BEYOND ANGULAR ERROR: GLOBAL AND LOCAL GEOMETRY OF FAILURE IN MONOCULAR SURFACE-NORMAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Per-pixel angular error measures how much a surface-normal prediction is wrong, but it hides the geometric form of that error. We show that failures of modern monocular surface-normal estimators exhibit reproducible structure at both global and local scales. Globally, we use continuous generalized bas-relief (GBR) fitting as an interpretable coordinate system for prediction–reference discrepancies. Across four architectures, matched pairs concentrate in a shared low-shear regime, while correspondence-destroying controls move the fitted transforms toward higher shear. The remaining relief coordinate λ also transfers beyond the pixels used to estimate it: on 100 randomly sampled NYUv2 validation images, cross-fitted $\log \lambda$ correlates with angular error on disjoint held-out pixels across all four models ($r \in [-0.70, -0.60]$). Locally, we derive a prediction-only curl score from a single estimated normal field. On 500 NYUv2 images per model, it identifies the worst 10% of predictions with AUROC 0.77–0.91, outperforming generic normal roughness and improving selective risk–coverage. Together, these results show that normal-estimation errors are organized by reusable geometric structure, providing both a diagnostic view of model failure and a practical signal for test-time reliability.

1 INTRODUCTION

Monocular surface-normal estimation is an ill-posed inverse problem: a single RGB image does not uniquely determine scene geometry, so modern systems rely heavily on learned visual priors. Recent predictors draw on increasingly strong generative and discriminative representations, including Marigold, StableNormal, Lotus, and MoGe-2 (Ke et al., 2025; Ye et al., 2024; He et al., 2025; Wang et al., 2025). Their progress is typically summarized by angular error and threshold accuracy. Prior work has also noted that these standard metrics can miss geometrically important errors (Chen et al., 2020).

These metrics measure *how much* a prediction is wrong, but collapse geometrically different failures to the same scalar. Two normal maps can have similar mean angular error even when one is a coherent global deformation of the reference surface and the other contains strong local inconsistency. The distinction matters both scientifically—for understanding what failure modes are shared across models—and practically, because different geometric failures can have different consequences for downstream reconstruction and reliability.

We study whether monocular normal error has a reusable *geometry of failure*. Our central finding is that it does. Across four model families, global prediction–reference discrepancies repeatedly occupy a restricted region of a simple geometric transformation family, and local inconsistency inside the prediction itself provides a strong signal of image-level error.

Globally, we use the classical generalized bas-relief (GBR) family (Belhumeur et al., 1999) as a coordinate system for relative shape distortion. For each prediction–reference pair, we fit the continuous GBR transformation that best aligns the two. The fitted transforms consistently concentrate in a low-shear region across architectures. This concentration is substantially weakened when prediction–reference correspondence is destroyed and becomes tighter on controlled orthographic GSO scenes, establishing a reproducible geometric organization of model error across data conditions.

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.

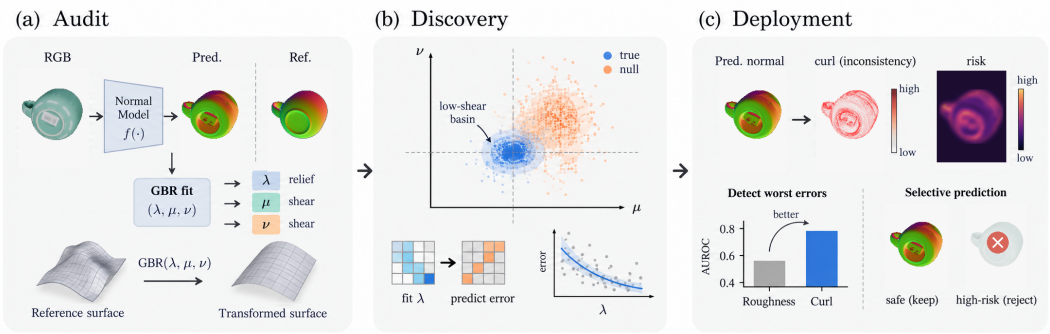

Figure 1: **From geometric audit to test-time reliability.** Continuous GBR fitting exposes the global geometry of monocular normal errors: matched prediction–reference pairs concentrate in a low-shear region, and the fitted relief coordinate λ generalizes to held-out pixels. A complementary prediction-only local curl score ranks high-risk outputs at test time.

The remaining relief coordinate, λ, carries information beyond the pixels used to fit it. In a pixel cross-fitting experiment, we estimate the GBR transform on one random half of the valid pixels and evaluate on the other half. The fitted $\log \lambda$ predicts held-out angular error across all four models, while the fitted transform itself transfers to reduce error on the held-out pixels. Thus the global coordinates capture image-level structure rather than only fitting the pixels used for optimization.

Locally, we ask whether a single prediction contains a useful signal of its own reliability. We derive a simple image-space curl statistic from the normal field, motivated by local consistency of the implied surface gradients. With one fixed coordinate convention per model calibrated on six images disjoint from evaluation, this prediction-only score identifies high-error outputs better than generic normal roughness and improves selective risk–coverage across all four architectures.

These global and local results connect two uses of failure geometry. GBR coordinates provide an interpretable audit of *how* predictions depart from reference geometry; local curl turns structure inside a prediction into a lightweight signal of *when* that prediction is likely to be unreliable. The main contributions are:

- **A global coordinate system for normal-estimation failure.** Continuous GBR fitting reveals a shared low-shear regime across four model families, with systematic shifts under correspondence-destroying controls and controlled orthographic rendering.
- **Held-out predictive structure.** The fitted relief coordinate λ predicts angular error on spatially disjoint pixels, and the fitted transform transfers to reduce their error.
- **Prediction-only test-time reliability.** A local curl-based score computed from one normal map ranks high-error outputs better than generic roughness and supports selective prediction without per-image reference geometry.

2 Related Work

Monocular geometry and surface normals. Single-image normal estimation has evolved from data-driven geometric primitives and task-specific CNNs (Fouhey et al., 2013; Wang et al., 2015; Eigen and Fergus, 2015; Bansal et al., 2016) to broader dense-prediction and multi-task representations (Zamir et al., 2018; Eftekhar et al., 2021; Ranftl et al., 2021). Recent geometry foundation models combine large-scale data, discriminative backbones, and generative priors (Fu et al., 2024; Hu et al., 2024). Marigold, StableNormal, Lotus, GenPercept, and MoGe-2 further advance diffusion- and representation-based monocular geometry (Ke et al., 2024; 2025; Ye et al., 2024; He et al., 2025; Xu et al., 2025; Wang et al., 2025). These works primarily improve prediction quality. We instead analyze the *shape of their errors* under a common geometric coordinate system.

Geometric ambiguities and error parameterizations. The generalized bas-relief ambiguity describes a family of surface transformations

$$Z'(x, y) = \lambda Z(x, y) + \mu x + \nu y, \tag{1}$$

that are classically ambiguous under orthographic Lambertian image formation with corresponding lighting transformations (Belhumeur et al., 1999). Its parameters separate relief scaling from directional slope offsets, making GBR a compact and interpretable coordinate system for coherent differences between predicted and reference normal fields.

Reliability and selective prediction. Selective prediction formalizes the risk–coverage trade-off of abstaining on uncertain examples (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; 2019). Predictive uncertainty is commonly obtained from Bayesian formulations, ensembles, learned uncertainty, or post-hoc calibration (Kendall and Gal, 2017; Lakshminarayanan et al., 2017; Guo et al., 2017). Dense normal estimators do not naturally expose a calibrated image-level confidence. Our goal is different: derive a training-free image-level risk score from the internal geometry of a single predicted normal field.

3 Geometric Failure Coordinates

We characterize failure geometry at two complementary scales. The *global* coordinate summarizes coherent distortion between prediction and reference geometry, while the *local* coordinate measures consistency inside the prediction itself. This separation gives the two analyses distinct roles: GBR provides a reference-based audit coordinate system, whereas curl provides a prediction-only reliability signal.

3.1 Global coordinate: continuous GBR fitting

For a height field $Z(x, y)$ with gradients $p = \partial_x Z$ and $q = \partial_y Z$, the generalized bas-relief (GBR) transformation induces

$$p' = \lambda p + \mu, \qquad q' = \lambda q + \nu. \tag{2}$$

The scalar $\lambda > 0$ controls relief scaling, while μ and ν introduce directional slope offsets. Given a calibrated reference normal field N and a prediction $\hat{N}$, we fit the continuous GBR transformation that best aligns the two:

$$(\hat{\lambda}, \hat{\mu}, \hat{\nu}) = \arg\min_{\lambda,\mu,\nu} \frac{1}{|\Omega|} \sum_{u \in \Omega} \angle\Big(\mathcal{T}_{\lambda,\mu,\nu}(N_u), \hat{N}_u\Big), \tag{3}$$

using bounded continuous optimization. We summarize the directional component of the fitted transformation by

$$s_{\text{shear}} = \sqrt{\hat{\mu}^2 + \hat{\nu}^2}. \tag{4}$$

A small value of s_{shear} means that the best-fitting coherent distortion lies primarily along the relief-scaling direction rather than requiring large directional slope offsets. We use $(\hat{\lambda}, s_{\text{shear}})$ as continuous coordinates that locate each prediction–reference pair within the GBR family. Because fitting Eq. (3) uses reference geometry, these coordinates serve as an audit of error structure.

3.2 Local coordinate: prediction-only curl risk

The global coordinate captures reference-relative distortion. For test-time reliability, we seek a complementary quantity computed directly from the prediction.

Under an orthographic height-field model, a normal field corresponds locally to a gradient field,

$$\hat{g}_x = -\frac{\hat{n}_x}{\hat{n}_z}, \qquad \hat{g}_y = -\frac{\hat{n}_y}{\hat{n}_z}. \tag{5}$$

For an integrable gradient field, equality of mixed partial derivatives implies $\partial_x \hat{g}_y - \partial_y \hat{g}_x = 0$ (Frankot and Chellappa, 1988). This motivates the image-level score

$$R_{\mathrm{curl}} = \mathrm{mean}_{u \in \Omega} \left| \partial_x \hat{g}_y(u) - \partial_y \hat{g}_x(u) \right|. \tag{6}$$

We use Eq. (6) as a *prediction-only local inconsistency score*: it measures departures from a locally consistent image-space gradient field and requires only a single predicted normal map at test time. This image-space formulation is lightweight, requires no camera intrinsics, and is applied uniformly across all four architectures.

To separate geometric consistency from generic spatial variation, we compare against R_{rough}, the mean spatial gradient magnitude of the three normal components, and additionally report partial correlations that control for roughness.

3.3 Models, data, and calibration

We evaluate four intrinsics-agnostic normal predictors spanning different training paradigms: Marigold-Normals, StableNormal-turbo, Lotus-normal-g, and MoGe-2-vitl (Ke et al., 2025; Ye et al., 2024; He et al., 2025; Wang et al., 2025). The first three are diffusion-based, whereas MoGe-2 uses discriminative ViT/DINOv2 geometry regression.

For controlled analysis, we render 10 Google Scanned Objects (GSO) (Downs et al., 2022) under three view and lighting configurations, yielding 30 orthographic scenes. For real-world evaluation, we use the NYUv2 validation set (Silberman et al., 2012). The detailed GBR diagnostic uses 30 NYUv2 images; the wrong-image control and pixel cross-fitting use the same 100 randomly sampled validation images across all four models. The prediction-only reliability experiments use 500 evenly spaced validation images for each model.

We normalize model-specific axis and sign conventions before evaluation. In the reference-dependent global audit, convention alignment is part of the audit; for pixel cross-fitting, both convention selection and GBR fitting are restricted to the fitting half before evaluation on the held-out half. For the deployment-style reliability experiments, we estimate a *single fixed coordinate convention per model* using six NYUv2 calibration images disjoint from the 500-image evaluation set and freeze that convention thereafter. The selected conventions are well separated from the runner-up alternatives, with calibration margins of $15.1°$–$21.1°$. The test-time curl score therefore uses neither per-image calibration nor reference normals from the evaluation set.

4 Global Errors Concentrate in a Low-Shear GBR Regime

We first ask where prediction–reference pairs lie in the global GBR coordinates. The evidence comes from three complementary settings: continuous fitting on real NYUv2 scenes, correspondence-destroying controls, and controlled orthographic GSO renders. Together they reveal a reproducible low-shear regime across architectures.

4.1 Continuous fitting reveals a low-shear concentration

Continuous optimization resolves the low-shear structure within the coarse audit grid. On the 30-image NYUv2 diagnostic subset, the median fitted shear is approximately 0.11–0.17 across the four architectures (Table 1), placing all four model families close to the relief axis across the $[-2, 2]$ shear range. The coarse audit independently shows the same concentration: 69.6–79.6% of evaluations land at zero for both shear coordinates and 98.2–99.6% at zero in at least one coordinate. Continuous fitting then resolves variation within this dominant cell, enabling quantitative comparisons across models, controls, and datasets.

To quantify how much of the GBR correction lies along the relief direction, we report η_λ, the per-image fraction of the full-GBR improvement explained by a λ-only fit.

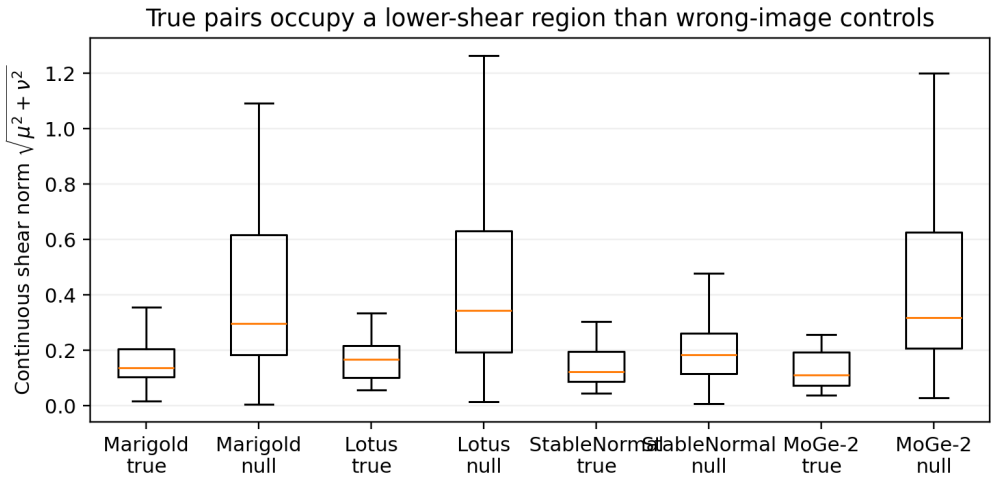

Figure 2: **True prediction–reference pairs occupy a lower-shear region than wrong-image controls.** Visualization on the 30-image diagnostic subset; Table 2 reports the $N = 100$ result.

Table 1: **Detailed continuous GBR diagnostic on NYUv2 ($N = 30$ maps/model).** True pairs have low continuous shear. η_λ is the per-image fraction of the full-GBR improvement already explained by a λ-only fit.

Model	Median shear	Median η_λ	Median base/full error
Marigold	0.139	0.718	35.19 / 32.71
Lotus	0.164	0.451	37.98 / 35.24
StableNormal	0.126	0.937	37.16 / 28.07
MoGe-2	**0.109**	0.737	33.74 / 31.41

The η_λ values reveal model-specific structure within this shared regime. StableNormal is strongly relief-scaling dominated, while Lotus gains more from the shear coordinates. Across models, however, the fitted transformations consistently remain near the low-shear part of the GBR family.

4.2 Breaking correspondence moves fits away from the low-shear regime

We next test whether the low-shear concentration is tied to matched prediction–reference geometry using two correspondence-destroying controls: (i) a wrong-image null, in which a prediction is fit to the reference geometry of another NYUv2 image, and (ii) a spatial block null, which disrupts within-image correspondence while preserving coarse field statistics.

Across all four models, destroying correspondence increases fitted shear on 100 randomly sampled NYUv2 images (Table 2). For the wrong-image null, the median shear is 1.74–2.61$\times$ the true-pair median, and image-bootstrap 95% confidence intervals for the ratio exclude one for every model. The spatial null, evaluated on the 30-image diagnostic subset, produces the same directional effect. The agreement across architectures and across two distinct null constructions ties the low-shear regime to matched prediction–reference geometry.

4.3 Controlled orthographic scenes strengthen the low-shear pattern

If perspective effects and real-scene complexity broaden the fitted transformation, cleaner orthographic geometry should lie closer to the pure-relief axis. This is what we observe on the controlled GSO renders. Median continuous shear ranges from 0.052 to 0.107 across all four models (Table 3), below the corresponding NYUv2 values.

Together, these experiments reveal a graded structure: controlled orthographic object geometry lies closest to the pure-relief axis, real perspective indoor scenes broaden the fitted optimum, and

Table 2: **True-vs-wrong-image continuous shear on NYUv2 ($N = 100$/model).** Each wrong-image value is the per-image median over ten mismatched reference maps; intervals bootstrap images.

Model	True med.	Wrong med.	Wrong/true	95% CI
Marigold	0.153	0.313	2.04×	[1.74, 2.41]
Lotus	0.190	0.330	1.74×	[1.49, 2.10]
StableNormal	0.124	0.216	1.75×	[1.45, 2.04]
MoGe-2	0.118	0.309	**2.61**×	[2.22, 3.09]

Table 3: **Continuous GBR structure on controlled GSO renders ($N = 30$/model).** Bootstrap intervals resample images.

Model	Median shear [95% CI]	Median η_λ
Marigold	0.107 [0.077, 0.134]	0.841
Lotus	0.079 [0.065, 0.114]	0.863
StableNormal	0.084 [0.069, 0.119]	0.775
MoGe-2	**0.052** [0.034, 0.068]	0.576

correspondence-destroying controls broaden it further. Thus the main finding is a reproducible low-shear concentration, supported by continuous fitting and strengthened by two complementary controls.

5 The Relief Coordinate Generalizes to Held-Out Pixels

The low-shear result identifies a shared region of GBR space. We next ask whether position along its relief direction carries information that transfers beyond the pixels used to estimate it. Pixel cross-fitting provides a direct test of this spatial generalization.

We test this directly with pixel cross-fitting on the same 100 randomly sampled NYUv2 images. For each image and model, we perform five random 50/50 pixel splits and evaluate both directions. Coordinate calibration and the continuous GBR fit are determined using only the fitting half, while angular error is measured on the disjoint held-out half. We then aggregate the fitted $\log \lambda$ and held-out error at the image level.

The relationship survives the held-out pixel split in all four architectures on the 100-image NYUv2 sample (Table 4). Correlations between cross-fitted $\log \lambda$ and held-out baseline error range from -0.598 to -0.699, with image-bootstrap 95% confidence intervals excluding zero for every model. On the original 30-image diagnostic subset, applying the transform estimated on the fitting pixels also reduces mean angular error on the held-out pixels by $2.09°$–$9.38°$. Together, these results establish the fitted global coordinate as an image-level property that generalizes to disjoint pixels.

The cross-fit result turns λ from a descriptive fit parameter into a transferable audit coordinate: its image-level variation predicts error in regions that played no role in estimating the transform.

6 Local Prediction Geometry Provides Test-Time Reliability

The global analysis above characterizes error relative to reference geometry. We now move to test-time reliability, where the prediction itself supplies the signal. We evaluate the prediction-only curl risk of Eq. (6) on 500 NYUv2 images per model, using the single fixed model-level coordinate convention described in Section 3.

6.1 Local curl predicts high-error outputs

Across all four models, the curl score tracks angular error and, more importantly, identifies high-error predictions substantially better than generic normal roughness. Table 5 reports Pearson correlation, partial correlation after controlling for roughness, AUROC for detecting the worst 10% of predictions, and area under the risk–coverage curve (AURC).

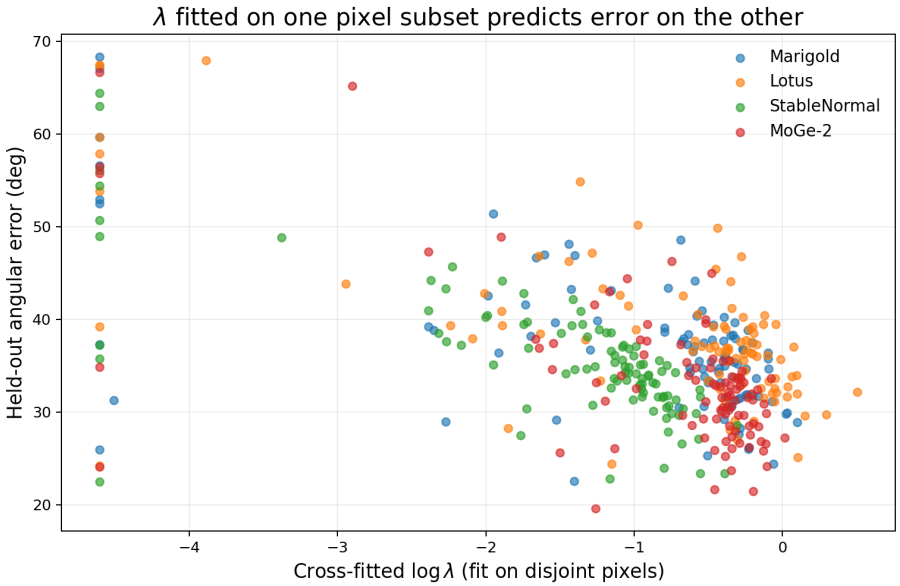

Figure 3: **The fitted relief coordinate generalizes to disjoint pixels.** Each point is an image-level cross-fit aggregate. Smaller $\log \lambda$ predicts larger angular error on pixels that were not used to estimate the GBR transform.

Table 4: **Pixel cross-fitting on NYUv2 ($N = 100$/model).** The GBR transform is estimated on one half of the valid pixels and evaluated on the disjoint half; intervals bootstrap images.

Model	$r(\log \lambda, \text{held-out error})$	95% CI
Marigold	-0.598	[-0.799, -0.319]
Lotus	-0.635	[-0.824, -0.341]
StableNormal	**-0.699**	[-0.876, -0.435]
MoGe-2	-0.608	[-0.821, -0.272]

Worst-10% AUROC ranges from 0.772 to 0.913 for curl, compared with 0.593 to 0.741 for roughness. The curl score also retains positive partial correlations with angular error after controlling for roughness, and image-bootstrap 95% confidence intervals for the AUROC improvement exclude zero for every model. Together, these results show that curl contributes reliability information beyond generic spatial roughness. The ranking remains strong under complementary operating points: worst-10% AUPRC ranges from 0.405 to 0.707, and worst-20% AUROC ranges from 0.771 to 0.863 across the four models (Appendix C).

For the deployment-style evaluation, each model uses a single coordinate convention calibrated once on six disjoint images and then frozen. All reported test-time scores therefore use only the predicted normal field, with no per-image reference geometry.

6.2 Geometric risk supports selective prediction

A useful reliability score should do more than correlate with error: rejecting predictions assigned high risk should improve the quality of the retained set. We evaluate this in the standard selective-prediction setting (Geifman and El-Yaniv, 2017; 2019).

Figure 5 shows risk–coverage curves obtained by ranking images with the curl score. Across all four models, retained mean angular error decreases as high-risk predictions are rejected, and curl-based rejection consistently outperforms roughness-based rejection at matched coverage. For example, MoGe-2's retained mean error falls to 33.82°, 33.07°, and 32.38° after rejecting the top 10%, 20%, and 30% highest-risk predictions, respectively.

Table 5: **Prediction-only reliability on NYUv2 ($N = 500$/model).** Partial r controls for normal roughness. Lower AURC is better. All 95% confidence intervals for the AUROC10 gain exclude zero.

Model	r / partial r	AUROC10 curl / rough	95% CI gain	AURC curl / rough
Marigold	0.580 / 0.437	0.813 / 0.741	[0.019, 0.130]	33.71 / 34.36
Lotus	0.557 / 0.526	0.772 / 0.593	[0.106, 0.237]	35.69 / 37.51
StableNormal	0.617 / 0.544	0.871 / 0.708	[0.096, 0.220]	35.07 / 35.91
MoGe-2	**0.763 / 0.709**	**0.913** / 0.740	[0.102, 0.244]	**31.13** / 32.93

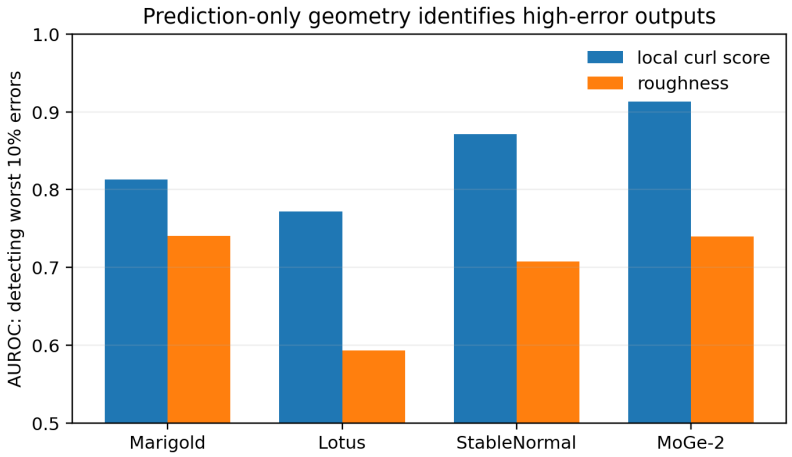

Figure 4: **Local geometry identifies the highest-error predictions.** AUROC for detecting the worst 10% of images by angular error. The prediction-only curl score outperforms generic normal roughness for every architecture.

The global and local analyses therefore play complementary roles. GBR fitting reveals a reproducible structure in prediction error when reference geometry is available for auditing, while local curl turns prediction geometry itself into a lightweight, reference-free ranking signal at test time.

7 Discussion

The experiments reveal a common geometric organization of monocular normal error across architectures that differ substantially in training paradigm and model design. At the global scale, fitted GBR transforms repeatedly concentrate near the relief direction; the same structure weakens when correspondence is destroyed and tightens on controlled orthographic scenes. At the local scale, curl exposes a complementary consistency signal directly from the prediction. Together, the two views turn "error geometry" into an object that can be measured, compared across models, and used for reliability.

Shared structure with model-specific signatures. The global coordinates expose both regularity and variation. StableNormal is especially relief-dominated, with a median η_λ of 0.937 on NYUv2, whereas Lotus receives substantially more benefit from the shear coordinates. Yet all four architectures occupy the same low-shear region relative to correspondence-destroying controls. This suggests that GBR coordinates can separate a shared mode of geometric error from architecture-specific mixtures within that mode, providing information that aggregate angular accuracy does not reveal.

From geometry to reliability. The held-out and local experiments make the geometric coordinates predictive as well as descriptive. Cross-fitted λ tracks error on spatially disjoint pixels and transfers to improve them, while prediction-only curl outperforms generic roughness; paired TTA/repeat-inference checks and iBims-1 transfer provide additional support (Appendix D). The local score is computed from a single normal map, requires no model retraining or repeated inference, and can

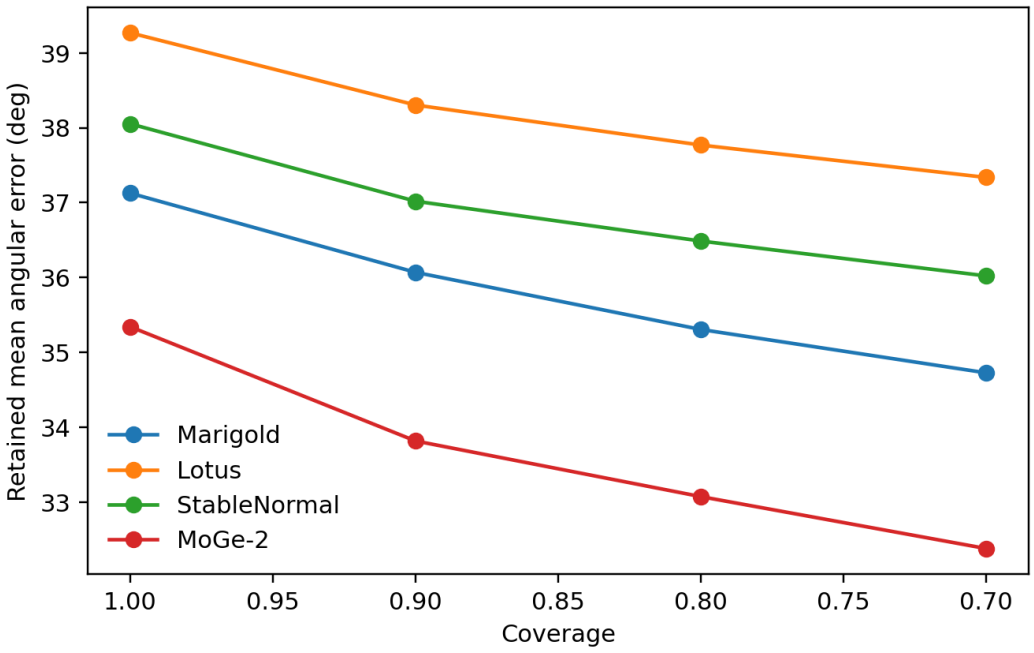

Figure 5: **Selective prediction from local geometry.** Retaining only lower-risk images reduces angular error for every model. Curves use the fixed-convention curl score and require no per-image reference geometry at inference.

therefore be attached directly to existing predictors. Its post-hoc, architecture-agnostic form also makes the reliability signal directly comparable across diffusion and discriminative systems without modifying their training objectives or prediction heads. These results point to a broader opportunity: the geometry already present in dense predictions can serve as a practical source of reliability.

Broader directions. The present results connect a reference-based global audit with a reference-free local score. A natural next step is to predict global failure coordinates directly from model representations or a single output field, extending reference-free reliability from local to global geometry. The same coordinate view can also be applied to depth, point maps, multi-view geometry, and broader camera and scene domains.

8 Conclusion

Angular error summarizes the magnitude of a surface-normal prediction error; our results show that the error also has a reproducible geometric form. Across four modern monocular normal estimators, global prediction–reference discrepancies concentrate in a shared low-shear GBR regime, the fitted relief coordinate generalizes to disjoint held-out pixels, and local inconsistency inside a single predicted normal field provides a strong test-time reliability signal. These findings make failure geometry useful at two levels: as an interpretable coordinate system for understanding model behavior and as a practical source of confidence for selective prediction. Because both coordinates are post-hoc, the framework applies across heterogeneous predictors without retraining: GBR supplies a common error coordinate when reference data are available, while curl supplies an output-only reliability layer at deployment. More broadly, the results suggest that dense geometric predictions should be evaluated not only by how far they are from reference geometry, but also by the structure of the deviations they make.

Reproducibility statement

The main text specifies the models, datasets, continuous GBR objective, null construction, cross-fit protocol, fixed model-level calibration protocol, and reliability metrics. The appendix records the exact fixed coordinate transforms, continuous-optimization details, null construction, and complementary reliability statistics.

AI use statement

Large language models were used to assist with manuscript organization and language editing. The authors verified and take responsibility for the scientific claims, experimental results, mathematical statements, citations, and final manuscript content.

References

Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5965–5974, 2016.

Peter N. Belhumeur, David J. Kriegman, and Alan L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.

Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

David F. Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3392–3399, 2013.

Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):439–451, 1988. doi: 10.1109/34.3909.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2151–2159, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330, 2017.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, 2025.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. doi: 10.1109/TPAMI.2024.3444912.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413, 2017.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025.

Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations*, 2025.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yutong Wu, Zilong Dong, Li Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics*, 43(6):1–18, 2024.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.

A Implementation and Reproducibility Details

Coordinate conventions. The fixed model-level transforms, written as `[swap_xy, sign_x, sign_y, sign_z]`, are: Marigold `[false,1,-1,-1]`, Lotus `[false,-1,-1,-1]`, StableNormal `[false,-1,-1,-1]`, and MoGe-2 `[false,1,1,1]`. They were selected on NYUv2 IDs 67, 135, 292, 360, 517, and 585, none of which appear in the 500-image reliability evaluation. The selected convention is frozen for every evaluation image of a model.

Continuous GBR optimization. We initialize from a discrete audit with 30 log-spaced values for $\lambda \in [0.01, 8]$ and a 9×9 shear grid over $[-2, 2]^2$, then run bounded continuous optimization of Eq. (3). The discrete stage provides a stable initialization, while the continuous stage resolves the fine structure of the low-shear region used throughout the main analysis.

Null construction and statistical inference. For the 100-image wrong-image null, each prediction is paired with ten reference normal maps from different NYUv2 images before continuous GBR fitting, and the per-image median shear is used for the summary. For the spatial null, block permutation disrupts within-image pixel correspondence while retaining coarse field statistics on the 30-image diagnostic subset. Paired tests use the image as the exchange unit. Confidence intervals reported for shear and reliability statistics are obtained by image bootstrap.

Pixel cross-fitting. For each of the 100 randomly sampled NYUv2 images, we draw five random 50/50 partitions of valid pixels and evaluate both directions for each model. Coordinate convention selection and GBR optimization are performed using only the fitting subset, after which λ, baseline error, and transformed error are evaluated on the disjoint subset. We aggregate the ten directional fits to one image-level value before computing the correlations in Table 4.

Reliability evaluation. The curl experiments use 500 evenly spaced NYUv2 validation images per model, excluding the six calibration images. Worst-$k\%$ labels are defined separately for each model from its angular-error distribution. AUROC and AUPRC evaluate ranking quality, while AURC summarizes the full risk–coverage curve. The roughness baseline is the mean spatial gradient magnitude of the three predicted normal components.

B Resolving the Low-Shear Regime

The initial discrete grid places 69.6–79.6% of NYUv2 evaluations at the zero grid point for both shear coordinates and 98.2–99.6% at zero in at least one coordinate. Continuous optimization resolves variation inside this coarse cell, yielding median shear norms of 0.109–0.164. The two views are consistent: the discrete audit reveals a strong concentration near the relief direction, while the continuous fit provides the resolution needed to compare models, controls, and datasets quantitatively.

This resolution also explains why continuous fitting is central to the main-text claim. The informative signal is the *location and spread* of prediction–reference pairs near the relief axis, not occupancy of one discrete grid point. The continuous coordinates expose that structure without quantization from the 0.5 shear spacing of the initialization grid.

C Additional Reliability Statistics

The fixed-calibration curl score also outperforms roughness on AUPRC for worst-10% failure detection. The AUROC10/AUPRC10 pairs are: Marigold 0.813/0.475, Lotus 0.772/0.405, StableNormal 0.871/0.568, and MoGe-2 0.913/0.707. Worst-20% AUROCs are 0.771, 0.788, 0.818, and 0.863, respectively. Across all four models, retained mean angular error decreases monotonically when rejecting the top 10%, 20%, and 30% of images ranked by curl risk.

Table 6: **Complementary reliability operating points.** All scores use the single fixed model-level coordinate convention.

Model	AUROC10	AUPRC10	AUROC20
Marigold	0.813	0.475	0.771
Lotus	0.772	0.405	0.788
StableNormal	0.871	0.568	0.818
MoGe-2	**0.913**	**0.707**	**0.863**

D Stronger Reliability Baselines and Cross-Dataset Transfer

Table 7 reports the paired robustness evaluation used to compare curl with stronger consistency baselines. TTA consistency uses flip/photometric augmentations and requires two additional forward passes. Repeat-inference consistency is reported where repeated model outputs are available. The iBims-1 evaluation keeps the NYUv2-calibrated coordinate convention fixed and uses 100 scenes. Because the consistency comparison requires matched augmented or repeated predictions, it uses a fresh paired inference run under the finalized model configurations; its absolute AUROC values can therefore differ from the primary reliability run in Table 5.

Table 7: **Stronger reliability baselines and cross-dataset transfer.** Entries are worst-10% AUROC. A dash indicates that repeated predictions were not available.

Model	NYU curl	NYU TTA	NYU repeat	iBims curl	iBims rough
Marigold	0.813	0.688	0.728	0.881	0.706
Lotus	0.772	0.734	0.659	0.896	0.672
StableNormal	0.848	0.689	0.703	**0.930**	0.534
MoGe-2	**0.858**	0.691	–	0.873	0.683

E Protocol Summary and Metric Definitions

Table 8 summarizes the role and scale of each experiment. The global audit and held-out analyses use reference geometry because their purpose is to characterize prediction–reference structure. The deployment-style reliability evaluation uses only the predicted normal map after the one-time six-image model-level calibration described in the main text.

Table 8: **Experimental protocol summary.** Counts are per model.

Experiment	Data	Scale	Role
Continuous GBR diagnostic	NYUv2	30 images	detailed global coordinates
Wrong-image null	NYUv2	100 images	correspondence control
Spatial null	NYUv2	30 images	complementary control
Controlled GBR audit	GSO	30 renders	orthographic geometry
Pixel cross-fitting	NYUv2	100 images	held-out transfer
Prediction-only reliability	NYUv2	500 images	test-time ranking

For a valid pixel u, angular error is

$$e_{\mathrm{ang}}(u) = \arccos\Big(\mathrm{clip}(\hat{N}(u)^\top N(u), -1, 1)\Big). \tag{7}$$

Image-level angular error is the mean over valid pixels. The continuous GBR shear magnitude is $s_{\mathrm{shear}} = \sqrt{\mu^2 + \nu^2}$. The relief fraction η_λ is computed per image as the improvement from the best λ-only fit divided by the improvement from the full continuous GBR fit. Pearson correlation is computed across images. AUROC and AUPRC treat the worst-$k\%$ images by angular error as positives, while AURC integrates retained angular risk over coverage.

F Full Fixed-Calibration Reliability Results

Table 9: **Fixed-calibration reliability metrics on NYUv2.** Curl uses one model-level coordinate convention calibrated on six disjoint images. Roughness uses the same evaluation images.

Model	Pearson r	Partial r	AUROC10 / rough	AUROC20 / rough
Marigold	0.580	0.437	0.813 / 0.741	0.771 / 0.693
Lotus	0.557	0.526	0.772 / 0.593	0.788 / 0.602
StableNormal	0.617	0.544	0.871 / 0.708	0.818 / 0.691
MoGe-2	0.763	0.709	0.913 / 0.740	0.863 / 0.673

G Unified Global Geometry Summary

The main text presents the global experiments in the order needed for the argument. Table 10 collects their principal statistics in one place. The same pattern is visible across all four architectures: low continuous shear on matched NYUv2 pairs, a substantial increase after wrong-image pairing, lower shear on controlled orthographic GSO renders, and a relief coordinate that transfers to held-out pixels.

Table 10: **Global geometry summary across datasets and controls.** NYU shear, Wrong/true, and Held-out r use the 100-image NYUv2 sample; Held-out Δ is the transfer improvement on the original 30-image diagnostic subset.

Model	NYU shear	Wrong/true	GSO shear	Held-out r	Held-out Δ
Marigold	0.153	2.04×	0.107	-0.598	2.89°
Lotus	0.190	1.74×	0.079	-0.635	2.57°
StableNormal	0.124	1.75×	0.084	**-0.699**	9.38°
MoGe-2	**0.118**	**2.61**×	**0.052**	-0.608	2.09°

The table also highlights complementary model signatures inside the shared pattern. StableNormal has the strongest held-out relief relationship, while MoGe-2 lies closest to the relief axis and separates matched from wrong-image pairs most strongly by shear. These differences are useful precisely because the coordinates are common across otherwise heterogeneous predictors.

H Matched Color Examples of Relief and Local Error

Figures 6–8 show color inputs, geometry-derived reference normals, and Marigold predictions for six matched views of five objects. The RGB, reference, and prediction in each row share the same camera view and foreground mask. Several mugs have smooth GT surfaces but textured or altered predicted normals, whereas the planar sink cabinet is reproduced more closely. The fitted λ values range from strong relief contraction to approximately identity, including two views of the same red mug on opposite sides of $\lambda = 1$. These supplementary perspective renders are selected for visual inspection and are separate from the orthographic GSO cohort used for the quantitative results in Table 3.

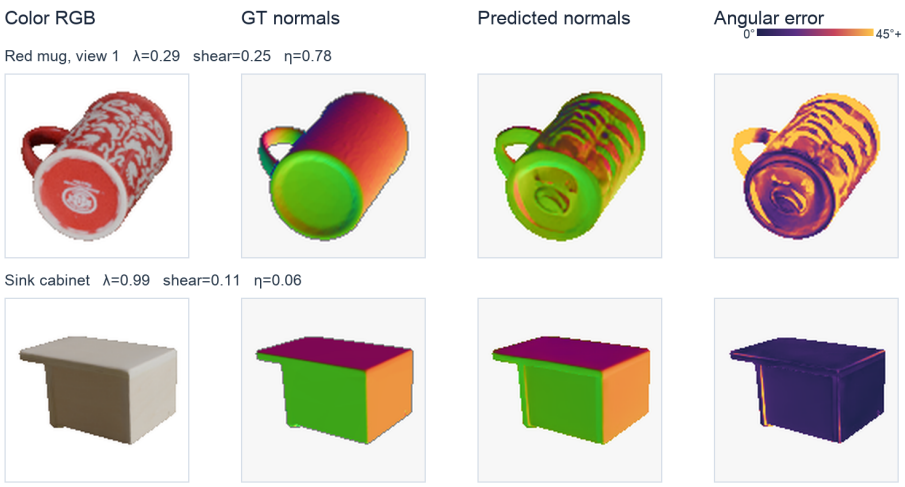

Figure 6: **Matched color examples: textured mug and planar cabinet.** Each row shares a camera view across RGB, GT normals, Marigold normals, and angular error. All panels use the same normal encoding and 0–45° error scale; invalid GT pixels are hidden. Labels give fitted relief λ, shear s_{shear}, and relief-only gain fraction η_λ.

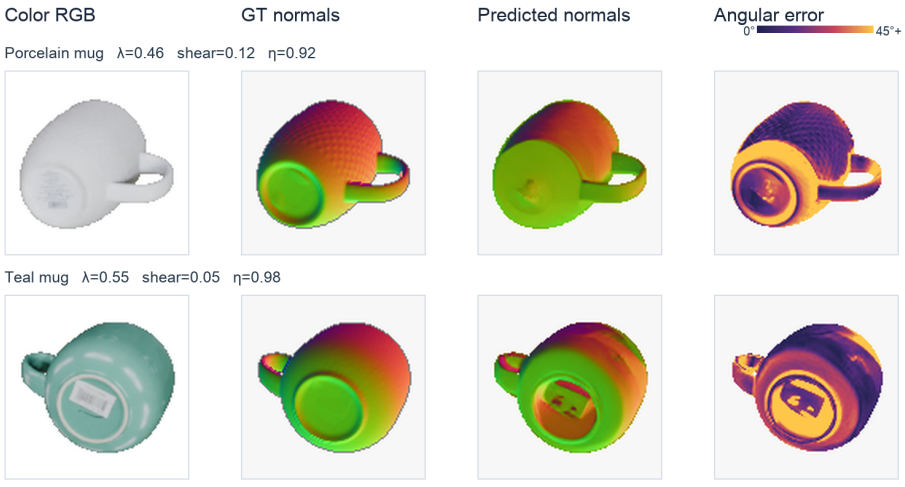

Figure 7: **Matched color examples: two other mugs.** Both fitted relief coordinates are below one. The GT and prediction use identical views and the same display scales as Fig. 6.

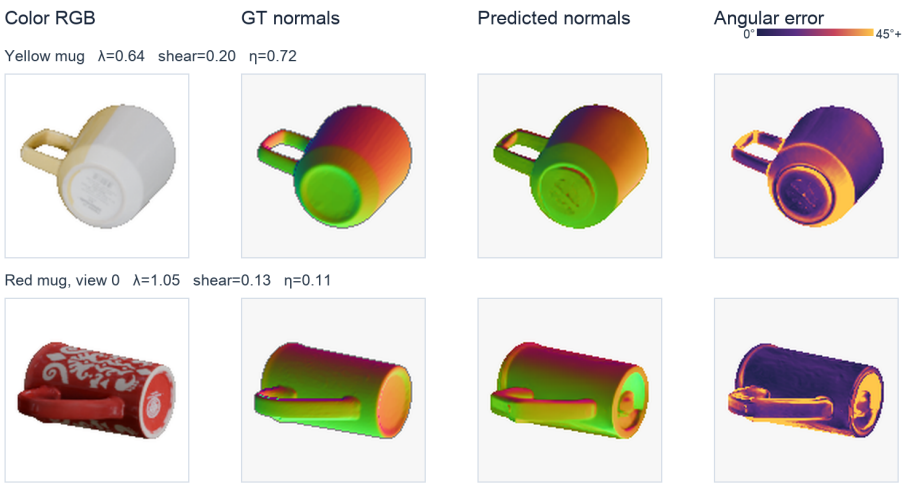

Figure 8: **Matched color examples: a second view of the red mug.** The yellow mug has $\lambda < 1$, while the red mug's second view has $\lambda > 1$. These selected examples illustrate variation rather than its frequency.

I Model and Evaluation Coverage

Table 11: **Coverage of the four evaluated model families.** The same released predictor is used throughout each model's global, held-out, and reliability analyses.

Model	Paradigm	GBR audit	Cross-fit	Reliability
Marigold-Normals	diffusion	30 NYU + 100 null + 30 GSO	100 NYU	500 NYU
StableNormal-turbo	diffusion	30 NYU + 100 null + 30 GSO	100 NYU	500 NYU
Lotus-normal-g	diffusion	30 NYU + 100 null + 30 GSO	100 NYU	500 NYU
MoGe-2-vitl	discriminative	30 NYU + 100 null + 30 GSO	100 NYU	500 NYU

This coverage is designed to test whether the failure coordinates persist across different modern prediction paradigms rather than within one architecture family. All four models participate in every principal analysis, so the global and local comparisons use a matched set of predictors throughout the paper.

J Coordinate Interpretation and Reporting Conventions

The identity GBR transform is $(\lambda, \mu, \nu) = (1, 0, 0)$. The relief coordinate λ therefore measures multiplicative scaling away from unit relief, while $s_{\text{shear}} = \sqrt{\mu^2 + \nu^2}$ summarizes the directional part of the fitted transform. The norm is unchanged by swapping the image axes or flipping their signs, which makes it a convenient scalar summary after model-specific coordinate conventions have been normalized.

For the held-out analysis we use $\log \lambda$, so the identity transform maps to zero and reciprocal relief changes lie on opposite sides of the same additive coordinate. For the contribution of the relief axis itself, let E_0 denote baseline angular error, E_λ the error after the best λ-only fit, and E_{full} the error after the full GBR fit. We report

$$\eta_\lambda = \frac{E_0 - E_\lambda}{E_0 - E_{\text{full}}}, \tag{8}$$

computed independently for each image before summarizing across the dataset. Values near one indicate that most of the available GBR improvement is already captured by movement along the relief axis.

Table 12: **Interpretation of the principal geometric quantities.**

Quantity	Reference value	Role in the paper
λ	1	relief scaling
$\log \lambda$	0	held-out image-level coordinate
s_{shear}	0	directional GBR displacement
η_λ	1	fraction of GBR gain from relief-only fitting
R_{curl}	lower is more consistent	prediction-only local risk

All inferential summaries use the image as the unit of analysis. Continuous-shear comparisons use per-image fitted coordinates, pixel cross-fitting is aggregated to one value per image after the ten directional splits, and reliability statistics are computed across the 500 evaluation images. This keeps the reported global and local statistics aligned with the image-level claims made in the main text.

K Additional Geometric Checks

We also compared the image-space curl score with perspective-correct and projection-based geometric variants on the auxiliary 30-image cache. Across these formulations, the image-space score used in the main paper gave the strongest and most consistent cross-model ranking signal. This supports the simple output-space construction in Eq. (6), which is shared across all four architectures and does not require camera intrinsics.

The global analysis provides a complementary consistency check. The discrete audit, continuous fitting, correspondence-destroying nulls, controlled orthographic renders, and held-out pixel transfer all probe different consequences of the same global coordinate system. Their agreement motivates treating the low-shear GBR region as a reusable description of prediction–reference error across models and data conditions.