcal{I}_{\text{shift}} = \{I_{\text{shift},1}, \ldots, I_{\text{shift},N}\}. \tag{3}$$

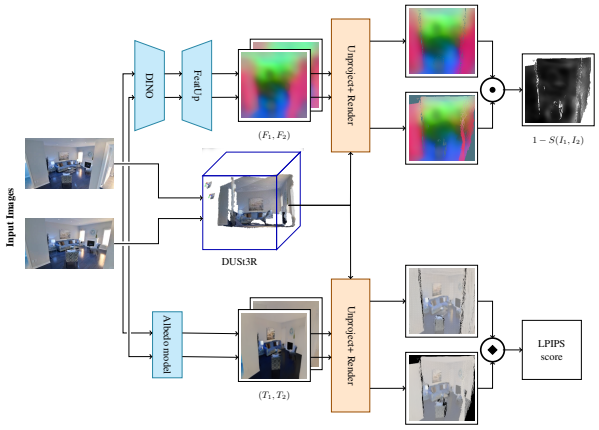

Figure 1. Overview of the dual-stream evaluation pipeline. Input images are processed in parallel by a geometry branch (DINO features) to assess semantic consistency, and a texture branch (albedo model) for texture-sensitive comparison. Both branches reproject features onto 3D geometry to compute separate error maps for geometric and textural discrepancies.

This procedure ensures that the geometric structure of each image remains unchanged, while the overall appearance and color distribution are systematically altered.

We compute the DINO feature representations for each original and hue-shifted image using the pretrained DINO backbone:

$$F_{\text{orig},i} = \text{DINO}\big(I_{\text{orig},i}\big), \quad F_{\text{shift},i} = \text{DINO}\big(I_{\text{shift},i}\big).$$

To quantify the similarity between feature maps, we measure the cosine similarity and L2 distance between corresponding pairs $\big(F_{\text{orig},i}, F_{\text{shift},i}\big)$. Empirically, we observe consistently high similarity across all samples, indicating that the learned representations are largely invariant to hue-based appearance changes. This provides evidence that DINO features emphasize object shapes and geometric relationships rather than texture.

While MEt3R [1] provides a robust and efficient framework for assessing multi-view consistency based on DINO features, it inherits this focus on geometric alignment. To enable separate evaluation of geometric and textural consistency in multi-view generation, we extend the MEt3R pipeline by introducing a dual-stream analysis framework. The full procedure is illustrated in Figure 1 and comprises the following components:

Given input images $\{I_1, I_2\}$, we first extract semantic feature maps using a pretrained DINO encoder E:

$$F_i^{\text{low}} = E(I_i), \quad i \in \{1, 2\}.$$

Since these feature maps are of relatively low resolution and may lack fine structural detail, we upsample them via FeatUp [11], which leverages a stack of learned Joint Bilateral Upsamplers (JBUs) that utilize the high-resolution RGB image to transfer high-frequency information. This process yields higher-resolution features:

$$F_i = \text{FeatUp}\big(F_i^{\text{low}}, I_i\big).$$

Next, we unproject the upsampled features into 3D space using point maps produced by DUSt3R [11]. Each 3D point is assigned the corresponding feature vector from F_i, yielding a feature point cloud P_i. We then reproject and render the point clouds into the canonical camera frame of I_1 using the PyTorch3D point rasterizer, producing the canonical-view feature maps:

$$\hat{F}_i = \textit{Render}(P_i, C_1), \tag{4}$$

where C_1 is the camera if I_1 generated by DUSt3R. Finally, we quantify geometric consistency by computing the similarity score $S(I_1, I_2)$, defined as the mean cosine similarity over all pixels in the overlapping region:

$$S(I_1, I_2) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{W} \sum_{j=1}^{H} m_{ij} \, \frac{\hat{\mathbf{f}}_{ij}^{(1)} \cdot \hat{\mathbf{f}}_{ij}^{(2)}}{\big\|\hat{\mathbf{f}}_{ij}^{(1)}\big\| \, \big\|\hat{\mathbf{f}}_{ij}^{(2)}\big\|},$$

where $m_{ij} \in \{0, 1\}$ indicates whether pixel (i, j) is in the overlapping region, and $\hat{\mathbf{f}}_{ij}^{(k)} = \hat{F}_k[i, j]$ denotes the feature vector rendered at pixel (i, j) for view k.

In parallel, the texture branch applies an albedo estimation module to extract appearance-specific representations that remain sensitive to color and fine texture details. These texture features are likewise unprojected and re-rendered onto the same 3D geometry, enabling pixel-level comparison. We then compute texture consistency using Learned Perceptual Image Patch Similarity (LPIPS) [14].We have chosen this metric because standard measures such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) operate strictly on a per-pixel basis, making them highly sensitive to small spatial misalignments or minor artifacts introduced by the point cloud rendering process. Such pixel-wise discrepancies, often manifesting as slight aberrations or local distortions, do not necessarily reflect true perceptual differences in texture or appearance. In contrast, LPIPS compares local image patches using deep feature embeddings, capturing perceptual similarity in a way that is more robust to these low-level variations and better aligned with human judgment of visual fidelity.

This dual-pathway design, illustrated in Figure 1, allows a disentangled assessment of 3D-aware generative models, facilitating a nuanced understanding of how well they preserve object geometry versus visual appearance across novel views.

This extension allows decoupled evaluation of:

1. Geometric consistency (captured by DINO feature similarity)

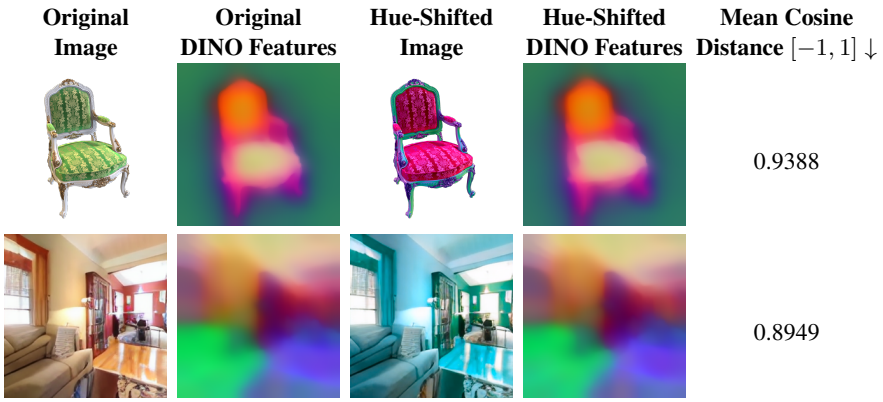

Figure 2. Evaluation of DINO features under hue shifts. Columns show images, features, distance maps, and mean cosine distances.

2. Textural consistency (captured by reprojected image similarity)

enabling a more nuanced and interpretable assessment of 3D-aware image generation pipelines.

Through this methodology, we systematically verify the geometry-centric nature of DINO features and introduce an enhanced evaluation framework that jointly assesses structural and appearance consistency. This approach ensures that models are not only spatially coherent but also faithful in reproducing texture across synthesized views.

4. Experiments

In this section, we present experiments designed to evaluate the robustness and accuracy of our approach. First, we investigate the sensitivity of DINO features to hue perturbations, highlighting potential vulnerabilities in feature representations under color shifts. Second, we validate MEt3R-Tex, our proposed metric for quantifying texture consistency, by comparing its outputs against perceptual expectations across diverse generative models.

4.1. Sensitivity of DINO Features to Hue Perturbations

To systematically evaluate the invariance properties of DINO features with respect to color transformations, we constructed a controlled dataset comprising natural images and their hue-shifted counterparts. For each original image, a set of synthetic variants was generated by applying uniform hue rotations in HSV color space. This procedure enables the isolation of appearance changes while preserving the underlying geometric and semantic content.

Following the methodology described in Section 3, dense DINO feature maps were extracted for each image pair. We then computed the per-pixel cosine distance between corresponding feature embeddings to obtain spatial sensitivity maps. The mean cosine distance across all pixels was used as a quantitative measure of feature shift induced by hue modification.

Figure 2 presents representative results for four example images, including the original input, the hue-shifted version, the corresponding cosine distance map, and the aggregated mean distance. Across the dataset, the observed mean distances remain consistently low, indicating limited sensitivity of DINO representations to hue changes. These findings support the hypothesis that DINO features predominantly capture shape and semantic structure, motivating the integration of a dedicated texture-sensitive branch in our evaluation pipeline to complement their invariance.

These results empirically validate the suitability of DINO features for capturing structural consistency under varying illumination or color conditions. In subsequent experiments, we evaluate the complementary sensitivity of albedo-based descriptors and the effectiveness of the proposed dual-stream framework.

4.2. Validating MEt3R-Tex

To assess the effectiveness of MEt3R-Tex in quantifying geometric and textural consistency, we evaluated three generative models: MV-LDM [1], PhotoNVS [13], and GenWarp [10]. For each model, two novel views were generated from a given input image out of the RealEstate10K dataset. The MEt3R-Tex pipeline was then applied to compute both the MEt3R score (reflecting geometric alignment) and the texture score based on LPIPS distance.

As shown in Table 3, the first interior scene generated by PhotoNVS achieves a low MEt3R score, indicating high geometric consistency across views. However, visual inspection reveals a change in the carpet's pattern, demonstrating a loss of textural fidelity. This discrepancy is captured by the elevated texture score.

Although GenWarp generally produces outputs of lower perceptual quality, the first interior scene illus-

MEt3R ↓	0.4024	0.0994	0.3952	0.3798	0.2291	0.7384
Texture ↓	0.7486	0.3302	0.6710	0.6452	0.3422	0.7338

Figure 3. Geometry and Texture Scores of various multi-view image generation models, where I_{in} is the input image of the models and I_{novel} the generated novel view.

trates comparatively accurate geometric alignment, despite significant texture deviations. In this case, the red carpet is transformed to a brown tone, which is reflected in a low MEt3R score but a high texture score.

For the second interior scene (a living room), PhotoNVS again yields superior geometric consistency, evidenced by lower MEt3R values. However, texture consistency remains limited, with observable changes in the chair's pattern across views, corresponding to higher LPIPS-based texture scores.

Overall, these results demonstrate that MEt3R-Tex effectively distinguishes geometric misalignment from textural inconsistencies, highlighting its utility for the comprehensive evaluation of multi-view image generation models.

5. Discussion

In this work, we introduced an extension of the existing MET3R pipeline to enable a more granular and separate evaluation of geometry and texture consistency in multi-view image generation without requiring ground truth geometry. This distinction is essential because, as demonstrated by our results, metrics focusing solely on geometric alignment fail to capture the full capability of generative models to produce consistent novel views. Specifically, while geometry consistency ensures accurate spatial correspondence across views, it does not account for discrepancies in texture appearance that can arise from model limitations or view-dependent effects.

Our experiments further revealed that relying exclusively on DINO features for assessing consistency is insufficient, as these features are predominantly optimized for geometric cues and lack sensitivity to fine-grained texture deviations. To address this limitation, we developed a parallel evaluation stream dedicated to texture consistency. However, evaluating texture similarity across views is inherently challenging due to view-dependent phenomena, such as reflections and specular highlights, which are particularly pronounced in non-Lambertian materials. These effects can distort the appearance of reprojected images, thereby introducing artifacts into consistency measurements.

To mitigate this problem, we incorporated an intrinsic image decomposition approach based on a neural network to extract albedo representations of the generated views. The resulting albedo maps inherently suppress view-dependent lighting variations, allowing for more robust and reliable comparisons of texture consistency between different viewpoints. This strategy proved effective in isolating intrinsic surface appearance from extraneous lighting artifacts, improving the interpretability and accuracy of consistency metrics.

While the proposed framework demonstrates notable improvements over prior approaches, it is subject to several limitations. First, the evaluation is restricted to pairwise view comparisons, which constrains its applicability in scenarios where consistency across larger view ensembles must be assessed comprehensively. Second, the reliance on rendering unprojected point clouds inherently introduces artifacts such as incomplete surfaces, floating pixels, and holes that can adversely affect both geometric and texture consistency evaluations. Third, the intrinsic image decomposition itself is a learned process whose accuracy depends on the underlying neural model. Consequently, decomposition errors or biases can propagate into the consistency measurements, particularly in scenes exhibiting complex illumination or non-Lambertian reflectance. Furthermore, the consistency of the intrinsic image decomposition model itself must be carefully assessed, as variations across views or

scenes can further undermine the reliability of the derived consistency metrics. In addition, the current evaluation is limited by the relatively small number of test samples, both for assessing the sensitivity of DINO features to hue perturbations and for analyzing their robustness against purely textural changes. This aspect warrants further investigation to confirm that the extracted features genuinely reflect geometric or semantic structures rather than texture-specific cues. Similarly, the Met3r-Tex evaluation was conducted on a limited set of manually selected scenes, and future work should aim to extend this analysis to a more comprehensive and diverse set of samples to ensure broader generalizability.

Overall, the proposed extension provides an integrated framework that separately quantifies geometric and texture consistency, enabling a more comprehensive assessment of generative models and supporting a deeper understanding of their strengths and limitations in synthesizing plausible and coherent novel views.

6. Conclusion

This work presented an enhanced evaluation framework for benchmarking multi-view image generation with a focus on disentangling geometric and texture consistency in the absence of ground truth geometry. By extending the MET3R pipeline to incorporate a parallel texture evaluation stream, we demonstrated that assessing geometry alone is insufficient to characterize the fidelity of generative models in producing consistent novel views. Our findings showed that commonly used geometric features such as DINO embeddings fail to capture fine-grained texture discrepancies, motivating the integration of intrinsic image decomposition to isolate view-independent albedo representations.

The proposed approach effectively reduces the confounding influence of view-dependent effects, such as reflections and specular highlights, thereby enabling more robust texture consistency measurements. Despite certain limitations, most notably the restriction to pairwise comparisons, the susceptibility to rendering artifacts from point cloud projections, and the dependency on decomposition accuracy, this framework provides a more comprehensive and interpretable evaluation of generative model performance.

References

[1] M. Asim, C. Wewer, T. Wimmer, B. Schiele, and J. E. Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] C. Careaga and Y. Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Trans. Graph.*, 43(1), 2023.

[3] C. Careaga and Y. Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Trans. Graph.*, 43(6), 2024.

[4] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *CoRR*, abs/2103.15595, 2021.

[5] Y. Chen, X. Chen, A. Chen, G. Pons-Moll, and Y. Xiu. Feat2gs: Probing visual foundation models with gaussian splatting, 2024.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, Dec. 2021.

[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[9] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] J. Seo, K. Fukuda, T. Shibuya, T. Narihira, N. Murata, S. Hu, C.-H. Lai, S. Kim, and Y. Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping, 2024.

[11] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy, 2024.

[12] J. Yang, Z. Cheng, Y. Duan, P. Ji, and H. Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. *arXiv*, 2023.

[13] J. J. Yu, F. Forghani, K. G. Derpanis, and M. A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models, 2023.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.