Complexity-Stratified Evaluation Reveals Shortcut Regimes in Rotational Novel View Synthesis

Rohan Keyur Dalal [1] **Han Lee** [2] **Irene Tang** [3]

Abstract

Standard novel-view synthesis benchmarks report aggregate metrics across heterogeneous object sets, obscuring systematic performance differences tied to object appearance complexity. We introduce a *view-change complexity* score (the mean VGG perceptual distance between views of the same object separated by a fixed angle) and use it to stratify the 100-object COIL-100 turntable dataset into four quartiles spanning $C \in [0.021, 0.629]$. Evaluating two training-free baselines reveals a sharp *regime crossover* under VGG-16 distance: on low-complexity objects (Q1, $\bar{C} = 0.113$), copying the source image achieves VGG distance 0.157 while nearest-neighbour retrieval scores 0.353 (+125% worse); on high-complexity objects (Q4, $\bar{C} = 0.535$), this reverses, with retrieval outperforming copy-source by 45%. A conditional DDPM trained for 30,000 steps confirms the stratified picture: the model consistently underperforms copy-source on Q1 while outperforming it on Q4, and conditioning on the rotation angle provides no measurable benefit at intermediate training checkpoints. These findings demonstrate that stratified complexity evaluation is essential for exposing and thus avoiding shortcut regimes in rotational NVS.

1. Introduction

Novel view synthesis (NVS) asks a model to generate an image of an object or scene from a novel viewpoint, given one or more reference views. Recent work ranging from neural radiance fields (Mildenhall et al., 2020) to geometry-aware GANs (Chan et al., 2022) and diffusion-based approaches (Watson et al., 2023; Liu et al., 2023) has driven rapid progress on standard benchmarks. Yet a recurring concern is that aggregate metrics (LPIPS, PSNR, SSIM averaged across all test objects) can be dominated by "easy" instances where the model need only copy the source image with minor modifications (Geirhos et al., 2020).

The COIL-100 dataset (Nene et al., 1996) provides a canonical testbed for rotational NVS: 100 household objects photographed at 72 evenly-spaced azimuths under controlled lighting. Objects in COIL-100 span a wide range of appearance variation with rotation—a featureless sphere changes almost imperceptibly over 90°, while a multi-coloured toy car looks entirely different. Despite this diversity, published work routinely aggregates performance across all objects, collapsing this meaningful variation into a single number. This is closely related to hidden stratification in medical imaging, where aggregate scores can conceal subgroup-specific failures (Oakden-Rayner et al., 2020). Our contribution is to make that idea concrete for rotational NVS by defining a simple object-level complexity axis.

We argue that this aggregation conceals fundamentally different *learning regimes*. For objects with low view-change complexity, copying the source view is nearly optimal: the source and target look nearly identical. For high-complexity objects, a model must learn genuine rotational structure. A system that achieves a good aggregate score may be doing so by exploiting the former shortcut regime while failing entirely on the latter.

Contributions.

1. We introduce a *view-change complexity* score $C(o; \alpha)$ based on mean VGG perceptual distance, and show it cleanly stratifies COIL-100 objects into four distinct performance regimes.

2. We document a *regime crossover* in two training-free baselines: copy-source wins by 125% over NN retrieval on Q1 objects, while NN retrieval wins by 45% over copy-source on Q4 objects (Table 1).

3. We evaluate a conditional DDPM under stratified metrics and find an honest picture: the model improves over copy-source only on high-complexity objects, and

[1]Georgia Institute of Technology [2]Stanford University [3]Massachusetts Institute of Technology. Correspondence to: Rohan Dalal <rohan.dalal@gatech.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

rotation-angle conditioning provides no measurable benefit at 15,000 training steps.

4. We release a fully reproducible evaluation codebase with deterministic pair manifests, object-disjoint splits, and auditable result CSVs.

2. Related Work

Novel view synthesis. NeRF (Mildenhall et al., 2020) introduced differentiable volume rendering for view synthesis; subsequent work has scaled to real-time rendering and general object categories. Geometry-free methods (Watson et al., 2023; Liu et al., 2023) use diffusion models or other generative priors to predict novel views from a single image without explicit 3D reconstruction. EG3D (Chan et al., 2022) combines GAN training with a tri-plane 3D representation. All of these works report aggregate evaluation metrics; none stratify results by object-level view-change difficulty. NerfBaselines (Kulhánek & Sattler, 2025) addresses reproducibility of NVS evaluation but not per-object complexity stratification.

Shortcut learning and memorization. Geirhos et al. (Geirhos et al., 2020) systematically document how deep networks exploit spurious correlations ("shortcuts") in training data. Feldman (Feldman, 2020) shows that memorizing long-tail examples is necessary for good generalisation in certain regimes. Carlini et al. (Carlini et al., 2023) demonstrate that diffusion models can memorise and reproduce training images verbatim. Our work asks a related but distinct question: does *evaluating* a generative model on a heterogeneous dataset inadvertently reward copy-source shortcut behaviour on easy subsets?

Complexity-aware evaluation. Dataset difficulty stratification has been studied in object recognition (e.g., calibration across frequency bins) and language modelling (easy/hard subsets), but has received little attention in NVS. Prior work on *grokking* (Power et al., 2022) studies delayed generalisation on algorithmic tasks; that phenomenon involves a sudden phase transition not observed here. We see only a persistent Q4>Q1 gap that narrows gradually throughout training. Stein et al. (2023) show that standard generative evaluation metrics can produce systematically misleading rankings; our stratified approach surfaces a related failure mode specific to NVS.

Hidden stratification. Oakden-Rayner et al. (Oakden-Rayner et al., 2020) document hidden stratification in medical imaging, where aggregate metrics obscure clinically important subgroup failures. Our setting is different in domain but similar in structure: we use a label-free complexity score to expose an NVS-specific stratification axis rather than claiming novelty over the general phenomenon itself.

Concurrent work. Stern et al. (2026) propose task-aware evaluation for NVS by matching metrics to user-defined task requirements; our work is complementary, focusing on object-level complexity stratification in the turntable setting and demonstrating how aggregate metrics conceal shortcut regimes.

3. View-Change Complexity

Dataset. COIL-100 (Nene et al., 1996) contains 100 objects photographed at 72 azimuths at 5° spacing on a black background. We resize all images to 64×64 pixels (downsampled from the original 128×128 for computational tractability). The dataset has a clean turntable structure—only azimuthal rotation varies—making it ideal for controlled complexity analysis.

Complexity metric. For object o and angular offset α, we define view-change complexity as

$$C(o;\,\alpha) \;=\; \frac{1}{K}\sum_{\theta=1}^{K} d_{\mathrm{VGG}}\big(I_{o,\theta},\, I_{o,\theta+\alpha}\big), \qquad (1)$$

where d_{VGG} is the cosine distance in the VGG-16 `relu3_3` feature space (Simonyan & Zisserman, 2015), $I_{o,\theta}$ is the image of object o at angle θ, and $K = 72$ views are summed modulo the full circle. We fix $\alpha = 90°$ as our primary evaluation angle unless otherwise noted.

Lower C indicates that views *look similar* across the offset, so copying the source view is a near-optimal strategy. Higher C indicates that the appearance changes substantially, demanding genuine geometric reasoning.

Quartile assignment. We compute $C(o; 90°)$ for all 100 COIL-100 objects and assign them to four equal-size quartiles. Q1 (simplest) has $C \leq 0.193$ with mean $\bar{C}_{\mathrm{Q1}} = 0.113$; Q4 (most complex) has $C > 0.483$ with mean $\bar{C}_{\mathrm{Q4}} = 0.535$. Figure 1 shows representative objects at four azimuths: Q1 objects (row 1) are nearly rotationally symmetric, while Q4 objects (row 4) exhibit dramatic appearance changes across the same 90° steps.

4. Experimental Setup

Splits and pairs. We use an object-disjoint split: 40 objects for training, 20 for evaluation, drawn deterministically from all four complexity quartiles. This split applies to the DDPM, which requires training data. The training-free baselines (copy-source and NN retrieval) do not depend on a training set, so their regime crossover (Table 1) is evaluated on all 100 objects (25 per quartile) to give a cleaner picture of the full complexity range. Each evaluation sample is a

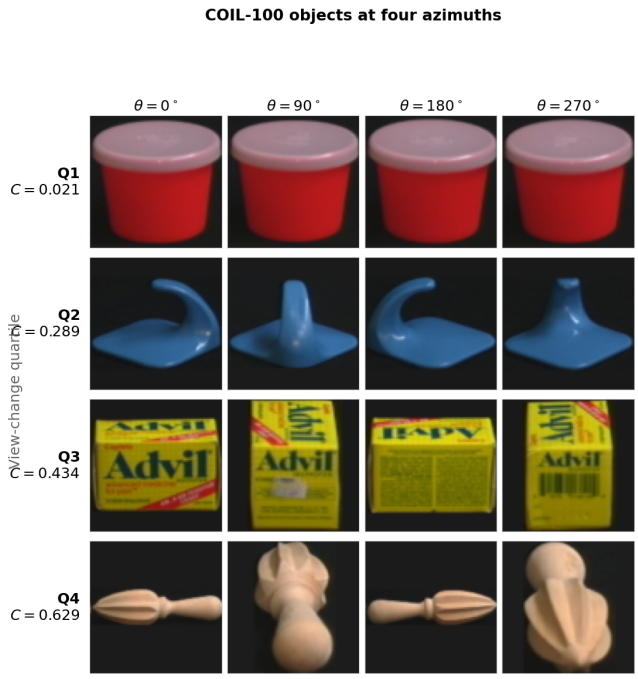

Figure 1. COIL-100 objects at azimuths $0°, 90°, 180°, 270°$. Rows correspond to complexity quartiles Q1–Q4 (top to bottom) with view-change complexity C indicated. Q1 objects are near-rotationally-symmetric; Q4 objects change dramatically across 90° steps.

(source $I_{o,\theta}$, target $I_{o,\theta+90°}$) pair; we evaluate on *all* such pairs for each test object (72 per object, $20 \times 72 = 1{,}440$ pairs total for the DDPM).

Evaluation metric. Our primary metric is the VGG cosine distance d_{VGG} (lower is better), which corresponds to perceptual dissimilarity in deep feature space and is more sensitive to structural differences than pixel-level metrics.

Baselines. *Copy-source* simply outputs the source image as the prediction. This is a strong baseline for low-complexity objects where source and target look nearly identical. *NN retrieval* finds the nearest training image at the target angle, using DINOv2-ViT-S/14 (Oquab et al., 2023) features to embed source images and retrieve the closest training example; the retrieved image at the target $\theta{+}90°$ is then returned. This baseline is an *oracle-like diagnostic*: it has direct access to training-set images at the exact target angle, which a deployed model would not. It bounds what appearance-based retrieval can achieve rather than representing a practical NVS system.

Conditional DDPM. Our generative model is a conditional DDPM (Ho et al., 2020; Sohl-Dickstein et al., 2015) based on a 3-level U-Net with $\sim$4.4M parameters.

- **Input**: 6-channel tensor — noisy target image concatenated with source image along the channel dimension.
- **Angle conditioning**: rotation offset encoded as $[\sin \Delta\theta, \cos \Delta\theta]$ and injected at every residual block via adaptive group normalisation (AdaGN) (Nichol & Dhariwal, 2021).
- **Schedule**: cosine noise schedule, $T = 1{,}000$ diffusion steps.
- **Training**: AdamW, lr $= 10^{-4}$, batch 64, 30,000 gradient steps, 3 independent random seeds (42, 7, 123).
- **Evaluation**: DDIM sampling (Song et al., 2021) with 50 steps every 1,000 training steps on 20 test objects.

Angle utilization ablation. To test whether the model actually uses the angle signal, we train a second model with the same architecture but with $\Delta\theta$ labels randomly shuffled within each batch. We compare both variants at 15,000 steps on the same test set.

5. Results

5.1. Baseline Regime Crossover

Table 1 and Figure 2 present the central finding of this paper. On low-complexity objects (Q1), copy-source achieves VGG distance 0.157 while NN retrieval scores 0.353, which is +125% worse. On high-complexity objects (Q4), this ordering completely reverses: copy-source scores 0.520 while NN retrieval achieves 0.288, a 45% improvement.

The crossover occurs between Q2 and Q3. For Q1 objects, source and target views are so perceptually similar ($\bar{C} = 0.113$) that the source image is already a near-optimal prediction; NN retrieval fails because it returns images from *other* training objects that do not match the test object's appearance. For Q4 objects, the source is a poor predictor of the target ($\bar{C} = 0.535$), and a retrieved training image at the correct target angle turns out to be a better proxy than the source view.

Aggregate metrics hide the crossover. Averaged uniformly over all four quartiles, copy-source achieves 0.325 and NN retrieval 0.347 showing that copy-source *wins* in aggregate, masking the fact that NN retrieval is substantially superior on Q4. This is precisely the failure mode aggregate benchmarking enables: a model that copies the source on easy objects and fails on hard ones can score well overall (Stein et al., 2023).

Cross-metric validation. As a robustness check, we re-evaluate both baselines using LPIPS-AlexNet (Zhang et al., 2018), an entirely different perceptual backbone. Copy-source: Q1 $= 0.071$, Q4 $= 0.307$; NN retrieval: Q1 $=$

Table 1. VGG cosine distance (lower ↓) by view-change complexity quartile and aggregate (equal-weight mean) for copy-source and NN retrieval. $\Delta\theta = 90°$, all 100 COIL-100 objects (25 per quartile). NN retrieval is an oracle-like diagnostic with direct access to training-set images at the exact target angle—not a practical NVS system. The aggregate column shows that copy-source wins overall (**bold**), hiding the Q4 crossover.

	View-change quartile				
Baseline	Q1	Q2	Q3	Q4	Agg.
Copy-source	0.157	0.242	0.381	0.520	**0.325**
NN retrieval	0.353	0.349	0.397	0.288	0.347
NN vs. copy-source	+125%	+44%	+4%	−45%	+7%

0.338, Q4 = 0.397. The Q1<Q4 complexity ordering is confirmed by both methods. LPIPS-AlexNet does not reproduce the same crossover, so we do not treat the VGG result as backbone-invariant. Instead, we interpret the disagreement as a reminder that the observed crossover is metric-dependent and therefore should be read as evidence about the chosen perceptual geometry, not as a universal property of all similarity measures. The Pearson correlation between VGG-16 complexity $C(o; 90°)$ and LPIPS-AlexNet evaluation distance is $r = 0.906$ ($p < 10^{-20}$), confirming that complexity stratification itself is not a VGG artifact.

Complexity metric robustness. To verify that VGG-16 complexity rankings are not circular, we compute $C(o; 90°)$ using DINOv2-ViT-S/14 (Oquab et al., 2023) features. The Spearman rank correlation between VGG and DINOv2 complexity scores across all 100 objects is $r_s = 0.782$ ($p < 10^{-20}$), and 69.0% of objects receive the same Q1/Q4 assignment under both backbones. This confirms that the complexity ordering captures a genuine property of object appearance rather than a VGG-specific feature-space structure.

Sensitivity to view-change angle α. To test whether the crossover is specific to $\Delta\theta = 90°$, we repeat the baseline evaluation at $\alpha \in \{30°, 45°, 60°, 120°, 180°\}$, re-stratifying objects by $C(o; \alpha)$ at each angle. The crossover (NN retrieval wins Q4, copy-source wins Q1) holds at 45°, 60°, 90°, and 120°, but not at the extremes. At 30°, the view change is small even for complex objects, so copy-source remains competitive on Q4 (0.282 vs. 0.312 for NN retrieval). At 180°, the half-rotation is maximally ambiguous and copy-source again outperforms NN retrieval on Q4 (0.285 vs. 0.334). The crossover thus operates in the intermediate regime where view changes are large enough to penalise copy-source yet structured enough for NN retrieval to exploit object similarity.

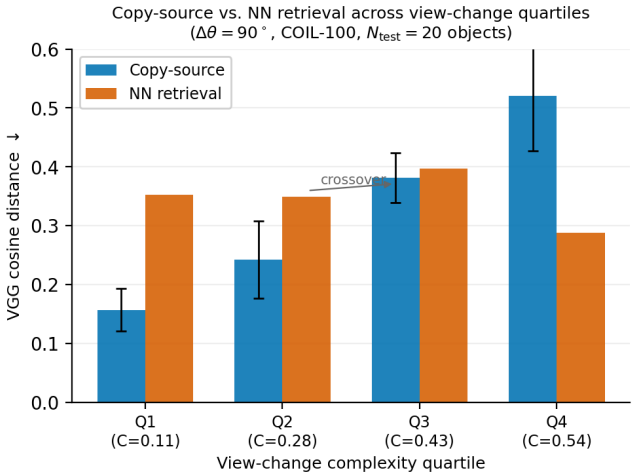

Figure 2. VGG cosine distance (lower ↓) for copy-source and NN retrieval across complexity quartiles ($\Delta\theta = 90°$). Copy-source dominates on low-complexity objects (Q1–Q2); NN retrieval is superior on high-complexity objects (Q3–Q4). Evaluating only aggregate averages would conceal this crossover.

5.2. DDPM Training Dynamics

The left panel of Figure 4 shows the training trajectory of the conditional DDPM over 30,000 steps across three seeds. Two features are consistent across all seeds:

Persistent Q4>Q1 gap. At convergence, Q1 performance is 0.359 and Q4 is 0.434 for seed 42, a gap of 0.074. Averaging across all three seeds, the gap is 0.092 ± 0.033 (mean $\pm$ s.d.), confirming that the model consistently finds Q4 objects harder.

Shortcut underexploitation on Q1. Despite Q1 objects being near-rotationally symmetric, the DDPM achieves only 0.359 on Q1 at convergence, far worse than the copy-source baseline (0.157). The model is not exploiting the obvious copy-source shortcut, despite that shortcut being optimal for these objects. This gap arises from the generative process itself: the DDPM generates images by iteratively denoising from Gaussian noise guided by learned training-object statistics, and even for Q1 objects where the target view is nearly identical to the source, the model produces outputs that reflect its training distribution rather than passing through the source image. Copy-source trivially achieves near-zero error by design. On Q4, the DDPM (0.434) does outperform copy-source (0.520), showing the model has learned some rotational structure for harder objects.

5.3. Angle Utilization Ablation

The right panel of Figure 4 shows VGG distances at 15,000 steps for the correctly-conditioned model and the shuffled-label ablation. On Q1, correct conditioning achieves 0.385 versus 0.371 with shuffled labels. On Q4, correct condi-

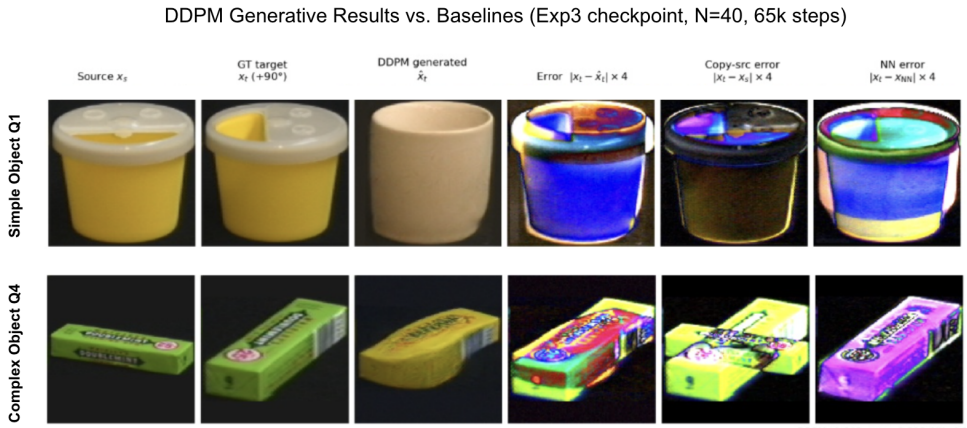

Figure 3. Qualitative synthesis ($\Delta\theta = 90°$, $N = 40$, 65,000 steps). Each row shows source | target | **DDPM output** (blue border) | error | copy-src error | NN error for one test object. **Q1 objects** (row 1, $C \leq 0.193$): source and target show a generally similar shape, yet the DDPM generates a noticeably different image, confirming that the model does not exploit the copy-source shortcut even when it is optimal. **Q4 objects** (row 2 $C > 0.483$): source and target differ substantially; the DDPM output partially captures the new viewpoint but lacks fine detail, consistent with the persistent Q4>Q1 gap.

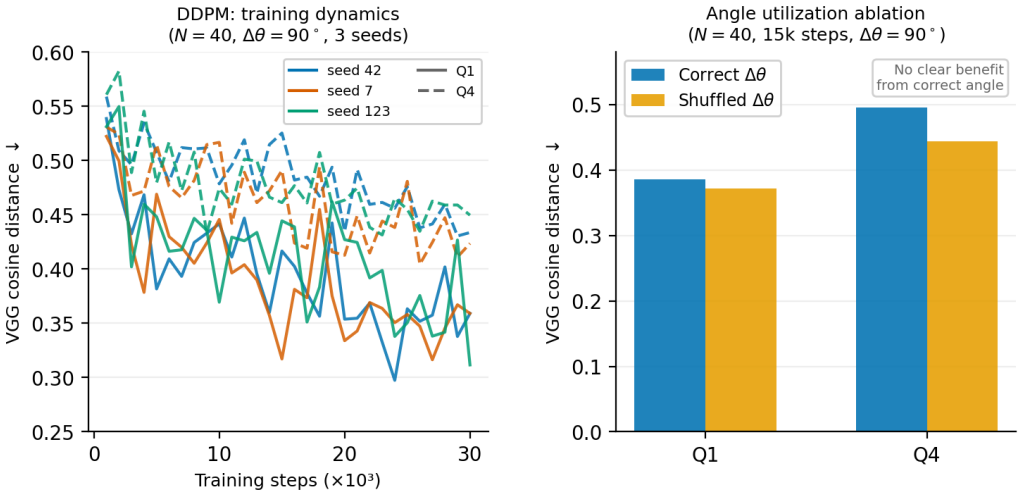

Figure 4. **Left:** DDPM training dynamics for Q1 and Q4 test objects over 30,000 steps, across 3 seeds (colours). Solid lines: Q1; dashed lines: Q4. The Q4>Q1 gap (0.092 ± 0.033) is consistent across all seeds. **Right:** Angle utilization ablation at 15,000 steps. Shuffled $\Delta\theta$ labels (orange) perform indistinguishably from correct labels (blue), indicating the model has not yet learned to use the rotation-angle signal at this training stage.

tioning achieves 0.496 versus 0.444 with shuffled labels. In both quartiles, shuffled conditioning is *not measurably worse* and, indeed, is marginally better in absolute terms.

This is an honest negative result: at 15,000 training steps, the conditional DDPM does not meaningfully exploit the rotation-angle signal. The model appears to generate plausible outputs primarily through learned visual patterns from training objects, rather than explicit geometric conditioning. This is consistent with the shortcut-regime picture: on low-complexity objects, generating something that looks like the source is sufficient; on high-complexity objects, more training or a stronger inductive bias would be needed to leverage the geometric signal. We replicate this finding with two additional random seeds. Seed 7: normal Q1/Q4 $= 0.411/0.429$, shuffled $= 0.342/0.435$; seed 7 shuffled wins Q1 by a margin of 0.069, consistent with the main seed. Seed 123: normal Q1/Q4 $= 0.405/0.464$, shuffled $= 0.444/0.497$; here normal conditioning performs better on both quartiles (margins of 0.039 and 0.033). The results are mixed: two of three seeds show shuffled winning Q1, while seed 123 reverses this, and no seed shows a large, consistent advantage for correct angle conditioning on either quartile. The overall pattern suggests the model has not learned to robustly exploit the angle signal at 15,000 steps, though the evidence is not uniformly null across seeds.

5.4. Effect of Training Scale

Figure 5 shows how performance changes as the number of training objects N grows from 10 to 80. Both quartiles improve with more training data, but the improvement is larger for Q1 than for Q4. Q1 improves from 0.433 at $N = 10$ to 0.361 at $N = 80$ (improvement of 0.072); Q4 improves from 0.495 to 0.467 (improvement of 0.028). The Q4 performance gap relative to Q1 narrows only modestly with more data, suggesting that the difficulty of high-complexity objects is not primarily a data-quantity limitation.

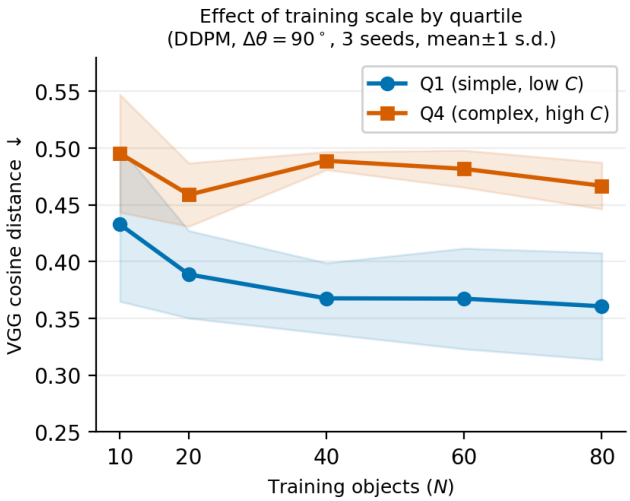

Figure 5. Effect of training scale (N objects) on VGG cosine distance for Q1 and Q4 test objects (3 seeds, mean $\pm$ 1 s.d.). Both quartiles improve with more data, but Q4 consistently remains harder and benefits less from additional training objects.

6. Discussion

What the crossover tells us. The regime crossover in Table 1 arises from a fundamental property of turntable datasets: view-change complexity is object-specific and spans orders of magnitude. When a benchmark reports an aggregate metric, a single number is averaging over qualitatively different tasks. A model that performs well in aggregate may be scoring low (better) on Q1 via trivial copy-source behaviour while performing poorly on Q4 objects that actually require geometric synthesis. Critically, the crossover is a claim about *which method wins within each complexity stratum*, not about the absolute magnitude of the Q4>Q1 gap—so the VGG-VGG circularity acknowledged in Limitations does not undermine it. Two methods evaluated under the same metric at the same complexity level stand on equal footing regardless of how the strata were defined: copy-source and NN retrieval are both penalised by the same perceptual geometry, and the reversal in their relative ordering across quartiles is a genuine empirical finding. The DINOv2 cross-check ($r_s = 0.782$, 69.0% Q1/Q4 agreement) confirms that the strata themselves reflect a backbone-independent property of object appearance.

Implications for the DDPM findings. The DDPM results illustrate this tension concretely. The model fails to replicate the copy-source shortcut on Q1 (DDPM 0.359 vs. copy-source 0.157), but succeeds in generating useful predictions for Q4 (DDPM 0.434 vs. copy-source 0.520). Rather than interpreting the persistent Q4>Q1 gap as a training failure, we view it as evidence that the model has learned something real for high-complexity objects but has not converged to the trivial solution for easy ones. Whether this is a feature (robustness to easy shortcuts) or a bug (suboptimal convergence) depends on the downstream application.

The angle signal. The null result in Section 5.3 is worth taking seriously. The rotation angle $\Delta\theta$ is encoded via AdaGN at every residual block. Yet at 15,000 steps, this signal provides no measurable benefit. We note two competing interpretations: perceptual feature matching (visual similarity to training objects) may simply dominate the geometric signal at this scale, *or* the model may merely be undertrained—40 objects and 15,000 steps may be insufficient for angle to become a useful cue regardless of shortcut regime. We cannot rule out the latter, and we flag it as an honest limitation of the current setup. Resolving the two accounts would require either substantially more training

steps, a larger training set, or controlled ablations varying each independently.

Limitations. This study focuses on COIL-100's clean turntable setting; other datasets with lighting variation, background clutter, or non-planar trajectories may show different regime structures. The DDPM is relatively small ($\sim$4.4M parameters) and trained with limited data (40 objects); larger models may close the Q4>Q1 gap. Both our complexity metric $C(o; \alpha)$ and our evaluation metric use VGG-16 cosine distance, introducing a degree of circularity: by construction, objects with high C will incur higher evaluation distances. That makes the absolute Q4>Q1 gap a self-consistency result rather than an independent geometric certificate. Relative comparisons *between methods* at the same complexity level remain valid, but the observed correlation between complexity and task difficulty is partly tautological rather than purely empirical. In other words, the key claim is that stratification makes the benchmark behavior visible; it is not that the VGG-based crossover alone proves true 3D understanding. We also do not claim that these toy-scale results directly transfer to large foundation models or unconstrained real-world scenes.

Recommendations. Based on our findings, we recommend that future NVS evaluations: (1) report stratified metrics by per-object view-change complexity; (2) include copy-source and NN retrieval as baseline comparisons stratified by complexity; and (3) explicitly test whether angle/viewpoint conditioning is being exploited via shuffled-label ablations.

7. Conclusion

We have introduced view-change complexity as a principled stratification criterion for rotational NVS evaluation and demonstrated that COIL-100 objects span qualitatively different learning regimes. The central empirical finding is a baseline regime crossover: copy-source outperforms NN retrieval by +125% on low-complexity objects while NN retrieval outperforms copy-source by 45% on high-complexity objects. A conditional DDPM confirms this picture by performing worse than copy-source on easy objects and better on hard ones, while providing a clear negative result on angle utilization. Together, these results make a concrete case for stratified evaluation as a diagnostic tool in generative NVS research.

References

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L. J., Tremblay, J., Khamis, S., et al. Efficient geometry-aware 3D generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Feldman, V. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, pp. 954–959, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zeiler, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Kulhánek, J. and Sattler, T. NerfBaselines: Consistent and reproducible evaluation of novel view synthesis methods. In *Advances in Neural Information Processing Systems*, 2025.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. Zero-1-to-3: Zero-shot one image to 3D object. In *IEEE International Conference on Computer Vision*, pp. 9298–9309, 2023.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020.

Nene, S. A., Nayar, S. K., and Murase, H. Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, 1996.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Stein, G., Cresswell, J. C., Hosseinzadeh, R., Sui, Y., Ross, B. L., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, J. E. T., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Stern, S., Sobol, I., and Litany, O. Appreciate the view: A task-aware evaluation framework for novel view synthesis. In *International Conference on 3D Vision*, 2026.

Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., and Norouzi, M. Novel view synthesis with diffusion models. In *International Conference on Learning Representations*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.