PHYSPROBE: PROBING PHYSICAL-MAP RECOVERY IN FROZEN VISION GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

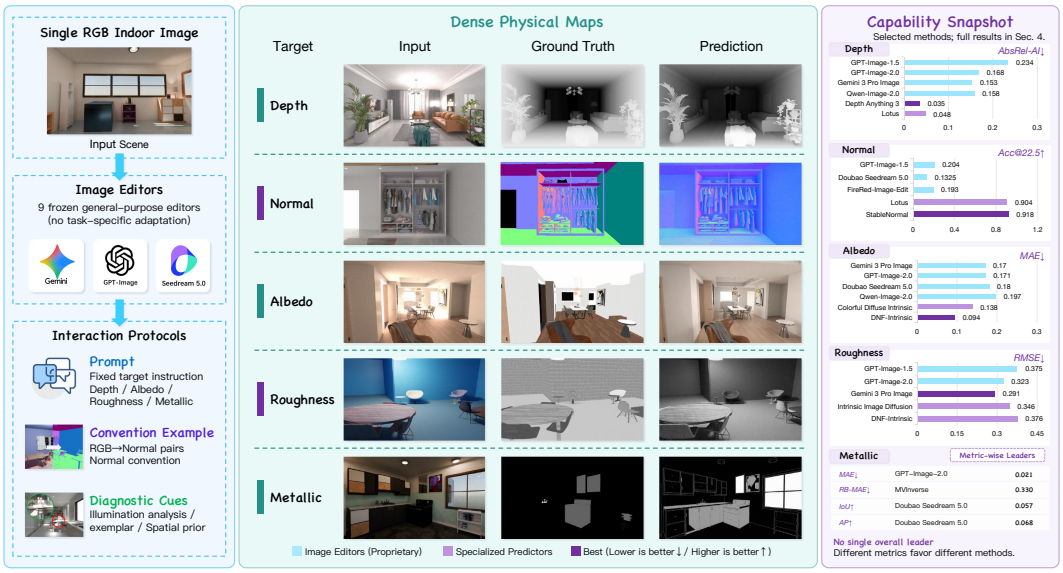

Figure 1: **Motivation and result snapshot of PhysProbe.** PhysProbe evaluates whether frozen general-purpose image editors can recover dense physical maps from a single RGB indoor image without task-specific training. **Left:** target-specific access protocols and separately evaluated diagnostic cues. **Middle:** representative input, ground-truth, and predicted maps for depth, surface normal, albedo, roughness, and metallic. **Right:** selected main-axis results, including the primary metrics for depth, normal, albedo, and roughness and a multi-metric metallic profile; the metallic MAE entry reflects full-map background fidelity and is not evidence of metallic-region recovery. Complete results are reported in Section 4.1, and the full benchmark design is summarized in Figure 2.

ABSTRACT

Physical perception requires recovering the scene factors behind an image, beyond producing a plausible RGB transformation. Can frozen general-purpose image editors expose these factors through their native interfaces? We introduce PhysProbe, a protocol-aware framework that evaluates depth, surface normals, albedo, roughness, and metallicity against dense physical references. Each score describes a model-protocol pair with fixed inputs, output conventions, metrics, and aggregation. The results reveal three separations: local depth-boundary recovery does not imply accurate relative geometry; low full-map material error does not imply successful localization of sparse metallic regions; and additional context does not uniformly improve physical recovery. Specialists retain large advantages in relative geometry, surface orientation, and albedo. Gemini achieves the strongest reported roughness metrics under the evaluated protocol, while metallic evaluation reveals that full-map fidelity and sparse-region localization can favor different models. Normal exemplars and auxiliary cues can further change observed rankings across models, showing that physical recovery depends on both target properties and access protocols rather than a single model ordering. PhysProbe probes the physical-perception component of Physical AI while separating observed output fidelity from hypotheses about internal representations or training.

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.

1 INTRODUCTION

General-purpose image models, now combining understanding and generation, have recently demonstrated strong scene manipulation capabilities. Modern image editors can preserve scene layout, modify appearance, and produce visually coherent transformations from images and text instructions Brooks et al. (2023); Zhang et al. (2023); Sheynin et al. (2024); OpenAI (2026); Google (2025). These capabilities raise a broader question: whether general-purpose image models already encode physical factors underlying image formation, including scene geometry, surface orientation, and material properties Wang et al. (2024); Liu et al. (2025); Gabeur et al. (2026). However, such physical knowledge is difficult to evaluate directly. Unlike task-specific predictors with explicit output heads, generative models expose their capabilities through instructions, editing behaviors, and generated representations rather than predefined physical interfaces. Dense physical prediction provides a quantitative probe because it requires recovering spatially aligned variables that explain image formation Kocsis et al. (2024); Lin et al. (2025a); Zheng et al. (2025). We therefore ask: which physical scene factors can off-the-shelf image editors (frozen parameters) recover without task-specific adaptation?

This question matters for the perception component of Physical AI, where recovered geometry and material properties support scene reconstruction and reasoning about appearance formation. A plausible RGB edit need not separate these factors correctly. We therefore examine whether visual structure translates into accurate physical values, whether material predictions go beyond dominant-background fidelity, and whether these behaviors remain consistent when the access protocol changes.

Measuring editor-accessible physical signals is challenging because these models are not accessed through fixed prediction interfaces. Unlike task-specific predictors that map RGB inputs directly to target representations, editors expose their capabilities through generative interfaces involving instructions, exemplars, auxiliary cues, and output conventions Brooks et al. (2023); Zhang et al. (2023); Sheynin et al. (2024). Recent efforts have adapted image editing models into dense predictors through task-specific training or additional supervision Wang et al. (2026); Shi et al. (2025), demonstrating that editing models contain useful representations after adaptation but leaving open what physical knowledge is accessible from frozen models. Evaluating these capabilities therefore requires specifying the interaction protocol through which the model is queried. Each score characterizes a model-protocol pair rather than a prompt-independent property of the model.

To answer this question, we introduce **PhysProbe**, a protocol-aware framework that uses dense recovery of physical variables to measure capabilities exposed by frozen image editors. PhysProbe evaluates five physical targets under fixed access protocols, output conventions, and scoring rules. Prompt, exemplar, auxiliary-cue, and photometric-stress variants are analyzed separately as diagnostics of capability elicitation rather than used to select favorable evaluation settings. PhysProbe builds on aligned inverse-rendering sources with reliable annotations, including OpenRooms-FF Li et al. (2021); Choi et al. (2023) and InteriorVerse Zhu et al. (2022), and introduces PBR-Rooms as a companion source for metallic evaluation.

Our study reveals separable physical-recovery behaviors. Gemini combines strong local depth-boundary scores with substantially larger affine-aligned depth errors than specialists, distinguishing contour recovery from accurate scene geometry. GPT-Image-2.0 improves on GPT-Image-1.5 on all reported depth metrics, but its low metallic MAE does not translate into the best localization. Gemini's roughness advantage also appears in range- and region-balanced metrics, motivating closer study of the material cues available to frozen editors. Finally, exemplars and auxiliary analysis can help one model while harming another, including reversing normal-model rankings. These observations identify what each interface exposes; they do not isolate architectural causes or establish which physical tasks appeared in a model's training data.

This paper makes the following contributions:

- We introduce **PhysProbe**, a controlled probing framework that uses dense recovery of depth, surface normals, albedo, roughness, and metallicity to measure physical signals accessible through frozen image editors, with task-specific predictors serving as specialist references.
- We identify separations between local boundary recovery and relative-geometry accuracy, and between full-map material fidelity and sparse-region localization, while finding competitive editor roughness estimates under the evaluated protocol.

- We document model-dependent responses to prompts, exemplars, and photometric conditions, including protocol-induced rank reversals, motivating evaluation of explicitly specified model-protocol pairs rather than a prompt-independent capability hierarchy.

2 Related Work

Generative models as visual and physical perceivers. Recent work studies 3D awareness, material reasoning, and perception-oriented behavior in generative or multimodal models Wang et al. (2024); Guo et al. (2025); Liu et al. (2025); Gabeur et al. (2026). PhysProbe focuses on frozen image editors and measures dense physical-map recovery through explicitly fixed interfaces, complementing work on editing quality and instruction following Brooks et al. (2023); Zhang et al. (2023); Sheynin et al. (2024); OpenAI (2026); Google DeepMind (2025); Labs et al. (2025); Meituan LongCat Team et al. (2025).

Physical scene understanding with specialized predictors. Dense physical prediction has been extensively studied using datasets with pixel-aligned supervision for geometry, intrinsic appearance, and material properties. Representative resources, including OpenRooms, OpenRooms-FF, InteriorVerse, and Hypersim, provide physically grounded annotations for inverse-rendering tasks Li et al. (2021); Choi et al. (2023); Zhu et al. (2022); Roberts et al. (2021). Building on these resources, specialized predictors achieve strong performance in depth, surface-normal estimation, intrinsic decomposition, and material recovery Lin et al. (2025a); Zeng et al. (2024); Wang et al. (2025); Zheng et al. (2025); He et al. (2025); Ye et al. (2024); Bae & Davison (2024); Kocsis et al. (2024); Liu et al. (2026a;b); Lin et al. (2025b); Careaga & Aksoy (2024); Wu et al. (2025). However, these methods rely on task-specific supervision and predefined RGB-to-map interfaces, and therefore do not measure the physical signals accessible through frozen general-purpose models. In PhysProbe, specialized predictors serve as references for physical recovery quality rather than the primary target of evaluation.

Probing frozen generative models through controlled interfaces. Prior work adapting editors for dense prediction measures behavior after task-specific training or supervision Wang et al. (2026); Shi et al. (2025). Other evaluations study 3D awareness and visual reasoning in generative or multimodal models Liu et al. (2025); Gabeur et al. (2026). Together, these lines leave the physical signals accessible from frozen image editors under controlled generative interfaces insufficiently characterized. PhysProbe treats allowed inputs, output conventions, and scoring as part of each model–protocol measurement and probes five physical factors through dense recovery.

3 PhysProbe: Protocol-Conditioned Physical-Map Recovery

3.1 Evaluation formulation

PhysProbe treats dense physical-map recovery as a protocol-conditioned output measurement. Given a single RGB observation, an editor is queried to produce a target representation without task-specific training, fine-tuning, or parameter updates. The resulting prediction is compared against aligned physical annotations to measure recovery fidelity through that interface.

The five targets probe geometry, surface orientation, intrinsic appearance, and material properties. Each has a fixed access contract specifying allowed inputs, output conventions, and scoring; separate diagnostics test how interaction context changes the observed capability.

Here, *frozen-interface recovery* means output obtained without adaptation in this study; scores jointly reflect scene inference, interpretation of the requested representation, and numerical image generation rather than any one component alone.

3.2 Controlled access protocols

PhysProbe evaluates frozen editors through controlled generative interfaces. All official protocols provide only the query RGB image and target definition, while excluding query-specific supervision, test feedback, or additional scene information. Depth, albedo, roughness, and metallic use the query

RGB image together with a target-specific instruction because their output representations can be directly specified.

For surface normals, visually similar maps can encode incompatible coordinate and RGB conventions. PhysProbe therefore provides three fixed, scene-disjoint RGB–normal exemplars to define the output convention rather than provide query-specific information. Each query is processed in three independent one-shot runs, one for each exemplar, and the official normal score averages these runs to reduce dependence on any single example.

Prompt-only normal variants and other access cues are analyzed separately as diagnostics; they do not change the fixed official protocols.

3.3 PHYSICAL PROBES AND EVALUATION

PhysProbe evaluates both proprietary and open-source general-purpose image editors together with target-specific dense predictors as specialist references. Editors are evaluated through their released inference interfaces without task-specific adaptation, while specialist predictors provide reference points for supervised physical recovery. Because generative editors may fail to produce complete or scoreable outputs under particular target protocols, the main comparisons include only complete official runs. Target-wise completion status, incomplete outputs, and representation failures are documented in Appendix Table 31.

No retained public source jointly provides benchmark-suitable supervision for all five targets under the required PBR semantics. Hypersim provides rich geometry, diffuse-reflectance, and illumination-related annotations, but not paired roughness and metallic channels in the PBR parameterization used here Roberts et al. (2021). OpenRooms and OpenRooms-FF cover depth, normal, albedo, and roughness but do not provide metallic ground truth Li et al. (2021); Choi et al. (2023). InteriorVerse has broader channel coverage, yet the presence of a channel alone is insufficient when its physical meaning is ambiguous for the intended metric; consistent with prior work and our screening, we retain it for depth, normal, and albedo but not roughness or metallic Zhu et al. (2022); Zeng et al. (2024); Guo et al. (2025).

PBR-Rooms complements rather than replaces these sources. It exports all five maps from a shared PBR scene representation and is used here only for metallic, where reliable parameter-level supervision, targeted positive-region sampling, and matched scene–camera lighting variations are otherwise unavailable. Source eligibility is determined per target and frozen before scoring; Appendix B.1.2 documents the construction and scope.

Table 1: Retained physical reference sources and aggregation rules for the official PhysProbe evaluation.

Target	Source	Scenes	Images	Aggregation
Depth / Normal / Albedo	OpenRooms-FF InteriorVerse	48 167	1125 2109	Source-balanced
Roughness	OpenRooms-FF	48	1125	Single-source
Metallic	PBR-Rooms	40	960	Single-source

Counts refer to the official main evaluation split. For depth, normal, and albedo, source-specific scores are computed separately on OpenRooms-FF and InteriorVerse and then macro-averaged. Roughness and metallic use one retained reference source each.

We construct scene-level evaluation subsets to reduce view redundancy while preserving source coverage. The main-axis split is used for official target-wise comparisons.

For depth, surface normal, albedo, and roughness, PhysProbe adopts target-specific metrics designed to capture different aspects of physical recovery. Primary metrics include AbsRel-AI for depth, angular errors and threshold accuracies for normals, and MAE/RMSE-based measures for albedo and roughness. Metallic prediction is evaluated using a joint MAE/RB-MAE/IoU/AP profile because sparse material regions create a mismatch between global numerical fidelity and region-level material recognition. The four metallic metrics are therefore reported separately rather than collapsed into a single scalar summary.

All metrics use fixed target-specific normalization, masking, aggregation, and failure-handling rules. Depth permits per-image scale, shift, and polarity alignment because the requested field is relative. Albedo, roughness, and metallic are absolute bounded outputs, so global brightness or scale offsets count as errors. Lotus and StableNormal use one shared normal-convention adapter frozen on a scene-disjoint calibration subset (Appendix D.5.2). Editors receive no normal-axis alignment because exemplars specify the encoding; their angular errors can therefore include convention errors as well as orientation errors.

3.4 Protocol diagnostics

Access-cue diagnostics vary prompts, exemplars, and auxiliary cues to measure protocol sensitivity; photometric stress analysis measures conditional performance while preserving the official access protocol. OpenRooms-FF and InteriorVerse use RGB-derived lighting slices, and PBR-Rooms uses renderer-defined lighting variations. Split construction and diagnostic settings are summarized in Appendix B.3 and Appendix E.

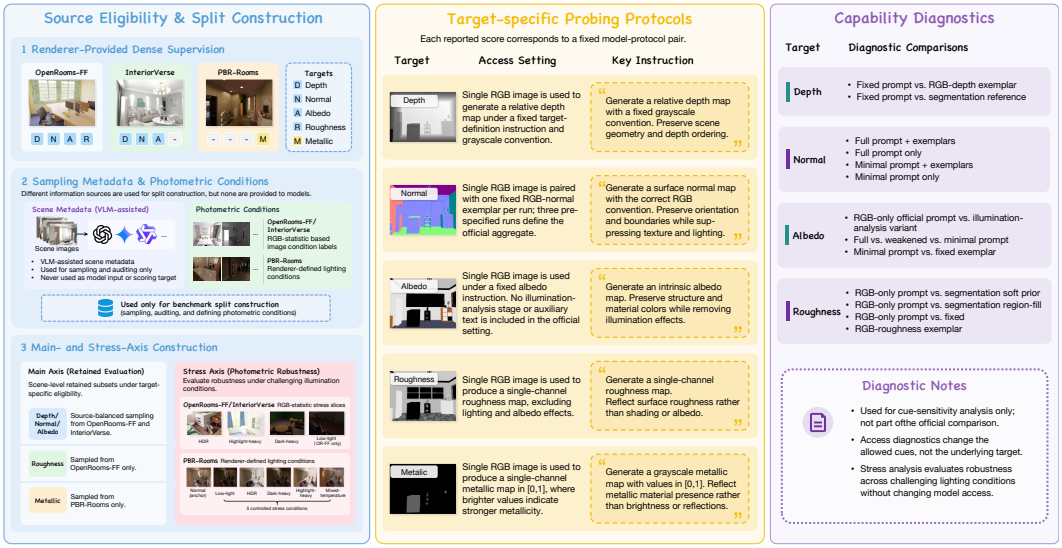

Figure 2: **Overview of PhysProbe.** PhysProbe measures physical-map recovery through the released interfaces of frozen image editors under controlled protocols. **Left:** probe construction and evaluation axes. Renderer-provided dense maps serve as physical references, while auxiliary metadata are used only for sampling and auditing. **Middle:** target-specific probing protocols define the allowed interaction context, output representation, and evaluation contract; normal uses fixed RGB–normal convention exemplars to specify representation semantics rather than query-specific evidence. **Right:** capability elicitation diagnostics analyze how prompts and auxiliary cues influence recovered physical representations without changing the official evaluation protocol.

4 Protocol-Conditioned Recovery in General-Purpose Editors

We first examine geometry and intrinsic appearance, followed by material properties and controlled diagnostics. Methods without complete, scoreable official runs are omitted from the corresponding comparisons rather than assigned artificial scores. Unless otherwise stated, boldface marks the best value using unrounded evaluator scores; displayed rounding may therefore produce apparent ties.

4.1 Geometry and Appearance Probes

Depth. Table 2 shows that the specialist gap remains after per-image scale, shift, and polarity alignment: Depth Anything 3 reaches AbsRel-AI 0.035 versus Gemini's 0.153. Yet Gemini's Boundary F1 is 0.322, close to Lotus's 0.318, whereas Qwen has a similar scalar error to Gemini (0.158) but much lower Boundary F1 (0.016). These behaviors separate recovery of local depth transitions from overall relative-depth accuracy; neither metric alone describes spatial fidelity. Boundary F1 is sensitive to

Table 2: Source-balanced relative-depth results under the affine-invariant single-image evaluation protocol.

Method	AbsRel-AI↓	RMSE-AI↓	MAE-AI↓	δ_1-AI↑	δ_2-AI↑	Boundary F1↑
GPT-Image-1.5	0.234	0.505	0.388	0.760	0.932	0.027
GPT-Image-2.0	0.168	0.370	0.288	0.838	0.952	0.272
Wan2.7-Image	0.166	0.359	0.278	0.862	0.962	0.182
Doubao Seedream 5.0	0.171	0.458	0.352	0.796	0.951	0.136
FLUX.1-Kontext-dev	0.389	0.836	0.678	0.523	0.811	0.003
FireRed-Image-Edit	0.357	0.811	0.661	0.532	0.819	0.063
Gemini 3 Pro Image	0.153	0.383	0.299	0.834	0.955	**0.322**
LongCat-Image-Edit	0.228	0.481	0.374	0.777	0.933	0.196
Qwen-Image-2.0	0.158	0.391	0.290	0.854	0.952	0.016
Depth Anything 3	**0.035**	**0.117**	**0.062**	**0.981**	**0.994**	0.240
Lotus	0.048	0.143	0.091	0.978	0.994	0.318

contour shifts and smoothing, and the small Gemini–Lotus difference is a point-estimate comparison rather than an established significant advantage (Appendix F.3).

Table 3: Source-balanced surface-normal recovery results under the official three-exemplar protocol.

Method	Mean↓	Median↓	Acc@11.25↑	Acc@22.5↑	Acc@30↑
GPT-Image-1.5	44.791	42.507	0.056	0.2042	0.326
Doubao Seedream 5.0	56.294	50.210	0.035	0.1325	0.227
FireRed-Image-Edit	61.371	64.417	0.096	0.1931	0.240
Lotus	9.590	**5.005**	0.810	0.9042	0.930
StableNormal	**9.115**	5.455	**0.818**	**0.9183**	**0.942**

Normal. Under the fixed three-exemplar protocol in Table 3, StableNormal reaches Acc@22.5 of 0.9183, compared with 0.2042 for the strongest editor with a complete official run. Of the nine evaluated editors, three have complete scoreable runs, GPT-Image-2.0 and Gemini are marked incomplete, and four have recorded run-level normal-representation failures (Appendix Tables 31 and 32). The completion rates and reasons for the two incomplete runs are not established by the available run-status summary. The numerical specialist gap therefore applies only to the three completely scored editors. Exemplar diagnostics further reverse some scored-editor rankings (Section 4.3).

Table 4: Source-balanced albedo results under the controlled RGB-only single-image evaluation protocol.

Method	MAE↓	PSNR↑	SSIM↑	LPIPS↓
GPT-Image-1.5	0.208	12.173	0.532	0.421
GPT-Image-2.0	0.171	13.594	0.555	0.417
Wan2.7-Image	0.223	12.014	0.602	0.390
Doubao Seedream 5.0	0.180	13.735	0.631	0.335
FLUX.1-Kontext-dev	0.247	10.672	0.493	0.583
FireRed-Image-Edit	0.209	12.432	0.635	0.437
Gemini 3 Pro Image	0.170	14.393	0.708	0.300
LongCat-Image-Edit	0.239	11.200	0.590	0.516
Qwen-Image-2.0	0.197	12.459	0.525	0.425
Colorful Diffuse Intrinsic	0.138	16.137	0.744	0.337
DNF-Intrinsic	**0.094**	**19.217**	**0.758**	**0.175**

Albedo. In Table 4, DNF-Intrinsic achieves source-balanced MAE 0.094, while Gemini and GPT-Image-2.0 have editor MAEs of 0.170 and 0.171 with overlapping scene-cluster intervals (Appendix Table 29); their ordering is therefore unresolved by the reported intervals. Their structural and perceptual scores differ more: Gemini has SSIM/LPIPS of 0.708/0.300 versus 0.555/0.417 for GPT-Image-2.0. Similar average color error can therefore conceal different spatial and perceptual fidelity. The source-specific results also reject a single specialist ordering: Colorful Diffuse Intrinsic leads most OpenRooms-FF metrics (MAE 0.126 versus DNF's 0.132), whereas DNF leads all InteriorVerse metrics (MAE 0.056 versus Colorful Diffuse Intrinsic's 0.149; Appendix F.2). The evaluator treats both reference and predicted albedo maps as normalized sRGB values and does not linearize them.

SSIM and LPIPS are computed in the valid-mask bounding box after invalid prediction pixels are filled with ground truth. Comparisons therefore condition on this display-space convention and masked-crop rule; they do not establish equivalence of the methods' original albedo parameterizations. Explicit illumination analysis also helps some editors and harms others (Section 4.3), so supplying a verbal account of lighting is not sufficient for accurate dense reflectance recovery.

DNF's albedo source dependence also tracks disclosed training exposure: it was trained only on InteriorVerse, whereas Colorful Diffuse Intrinsic used a broader mixture including that source (Appendix Table 12). This is consistent with source or annotation matching, without establishing causality. Equal-source averaging ensures balanced reporting but cannot make these references domain-neutral.

4.2 Material Recovery Across Error and Localization Metrics

Roughness and metallic probe complementary aspects of material inference: continuous surface variation and sparse region identification.

Table 5: Controlled RGB-to-roughness results on the OpenRooms-FF main-axis subset.

Method	RMSE↓	Bin-BMAE↓	Region-Macro MAE↓	SSIM↑
GPT-Image-1.5	0.375	0.333	0.314	0.399
GPT-Image-2.0	0.323	0.271	0.280	0.444
Qwen-Image-2.0	0.384	0.339	0.316	0.397
Doubao Seedream 5.0	0.364	0.288	0.333	0.348
FLUX.1-Kontext-dev	0.337	0.325	0.278	0.436
FireRed-Image-Edit	0.367	0.286	0.322	0.441
Gemini 3 Pro Image	**0.291**	**0.251**	**0.258**	**0.525**
LongCat-Image-Edit	0.391	0.331	0.339	0.420
Intrinsic Image Diffusion	0.346	0.264	0.309	0.513
DNF-Intrinsic	0.376	0.286	0.349	0.478

Roughness. On the OpenRooms-FF main-axis split in Table 5, Gemini has the lowest RMSE (0.291 versus 0.346 for Intrinsic Image Diffusion), Bin-BMAE, and Region-Macro MAE, and the highest SSIM. Its scene-bootstrap RMSE interval lies below that of the next-best editor (Appendix Table 29). Model ordering still depends on the evaluated aspect: GPT-Image-2.0 has lower RMSE than FLUX (0.323 versus 0.337), while their region-macro errors are nearly equal (0.280 versus 0.278). All outputs are evaluated in the common $[0, 1]$ grayscale convention without method-specific calibration. Agreement across these metrics strengthens a specific interpretation: Gemini's lead persists when frequent roughness values and large regions receive less weight, so it is not solely an advantage on dominant pixels. It does not show which cues produce the gain. Constant and material-category baselines are still needed to distinguish image-conditioned estimation from typical-material priors. Moreover, both roughness specialists use InteriorVerse supervision and are evaluated here on OpenRooms-FF (Appendix Table 12); the result compares transferred specialist references with frozen editors under this protocol, rather than isolating an architectural advantage.

Metallic. Only 0.83% of evaluable PBR-Rooms pixels have $M > 0.5$, so full-map error is dominated by non-metallic pixels. An all-zero prediction obtains MAE 0.019, below every evaluated method, including GPT-Image-2.0 (0.021). Low full-map MAE therefore mainly reflects background fidelity; it does not establish metallic recovery. Qwen, FLUX, FireRed, and LongCat fail to improve on the all-zero RB-MAE of 0.480 (their scores range from 0.482 to 0.507), while their AP remains within 0.004 of its 0.005. DNF-Intrinsic improves only marginally (RB-MAE 0.459, AP 0.007). The remaining methods localize some metal, but the highest IoU is only 0.057. Doubao, Gemini, and GPT-Image-2.0 have overlapping scene-bootstrap intervals for IoU and AP (Appendix Table 30); these summaries do not establish a stable localization winner. MVInverse has the lowest RB-MAE (0.330) but only 0.027 IoU, whereas Doubao has the highest IoU/AP point estimates (0.057/0.068). Balanced value accuracy, thresholded overlap, and ranking quality therefore expose different output behavior.

Table 6: Controlled RGB-to-metallic estimation results on the PBR-Rooms main-axis subset.

Method	MAE↓	RB-MAE↓	IoU↑	AP↑
GPT-Image-1.5	0.026	0.441	0.031	0.022
GPT-Image-2.0	0.021	0.419	0.052	0.042
Wan2.7-Image	0.096	0.389	0.011	0.014
Qwen-Image-2.0	0.496	0.503	0.004	0.006
Doubao Seedream 5.0	0.044	0.351	0.057	0.068
FLUX.1-Kontext-dev	0.421	0.486	0.004	0.006
FireRed-Image-Edit	0.277	0.507	0.004	0.005
Gemini 3 Pro Image	0.040	0.375	0.055	0.051
LongCat-Image-Edit	0.593	0.482	0.005	0.009
MVInverse	0.197	**0.330**	0.027	0.031
DNF-Intrinsic	0.121	0.459	0.008	0.007
All-zero prediction	*0.019*	*0.480*	*0.000*	*0.005*

All four metrics are macro-averaged equally over the same 40 retained scenes, each of which contains metallic-positive ground-truth support. IoU thresholds ground truth and predictions at 0.5, whereas AP ranks continuous prediction scores. Lower MAE/RB-MAE and higher IoU/AP are better. Boldface marks the RB-MAE point-estimate leader. The all-zero baseline uses the same mask and evaluator and obtains MAE 0.019, below every evaluated method; its AP under tied scores reflects positive prevalence. IoU and AP leaders are not emphasized because the leading scene-bootstrap intervals overlap.

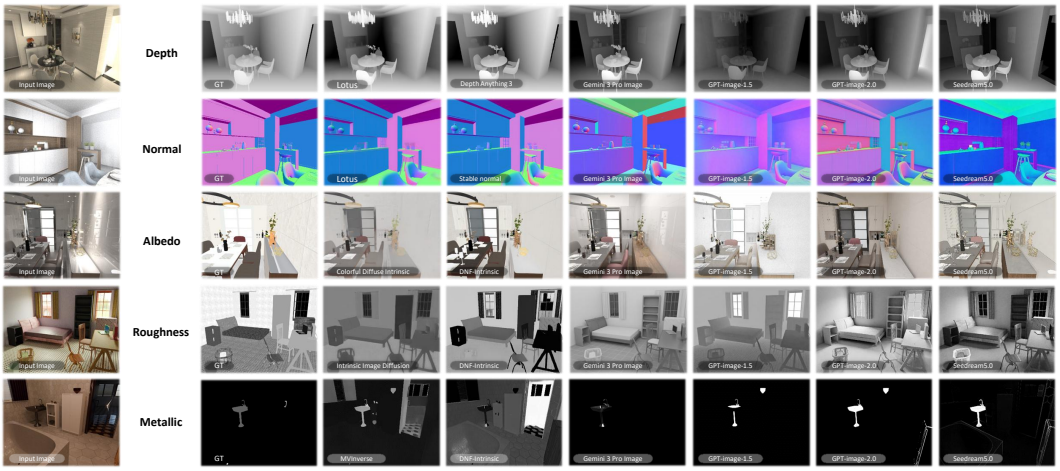

Figure 3: **Qualitative examples of physical map recovery across five targets.** Each row shows the input RGB image, ground-truth physical map, and recovered representations from selected specialized predictors and general-purpose image editors for one target. These examples complement the quantitative comparisons by illustrating differences in spatial structure, appearance decomposition, material consistency, and target-specific failure modes across depth, normal, albedo, roughness, and metallic prediction.

4.3 Prompts and Exemplars Change the Ranking

The depth diagnostic shows that additional context can change recovered spatial relations. A fixed, scene-disjoint RGB–depth exemplar lowers Qwen's Spearman correlation from 0.944 to 0.467 while increasing AbsRel-AI from 0.141 to 0.317. Correlations are computed before affine fitting and after polarity selection, so a simple global rescaling or polarity reversal cannot explain this loss of depth ordering. The protocol affects the structure recovered from the query, beyond the requested map's visual format.

The response is model-dependent. In the normal diagnostic, exemplars raise GPT-Image-1.5's Acc@22.5 from 0.1826 to 0.2016 but lower FireRed's from 0.3192 to 0.1938, reversing their ordering under the full prompt. Both remain below 0.33; the reversal illustrates sensitivity at low accuracy. Sharing a convention cue therefore does not guarantee equally effective elicitation across models.

Albedo illumination analysis improves all four metrics for four of six editors, but worsens all four for GPT-Image-2.0 and Qwen. Depth exemplars likewise improve Doubao's scalar accuracy

while reducing its Boundary F1, and degrade Wan and Qwen. Thus richer instructions or evidence can interfere with useful existing behavior. These single-generation comparisons reveal protocol sensitivity, rather than causal effects separated from generation randomness; detailed results appear in Appendix E.2.

4.4 Photometric Stress Diagnostics

Specialists retain their relative-geometry advantage across the stress slices, while Gemini leads the reported roughness metrics in most slices. In paired PBR-Rooms lighting conditions, GPT-Image-2.0 has almost unchanged metallic MAE (0.018), near the all-zero baseline (0.017), but IoU ranges from 0.036 to 0.103. Stable full-map error therefore does not imply stable material localization. OpenRooms-FF and InteriorVerse slices contain different scenes and therefore do not isolate a causal lighting effect (Appendix E.1).

For the paired metallic comparison, scene, camera, and material assignments remain fixed as lighting changes. The IoU variation therefore occurs without a change in the target material field or scene composition. This exposes appearance dependence that an aggregate full-map error conceals; repeated generations are still needed to quantify its magnitude relative to sampling variability.

5 Limitations

Our evaluation does not exhaust the possible access space of prompts, exemplars, and photometric conditions. Moreover, PhysProbe relies on public synthetic indoor datasets to obtain dense, pixel-aligned references for multiple physical factors. This enables controlled comparisons, especially for albedo, roughness, and metallicity, but does not establish robustness to real-world imaging conditions or downstream deployment scenarios. Here, frozen refers to the absence of task-specific adaptation during evaluation, rather than the absence of related pretraining exposure.

Official and diagnostic evaluations use one generation per query-setting pair. Scene-cluster intervals measure variation across retained scenes rather than provider-side generation randomness, so prompt or exemplar effects should not be interpreted as stable causal effects without repeated-generation studies. The roughness evaluation still lacks constant and material-category baselines, while metallic evaluation lacks scene-matched constant baselines.

Finally, recovering physical factors from a single RGB image remains ambiguous. Future work should use controlled interventions to separate image cues, material priors, and output-representation limitations.

6 Conclusion

We introduced **PhysProbe**, a controlled framework for measuring protocol-conditioned dense physical-map recovery from frozen general-purpose image editors. Specialist predictors provide reference points for identifying which maps remain difficult to reproduce through generative interfaces.

The findings distinguish local depth-boundary recovery from accurate relative geometry, full-map material fidelity from sparse-region identification, and additional context from effective use of that context. On OpenRooms-FF roughness, Gemini leads the evaluated editors and specialists on RMSE, Bin-BMAE, and Region-Macro MAE; its advantage over constant or category-level predictors remains unestablished. Normal exemplars reverse the FireRed–GPT-Image-1.5 ordering, showing that model comparisons depend on the specified access setting.

Visually structured outputs and low aggregate errors do not guarantee faithful recovery of scene factors. PhysProbe makes these protocol-conditioned differences measurable and motivates controlled studies of illumination invariance, material priors, output conventions, and generation variability.

References

Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics*, 43(6), 2024. doi: 10.1145/3687984.

JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. Mair: Multi-view attention inverse rendering with 3d spatially-varying lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Introduces the OpenRooms Forward Facing (OpenRooms FF) dataset.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners, 2026. Accepted by CVPR 2026.

Google. Introducing gemini 2.5 flash image, google's state-of-the-art image generation and editing model. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`, 2025. Ac-imagecessed: 2026-05-06.

Google DeepMind. Gemini 3 pro image (nano banana pro). `https://deepmind.google/models/gemini-image/pro/`, 2025. Accessed: 2026-09-13.

Yu Guo, Zhiqiang Lao, Xiyun Song, Yubin Zhou, Zongfang Lin, and Heather Yu. epbr: Extended pbr materials in image synthesis. *arXiv preprint arXiv:2504.17062*, 2025.

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 (ICLR)*, 2025.

Peter Kocsis, Vincent Sitzmann, and Matthias Nieß ner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Luke Smith, and Axel Sauer. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL `https://arxiv.org/abs/2506.15742`.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Bharadwaj Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7190–7199, 2021. doi: 10.1109/CVPR46437.2021.00711.

Chih-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. Iris: Inverse rendering of indoor scenes from low dynamic range images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 465–474, 2025a. doi: 10.1109/CVPR52734.2025.00052.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025b. URL `https://arxiv.org/abs/2511.10647`.

Hengyu Liu, Chenxin Li, Zhengxin Li, Yipeng Wu, Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yunlong Lin, Sirui Han, and Brandon Y. Feng. Ir3d-bench: Evaluating vision-language model scene understanding as agentic inverse rendering. In *Advances in Neural Information Processing Systems, Track on Datasets and Benchmarks*, 2025.

Muxin Liu, Tianbo Liu, Jing Xia, Xiaoyang Lyu, Xiaoshan Wu, Bo Wang, Peng Dai, Zhongrui Wang, Shaoshuai Shi, and Xiaojuan Qi. Optigeo: Efficient monocular geometry for embodied perception in optically challenging scenes. In *Conference on Robot Learning (CoRL)*, 2026a.

Muxin Liu, Xiaoyang Lyu, Tianhe Ren, Peng Dai, Xiaoshan Wu, Zhiyue Zhang, Jiaqi Zhang, Jiehong Lin, Shaoshuai Shi, and Xiaojuan Qi. Foundationgeo: Learning spatial pixel-wise fields for monocular metric geometry. In *European Conference on Computer Vision*, pp. 353–371. Springer, 2026b.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report, 2025. URL https://arxiv.org/abs/2512.07584.

OpenAI. Gpt image 2 model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. Accessed: 2026-05-06.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12630–12641, 2023.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yiqing Shi, Yiren Song, and Mike Zheng Shou. Edit2perceive: Image editing diffusion models are strong dense perceivers, 2025. URL https://arxiv.org/abs/2511.18673.

JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, and Xiangxiang Chu. From editor to dense geometry estimator, 2026. URL https://arxiv.org/abs/2509.04338.

Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering, 2025. URL https://arxiv.org/abs/2501.03717.

Yao Wang et al. Boosting 3d object generation through pbr materials, 2024. URL https://arxiv.org/abs/2411.16080.

Xiangzuo Wu, Chengwei Ren, Jun Zhou, Xiu Li, and Yuan Liu. MVInverse: Feed-forward multi-view inverse rendering in seconds. *arXiv preprint arXiv:2512.21003*, 2025.

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024. doi: 10.1145/3687971.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657445.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

Rongjia Zheng, Qing Zhang, Chengjiang Long, and Wei-Shi Zheng. Dnf-intrinsic: Deterministic noise-free diffusion for indoor inverse rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. doi: 10.1145/3550469.3555407.

A Appendix

B Evaluation Data and Split Design

PhysProbe uses three synthetic inverse-rendering sources with complementary target coverage. OpenRooms-FF and InteriorVerse provide depth, surface-normal, and albedo evaluation data; roughness uses OpenRooms-FF, while metallic uses PBR-Rooms. Source eligibility and retained samples were fixed before model evaluation based on channel availability and annotation reliability. Main-axis comparisons use source-specific scene splits, and stress analyses use fixed photometric conditions; released manifests record the retained samples and audit metadata (Appendix H).

B.1 Data Sources

B.1.1 Public Inverse-Rendering Sources

OpenRooms-FF supplies aligned depth, normal, albedo, and roughness annotations, while InteriorVerse supplies a complementary source for depth, normal, and albedo. Exact source files and retained sample identifiers are recorded in the released manifests. InteriorVerse material-channel exclusions are documented in Appendix B.2.1.

B.1.2 PBR-Rooms

Positioning relative to existing sources. PBR-Rooms is a targeted companion to public indoor inverse-rendering datasets, not a replacement for them. Hypersim provides large-scale synthetic indoor imagery with depth, normals, diffuse reflectance, and illumination decomposition, making it valuable for geometry and intrinsic-image research, but it does not jointly expose roughness and metallic in the PBR parameterization used by PhysProbe Roberts et al. (2021). OpenRooms and OpenRooms-FF provide high-quality depth, normal, diffuse-reflectance, and roughness supervision, but no metallic ground truth Li et al. (2021); Choi et al. (2023). InteriorVerse releases broader material channels, including roughness and metallic, but channel availability alone does not guarantee suitability as an evaluation reference. As detailed in Appendix B.2.1, prior work and our source screening indicate that those channels can contain patterns whose relationship to the intended intrinsic PBR parameters is ambiguous Zhu et al. (2022); Zeng et al. (2024); Guo et al. (2025).

This gap is consequential for metallic evaluation. Non-metallic walls, floors, wood, plastics, fabrics, and glass dominate typical indoor images, while exposed metal occupies little area. Under this imbalance, full-map MAE or RMSE can reward conservative near-zero output without demonstrating metallic-region recovery. A suitable companion source must therefore provide parameter-level metallic ground truth, support deliberate retention of scenes with positive metallic regions, and permit region-aware metrics. Controlled camera and lighting variation is additionally useful for testing whether a fixed material prediction changes with appearance. PBR-Rooms was designed for these requirements; positive pixels nevertheless remain sparse, so the benchmark reports their prevalence and does not treat source construction as a substitute for constant baselines.

Construction and benchmark role. **PBR-Rooms** is introduced and released with PhysProbe. It is generated with an Infinigen-based indoor scene and rendering pipeline Raistrick et al. (2023; 2024) that procedurally creates geometry, layouts, material assignments, illumination, and camera views. The complete release contains 1,000 scenes and 39,000 RGB images at 1280×720, with aligned depth, camera-space normal, albedo, roughness, metallic, and object-segmentation channels. These channels are exported from the shared scene representation. In particular, metallic is the renderer-exported bounded PBR material parameter in $[0, 1]$, where 0 denotes dielectric and 1 denotes metallic material; it is not derived from brightness, highlights, or reflections.

The official metallic main axis retains 40 scenes and 24 views per scene (960 images). The diagnostic lighting axis retains 25 separate scenes, three cameras per scene, and six renderer-defined lighting conditions with fixed material parameters. Invalid or incomplete renders are removed before sampling; retained records require an aligned metallic channel and dedicated valid evaluation mask, and preserve scene, camera, and lighting identifiers for scene-level sampling. PBR-Rooms is released under CC BY-NC 4.0. Metallic-positive pixels remain sparse: only 0.83% of evaluable pixels have $M > 0.5$. PhysProbe therefore reports full-map MAE together with region-balanced error, IoU, and AP so that performance on the dominant non-metallic background cannot substitute for metallic-region recovery. Main-axis sample counts appear in Table 1, and the diagnostic lighting split is summarized in Appendix B.4.

B.2 Target-Specific Source Eligibility

The three sources differ in channel availability and in whether a released channel is sufficiently reliable for target-specific scoring. OpenRooms-FF provides official evaluation data for depth, normal, albedo, and roughness. InteriorVerse contributes to depth, normal, and albedo, but its roughness and metallic maps are excluded after source screening. PBR-Rooms releases every benchmark channel, but only its metallic channel is used for evaluation in this paper.

Table 7: Dataset-by-target channel availability and official evaluation status.

Source	Depth	Normal	Albedo	Roughness	Metallic	Benchmark role
OpenRooms-FF	Official	Official	Official	Official	Unavailable	Official D/N/A/R source
InteriorVerse	Official	Official	Official	Excluded	Excluded	Official D/N/A source
PBR-Rooms[†]	Available	Available	Available	Available	Official	Official metallic source

[†] PBR-Rooms releases all five aligned physical-map channels, but this paper evaluates only its metallic channel. For depth, normal, and albedo, OpenRooms-FF and InteriorVerse are both official sources and receive equal weight in source-balanced aggregation. "Available" denotes a released channel that is not evaluated in this paper; "Excluded" denotes a released channel omitted from scoring after source screening.

B.2.1 Target-Specific Material Channel Selection

We do not use InteriorVerse roughness and metallic maps for official scoring because they are not sufficiently suitable for our target-specific evaluation protocol. This is neither a claim that the channels are universally incorrect nor a dataset-wide rejection: InteriorVerse remains an official source for depth, normal, and albedo.

RGB↔X likewise omits InteriorVerse roughness and metallicity, while ePBR reports that metallicity-related channels require correction in its rendering setting Zeng et al. (2024); Guo et al. (2025). Our screening finds appearance-correlated or coarsely segmented patterns whose relation to the intended bounded material parameters is ambiguous. Under our scoring protocol, this ambiguity could make results reflect source-specific channel behavior. We therefore use OpenRooms-FF for roughness and PBR-Rooms for metallic; Figure 4 illustrates the selection rationale rather than auditing the dataset.

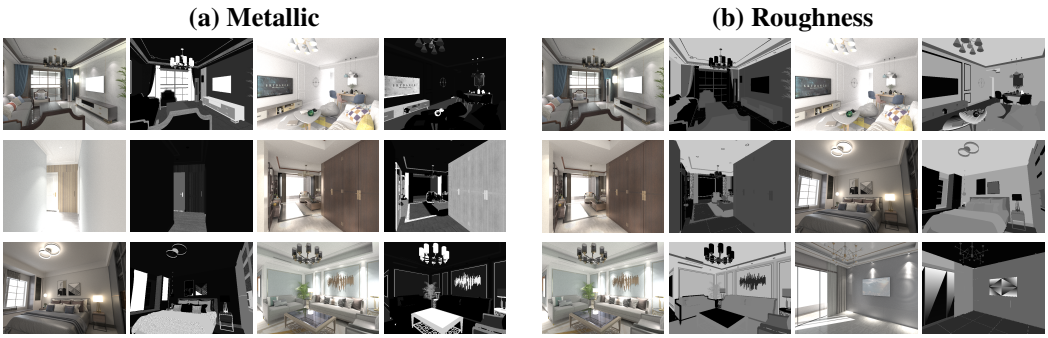

Figure 4: Representative InteriorVerse material-channel selection examples, arranged as RGB–channel pairs within each panel. **(a–b)** Some released material channels contain appearance-correlated or coarsely segmented patterns whose relation to the intended physical parameter is ambiguous. These illustrative cases support the conservative target-specific exclusion and do not constitute a dataset-wide audit.

B.3 Main-Axis Split Construction

We perform scene-level stratified sampling with fixed manifests, preserving dataset-native categories where available. All splits are frozen before model scoring, and official results use only the released scene and image identifiers. For OpenRooms-FF, each reference pose belongs to a local 3×3 view group, from which only the central view is retained to reduce near-duplicate views; exact identifiers and remaining source-specific rules are provided with the release. Every prediction uses a single RGB image without cross-view input, and the separately defined stress axis does not alter the main-axis manifests.

B.4 Photometric Stress-Split Definition

The photometric stress analysis is a diagnostic outside the main benchmark aggregation. It keeps each target's official access protocol and evaluator fixed, and all stress scenes are scene-disjoint from the corresponding main-axis scenes.

OpenRooms-FF and InteriorVerse use fixed, mutually exclusive photometric slices derived programmatically from RGB image statistics. Because the slices contain different scenes, their results characterize conditional performance rather than causal responses to illumination. PBR-Rooms instead uses matched scene–camera renders across controlled lighting conditions while holding material parameters fixed, enabling a paired diagnostic of full-map metallic error and positive-region localization. Exact retained assignments are provided in the released manifests; none of these analyses defines a separate leaderboard.

C Evaluation Protocol and Output Contracts

Official evaluation follows a controlled single-query-image protocol. For each target, the access setting fixes the query evidence, target instruction, output representation, generation policy, and failure handling before evaluation. The query RGB image is the only query-specific scene evidence in every official setting; normal estimation additionally uses fixed, query-independent convention exemplars. Main-axis evaluation and stress-axis analysis use the same official access settings, with the latter changing only the evaluated image conditions. Diagnostic access variants instead alter the prompt or available evidence and do not affect the official target-wise comparison. Model versions and run status are documented in Appendix H.

C.1 Official Access Settings and Output Contracts

Specialized predictors are evaluated through their standard single-image inference interfaces.

Table 8: Official access settings used in the main-axis and stress-axis evaluations.

Target	Official access	Output contract
Depth	Query RGB + fixed relative-depth instruction; no auxiliary input.	Spatially aligned grayscale map in $[0, 1]$; the requested convention is white-near and black-far. Scoring is polarity-invariant.
Normal	Query RGB + fixed normal instruction + one fixed RGB–normal exemplar pair per run; E1–E3 define three independent runs.	Spatially aligned RGB view-space normal map; RGB $[0, 1]$ decodes to vectors in $[-1, 1]^3$.
Albedo	Query RGB + fixed structure-preserving instruction; no auxiliary input.	Spatially aligned RGB reflectance map in $[0, 1]$.
Roughness	Query RGB + fixed roughness instruction; no auxiliary input.	Spatially aligned grayscale map in $[0, 1]$; black is smooth and white is rough.
Metallic	Query RGB + fixed metallic instruction; no auxiliary input.	Spatially aligned grayscale map in $[0, 1]$; black is non-metal and white is metallic.

Only Normal uses fixed additional input, with one exemplar supplied per independent run. For Depth, white-near is requested, but polarity selection and affine alignment are evaluator operations applied only after generation.

C.2 Fixed Target Instructions

All image-editor runs use target-specific instructions fixed before scoring. The instructions define target semantics, value convention, spatial-alignment requirements, and restrictions on appearance leakage; they do not provide benchmark labels, metric feedback, multi-view evidence, or ground-truth maps. For material maps, the instructions explicitly distinguish the requested physical parameter from image brightness, shadows, highlights, reflections, and edges. Target-specific value conventions are summarized in Table 8.

Model-specific prompt truncations were applied only when required by interface constraints, preserved the same target definition and output convention, and were fixed before evaluation; exact prompt strings are retained in the released run records.

C.3 Generation and Output Selection

Each query requests one image, and the protocol uses neither best-of-n nor manual candidate selection. If a provider-level retry is needed, the first valid response is retained without evaluator feedback; unresolved or invalid outputs follow Appendix D.2.1. Seeds, returned pixel dimensions, and interface-specific canvas settings are fixed in the released run records. Editors are allowed to return their interface-native canvas size and aspect ratio. The evaluator uses the full returned canvas, without content-aware cropping or padding, and resizes it directly to the target frame: nearest-neighbor for normals, bicubic for albedo, and bilinear for depth, roughness, and metallic; masks use nearest-neighbor interpolation. Consequently, aspect-ratio or framing mismatch remains part of measured interface behavior and can reduce pixel-aligned metrics, especially the zero-tolerance depth Boundary F1.

C.4 Target-Specific Exceptions and Clarifications

C.4.1 Normal Convention Stabilization

Motivation. Normal-map prediction requires a stable RGB encoding convention: colors must represent surface orientation rather than material appearance, shading, or semantic regions. Because prompt-only instructions may not communicate this convention consistently across image editors, the official normal setting uses fixed RGB–normal exemplar pairs as convention cues.

The official setting uses three shared RGB–normal exemplar pairs to define the RGB encoding convention and output format. Each query is processed independently with one pair per run. The exemplars are scene-disjoint from the evaluation data, fixed before evaluation, and shared across all

models; they contain no information about the evaluated query scene. Because generated outputs can still depend on the supplied context, these scores are interpreted as protocol-conditioned results.

Exemplar Selection. The exemplar pairs come from OpenRooms-FF and use its dataset-provided normal ground truth. All three pairs are disclosed in Figure 5.

Aggregation. Each retained query is processed independently once under each exemplar pair. Aggregation follows a fixed three-stage order: for each exemplar, we first compute a complete score within each source; we then macro-average the OpenRooms-FF and InteriorVerse source scores with equal weight; finally, we take the arithmetic mean of the three resulting E1–E3 source-balanced scores. Per-exemplar results are also reported to expose residual sensitivity to the convention cue. The three predictions are neither selected nor fused at the image level; aggregation is performed only at the score level.

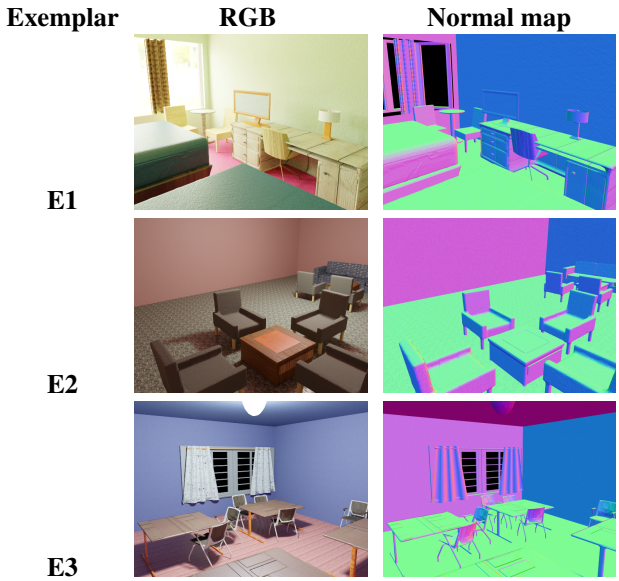

Figure 5: The three fixed RGB–normal exemplar pairs used in the official normal-evaluation protocol. Each row shows one exemplar pair, with the RGB image on the left and the corresponding normal map on the right. These exemplar pairs are shared across all evaluated image editors and benchmark images and are used only to communicate the normal-map convention.

C.4.2 Albedo: RGB-Only Official Setting

The official albedo setting follows the same RGB-only access rule as depth, roughness, and metallic. Each editor receives the query RGB image and the fixed albedo instruction, with no illumination-analysis text, segmentation mask, ground-truth annotation, benchmark label, multi-view input, or external reference. Query-image illumination analysis is evaluated only as a diagnostic access variant in Appendix E.2.1; it is not part of either the main-axis or stress-axis official access setting.

C.5 Diagnostic Access Variants

Diagnostic experiments vary prompts, convention exemplars, query-image analysis, and spatial or paired-reference priors. They use fixed diagnostic subsets and never replace or redefine the official settings. Table 9 maps the setting identifiers used in Appendix E to their evidence conditions.

Diagnostic cue provenance. Albedo-A1 adds cached query-image-only illumination analysis generated once per query by Doubao Seed 2.0 Pro and then shared unchanged across the evaluated editors. It uses no ground-truth map or multi-view evidence. The remaining diagnostic variants use the fixed exemplar or query-aligned spatial-prior assets named in Table 9; their exact asset identifiers, source paths, input order, and configured seeds are retained in the diagnostic run records. Each

Table 9: Official access settings and separately evaluated diagnostic variants.

Target	Official setting	Diagnostic variants
Depth	A0: query RGB + fixed prompt	A1: + fixed RGB–depth exemplar; A2: + segmentation reference.
Normal	A0: full prompt + three independent fixed-exemplar runs	A1: full prompt only; A2: minimal prompt + three fixed-exemplar runs; A3: minimal prompt only.
Albedo	A0: query RGB + fixed albedo instruction	A1: + query-image illumination analysis; A2: weakened prompt; A3: minimal prompt; B0: minimal prompt + fixed external exemplar.
Roughness	A0: query RGB + fixed prompt	A1: + soft segmentation; A2: + segmentation region fill; A3: + fixed RGB–roughness exemplar.

Here, "prompt" denotes either the canonical target instruction or the disclosed compressed form that preserves the same target definition, value convention, spatial-alignment requirement, and appearance-leakage restrictions. Setting identifiers are target-local: for example, Depth-A1, Normal-A1, and Albedo-A1 denote different interventions. A1–A3 and B0 alter the prompt or evidence condition and are diagnostic only. They are not included in the official target-wise comparison.

query–setting pair contributes one retained generation, so these comparisons characterize protocol sensitivity under the recorded runs rather than causal effects isolated from provider-side generation stochasticity.

The diagnostic image subsets for depth, normal, albedo, and roughness are stratified random samples from the corresponding retained main-axis and stress-axis images. They are therefore part of the retained evaluation data and are not scene-disjoint auxiliary test sets. All variants for a target use the same sampled images, and their results are reported separately from the official aggregation.

D Metrics, Aggregation, and Failure Handling

This appendix specifies target-specific output-validity checks, post-processing rules, failure handling, metrics, masking rules, normalization conventions, and source-balanced aggregation. It defines how a generated prediction is converted into benchmark scores after the access protocol has produced an output.

D.1 Common Evaluator Rules

All official predictions are matched to ground truth by scene, view, and frame identifier. Before scoring, each target entry point checks duplicate predictions, missing ground-truth or mask files, and target-specific size requirements. Masks are read as grayscale, normalized to $[0, 1]$, and thresholded at ≥ 0.5; when a mask must be resized, nearest-neighbor interpolation is used.

Unless a target-specific rule below states otherwise, metrics are first computed per image and their reported macro summary gives every evaluated image equal weight. Evaluators may additionally emit micro summaries in which every valid pixel has equal weight; these are recorded separately and are not substituted for the image-macro values in the main tables. For depth, normal, and albedo, source-specific image-macro scores are computed first and OpenRooms-FF and InteriorVerse then receive equal weight in the source-balanced result. Target-specific interpolation, validity, and aggregation exceptions are stated below.

D.2 Output Validity and Failure Handling

Before metric computation, each prediction must be readable, matched to a retained sample, spatially aligned or resizable, and convertible to the required target representation. Table 10 summarizes the deterministic checks and pre-scoring conversions; failures follow the common policy in Appendix D.2.1.

Table 10: Deterministic validity checks and pre-scoring processing by target.

Target	Validity checks and deterministic pre-scoring processing
Depth	Require a readable matched grayscale map and mandatory mask; convert ground truth to meters, resize maps to 640×480, resize the mask with nearest-neighbor interpolation, and apply valid-region normalization, polarity selection, and affine fitting.
Normal	Require a matched dense RGB normal map and same-frame mask; resize with nearest-neighbor interpolation, decode $\mathbf{n} = 2\mathbf{rgb} - 1$, apply the frozen Lotus/StableNormal convention adapter, and L2-normalize valid vectors.
Albedo	Require matched RGB prediction, ground truth, and mask; resize predictions bicubically and evaluate normalized RGB values in the original sRGB space without linearization.
Roughness	Require matched grayscale prediction, ground truth, mask, and segmentation; normalize to $[0, 1]$ and resize predictions bilinearly.
Metallic	Require a matched grayscale prediction, 16-bit single-channel ground truth, and dedicated metallic mask; normalize to $[0, 1]$ and resize predictions bilinearly. No baseline substitution or mask fallback is used.
All targets	Require a readable, uniquely matched output from the correct target protocol; intersect the mandatory mask with target-specific finite-value checks. Missing, duplicate, unreadable, or unmatched files are logged rather than replaced.

The following subsections define the target-specific masks, metrics, aggregation, and shared failure policy.

D.2.1 Failure Handling

Invalid or incomplete generations are logged and excluded rather than assigned artificial scores. Official comparisons therefore include only complete, scoreable retained-split runs. Within those runs, frozen target-specific masks are applied uniformly, and no retained image has empty valid-pixel support.

D.3 Metric Summary

Table 11 summarizes the primary metric or metallic metric profile and complementary metrics used for each target. The benchmark uses target-specific metrics because depth, normal, albedo, roughness, and metallic represent different physical quantities and have different validity and aggregation requirements.

Table 11: Metric summary and main-axis aggregation in PhysProbe.

Target	**Primary metric / profile**	**Complementary metrics**	**Score-time invariance**	**Main-axis aggregation**
Depth	AbsRel-AI ↓	RMSE/MAE-AI ↓; $\delta_{1,2}$-AI and Boundary F1 ↑	Per-image scale, shift, and polarity	Image macro within source; equal-source macro
Normal	Acc@22.5 ↑	Mean/median angular error ↓; Acc@11.25/30 ↑	None for editors; one fixed shared adapter for Lotus/StableNormal	Image macro, then equal-source and fixed-exemplar macro
Albedo	MAE ↓	PSNR/SSIM ↑; LPIPS ↓	None	Image macro within source; equal-source macro
Roughness	RMSE ↓	Bin-BMAE/Region-Macro MAE ↓; SSIM ↑	None	OpenRooms-FF; image-macro RMSE and metric-specific diagnostics
Metallic	MAE/RB-MAE ↓; IoU/AP ↑	–	None	Equal-scene macro over 40 retained scenes; all four metrics use all scenes

AI denotes affine-invariant depth scoring. Metallic uses a four-metric profile whose components are reported separately. Stress-axis aggregation is specified separately in each stress table and in Appendix D.6.

We use affine-invariant AbsRel (AbsRel-AI) as the primary metric for depth, Acc@22.5 as the primary directional-accuracy metric for normal, RMSE as the primary scalar error metric for roughness, and MAE as the primary scalar distortion metric for albedo. Metallic is intentionally evaluated with the joint MAE/RB-MAE/IoU/AP profile because global continuous error and sparse-region recovery expose different behavior; the four components are not reduced to one scalar score.

D.4 Specialist Training-Data Audit

Table 12 records the training sources stated in the cited papers or released documentation. The audit concerns explicit dataset-level overlap only; it does not establish image-level duplication, and "none identified" does not rule out overlap through undisclosed pretraining data.

Table 12: Reported training data for the specialist reference models and relation to PhysProbe evaluation sources. Lists are condensed to the datasets relevant to distribution and overlap auditing; see the cited papers for complete training details.

Method	Evaluated target(s)	Reported training data	Relation to PhysProbe sources
Depth Anything 3 Lin et al. (2025b)	Depth	More than twenty public geometry and depth datasets, including Hypersim, Structured3D, and Replica, under a teacher–student training pipeline	No explicit OpenRooms-FF, InteriorVerse, or PBR-Rooms exposure identified in the cited report.
Lotus He et al. (2025)	Depth, normal	Hypersim and Virtual KITTI synthetic geometry data	Synthetic indoor training, but no direct overlap with the retained OpenRooms-FF or InteriorVerse evaluation sources identified.
StableNormal Ye et al. (2024)	Normal	Hypersim, Replica, 3D Ken Burns, MatrixCity, and Objaverse	Synthetic indoor and object-centric training; no direct overlap with the retained official sources identified.
Colorful Diffuse Intrinsic Careaga & Aksoy (2024)	Albedo	A multi-dataset synthetic intrinsic mixture that includes InteriorVerse albedo, together with real multi-illumination data	Direct dataset-level exposure to InteriorVerse albedo within a broader mixture; no OpenRooms-FF or PBR-Rooms exposure identified.
Intrinsic Image Diffusion Kocsis et al. (2024)	Roughness	The 85-degree-field-of-view InteriorVerse subset	Uses InteriorVerse material supervision, although official roughness scoring uses OpenRooms-FF.
DNF-Intrinsic Zheng et al. (2025)	Albedo, roughness, metallic	InteriorVerse only (45,073 reported training images)	Direct source exposure for InteriorVerse albedo evaluation; material channels are transferred to OpenRooms-FF roughness and PBR-Rooms metallic evaluation.
MVInverse Wu et al. (2025)	Metallic	Hypersim, InteriorVerse, PRID, Structured3D, Amazon–Berkeley Objects, MatrixCity, and CGIntrinsics, followed by real-video consistency fine-tuning; material channels come from InteriorVerse, MatrixCity, and ABO	Includes InteriorVerse material supervision but not the official PBR-Rooms metallic source.

Dataset-level exposure can advantage a specialist through in-distribution training even when retained evaluation scenes are disjoint. The table records disclosed sources, not image-level duplication. In particular, the retained InteriorVerse scenes have not yet been reconciled against every specialist's original train/validation/test manifest. All evaluated roughness and metallic specialists include InteriorVerse material supervision, whereas official material scoring uses OpenRooms-FF or PBR-Rooms.

This audit changes how the reference results should be read. The evaluated depth and normal specialists disclose synthetic geometry supervision but not OpenRooms-FF or InteriorVerse training,

and their large advantages persist on both evaluation sources. The albedo references instead have unequal exposure: DNF-Intrinsic is trained only on InteriorVerse, while Colorful Diffuse Intrinsic includes it within a broader mixture. For roughness and metallic, every evaluated specialist transfers material supervision across both dataset and annotation conventions because InteriorVerse material maps are not used as official references in PhysProbe. These observations qualify domain neutrality; they do not identify whether a gain comes from scene similarity, rendering statistics, or closer agreement with a source's encoding convention.

D.5 Target-Specific Evaluation

D.5.1 Depth Affine-Invariant Evaluation

Depth is evaluated primarily as affine-invariant relative-geometry recovery rather than calibrated metric-depth prediction. This protocol is used because image editors and some dense-depth systems may output relative-depth-like maps with different scale, shift, or polarity conventions.

InteriorVerse EXR and OpenRooms-FF DAT ground truth are converted to meters. Ground truth and predictions are resized to 640×480 with bilinear interpolation, while the mandatory mask uses nearest-neighbor interpolation. The valid set Ω contains mask-selected pixels with finite positive ground-truth depth and finite prediction values.

For each image, the relative-depth prediction is min–max normalized over Ω, and the white-is-far or white-is-near polarity with higher Spearman correlation is retained. A zero-range prediction is mapped to zero and follows the evaluator's fixed tie rule; fewer than two valid pixels make the item invalid. Spearman's ρ and Kendall's τ are computed before affine fitting. The selected prediction is then least-squares aligned independently for each image:

$$\hat{d}_{\mathrm{AI}} = a\hat{d} + b,$$

where a and b are fitted independently for each evaluated image over valid pixels Ω:

$$(a, b) = \arg\min_{a,b} \sum_{p\in\Omega} (a\hat{d}_p + b - d_p)^2.$$

The frozen evaluator uses an unconstrained least-squares solution; rank-deficient inputs therefore receive the solver's minimum-norm solution without a separate fallback.

Let $d_p^+ = \max(d_p, 10^{-6})$. The aligned scalar metrics are

$$\text{AbsRel-AI} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \frac{|\hat{d}_{\mathrm{AI},p} - d_p|}{d_p^+}, \quad \text{MAE-AI} = \frac{1}{|\Omega|} \sum_{p\in\Omega} |\hat{d}_{\mathrm{AI},p} - d_p|,$$

$$\text{RMSE-AI} = \sqrt{\frac{1}{|\Omega|} \sum_{p\in\Omega} (\hat{d}_{\mathrm{AI},p} - d_p)^2}.$$

The aligned prediction is not clipped in the AbsRel-AI numerator, MAE-AI, or RMSE-AI. For ratio and logarithmic metrics, the evaluator additionally defines $\hat{d}^+_{\mathrm{AI},p} = \max(\hat{d}_{\mathrm{AI},p}, 10^{-6})$; non-positive aligned depths therefore participate as 10^{-6} rather than being removed. For $k \in \{\frac{1}{2}, 1, 2, 3\}$, threshold accuracy is

$$\delta_k\text{-AI} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \mathbb{1}\left[\max\left(\frac{\hat{d}^+_{\mathrm{AI},p}}{d_p^+}, \frac{d_p^+}{\hat{d}^+_{\mathrm{AI},p}} \right) < 1.25^k \right].$$

The evaluator also emits affine-aligned $\log_{10}$ error in addition to AbsRel-AI, MAE-AI, and RMSE-AI.

The primary depth metric is AbsRel-AI. The main tables additionally report RMSE-AI, MAE-AI, δ_1-AI, δ_2-AI, and Boundary F1. The evaluator also records $\delta_{1/2}$-AI, δ_3-AI, $\log_{10}$-AI, Spearman's ρ, and Kendall's τ for diagnostic use. Boundary F1 is computed from the per-image affine-aligned prediction derived from the polarity-selected, min–max-normalized output. Before conversion to inverse depth, aligned prediction and ground truth are clipped from below at 10^{-6}. The resulting inverse-depth maps

define a DepthPro-style scale-invariant boundary diagnostic based on adjacent-pixel ratio contours. The ten thresholds are the linear grid

$$t_j = 1.05 + j\frac{1.25 - 1.05}{9}, \qquad j = 0, \ldots, 9,$$

and their aggregation weights are $w_j = t_j / \sum_{k=0}^{9} t_k$. At threshold t_j, the evaluator constructs left, right, top, and bottom binary contour relations by testing whether the inverse-depth ratio of the corresponding adjacent pixels exceeds t_j. A pair is eligible only when both endpoints satisfy the valid mask. Predicted and ground-truth contours match only when they occur on the same adjacent pair and in the same direction; no spatial dilation, displacement tolerance, or nearest-boundary matching is used. Precision and recall are each the equal-weight mean over the four directions, and their harmonic mean gives the per-threshold F1. The reported Boundary F1 is $\sum_j w_j \mathrm{F1}(t_j)$. It evaluates local depth-discontinuity recovery rather than global scalar alignment. Because affine alignment and polarity selection use ground truth only at scoring time, these numbers should not be interpreted as calibrated metric-depth performance. They measure whether the output preserves relative scene geometry after convention-level scale, shift, and polarity ambiguities are removed. This choice is intentionally favorable to systems that output only visually plausible or relative-depth-like maps: it removes global convention errors before scoring and therefore tests whether the remaining spatial ordering and depth structure are physically meaningful.

The auxiliary outputs separate evaluator stages: rank correlations measure ordering before affine fitting, whereas Boundary F1 is computed from the affine-aligned depth prediction and measures local discontinuity structure.

D.5.2 Normal Evaluation

Normal evaluation uses matched same-frame predictions, ground truth, and masks from InteriorVerse and OpenRooms-FF. Predictions are resized to ground-truth size with nearest-neighbor interpolation, mapped to $[0, 1]$, decoded as $\mathbf{n} = 2\mathbf{rgb} - 1$, and L2-normalized per pixel.

The target convention is image-right, image-up, and camera-facing $(+x, +y, +z)$, consistent with the retained annotations of OpenRooms-FF and InteriorVerse Choi et al. (2023); Zhu et al. (2022); editors receive fixed exemplars defining it. Because released specialist conventions are not standardized, Lotus and StableNormal use one shared adapter, $(x, y, z) \mapsto (-x, y, z)$, selected on an independent calibration subset that is scene-disjoint from every evaluation split. The adapter was frozen before evaluation and applied uniformly, with no per-model tuning or score-time alignment. Editors receive no score-time axis alignment because the exemplars specify the requested convention and convention adherence is part of the interface being evaluated. Their angular errors can therefore combine orientation error with encoding-convention error, including axis flips; the reported editor scores measure orientation recovery under the specified encoding rather than convention-invariant surface-normal accuracy.

The valid set intersects the thresholded mandatory mask with finite, nonzero ground-truth and predicted vectors and excludes pixels whose pre-decoding RGB triplet is all black under the threshold $4/255$. Consequently, normal support is prediction-dependent rather than fixed solely by the mask and ground truth. If Ω_{GT} denotes mask-valid, legal ground-truth pixels and Ω_m additionally requires a legal prediction from method m, prediction-valid coverage is

$$\mathrm{Coverage}_m = \frac{|\Omega_m|}{|\Omega_{\mathrm{GT}}|}.$$

Invalid predicted pixels are excluded rather than imputed as angular errors. Table 13 reports this support diagnostic for every scored normal method. The default evaluator neither erodes the mask nor removes back-facing normals.

Angular error is defined as

$$\theta = \arccos(\mathrm{clip}(\hat{\mathbf{n}} \cdot \mathbf{n}, -1, 1)) \cdot \frac{180}{\pi}.$$

For a threshold τ, directional accuracy is

$$\mathrm{Acc@}\tau = \frac{1}{|\Omega|} \sum_{p \in \Omega} \mathbb{1}(\theta_p < \tau).$$

The evaluator emits $\tau \in \{5^\circ, 7.5^\circ, 11.25^\circ, 22.5^\circ, 30^\circ\}$, together with mean, median, and root-mean-square angular error. The paper's main tables report the conventional 11.25, 22.5, and 30 degree accuracies plus mean and median error; the remaining fields are retained in evaluator outputs. All Acc@ thresholds are ratios in $[0, 1]$. Summaries include equal-image macro values and valid-pixel micro values; the official cross-source and E1–E3 aggregation follows Appendix C.4.1.

Table 13: Prediction-valid coverage for surface-normal evaluation.

Split	Method	E1	E2	E3	Mean
Main axis	Doubao Seedream 5.0	99.74%	100.00%	99.95%	99.90%
	FireRed-Image-Edit	100.00%	100.00%	99.99%	100.00%
	GPT-Image-1.5	99.96%	99.97%	99.98%	99.97%
	Lotus	–	–	–	100.00%
	StableNormal	–	–	–	100.00%
Stress axis	Doubao Seedream 5.0	99.76%	99.99%	99.93%	99.90%
	FireRed-Image-Edit	100.00%	100.00%	99.99%	100.00%
	GPT-Image-1.5	99.89%	99.99%	99.96%	99.95%
	Lotus	–	–	–	100.00%
	StableNormal	–	–	–	100.00%

Coverage is computed relative to ground-truth-valid support, not the full image canvas. For each split, prediction-valid and ground-truth-valid pixel counts are first aggregated over the complete retained pool within each source, and coverage is computed from their ratio. OpenRooms-FF and InteriorVerse coverage values are then averaged with equal source weight. For image editors, E1–E3 are computed independently using this procedure, and Mean is their arithmetic average.

D.5.3 Albedo Evaluation

Albedo evaluation uses matched predictions, ground truth, and mandatory masks from InteriorVerse and OpenRooms-FF. Predictions are resized bicubically to ground-truth size; masks use nearest-neighbor interpolation when needed and are valid at ≥ 0.5.

Images are converted to RGB and normalized to $[0, 1]$. Official scoring remains in normalized sRGB space without linearization. For RGB channel set C, masked MAE and MSE are

$$\mathrm{MAE} = \frac{1}{|\Omega||C|} \sum_{p\in\Omega} \sum_{c\in C} |\hat{A}_{p,c} - A_{p,c}|, \quad \mathrm{MSE} = \frac{1}{|\Omega||C|} \sum_{p\in\Omega} \sum_{c\in C} (\hat{A}_{p,c} - A_{p,c})^2,$$

with $\mathrm{PSNR} = 10 \log_{10}(1/\mathrm{MSE})$ for unit data range. SSIM is computed in RGB on the valid-mask bounding box after replacing invalid prediction pixels with ground truth. AlexNet LPIPS uses the same filled crop after mapping RGB to $[-1, 1]$. Summaries contain equal-image macro metrics and separate valid-pixel micro summaries; the main tables use the image-macro/source-balanced values.

D.5.4 Roughness Evaluation

Roughness is evaluated only on OpenRooms-FF using matched predictions, ground truth, mandatory masks, and segmentation maps. Predictions use bilinear resizing and masks use nearest-neighbor interpolation. Ground truth and predictions are normalized grayscale values in $[0, 1]$, evaluated where the mask is ≥ 0.5.

RMSE is the primary roughness scalar error metric:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\Omega|} \sum_{p\in\Omega} (\hat{r}_p - r_p)^2},$$

where r_p and $\hat{r}_p$ denote the ground-truth and predicted roughness values at valid pixel p, and Ω is the valid evaluation region.

The main-axis RMSE is the arithmetic mean of the per-image RMSE values. For N retained images, it is

$$\mathrm{RMSE}_{\mathrm{macro}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{|\Omega_i|} \sum_{p\in\Omega_i} (\hat{r}_{i,p} - r_{i,p})^2}.$$

MAE is also emitted:

$$\mathrm{MAE} = \frac{1}{|\Omega|} \sum_{p\in\Omega} |\hat{r}_p - r_p|.$$

To avoid domination by frequent roughness values, the released range is partitioned into $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1]$. If $\mathcal{B}_+$ denotes the non-empty bins, then

$$\text{Bin-BMAE} = \frac{1}{|\mathcal{B}_+|} \sum_{b \in \mathcal{B}_+} \frac{1}{|\Omega_b|} \sum_{p \in \Omega_b} |\hat{r}_p - r_p|.$$

Region metrics do not require semantic class names: each distinct segmentation RGB triplet defines one region. The evaluator computes per-region MAE and RMSE and macro-averages regions equally. For the set $\mathcal{R}$ of color-ID regions,

$$\text{Region-Macro MAE} = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{1}{|R|} \sum_{p \in R} |\hat{r}_p - r_p|.$$

The analogous region-macro RMSE and the number of parsed regions are retained in evaluator outputs. Every retained roughness image in the reported main-axis and stress-axis runs has a matched segmentation map, and all reported runs successfully parsed the region IDs. The evaluator retains both Region-Macro MAE and Region-Macro RMSE, while the paper's main-axis and stress-axis tables report Region-Macro MAE. The base evaluator also emits PSNR, SSIM, and required AlexNet LPIPS. For LPIPS, grayscale maps are repeated to three channels, mapped to $[-1, 1]$, and evaluated on the valid-mask bounding box after invalid prediction pixels are filled from ground truth; SSIM uses the corresponding masked-crop strategy.

Per-image MAE and RMSE use all valid pixels in that image; the reported main-axis RMSE then averages the per-image values equally. SSIM and LPIPS are averaged over valid image entries, Bin-BMAE macro-averages the non-empty ground-truth bins, and region metrics macro-average the parsed color-ID regions.

D.5.5 Metallic Evaluation

Metallic is evaluated on PBR-Rooms within the dedicated same-view mask Ω, using ground truth and grayscale predictions in $[0, 1]$ without per-image normalization. Full-map error is

$$\text{MAE} = \frac{1}{|\Omega|} \sum_{p \in \Omega} |M_p - \hat{M}_p|.$$

Let $P = \{p \in \Omega : M_p \geq 0.5\}$ denote metallic-positive pixels and $N = \Omega \setminus P$. Region-balanced error is

$$\text{RB-MAE} = \frac{1}{2} \left(\frac{1}{|P|} \sum_{p \in P} |M_p - \hat{M}_p| + \frac{1}{|N|} \sum_{p \in N} |M_p - \hat{M}_p| \right),$$

which prevents the dominant non-metallic region from determining the score.

For localization, $B_p = \mathbb{1}(M_p \geq 0.5)$ and $\hat{B}_p = \mathbb{1}(\hat{M}_p \geq 0.5)$, with

$$\text{IoU} = \frac{\sum_{p \in \Omega} B_p \hat{B}_p}{\sum_{p \in \Omega} \mathbb{1}(B_p + \hat{B}_p > 0)}.$$

AP instead summarizes the precision–recall trade-off obtained by ranking pixels with the continuous prediction $\hat{M}_p$; the released evaluator uses the non-interpolated precision–recall sum and groups tied scores before each update. AP therefore complements IoU at the fixed prediction threshold of 0.5.

RB-MAE, IoU, and AP are omitted only when P is empty, while full-map MAE remains defined. If P is non-empty but the prediction contains no positive pixels, IoU is zero and the other positive-support metrics remain in the aggregation. Although 33 of the 960 main-axis images are negative-only, every retained scene contains positive support.

For the official main-axis profile, valid pixels are pooled across the 24 images in each scene before the 40 scene scores are macro-averaged equally. All four metrics therefore use the same scene support. The components are reported separately: low full-map error can coexist with poor balanced error or sparse-region localization and is therefore not treated as a sufficient summary of metallic recovery.

D.6 Source-Balanced and Target-Specific Aggregation

For depth, normal, and albedo, let $m_{s,i}$ be an image-level metric for image i in source $s \in \{\mathrm{OR}, \mathrm{IV}\}$. The source macro is

$$\bar{m}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} m_{s,i},$$

and the reported source-balanced value is

$$\bar{m}_{\mathrm{SB}} = \tfrac{1}{2}(\bar{m}_{\mathrm{OR}} + \bar{m}_{\mathrm{IV}}).$$

Thus every image has equal weight within its source and the two sources have equal weight regardless of image count. Evaluator-emitted micro statistics instead pool valid pixels and are labeled separately.

For normal editors, this procedure is applied independently to each of E1–E3, after which the three source-balanced scores are averaged with equal weight. Depth, normal, and albedo stress slices use the same image-macro-then-source-macro order when both sources support the slice.

Roughness uses one source and reports image-macro averages over the retained OpenRooms-FF images on both axes. Main-axis and stress-axis values nevertheless summarize different retained subsets.

Metallic uses equal-scene aggregation on the main axis. Within each stress lighting condition, MAE averages all matched scene–camera groups, while RB-MAE, IoU, and AP use only groups with metallic-positive ground-truth support; the two axes are therefore interpreted separately.

Sampling strata are documented in Appendix B.3 and are not used as metric weights.

E Robustness and Access-Protocol Diagnostics

This appendix reports two forms of diagnostic evidence outside the official main-axis comparison: photometric stress slices and controlled access-setting ablations. Diagnostic analyses are post-hoc analyses on fixed evaluation subsets and do not affect official model comparison. They characterize protocol sensitivity rather than define alternative leaderboards. Source-specific official results and confidence intervals are kept separate in Appendix F.

E.1 Stress-Subset Results

All stress analyses retain the official target-specific evaluator, valid masks, and access protocol. Table notes state the aggregation used for each target. Roughness uses image-macro aggregation on both axes but evaluates different retained subsets; metallic also changes aggregation units. The stress summaries therefore support comparisons within the stress analysis rather than direct numerical comparison with main-axis scores.

E.1.1 Albedo Stress-Subset Results

Table 14: Stress-test albedo estimation results under the official controlled RGB-only single-image protocol.

Method	Slice	MAE↓	PSNR↑	SSIM↑	LPIPS↓
GPT-Image-1.5	Dark-heavy	0.199	12.616	0.502	0.421
	HDR	0.191	12.727	0.548	0.408
	Highlight-heavy	0.225	11.422	0.554	0.422
	Low-light	0.163	14.256	0.428	0.387
GPT-Image-2.0	Dark-heavy	0.161	14.041	0.535	0.419
	HDR	0.169	13.748	0.580	0.406
	Highlight-heavy	0.176	13.291	0.571	0.408
	Low-light	0.149	14.586	0.464	0.401

Continued on next page

Table 14 (continued)

Method	Slice	MAE↓	PSNR↑	SSIM↑	LPIPS↓
Doubao Seedream 5.0	Dark-heavy	0.173	14.164	0.608	0.344
	HDR	0.185	13.673	0.640	0.338
	Highlight-heavy	0.194	13.019	0.644	0.325
	Low-light	0.135	15.935	0.580	0.300
Wan2.7-Image	Dark-heavy	0.228	11.772	0.568	0.401
	HDR	0.237	11.694	0.606	0.405
	Highlight-heavy	0.214	12.286	0.631	0.371
	Low-light	0.216	11.975	0.497	0.368
Qwen-Image-2.0	Dark-heavy	0.205	12.167	0.490	0.438
	HDR	0.205	12.234	0.535	0.432
	Highlight-heavy	0.193	12.515	0.553	0.405
	Low-light	0.207	12.038	0.408	0.429
FLUX.1-Kontext-dev	Dark-heavy	0.257	10.458	0.458	0.607
	HDR	0.253	10.454	0.501	0.583
	Highlight-heavy	0.248	10.564	0.517	0.559
	Low-light	0.262	10.337	0.378	0.633
Gemini 3 Pro Image	Dark-heavy	0.156	15.210	0.701	0.291
	HDR	0.168	14.758	0.724	0.290
	Highlight-heavy	0.173	14.151	0.719	0.291
	Low-light	0.134	16.514	**0.679**	**0.239**
FireRed-Image-Edit	Dark-heavy	0.214	12.340	0.609	0.453
	HDR	0.226	11.895	0.637	0.442
	Highlight-heavy	0.204	12.505	0.653	0.412
	Low-light	0.224	11.945	0.529	0.502
LongCat-Image-Edit	Dark-heavy	0.259	10.675	0.555	0.560
	HDR	0.258	10.795	0.589	0.516
	Highlight-heavy	0.237	11.173	0.612	0.488
	Low-light	0.284	9.877	0.461	0.662
Colorful Diffuse Intrinsic	Dark-heavy	0.145	15.825	**0.718**	0.364
	HDR	0.141	15.978	**0.753**	0.322
	Highlight-heavy	0.141	15.859	0.748	0.324
	Low-light	**0.124**	**16.516**	0.660	0.392
DNF-Intrinsic	Dark-heavy	**0.106**	**18.214**	0.701	**0.200**
	HDR	**0.093**	**19.365**	0.740	**0.180**
	Highlight-heavy	**0.089**	**19.534**	**0.800**	**0.161**
	Low-light	0.189	13.582	0.519	0.272

For shared slice names, source-specific slice scores are first computed within each dataset and then macro-averaged across OpenRooms-FF and InteriorVerse. The low-light slice is defined only for OpenRooms-FF and is reported as a source-specific stress diagnostic. Only methods with a complete, scoreable stress-axis run under this protocol are included; the main-axis status in Table 31 does not imply stress-axis coverage. Masked MAE is the primary scalar distortion metric; SSIM and LPIPS are required diagnostics for structure preservation and perceptual consistency. Lower MAE/LPIPS and higher PSNR/SSIM are better. Boldface marks the best value within each slice and metric using unrounded scores; displayed rounding may produce apparent ties.

Across the three source-balanced stress slices, task-specific intrinsic methods retain the overall metric leads: DNF-Intrinsic gives the lowest MAE and LPIPS point estimates, while the highest SSIM is obtained by Colorful Diffuse Intrinsic on dark-heavy and HDR and by DNF-Intrinsic on highlight-heavy. Among image editors, however, **Gemini 3 Pro Image** gives the strongest point estimate on all four reported metrics for every stress slice.

The OpenRooms-FF-only low-light slice yields a mixed overall ordering. Colorful Diffuse Intrinsic gives the lowest MAE (0.124) and highest PSNR (16.516), while Gemini gives the highest SSIM (0.679) and lowest LPIPS (0.239); Gemini's PSNR of 16.514 is also nearly identical to the best point estimate. These results preserve the specialist advantage on the shared cross-source slices while showing that an image editor can become competitive or leading on selected complementary metrics under a source-specific extreme condition. The distinction matters for intrinsic perception: Gemini's

low-light structural and perceptual scores do not imply the lowest reflectance-value error. Conversely, DNF-Intrinsic's strong shared-slice MAE does not transfer to this low-light subset, where its MAE is 0.189. Because the slices differ in scene composition and source coverage, this is a conditional generalization pattern, not evidence that reduced illumination alone benefits or harms a particular architecture.

E.1.2 Normal Stress-Subset Results

Table 15: Stress-test normal results under the official multi-exemplar protocol.

Method	Slice	Mean↓	Median↓	Acc@11.25↑	Acc@22.5↑	Acc@30↑
GPT-Image-1.5	Dark-heavy	44.651	42.078	0.061	0.213	0.330
	HDR	44.559	42.327	0.058	0.205	0.329
	Highlight-heavy	44.931	42.955	0.055	0.201	0.322
	Low-light	40.335	37.085	0.072	0.251	0.386
Doubao Seedream 5.0	Dark-heavy	56.115	49.696	0.031	0.121	0.215
	HDR	56.192	50.368	0.038	0.139	0.232
	Highlight-heavy	55.714	50.058	0.041	0.145	0.240
	Low-light	48.824	41.455	0.043	0.161	0.284
FireRed-Image-Edit	Dark-heavy	61.637	64.623	0.103	0.199	0.247
	HDR	61.414	64.817	0.101	0.194	0.238
	Highlight-heavy	62.108	65.409	0.090	0.182	0.232
	Low-light	56.834	61.385	0.143	0.260	0.303
Lotus	Dark-heavy	10.158	**4.968**	0.792	0.891	0.918
	HDR	8.923	**4.463**	**0.830**	0.914	0.936
	Highlight-heavy	9.093	**4.724**	0.817	0.912	0.936
	Low-light	**8.285**	**4.382**	**0.823**	0.923	0.946
StableNormal	Dark-heavy	**9.804**	5.983	**0.798**	**0.909**	**0.936**
	HDR	**8.914**	5.541	0.827	**0.928**	**0.949**
	Highlight-heavy	**8.710**	5.246	**0.827**	**0.926**	**0.947**
	Low-light	9.812	6.938	0.791	**0.928**	**0.952**

For general-purpose image editors, scores are averaged over the three fixed exemplar-conditioned runs. Only methods with complete, scoreable stress-axis outputs for every required retained image and, for editors, every required exemplar-conditioned run are included. For shared slice names, source-specific slice scores are first computed within each dataset and then macro-averaged across OpenRooms-FF and InteriorVerse. The low-light slice is defined only for OpenRooms-FF and is not directly comparable to the cross-source stress slices. Acc@22.5 is the primary directional-accuracy metric. Lower Mean and Median are better; higher Acc@11.25, Acc@22.5, and Acc@30 are better. Boldface marks the best value within each slice and metric using unrounded scores; displayed rounding may produce apparent ties.

The task-trained normal predictors remain far ahead of the image editors under every stress condition. StableNormal gives the highest reported Acc@22.5 point estimate on all four slices, reaching 0.928 on both HDR and low-light, whereas Lotus gives the lowest reported median angular error on every slice and the lowest mean angular error on low-light. The strongest editor Acc@22.5 values remain between 0.213 and 0.260 across the slices, compared with 0.891–0.928 for Lotus and StableNormal. Although several editors obtain better scores on the OpenRooms-FF-only low-light slice than on the shared slices, the large directional-recovery gap persists. Mean source-balanced prediction-valid coverage remains between 99.90% and 100.00% at the displayed precision over the complete stress pools (Table 13), so this gap is likewise not attributable to materially different valid-pixel support. This persistence suggests that the observed interface-level orientation errors are not confined to a single challenging appearance regime. It does not locate their cause: illumination-dependent inference, coordinate encoding, and generated-map fidelity remain entangled. In particular, higher low-light scores should not be read as improved geometric reasoning in darkness, since that slice uses a different, single-source scene population.

E.1.3 Roughness Stress-Subset Results

Table 16: Stress-test roughness estimation results on OpenRooms-FF under the controlled direct RGB-to-map track.

Method	Slice	Macro RMSE↓	Bin-BMAE↓	Region-Macro MAE↓	SSIM↑
GPT-Image-1.5	Dark-heavy	0.386	0.335	0.329	0.360
	HDR	0.360	0.315	0.295	0.412
	Highlight-heavy	0.378	0.358	0.306	0.428
	Low-light	0.374	0.313	0.312	0.370
GPT-Image-2.0	Dark-heavy	0.333	0.289	0.295	0.408
	HDR	0.298	0.262	0.245	0.468
	Highlight-heavy	0.292	0.267	0.254	0.477
	Low-light	0.310	0.268	0.263	0.440
Qwen-Image-2.0	Dark-heavy	0.398	0.347	0.329	0.364
	HDR	0.393	0.347	0.314	0.390
	Highlight-heavy	0.351	0.339	0.294	0.430
	Low-light	0.385	0.342	0.315	0.370
Doubao Seedream 5.0	Dark-heavy	0.335	0.296	0.321	0.347
	HDR	0.343	0.290	0.290	0.373
	Highlight-heavy	0.359	0.283	0.320	0.362
	Low-light	0.332	0.284	0.303	0.366
FLUX.1-Kontext-dev	Dark-heavy	0.345	0.324	0.284	0.402
	HDR	0.341	0.321	0.280	0.433
	Highlight-heavy	0.321	0.333	0.268	0.453
	Low-light	0.315	0.305	0.259	0.420
FireRed-Image-Edit	Dark-heavy	0.345	0.292	0.313	0.434
	HDR	0.337	0.289	0.277	0.463
	Highlight-heavy	0.344	0.276	0.306	0.457
	Low-light	0.336	0.286	0.292	0.446
Gemini 3 Pro Image	Dark-heavy	**0.295**	**0.261**	**0.258**	0.493
	HDR	**0.280**	**0.239**	**0.232**	**0.540**
	Highlight-heavy	**0.286**	**0.255**	**0.242**	**0.540**
	Low-light	**0.284**	**0.245**	**0.233**	**0.528**
LongCat-Image-Edit	Dark-heavy	0.397	0.336	0.344	0.381
	HDR	0.372	0.319	0.304	0.437
	Highlight-heavy	0.371	0.334	0.301	0.463
	Low-light	0.396	0.341	0.340	0.406
Intrinsic Image Diffusion	Dark-heavy	0.307	0.265	0.308	**0.502**
	HDR	0.297	0.257	0.245	0.534
	Highlight-heavy	0.337	0.257	0.295	0.529
	Low-light	0.291	0.246	0.271	0.519
DNF-Intrinsic	Dark-heavy	0.395	0.315	0.381	0.422
	HDR	0.375	0.314	0.333	0.458
	Highlight-heavy	0.341	0.279	0.302	0.508
	Low-light	0.407	0.318	0.376	0.409

Scores are reported as per-image averages within each OpenRooms-FF slice, with slice assignment determined by `openroomsff_stresstest.csv`. Only methods with a complete, scoreable stress-axis run under this protocol are included. The evaluated slices are dark-heavy, HDR, highlight-heavy, and low-light. Image-macro RMSE is the scalar error summary on both axes; Bin-BMAE, Region-Macro MAE, and SSIM provide range-balanced, region-balanced, and structural diagnostics. Stress-slice values summarize different retained subsets and should not be directly compared with the main-axis aggregate. Lower RMSE/Bin-BMAE/Region-Macro MAE and higher SSIM are better. Boldface marks the best value within each slice and metric using unrounded scores; displayed rounding may produce apparent ties.

Gemini 3 Pro Image gives the strongest reported roughness point estimates across the stress slices. It gives the lowest three errors and highest SSIM on HDR, highlight-heavy, and low-light, and it also gives the lowest three errors on dark-heavy. Intrinsic Image Diffusion provides the only metric-wise exception on dark-heavy, where its SSIM of 0.502 exceeds Gemini's 0.493. Thus Gemini's point-estimate advantage is not confined to a single lighting condition, while the remaining methods exhibit larger metric- and slice-dependent trade-offs. Agreement across global, range-balanced, and region-balanced errors makes this a broader pattern than success on dominant pixels alone. However, strong

performance across scene slices can arise from either useful appearance cues or stable category-level material priors. Since the same surfaces are not tracked across controlled lighting changes here, these results do not establish illumination-invariant roughness inference.

E.1.4 DEPTH STRESS-SUBSET RESULTS

Table 17: Stress-test relative-depth results under the official depth protocol.

Method	Slice	AbsRel-AI↓	RMSE-AI↓	MAE-AI↓	δ_1-AI↑	δ_2-AI↑	Boundary F1↑
GPT-Image-2.0	Dark-heavy	0.188	0.385	0.299	0.824	0.944	0.266
	HDR	0.154	0.379	0.293	0.824	0.947	0.280
	Highlight-heavy	0.158	0.372	0.286	0.834	0.947	0.256
	Low-light	0.095	0.342	0.267	0.916	0.983	0.340
GPT-Image-1.5	Dark-heavy	0.251	0.524	0.403	0.733	0.919	0.030
	HDR	0.190	0.512	0.391	0.750	0.928	0.031
	Highlight-heavy	0.196	0.484	0.368	0.766	0.933	0.025
	Low-light	0.131	0.507	0.387	0.840	0.967	0.039
Wan2.7-Image	Dark-heavy	0.163	0.363	0.281	0.855	0.956	0.193
	HDR	0.134	0.360	0.278	0.856	0.960	0.194
	Highlight-heavy	0.146	0.356	0.273	0.853	0.958	0.180
	Low-light	0.097	0.366	0.286	0.913	0.979	0.220
Qwen-Image-2.0	Dark-heavy	0.173	0.416	0.312	0.832	0.944	0.018
	HDR	0.144	0.406	0.302	0.840	0.946	0.018
	Highlight-heavy	0.141	0.360	0.264	0.862	0.954	0.018
	Low-light	0.122	0.464	0.355	0.857	0.962	0.020
Doubao Seedream 5.0	Dark-heavy	0.181	0.462	0.353	0.782	0.946	0.131
	HDR	0.161	0.452	0.346	0.795	0.951	0.144
	Highlight-heavy	0.165	0.449	0.344	0.792	0.949	0.133
	Low-light	0.123	0.474	0.365	0.864	0.976	0.163
Gemini 3 Pro Image	Dark-heavy	0.194	0.390	0.303	0.819	0.951	**0.338**
	HDR	0.145	0.378	0.293	0.834	0.955	**0.342**
	Highlight-heavy	0.150	0.370	0.286	0.836	0.955	**0.324**
	Low-light	0.098	0.356	0.278	0.919	0.986	**0.396**
FLUX.1-Kontext-dev	Dark-heavy	0.422	0.850	0.690	0.498	0.790	0.004
	HDR	0.339	0.846	0.684	0.512	0.800	0.003
	Highlight-heavy	0.355	0.786	0.631	0.546	0.816	0.004
	Low-light	0.281	0.957	0.781	0.557	0.840	0.003
FireRed-Image-Edit	Dark-heavy	0.413	0.827	0.673	0.507	0.801	0.066
	HDR	0.338	0.840	0.686	0.508	0.803	0.056
	Highlight-heavy	0.336	0.761	0.613	0.556	0.828	0.069
	Low-light	0.273	0.923	0.754	0.573	0.851	0.074
LongCat-Image-Edit	Dark-heavy	0.225	0.473	0.362	0.777	0.932	0.209
	HDR	0.182	0.470	0.361	0.780	0.932	0.210
	Highlight-heavy	0.206	0.495	0.386	0.753	0.925	0.185
	Low-light	0.119	0.441	0.345	0.869	0.975	0.254
Depth Anything 3	Dark-heavy	**0.044**	**0.130**	**0.070**	0.973	0.990	0.248
	HDR	**0.034**	**0.121**	**0.065**	**0.981**	0.993	0.259
	Highlight-heavy	**0.038**	**0.122**	**0.066**	0.978	0.992	0.232
	Low-light	**0.012**	**0.085**	**0.034**	**0.996**	**0.999**	0.315
Lotus	Dark-heavy	0.052	0.151	0.096	**0.975**	**0.991**	0.325
	HDR	0.045	0.147	0.094	0.978	**0.995**	0.340
	Highlight-heavy	0.048	0.143	0.091	**0.978**	**0.994**	0.311
	Low-light	0.031	0.131	0.086	0.992	0.999	0.384

Only methods with a complete, scoreable stress-axis run under this protocol are included. For dark-heavy, HDR, and highlight-heavy, we first average per-image metrics within each source and slice, then macro-average the two source-wise means to obtain the reported source-balanced score. The low-light slice is available only in OpenRooms-FF; its reported values are therefore OpenRooms-FF-only per-image means. Lower AbsRel-AI, RMSE-AI, and MAE-AI are better; higher δ_1-AI, δ_2-AI, and Boundary F1 are better. Boldface marks the best result within each slice and metric using unrounded scores; displayed rounding may produce apparent ties.

Depth Anything 3 gives the lowest reported AbsRel-AI, RMSE-AI, and MAE-AI point estimates on every stress slice, including 0.034/0.121/0.065 on HDR and 0.012/0.085/0.034 on the OpenRooms-FF-only low-light slice. Boundary recovery follows a different ordering: Gemini 3 Pro Image gives the highest reported Boundary F1 on all four slices, from 0.324 under highlight-heavy conditions to 0.396 under low light. Lotus remains competitive on threshold accuracy and Boundary F1 but does not give the lowest scalar errors. The stress results therefore separate affine-aligned scalar geometry from local depth-discontinuity recovery rather than identifying one method as uniformly best. The contrast with Qwen is informative: its scalar error can be competitive among editors, yet its Boundary

F1 stays between 0.018 and 0.020. Thus comparable scene-scale depth agreement need not imply comparable localization of discontinuities. Boundary displacement or smoothing could contribute to this pattern because matching allows no spatial tolerance; the present scores cannot identify the underlying mechanism. They do show why a scalar depth score alone is insufficient for uses that depend on precise occlusion boundaries.

E.1.5 METALLIC STRESS-SUBSET RESULTS

The four metallic metrics favor different methods across lighting conditions. The all-zero prediction has lower MAE than every evaluated method in all six conditions (0.017 at displayed precision, compared with GPT-Image-2.0's 0.018), indicating that the lowest learned-method full-map error remains close to a background-only prediction. MVInverse gives the lowest learned-method RB-MAE under normal-light, low-light, HDR, dark-heavy, and mixed-temperature lighting, while Doubao Seedream 5.0 is lowest under highlight-heavy lighting. Localization and detection produce another ordering: Gemini 3 Pro Image gives the highest reported IoU and AP under normal-light, HDR, dark-heavy, and mixed-temperature lighting; under low light it is highest on IoU while Doubao is highest on AP, and under highlight-heavy GPT-Image-2.0 is highest on IoU while Doubao is highest on AP. These metric-wise reversals reinforce reporting the MAE/RB-MAE/IoU/AP components separately. The matched lighting conditions expose a further distinction between error stability and physical invariance. GPT-Image-2.0's MAE rounds to 0.018 in all six conditions, while its IoU changes from 0.036 in low-light/HDR to 0.103 in highlight-heavy lighting. Gemini's IoU instead falls from 0.129 under normal light to 0.091 under low light. The material parameters are unchanged across the matched renders, so stable global error does not establish stable region recovery. These observations motivate a paired prediction-consistency analysis; the current condition summaries and single generations cannot separate reliance on lighting cues from sampling variability.

Table 18: Metallic results across the six renderer-defined PBR-Rooms lighting conditions under the controlled direct RGB-to-map track.

Method	Slice	MAE↓	RB-MAE↓	IoU↑	AP↑
GPT-Image-1.5	Normal light	0.024	0.455	0.014	0.020
	Low-light	0.022	0.451	0.024	0.030
	HDR	0.024	0.459	0.014	0.026
	Dark-heavy	0.022	0.460	0.006	0.014
	Highlight-heavy	0.023	0.452	0.013	0.018
	Mixed temperature	0.023	0.453	0.015	0.021
GPT-Image-2.0	Normal light	0.018	0.419	0.067	0.078
	Low-light	0.018	0.450	0.036	0.049
	HDR	0.018	0.447	0.036	0.050
	Dark-heavy	0.018	0.437	0.041	0.050
	Highlight-heavy	0.018	0.412	**0.103**	0.123
	Mixed temperature	0.018	0.409	0.067	0.080
Wan2.7-Image	Normal light	0.100	0.413	0.010	0.026
	Low-light	0.093	0.401	0.018	0.030
	HDR	0.114	0.416	0.006	0.018
	Dark-heavy	0.097	0.413	0.009	0.022
	Highlight-heavy	0.096	0.411	0.012	0.024
	Mixed temperature	0.098	0.406	0.018	0.024
Doubao Seedream 5.0	Normal light	0.040	0.358	0.080	0.125
	Low-light	0.038	0.386	0.073	**0.118**
	HDR	0.040	0.374	0.081	0.116
	Dark-heavy	0.039	0.356	0.067	0.114
	Highlight-heavy	0.040	**0.333**	0.096	**0.161**
	Mixed temperature	0.038	0.365	0.057	0.106
Qwen-Image-2.0	Normal light	0.518	0.490	0.004	0.009
	Low-light	0.538	0.497	0.005	0.008
	HDR	0.532	0.495	0.005	0.009
	Dark-heavy	0.521	0.493	0.004	0.008
	Highlight-heavy	0.549	0.494	0.004	0.008
	Mixed temperature	0.506	0.497	0.005	0.008
FLUX.1-Kontext-dev	Normal light	0.418	0.458	0.007	0.013
	Low-light	0.379	0.462	0.004	0.009
	HDR	0.420	0.466	0.006	0.012
	Dark-heavy	0.400	0.472	0.004	0.008
	Highlight-heavy	0.432	0.460	0.003	0.008
	Mixed temperature	0.437	0.470	0.004	0.011

Continued on next page

Table 18 (continued)

Method	Slice	MAE↓	RB-MAE↓	IoU↑	AP↑
FireRed-Image-Edit	Normal light	0.269	0.507	0.009	0.017
	Low-light	0.258	0.501	0.010	0.019
	HDR	0.255	0.523	0.003	0.011
	Dark-heavy	0.265	0.519	0.003	0.009
	Highlight-heavy	0.270	0.516	0.007	0.020
	Mixed temperature	0.278	0.529	0.012	0.021
Gemini 3 Pro Image	Normal light	0.052	0.364	**0.129**	**0.162**
	Low-light	0.047	0.401	**0.091**	0.104
	HDR	0.034	0.394	**0.093**	**0.118**
	Dark-heavy	0.059	0.403	**0.097**	**0.122**
	Highlight-heavy	0.044	0.389	0.102	0.119
	Mixed temperature	0.048	0.355	**0.092**	**0.118**
LongCat-Image-Edit	Normal light	0.582	0.506	0.006	0.018
	Low-light	0.628	0.493	0.010	0.020
	HDR	0.616	0.524	0.008	0.013
	Dark-heavy	0.655	0.487	0.010	0.019
	Highlight-heavy	0.508	0.500	0.009	0.015
	Mixed temperature	0.535	0.476	0.010	0.017
MVInverse	Normal light	0.173	**0.343**	0.024	0.052
	Low-light	0.189	**0.346**	0.023	0.050
	HDR	0.195	**0.349**	0.025	0.047
	Dark-heavy	0.182	**0.346**	0.019	0.038
	Highlight-heavy	0.163	0.346	0.031	0.063
	Mixed temperature	0.178	**0.342**	0.029	0.049
DNF-Intrinsic	Normal light	0.135	0.466	0.021	0.031
	Low-light	0.103	0.471	0.012	0.025
	HDR	0.114	0.471	0.006	0.022
	Dark-heavy	0.107	0.465	0.007	0.018
	Highlight-heavy	0.159	0.469	0.024	0.027
	Mixed temperature	0.153	0.471	0.023	0.032
All-zero prediction	Normal light	*0.017*	*0.470*	*0.000*	*0.010*
	Low-light	*0.017*	*0.470*	*0.000*	*0.010*
	HDR	*0.017*	*0.470*	*0.000*	*0.010*
	Dark-heavy	*0.017*	*0.470*	*0.000*	*0.010*
	Highlight-heavy	*0.017*	*0.470*	*0.000*	*0.010*
	Mixed temperature	*0.017*	*0.470*	*0.000*	*0.010*

Only methods with complete, scoreable outputs across all six lighting conditions are included. The deterministic all-zero baseline uses the same evaluator and mandatory `MetallicEvalMask` with `–mask_fallback none`. Each condition uses the same matched scene–camera groups: MAE averages all groups, while RB-MAE, IoU, and AP use those with metallic-positive ground-truth support. Lower MAE/RB-MAE and higher IoU/AP are better. Boldface marks the best learned-method RB-MAE/IoU/AP values within each condition using unrounded scores; MAE is not emphasized because the all-zero baseline is lower. Displayed rounding may produce apparent ties.

E.2 Access-Protocol Ablations

Diagnostic subset sampling. Access ablations use fixed target-specific subsets sampled from the retained main-axis and stress-axis partitions but reported separately from official aggregation. Their scores are diagnostic and are not included as alternative official scores. Depth, normal, and albedo each use a stratified random sample of 96 images balanced across OpenRooms-FF and InteriorVerse and across main and stress partitions. Roughness uses a stratified random sample of 96 OpenRooms-FF images balanced between main and stress partitions. These diagnostic images already belong to the retained evaluation data and are not scene-disjoint from the retained splits. The same target-specific subset is shared by all methods and variants in the corresponding comparison. Where a diagnostic summary uses a different aggregation from the official evaluator, the distinction is stated in the table heading and note; such values support only within-table setting comparisons and are not numerically compared with official scores.

E.2.1 Albedo Protocol Diagnostics

We compare the official RGB-only instruction (A0) with cached query-image illumination analysis (A1), weaker prompts (A2–A3), and a fixed external exemplar (B0) on the subset in Appendix E.2. A1 uses no ground-truth or multi-view evidence but adds a separate Doubao Seed 2.0 Pro analysis component shared unchanged across editors; B0 changes the evidence condition by adding an external exemplar.

Table 19: Albedo access-protocol diagnostics on the 96-image diagnostic subset across six generative models.

Method	Setting	Balanced Macro MAE↓	PSNR↑	SSIM↑	LPIPS↓
Doubao Seedream 5.0	A0: structure-preserving prompt only	0.1703	14.0558	0.6330	0.3365
	A1: + query-image illumination analysis	0.1683	14.0757	0.6407	**0.3243**
	A2: weakened prompt only	0.1673	14.2763	0.6448	0.3350
	A3: minimal prompt only	**0.1564**	**14.7119**	0.6110	0.3600
	B0: minimal prompt + fixed exemplar	0.1573	14.6472	**0.6462**	0.3404
GPT-Image-1.5	A0: structure-preserving prompt only	0.2251	11.3238	0.5310	0.5124
	A1: + query-image illumination analysis	**0.2046**	**12.0354**	0.5325	**0.5070**
	A2: weakened prompt only	0.2096	11.8557	0.5293	0.5171
	A3: minimal prompt only	0.2348	10.9471	**0.5740**	0.5646
	B0: minimal prompt + fixed exemplar	0.2909	9.5920	0.5473	0.6121
GPT-Image-2.0	A0: structure-preserving prompt only	**0.1430**	**15.7205**	0.7279	**0.2707**
	A1: + query-image illumination analysis	0.1476	15.4696	0.7129	0.2865
	A2: weakened prompt only	0.1455	15.6345	**0.7369**	0.2753
	A3: minimal prompt only	0.1467	15.5931	0.7024	0.3081
	B0: minimal prompt + fixed exemplar	0.2146	12.0861	0.5677	0.5560
Gemini 3 Pro Image	A0: structure-preserving prompt only	0.1673	14.4864	0.7159	0.2969
	A1: + query-image illumination analysis	0.1596	14.8784	0.7227	**0.2839**
	A2: weakened prompt only	0.1643	14.7620	0.7229	0.2892
	A3: minimal prompt only	**0.1506**	**15.4877**	**0.7409**	0.2881
	B0: minimal prompt + fixed exemplar	0.1605	14.8553	0.7193	0.3254
Wan2.7-Image	A0: structure-preserving prompt only	0.2117	12.3950	0.6116	0.4386
	A1: + query-image illumination analysis	**0.2030**	**12.6482**	**0.6267**	**0.3826**
	A2: weakened prompt only	0.2217	11.9960	0.6073	0.4367
	A3: minimal prompt only	0.2211	12.1805	0.5981	0.4756
	B0: minimal prompt + fixed exemplar	0.2769	10.2075	0.4833	0.7581
Qwen-Image-2.0	A0: structure-preserving prompt only	**0.1996**	**12.2783**	**0.5276**	**0.4208**
	A1: + query-image illumination analysis	0.2044	11.9708	0.5172	0.4371
	A2: weakened prompt only	0.2073	12.0846	0.5262	0.4393
	A3: minimal prompt only	0.2466	10.7657	0.5168	0.4971
	B0: minimal prompt + fixed exemplar	0.2565	10.2429	0.4276	0.6745

A0 is the controlled RGB-only setting used in the official albedo evaluation. A1 adds cached illumination analysis generated once from the same query RGB image by Doubao Seed 2.0 Pro and shared unchanged across editors. A2 and A3 weaken the prompt constraints, while B0 adds a fixed external exemplar and therefore changes the evidence condition. Balanced Macro MAE first averages per-image MAE within each source and partition, then gives equal weight to OpenRooms-FF and InteriorVerse and to the main-axis and stress-axis partitions. It therefore preserves the official image-macro and equal-source aggregation within each partition while balancing the two diagnostic partitions. PSNR, SSIM, and LPIPS are averaged equally over images; because all four source–partition cells contain 24 images, these summaries are likewise balanced across cells. These results characterize protocol sensitivity and are not used for post-hoc selection of the main protocol. Boldface marks the numerical optimum within each model's diagnostic settings using unrounded scores; displayed rounding may produce apparent ties. It does not indicate an official comparison result.

A1 improves all four metrics for Doubao, GPT-Image-1.5, Gemini, and Wan, but worsens all four for GPT-Image-2.0 and Qwen. For example, GPT-Image-1.5's MAE decreases from 0.2251 to 0.2046, whereas GPT-Image-2.0's increases from 0.1430 to 0.1476. Thus supplying a shared lighting description is not sufficient to improve reflectance decomposition. The difference could reflect how editors integrate auxiliary text or how well that text matches their useful existing cues; the experiment does not identify the cause.

Prompt weakening is likewise informative. Doubao's minimal prompt lowers MAE relative to A0 but worsens SSIM and LPIPS, so numerical improvement need not preserve structure or perceptual fidelity. Relative to the minimal-prompt A3 condition, adding an exemplar in B0 worsens MAE for every tested editor. The corresponding structural effects are mixed: Doubao gains SSIM while GPT-Image-2.0 loses it. A reference can therefore change the output beyond simply clarifying the

target concept. These patterns argue against diagnosing low albedo scores as absent conceptual knowledge or absent task training. A1 adds an inference component and B0 adds evidence; both remain diagnostics, with no same-setting repeated generations to isolate their causal effects.

E.2.2 Normal Exemplar Diagnostics

We next examine the role of exemplar conditioning and prompt constraint strength under the fixed three-exemplar normal protocol.

Table 20: Normal elicitation diagnostics on the 96-image merged evaluation set across three generative models.

Method	Setting	Mean Angular↓	Acc@22.5↑	Acc@30↑
Doubao Seedream 5.0	A0: full prompt + three-exemplar aggregate	**58.56**	**0.1151**	**0.2030**
	A1: full prompt only	71.46	0.0912	0.1397
	A2: minimal prompt + three-exemplar aggregate	62.17	0.0886	0.1573
	A3: minimal prompt only	73.61	0.0694	0.1212
GPT-Image-1.5	A0: full prompt + three-exemplar aggregate	**46.64**	**0.2016**	**0.3171**
	A1: full prompt only	51.65	0.1826	0.2624
	A2: minimal prompt + three-exemplar aggregate	48.04	0.1852	0.2793
	A3: minimal prompt only	54.35	0.1305	0.2085
FireRed-Image-Edit	A0: full prompt + three-exemplar aggregate	61.69	0.1938	0.2445
	A1: full prompt only	57.81	0.3192	0.3922
	A2: minimal prompt + three-exemplar aggregate	61.39	0.1949	0.2481
	A3: minimal prompt only	**55.01**	**0.3801**	**0.4344**

Metrics are averaged equally over images. A0 and A2 additionally average scores equally across the three fixed exemplar-conditioned runs; A1 and A3 remove exemplar conditioning. Mean Angular is measured in degrees and reported as an error diagnostic; Acc@22.5 is the primary threshold metric for local directional correctness. This table diagnoses access-setting sensitivity and should not be interpreted as part of the official comparison. Boldface marks the numerical optimum within each model's diagnostic settings using unrounded scores; displayed rounding may produce apparent ties.

The A0–A1 and A2–A3 comparisons show opposite responses to exemplar conditioning. Doubao and GPT-Image-1.5 improve on all three metrics with exemplars under both prompt strengths. FireRed worsens on all three: under the full prompt, its Acc@22.5 falls from 0.3192 to 0.1938, while GPT-Image-1.5 rises from 0.1826 to 0.2016. Under the minimal prompt, FireRed's accuracy falls from 0.3801 to 0.1949. The official exemplar setting therefore reverses the FireRed–GPT ordering observed without exemplars.

The exemplars define the coordinate convention but may also change generated output behavior. Their opposite effects show that shared evidence is not equivalent to equally effective elicitation. Distinguishing convention clarification from context interference would require additional controls. Averaging the pre-specified E1–E3 scores reduces reliance on a single reference without establishing protocol-independent rankings, and these runs do not separate exemplar effects from generation randomness.

E.2.3 Roughness Prior-Utilization Diagnostics

Auxiliary priors have model- and metric-dependent effects. RGB-only A0 gives the best Balanced RMSE, Region-Macro MAE, and SSIM for Doubao and GPT-Image-2.0, and it leads all four metrics for Gemini. By contrast, GPT-Image-1.5 benefits most in Balanced RMSE and region error from soft segmentation, while Wan2.7-Image obtains its best values on all four metrics from the fixed exemplar. No added cue improves all models consistently. These diagnostics use pooled RMSE within each partition and equally weight the main-axis and stress-axis partitions; they remain separate from the official image-macro main-axis score.

The metric profile also separates different forms of benefit. For GPT-Image-1.5, soft segmentation reduces Balanced RMSE from 0.468 to 0.407 and Region-Macro MAE from 0.381 to 0.315, whereas the fixed exemplar produces its best Bin-BMAE and raises SSIM from 0.236 to 0.362. In contrast, every added prior worsens all four Gemini metrics. GPT-Image-2.0 and Doubao obtain only small Bin-BMAE gains from soft segmentation (0.266 to 0.262 and 0.312 to 0.311, respectively) while

Table 21: Roughness prior-utilization diagnostics across five generative models on the 96-image merged OpenRooms-FF diagnostic set.

Method	Setting	Balanced RMSE↓	Bin-BMAE↓	Region-Macro MAE↓	SSIM↑
Doubao Seedream 5.0	A0: prompt	**0.387**	0.312	**0.319**	**0.452**
	A1: prompt + segmentation soft	0.393	**0.311**	0.322	0.444
	A2: prompt + segmentation region-fill	0.423	0.336	0.336	0.412
	A3: prompt + fixed exemplar	0.416	0.334	0.339	0.420
GPT-Image-1.5	A0: prompt	0.468	0.356	0.381	0.236
	A1: prompt + segmentation soft	**0.407**	0.332	**0.315**	0.294
	A2: prompt + segmentation region-fill	0.421	0.339	0.344	0.279
	A3: prompt + fixed exemplar	0.426	**0.330**	0.335	**0.362**
GPT-Image-2.0	A0: prompt	**0.319**	0.266	**0.264**	**0.459**
	A1: prompt + segmentation soft	0.324	**0.262**	0.271	0.442
	A2: prompt + segmentation region-fill	0.326	0.267	0.275	0.434
	A3: prompt + fixed exemplar	0.323	0.272	0.275	0.435
Gemini 3 Pro Image	A0: prompt	**0.292**	**0.252**	**0.243**	**0.539**
	A1: prompt + segmentation soft	0.324	0.265	0.261	0.516
	A2: prompt + segmentation region-fill	0.322	0.266	0.267	0.524
	A3: prompt + fixed exemplar	0.321	0.260	0.268	0.511
Wan2.7-Image	A0: prompt	0.421	0.336	0.327	0.442
	A1: prompt + segmentation soft	0.415	0.331	0.325	0.446
	A2: prompt + segmentation region-fill	0.426	0.342	0.333	0.431
	A3: prompt + fixed exemplar	**0.396**	**0.320**	**0.315**	**0.452**

Balanced RMSE is the arithmetic mean of the valid-pixel pooled RMSE computed separately on the 48-image main-axis and 48-image stress-axis diagnostic partitions, which receive equal weight. It is a diagnostic definition and is not numerically interchangeable with the official image-macro main-axis RMSE. SSIM is averaged over images. Bin-BMAE pools valid pixels within each ground-truth roughness bin across the merged set and averages the non-empty bin errors equally. Region-Macro MAE averages region-level errors equally across the merged set. A0 is the controlled RGB-to-map setting. A1–A3 add soft segmentation, segmentation-based region filling, or a fixed RGB–roughness exemplar and therefore change the access setting. Lower RMSE/Bin-BMAE/Region-Macro MAE and higher SSIM are better. Boldface marks numerical optima within each model's diagnostic settings using unrounded scores; displayed rounding may produce apparent ties. It does not indicate an official comparison result.

their global, regional, and structural scores degrade. The region-fill setting is never the best variant for any evaluated model. Thus, more explicit spatial structure can help a model that does not exploit the RGB-only setting as effectively, but it can also interfere with already useful output behavior. Because each setting has one retained generation per image, the table cannot separate cue effects from generation randomness or determine whether an improvement reflects better within-region physical estimation or changed region-level material priors.

E.2.4 Depth Protocol Diagnostics

For **depth**, we additionally report protocol diagnostics for exemplar- and segmentation-assisted variants. These variants change the access setting and are therefore not used for the official controlled RGB-to-depth comparison.

The response to additional depth cues is strongly model-dependent. For Doubao, the fixed RGB–depth exemplar A1 improves all three affine-aligned scalar errors and both ordinal correlations relative to A0, but reduces Boundary F1 from 0.142 to 0.125. GPT-Image-1.5 changes only marginally in ordinal correlation under A1 while its scalar errors slightly worsen and Boundary F1 remains effectively unchanged. Wan2.7-Image degrades substantially under A1 on every reported metric. The segmentation-reference A2 setting also fails to provide a consistent gain: Doubao improves only AbsRel-AI while its remaining fields are tied or worse, and GPT-Image-1.5 and Wan2.7-Image are uniformly tied or worse than A0. Auxiliary depth cues therefore shift the recovered structure in a model-specific manner rather than uniformly improving the official RGB-only setting. Qwen provides the strongest exemplar-interference example: A1 increases AbsRel-AI from 0.141 to 0.317 and lowers Spearman's correlation from 0.944 to 0.467, whereas segmentation A2 leaves both almost unchanged. GPT-Image-2.0 responds differently: A2 improves scalar errors, ordinal correlations, and Boundary F1 relative to A0. These contrasts suggest distinct ways of using spatial and paired-

Table 22: Depth access-setting diagnostics on a 96-image subset. Spearman's ρ and Kendall's τ are computed before affine fitting, while AbsRel-AI/RMSE-AI/MAE-AI are computed after polarity selection and affine alignment.

Method	Setting	AbsRel-AI↓	RMSE-AI↓	MAE-AI↓	Spearman ρ ↑	Kendall τ ↑	B-F1↑
Doubao Seedream 5.0	A0	0.187	0.498	0.384	0.792	0.646	**0.142**
	A1	**0.141**	**0.433**	**0.328**	**0.839**	**0.699**	0.125
	A2	0.178	0.499	0.384	0.771	0.620	0.125
GPT-Image-1.5	A0	**0.217**	**0.587**	**0.444**	0.689	0.544	**0.004**
	A1	0.227	0.591	0.451	**0.692**	**0.550**	0.004
	A2	0.235	0.611	0.467	0.689	0.543	0.004
GPT-Image-2.0	A0	0.159	0.408	0.317	0.837	0.722	0.269
	A1	0.147	0.389	0.304	0.846	0.720	0.253
	A2	**0.143**	**0.362**	**0.279**	**0.898**	**0.781**	**0.280**
Gemini 3 Pro Image	A0	0.168	0.416	0.326	0.806	0.682	**0.333**
	A1	0.158	0.406	0.313	0.813	0.685	0.318
	A2	**0.135**	**0.361**	**0.279**	**0.851**	**0.725**	0.315
Wan2.7-Image	A0	**0.132**	**0.362**	**0.281**	**0.925**	**0.803**	**0.184**
	A1	0.187	0.524	0.406	0.729	0.599	0.145
	A2	0.135	0.367	0.283	0.916	0.795	0.172
Qwen-Image-2.0	A0	**0.141**	0.411	0.304	**0.944**	**0.856**	**0.018**
	A1	0.317	0.770	0.609	0.467	0.338	0.001
	A2	0.142	**0.409**	**0.303**	0.944	0.855	0.018

A0: RGB + prompt; A1: RGB + prompt + fixed RGB–depth exemplar; A2: RGB + prompt + segmentation reference. Each setting contains 96 evaluated images, with 24 from each dataset–split combination of OpenRooms-FF/InteriorVerse and main-axis/stress-test. Scores are averaged over images. AI denotes scoring after prediction min–max normalization, scoring-time polarity selection, and least-squares affine fitting. Spearman's ρ and Kendall's τ are computed on the selected-polarity prediction before affine fitting. B-F1 denotes the DepthPro-style inverse-depth boundary F1 diagnostic. Boldface marks the best setting within each model using unrounded scores; displayed rounding may produce apparent ties. These access-setting diagnostics are not part of the official comparison.

reference evidence, rather than a universal benefit from providing more context. The correlations here describe pixel-depth ordering, not stability of model rankings. Changes in reference-style transfer or conditioning could explain the pattern, but no architectural cause is established by these single-generation diagnostics.

F Additional Quantitative Results and Uncertainty

This appendix expands the official main-axis results without changing the access assumptions. It contains normal per-exemplar results, source-specific results for multi-source targets, and scene-cluster bootstrap uncertainty estimates. Stress slices and access-protocol ablations are reported only in Appendix E.

F.1 Normal Per-Exemplar Main-Axis Results

Table 23 reports the per-exemplar normal results corresponding to the averaged image-editor entries in the main normal table. The references change relative ordering as well as score magnitude: GPT-Image-1.5 exceeds FireRed on Acc@22.5 under E1 and E2, but falls below it under E3 (0.2050 versus 0.2092). The average therefore summarizes these three pre-specified conditions, not an invariant model ranking. Because each exemplar has only one retained generation per query, the variation cannot be partitioned into exemplar and generation effects. The three RGB–normal exemplar pairs are fixed before evaluation and are shared across all test images and evaluated image editors. The Avg row is the arithmetic mean over the three exemplar-conditioned runs and is the value reported in the main text. These per-exemplar rows are included to expose residual sensitivity to exemplar choice, not to select a best exemplar after observing test performance.

Table 23: Per-exemplar source-balanced surface normal estimation results on OpenRooms-FF and InteriorVerse.

Method	Exemplar	Mean↓	Median↓	RMSE↓	Acc@11.25↑	Acc@22.5↑	Acc@30↑
GPT-Image-1.5	E1	45.681	43.647	50.963	0.049	0.1886	0.311
	E2	43.913	41.569	49.383	0.063	0.2190	0.343
	E3	44.779	42.304	50.234	0.059	0.2050	0.325
	Avg	44.791	42.507	50.194	0.056	0.2042	0.326
Doubao Seedream 5.0	E1	56.463	50.376	63.919	0.040	0.1395	0.234
	E2	57.329	51.385	64.079	0.029	0.1177	0.211
	E3	55.092	48.870	62.764	0.037	0.1402	0.237
	Avg	56.294	50.210	63.587	0.035	0.1325	0.227
FireRed-Image-Edit	E1	63.377	67.100	71.504	0.085	0.1731	0.219
	E2	60.689	62.344	69.329	0.105	0.1972	0.243
	E3	60.048	63.807	68.964	0.101	0.2092	0.260
	Avg	61.371	64.417	69.932	0.096	0.1931	0.240
Lotus	–	9.590	**5.005**	16.680	0.810	0.9042	0.930
StableNormal	–	**9.115**	5.455	**15.530**	**0.818**	**0.9183**	**0.942**

Scores are source-balanced macro-averages, assigning equal weight to OpenRooms-FF and InteriorVerse regardless of source size. Each entry is computed from the full-precision source-specific scores rather than from the rounded values displayed in the two source-specific tables. E1–E3 denote the three fixed RGB–normal exemplar-conditioned runs. Avg is obtained from the full-precision source-specific values by equally averaging the three exemplar-conditioned runs within each source and then the two source scores. The displayed Avg rows match the corresponding entries in the main table; recomputing from the rounded E1–E3 entries may differ in the final displayed digit. Specialized RGB-to-normal predictors are evaluated in native inference mode without exemplars. Acc@22.5 is the primary metric. It is shown to four decimal places to match the bootstrap point-estimate table. Per-exemplar rows expose residual sensitivity to exemplar choice and are not used to select a best exemplar. Boldface marks the best value using unrounded scores; displayed rounding may produce apparent ties.

F.2 Source-Specific Main-Axis Results

F.2.1 Depth Source-Specific Results

Table 24: Relative-depth results on the OpenRooms-FF main-axis subset under the affine-invariant single-image evaluation protocol.

Method	AbsRel-AI↓	RMSE-AI↓	MAE-AI↓	δ_1-AI↑	δ_2-AI↑	Boundary F1↑
GPT-Image-2.0	0.102	0.340	0.268	0.901	0.978	0.315
GPT-Image-1.5	0.134	0.486	0.373	0.834	0.965	0.038
Wan2.7-Image	0.097	0.346	0.271	0.915	0.981	0.213
Doubao Seedream 5.0	0.121	0.447	0.344	0.866	0.977	0.166
FLUX.1-Kontext-dev	0.275	0.887	0.725	0.569	0.848	0.003
FireRed-Image-Edit	0.268	0.859	0.708	0.577	0.857	0.070
Gemini 3 Pro Image	0.104	0.355	0.278	0.905	0.983	0.380
LongCat-Image-Edit	0.142	0.485	0.385	0.823	0.957	0.239
Qwen-Image-2.0	0.113	0.417	0.316	0.876	0.964	0.019
Depth Anything 3	**0.012**	**0.081**	**0.033**	**0.996**	**0.999**	0.304
Lotus	0.031	0.121	0.078	0.993	0.999	**0.384**

Scores are per-image averages over the 1,125-image OpenRooms-FF main-axis subset. Lower AbsRel-AI, RMSE-AI, and MAE-AI are better; higher δ_1-AI, δ_2-AI, and Boundary F1 are better. Boldface marks the best value using unrounded scores; displayed rounding may produce apparent ties.

Depth source-specific interpretation. Both sources preserve the task-specific advantage on scalar relative-depth error and threshold accuracy, while Boundary F1 remains source-dependent and can favor an image editor. This supports the distinction between global relative geometry and local discontinuity recovery. For example, Gemini trails Lotus slightly in Boundary F1 on OpenRooms-FF (0.380 versus 0.384), but leads on InteriorVerse (0.263 versus 0.252). Its pooled boundary ordering is therefore not a source-uniform advantage. Qwen's low boundary scores persist on both sources despite competitive editor scalar errors, making the separation between these two recovery properties more consistent than any small first-place difference. The evaluated depth specialists do not report training on either OpenRooms-FF or InteriorVerse (Table 12). Their scalar advantage on both sources is therefore not explained by disclosed direct-source exposure, although their synthetic indoor geometry supervision remains a domain-level advantage over zero-shot editors.

Table 25: Relative-depth results on the InteriorVerse main-axis subset under the affine-invariant single-image evaluation protocol.

Method	AbsRel-AI↓	RMSE-AI↓	MAE-AI↓	δ_1-AI↑	δ_2-AI↑	Boundary F1↑
GPT-Image-2.0	0.234	0.399	0.307	0.776	0.925	0.228
GPT-Image-1.5	0.334	0.525	0.403	0.686	0.899	0.017
Wan2.7-Image	0.234	0.372	0.286	0.808	0.942	0.150
Doubao Seedream 5.0	0.222	0.469	0.359	0.727	0.924	0.106
FLUX.1-Kontext-dev	0.503	0.784	0.630	0.478	0.773	0.004
FireRed-Image-Edit	0.446	0.763	0.614	0.487	0.782	0.056
Gemini 3 Pro Image	0.201	0.410	0.321	0.764	0.927	**0.263**
LongCat-Image-Edit	0.314	0.477	0.363	0.731	0.909	0.152
Qwen-Image-2.0	0.203	0.365	0.264	0.832	0.941	0.012
Depth Anything 3	**0.058**	**0.154**	**0.092**	**0.966**	**0.989**	0.176
Lotus	0.065	0.165	0.104	0.964	0.988	0.252

Scores are per-image averages over the 2,109-image InteriorVerse main-axis subset. Lower AbsRel-AI, RMSE-AI, and MAE-AI are better; higher δ_1-AI, δ_2-AI, and Boundary F1 are better. Boldface marks the best value using unrounded scores; displayed rounding may produce apparent ties.

F.2.2 Normal Source-Specific Results

Table 26: Surface normal estimation results on OpenRooms-FF.

Method	Aggregate	Mean↓	Median↓	RMSE↓	Acc@11.25↑	Acc@22.5↑	Acc@30↑
GPT-Image-1.5	Avg	40.383	37.095	45.490	0.060	0.231	0.376
Doubao Seedream 5.0	Avg	47.124	40.455	54.188	0.049	0.179	0.301
FireRed-Image-Edit	Avg	57.590	62.216	67.124	0.135	0.251	0.296
Lotus	–	**8.697**	**5.187**	**14.223**	**0.815**	0.923	0.948
StableNormal	–	9.583	6.808	14.393	0.784	**0.926**	**0.953**

Scores are per-image averages within the source. The source contains 1,125 retained main-axis images, and only complete, scoreable official runs are included. For image editors, Avg denotes the official arithmetic mean over the three fixed RGB–normal exemplar-conditioned runs; the individual E1–E3 results are reported once in Table 23. Specialized predictors do not use exemplars. Acc@22.5 is the primary metric. Boldface marks the best value using unrounded scores; displayed rounding may produce apparent ties.

Table 27: Surface normal estimation results on InteriorVerse.

Method	Aggregate	Mean↓	Median↓	RMSE↓	Acc@11.25↑	Acc@22.5↑	Acc@30↑
GPT-Image-1.5	Avg	49.199	47.919	54.897	0.052	0.178	0.275
Doubao Seedream 5.0	Avg	65.465	59.965	72.986	0.021	0.086	0.154
FireRed-Image-Edit	Avg	65.152	66.617	72.740	0.058	0.135	0.185
Lotus	–	10.483	4.822	19.137	0.805	0.886	0.911
StableNormal	–	**8.648**	**4.103**	**16.667**	**0.851**	**0.911**	**0.932**

Scores are per-image averages within the source. The source contains 2,109 retained main-axis images, and only complete, scoreable official runs are included. For image editors, Avg denotes the official arithmetic mean over the three fixed RGB–normal exemplar-conditioned runs; the individual E1–E3 results are reported once in Table 23. Specialized predictors do not use exemplars. Acc@22.5 is the primary metric. Boldface marks the best value using unrounded scores; displayed rounding may produce apparent ties.

Normal source-specific interpretation. Under the same three-exemplar aggregate used in the main table, Lotus and StableNormal retain a large advantage in angular errors and threshold accuracies on both sources. The larger editor gap on InteriorVerse is reported descriptively rather than as a causal dataset claim. The editor ordering also changes: FireRed exceeds GPT-Image-1.5 in Acc@22.5 on OpenRooms-FF (0.251 versus 0.231), but falls below it on InteriorVerse (0.135 versus 0.178). Thus a source-balanced mean can conceal domain-dependent strengths. Since scene layout, orientation distribution, and appearance differ together, this comparison cannot attribute the reversal to any one factor or establish training-set overlap. Neither evaluated normal specialist reports training on OpenRooms-FF or InteriorVerse (Table 12). The specialist gap on both sources is therefore consistent with the benefit of task-specific geometry supervision rather than a disclosed home-source advantage.

F.2.3 ALBEDO SOURCE-SPECIFIC RESULTS

Table 28: Source-specific main-axis albedo results under the official RGB-only protocol.

(a) OpenRooms-FF

Method	MAE↓	PSNR↑	SSIM↑	LPIPS↓
GPT-Image-1.5	0.226	11.955	0.453	0.407
GPT-Image-2.0	0.154	14.421	0.488	0.406
Gemini 3 Pro Image	0.150	15.443	0.685	0.269
Wan2.7-Image	0.178	13.580	0.562	0.343
Qwen-Image-2.0	0.186	12.937	0.461	0.417
Doubao Seedream 5.0	0.178	13.995	0.576	0.319
FLUX.1-Kontext-dev	0.257	10.460	0.409	0.607
FireRed-Image-Edit	0.219	12.034	0.560	0.494
LongCat-Image-Edit	0.265	10.459	0.509	0.589
Colorful Diffuse Intrinsic	**0.126**	**16.690**	**0.695**	0.369
DNF-Intrinsic	0.132	16.389	0.654	**0.225**

(b) InteriorVerse

Method	MAE↓	PSNR↑	SSIM↑	LPIPS↓
GPT-Image-1.5	0.189	12.390	0.612	0.435
GPT-Image-2.0	0.188	12.767	0.622	0.429
Gemini 3 Pro Image	0.190	13.344	0.731	0.331
Wan2.7-Image	0.269	10.447	0.642	0.437
Qwen-Image-2.0	0.209	11.980	0.589	0.434
Doubao Seedream 5.0	0.182	13.475	0.686	0.351
FLUX.1-Kontext-dev	0.238	10.885	0.576	0.559
FireRed-Image-Edit	0.200	12.829	0.710	0.380
LongCat-Image-Edit	0.214	11.941	0.670	0.442
Colorful Diffuse Intrinsic	0.149	15.584	0.794	0.305
DNF-Intrinsic	**0.056**	**22.044**	**0.863**	**0.126**

Scores are equal-image macro-averages over the 1,125-image OpenRooms-FF and 2,109-image InteriorVerse main-axis subsets using the mandatory albedo mask. Only methods with complete, scoreable official outputs for the corresponding source are included. Lower MAE/LPIPS and higher PSNR/SSIM are better. Boldface marks the best value within each panel and metric using unrounded scores; displayed rounding may produce apparent ties.

Albedo source-specific interpretation. Both sources preserve the specialized-method advantage under the RGB-only protocol, but their internal ordering differs: Colorful Diffuse Intrinsic leads most OpenRooms-FF metrics, whereas DNF-Intrinsic leads all reported InteriorVerse metrics. Among editors, Gemini gives the strongest point estimates on all four OpenRooms-FF metrics and on SSIM and LPIPS in InteriorVerse, while Doubao Seedream 5.0 gives the lowest editor MAE and highest editor PSNR on InteriorVerse. On InteriorVerse, Doubao's lower MAE than Gemini (0.182 versus 0.190) coexists with lower SSIM (0.686 versus 0.731) and almost identical LPIPS (0.335 versus 0.331). This again separates average reflectance-value fidelity from structural agreement. The change in specialist and editor orderings shows that source-balanced reporting is useful for coverage, but does not make the underlying recovery behavior domain-independent. The exposure audit provides a concrete source-conditioned diagnostic. From OpenRooms-FF to InteriorVerse, the median MAE change across the nine editors is $+0.004$, whereas DNF-Intrinsic improves by 0.076 (from 0.132 to 0.056). Subtracting the median editor change gives a descriptive difference-in-differences of -0.080 MAE. Colorful Diffuse Intrinsic, which includes InteriorVerse within a broader training mixture, instead changes by $+0.023$. This pattern is consistent with DNF's exclusive InteriorVerse training conferring source or annotation-convention matching, but it is not a causal estimate: editor pretraining exposure is undisclosed, source difficulty changes jointly, and we have not completed a scene-level overlap audit.

F.3 DEPTH BOUNDARY DIAGNOSTICS

The main-axis depth comparison separates global relative-depth error from local discontinuity recovery. In a per-image diagnostic of Gemini 3 Pro Image, Boundary F1 correlates only weakly with AbsRel-AI (Pearson $r = -0.187$) and more strongly with depth-rank correlation ($r = 0.579$) across the 3,234 evaluated images. Within OpenRooms-FF, its Boundary F1 is 0.379 on 430 base-rendered images, 0.385 on 343 lighting-varied images, and 0.377 on 352 material-varied images. The similar scores across these subsets are consistent with reliance on persistent scene contours, although the metric alone cannot establish semantic object recognition.

Boundary F1 requires a directed depth transition at the same adjacent-pixel pair in prediction and ground truth, with no spatial tolerance (Appendix D.5.1). Small contour shifts or smoothed transitions can therefore score poorly even when coarse depth ordering is plausible. We interpret the editor's boundary advantage as evidence of local discontinuity recovery under this metric, not as a claim of uniformly accurate geometry.

F.4 MAIN-AXIS UNCERTAINTY ANALYSIS

This section reports scene-cluster bootstrap uncertainty analyses for the primary main-axis metrics and the metallic metric profile where complete scene-level grouping is available. All intervals use

1,000 resamples with random seed 42 and report percentile 95% intervals from the 2.5th and 97.5th percentiles using NumPy's linear quantile interpolation. For multi-source targets, OpenRooms-FF and InteriorVerse are resampled independently at the scene level and then combined using the same equal-source aggregation as the main results; within each source, all methods reuse the same frozen scene-resampling indices. Single-source targets resample scenes within their retained source subset. These intervals quantify scene-level variation in the reported point estimates and are not used to redefine metric selection or method ordering.

F.4.1 Primary-Metric Uncertainty

The reported depth, normal, and albedo intervals separate the leading specialist from the leading scored editor more clearly than they resolve several editor–editor comparisons. In particular, Gemini and GPT-Image-2.0 have overlapping albedo MAE intervals. For roughness, Gemini's interval lies below that of the next-best editor, GPT-Image-2.0, supporting its point-estimate advantage under scene resampling. Overlap alone is not a significance test; model ordering requires paired differences under the shared resampling indices. These intervals condition on the retained outputs and do not separately estimate generation variability or uncertainty over alternative prompts and exemplars.

Table 29: Scene-cluster bootstrap confidence intervals for the primary depth, normal, albedo, and roughness metrics on the main-axis subsets.

(a) Depth: AbsRel-AI ↓

Method	Estimate	95% CI
GPT-Image-1.5	0.2339	[0.1844, 0.3211]
GPT-Image-2.0	0.1682	[0.1439, 0.2083]
Gemini 3 Pro Image	0.1529	[0.1440, 0.1627]
Wan2.7-Image	0.1657	[0.1294, 0.2312]
Doubao Seedream 5.0	0.1713	[0.1577, 0.1877]
Qwen-Image-2.0	0.1578	[0.1308, 0.2008]
FLUX.1-Kontext-dev	0.3890	[0.3322, 0.4812]
FireRed-Image-Edit	0.3572	[0.3170, 0.4208]
LongCat-Image-Edit	0.2280	[0.1797, 0.3192]
Depth Anything 3	**0.0350**	[0.0310, 0.0400]
Lotus	0.0479	[0.0432, 0.0548]

(b) Normal: Acc@22.5 ↑

Method	Estimate	95% CI
GPT-Image-1.5	0.2042	[0.1985, 0.2099]
Doubao Seedream 5.0	0.1325	[0.1267, 0.1383]
FireRed-Image-Edit	0.1931	[0.1865, 0.1989]
Lotus	0.9042	[0.8981, 0.9101]
StableNormal	**0.9183**	[0.9128, 0.9236]

(c) Albedo: MAE ↓

Method	Estimate	95% CI
GPT-Image-1.5	0.2076	[0.1999, 0.2156]
GPT-Image-2.0	0.1714	[0.1677, 0.1756]
Gemini 3 Pro Image	0.1696	[0.1657, 0.1740]
Wan2.7-Image	0.2233	[0.2177, 0.2296]
Doubao Seedream 5.0	0.1799	[0.1747, 0.1852]
Qwen-Image-2.0	0.1972	[0.1925, 0.2018]
FLUX.1-Kontext-dev	0.2473	[0.2404, 0.2536]
FireRed-Image-Edit	0.2094	[0.2039, 0.2147]
LongCat-Image-Edit	0.2395	[0.2329, 0.2466]
Colorful Diffuse Intrinsic	0.1379	[0.1342, 0.1416]
DNF-Intrinsic	**0.0942**	[0.0905, 0.0984]

(d) Roughness: RMSE ↓

Method	Estimate	95% CI
GPT-Image-1.5	0.3750	[0.3612, 0.3901]
GPT-Image-2.0	0.3234	[0.3114, 0.3349]
Gemini 3 Pro Image	**0.2911**	[0.2820, 0.3010]
Doubao Seedream 5.0	0.3644	[0.3548, 0.3740]
Qwen-Image-2.0	0.3839	[0.3687, 0.3993]
FLUX.1-Kontext-dev	0.3369	[0.3176, 0.3565]
FireRed-Image-Edit	0.3667	[0.3543, 0.3771]
LongCat-Image-Edit	0.3912	[0.3770, 0.4050]
Intrinsic Image Diffusion	0.3462	[0.3313, 0.3610]
DNF-Intrinsic	0.3756	[0.3617, 0.3884]

All intervals use 1,000 scene-cluster bootstrap resamples. Depth, normal, and albedo resample scenes separately within OpenRooms-FF and InteriorVerse and then macro-average source-wise scores; their support is 1,125 images from 48 OpenRooms-FF scenes and 2,109 images from 167 InteriorVerse scenes. Normal recomputes the official three-exemplar aggregate within each replicate. Roughness resamples the 48 OpenRooms-FF scenes (1,125 images; seed 42), retains all images from each sampled scene, and recomputes the arithmetic mean of per-image valid-pixel RMSE values within each bootstrap replicate. Boldface marks the best point estimate in each panel using unrounded scores; displayed rounding may produce apparent ties.

F.4.2 Metallic Multi-Metric Uncertainty

The intervals qualify the metric-wise leaders in the main text. Doubao's IoU interval [0.0350, 0.0821] overlaps Gemini's [0.0266, 0.0908] and GPT-Image-2.0's [0.0283, 0.0796], so the separate intervals do not establish a stable localization winner. MVInverse's lower region-balanced point estimate and GPT-Image-2.0's lower full-map point estimate describe different error profiles, not a proven total ordering. Paired model differences and repeated generations would be needed for stronger comparative claims.

Table 30: Scene-cluster bootstrap confidence intervals for the metallic metric profile.

(a) Full-map and region-balanced error

Method	MAE↓	RB-MAE↓
GPT-Image-1.5	0.0257 [0.0176, 0.0355]	0.4406 [0.4208, 0.4580]
GPT-Image-2.0	0.0207 [0.0126, 0.0309]	0.4192 [0.3898, 0.4466]
Gemini 3 Pro Image	0.0404 [0.0308, 0.0522]	0.3753 [0.3405, 0.4046]
Wan2.7-Image	0.0959 [0.0882, 0.1042]	0.3895 [0.3652, 0.4145]
Doubao Seedream 5.0	0.0438 [0.0366, 0.0524]	0.3513 [0.3168, 0.3831]
Qwen-Image-2.0	0.4958 [0.4839, 0.5081]	0.5029 [0.4783, 0.5288]
FLUX.1-Kontext-dev	0.4213 [0.4043, 0.4381]	0.4859 [0.4678, 0.5041]
FireRed-Image-Edit	0.2775 [0.2670, 0.2898]	0.5073 [0.4890, 0.5247]
LongCat-Image-Edit	0.5929 [0.5680, 0.6159]	0.4815 [0.4647, 0.4973]
MVInverse	0.1968 [0.1821, 0.2130]	**0.3295** [0.3065, 0.3527]
DNF-Intrinsic	0.1205 [0.1049, 0.1396]	0.4586 [0.4351, 0.4826]

(b) Metallic-region localization

Method	IoU↑	AP↑
GPT-Image-1.5	0.0306 [0.0176, 0.0465]	0.0223 [0.0119, 0.0349]
GPT-Image-2.0	0.0524 [0.0283, 0.0796]	0.0420 [0.0199, 0.0692]
Gemini 3 Pro Image	0.0551 [0.0266, 0.0908]	0.0511 [0.0213, 0.0902]
Wan2.7-Image	0.0112 [0.0064, 0.0157]	0.0139 [0.0069, 0.0213]
Doubao Seedream 5.0	0.0569 [0.0350, 0.0821]	0.0679 [0.0375, 0.1014]
Qwen-Image-2.0	0.0042 [0.0013, 0.0084]	0.0056 [0.0017, 0.0112]
FLUX.1-Kontext-dev	0.0039 [0.0013, 0.0083]	0.0057 [0.0017, 0.0125]
FireRed-Image-Edit	0.0036 [0.0017, 0.0060]	0.0045 [0.0019, 0.0082]
LongCat-Image-Edit	0.0055 [0.0017, 0.0110]	0.0089 [0.0022, 0.0184]
MVInverse	0.0273 [0.0107, 0.0484]	0.0307 [0.0108, 0.0584]
DNF-Intrinsic	0.0077 [0.0039, 0.0119]	0.0069 [0.0037, 0.0107]

Each cell reports the point estimate above its percentile 95% confidence interval. Intervals use 1,000 scene-cluster bootstrap resamples (seed 42) over 40 PBR-Rooms scenes and retain all 24 images from each sampled scene. Metrics are computed per scene and macro-averaged equally over scenes. MAE uses all images; RB-MAE, IoU, and AP require metallic-positive support. IoU uses a 0.5 threshold, whereas AP ranks continuous prediction scores. Boldface is retained only for the RB-MAE point-estimate leader. The MAE leader is not emphasized because full-map MAE is dominated by non-metallic pixels, and localization leaders are not emphasized because their intervals overlap.

G Qualitative Results and Normal-Representation Failures

This appendix complements the quantitative tables with illustrative target-wise predictions and normal-representation failures. The figures are qualitative evidence and do not enter metric aggregation or claim statistical representativeness. Source-annotation audit examples are reported with the dataset eligibility discussion in Appendix B.2.1, rather than mixed with model failures here.

G.1 Target-Wise Prediction Examples

Figure 6 gives a compact five-target overview. Figures 7–9 provide additional target-specific comparisons; roughness and metallic examples follow separately because each uses one official source.

G.2 Roughness Prediction Examples

Roughness is evaluated on its single official source, OpenRooms-FF. Table 5 reports the quantitative results, and Figure 10 provides an illustrative comparison.

G.3 Metallic Prediction Examples

Metallic is evaluated on its single official source, PBR-Rooms. Because positive metallic pixels are sparse and their prevalence varies across room categories, retained scenes are sampled within each category. Table 6 reports the quantitative profile, and Figure 11 provides an illustrative comparison.

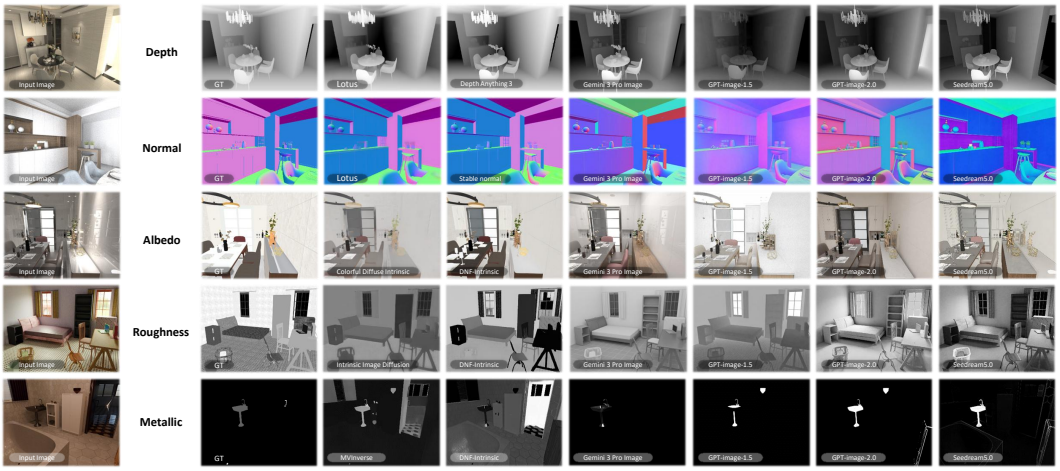

Figure 6: Qualitative examples of physical-map recovery across five targets. Each row shows an input RGB image, ground truth, and selected specialist and editor predictions.



Figure 7: Illustrative retained main-axis examples for depth under the official relative-depth protocol. Each row shows the query RGB image, ground-truth depth map, and predictions from the task-specific predictors and general-purpose image editors identified by the column labels.

G.4 Qualitative Normal-Representation Failure Cases

In addition to the officially scored systems reported in the main normal table, we show illustrative Wan2.7-Image outputs from the run recorded as a normal-representation failure in Appendix H.3. Other run-level normal-representation failures are listed in the same status table and are not repeated as additional image grids. These examples are not additional quantitative results and are not used as a post-hoc rule for excluding individual images. Quantitative inclusion follows the frozen output contract and run-level failure policy in Appendices D.2 and D.2.1.

The outputs in Figure 12 illustrate the recorded failure pattern: they preserve scene appearance, shading, or stylized colorization rather than instantiating the requested spatially aligned RGB directional encoding. This is a target-representation failure rather than merely a large angular error. Although any RGB image can be mapped numerically through $2\mathbf{rgb} - 1$, doing so here would reinterpret appearance colors as surface directions even though the output is not a normal-map estimate. The figure documents what the systematic run-level status looks like qualitatively; it does not define a visual quality threshold for otherwise scoreable normal predictions or select isolated low-quality images for exclusion.

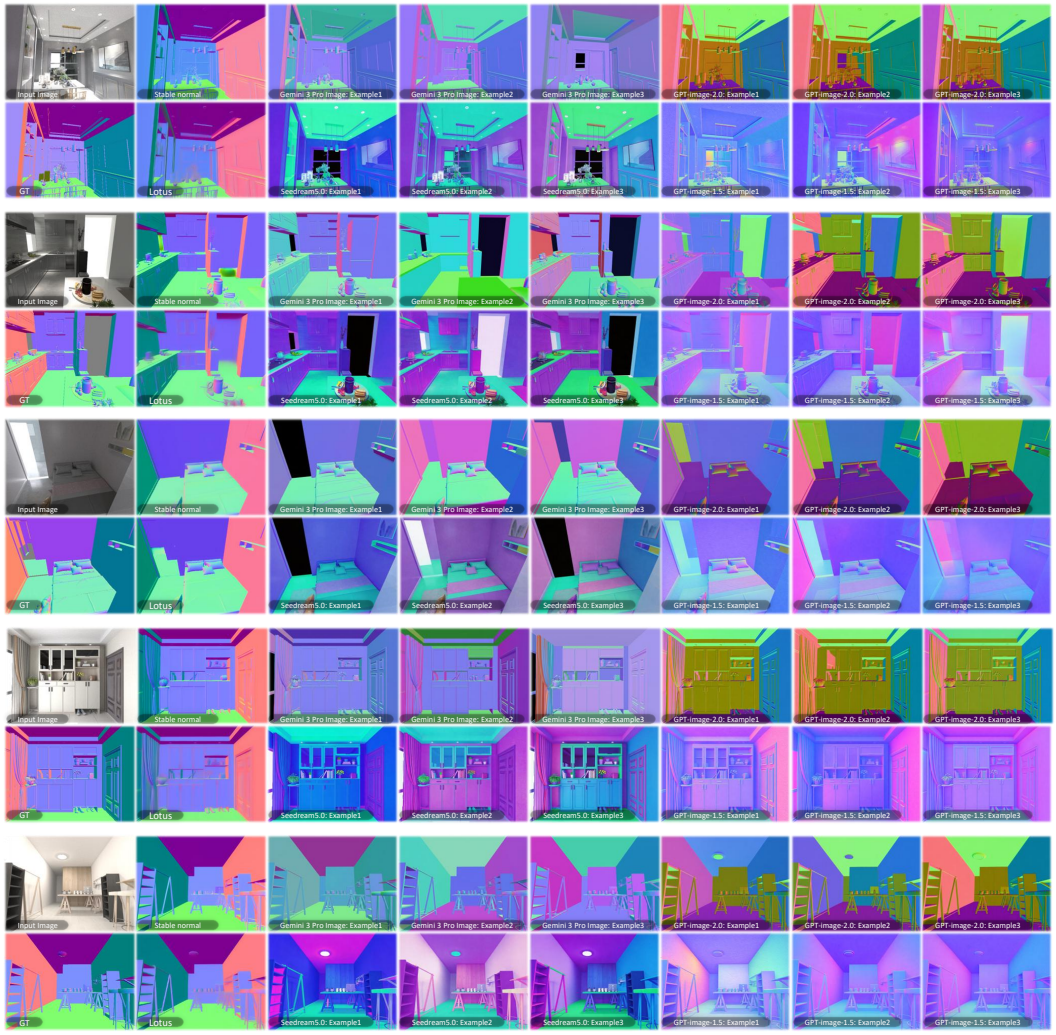

Figure 8: Illustrative retained main-axis examples for normal under the official convention-stabilized protocol. Each group shows the query RGB image, ground-truth normal map, and predictions from the specialized predictors and image editors identified in the figure.



Figure 9: Illustrative retained main-axis examples for albedo under the official RGB-only protocol. Each row shows the query RGB image, ground-truth albedo map, and predictions from the specialized predictors and image editors identified in the figure.

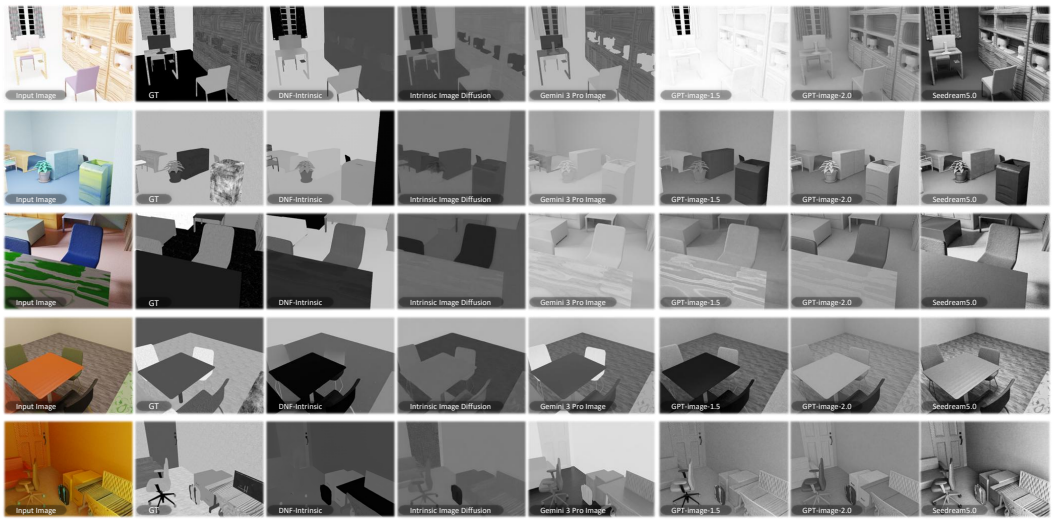

Figure 10: Illustrative retained OpenRooms-FF main-axis examples for roughness under the official direct RGB-to-map protocol. Each row shows the query RGB image, ground-truth roughness map, and predictions from the specialized predictors and image editors identified in the figure.

These examples illustrate a representation-level failure mode of image-editor access for normal prediction: the model can produce colored map-like or stylized outputs while failing to adopt the dense physical-map representation required for angular evaluation.

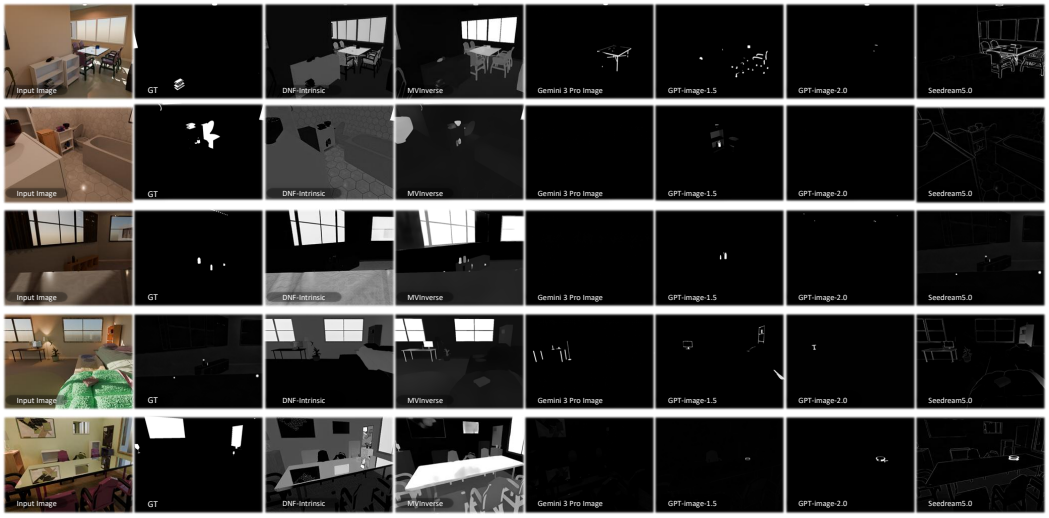

Figure 11: Illustrative retained PBR-Rooms main-axis examples for metallic under the official direct RGB-to-map protocol. Each group shows the query RGB image, ground-truth metallic map, and predictions from the specialized predictors and image editors identified in the figure.

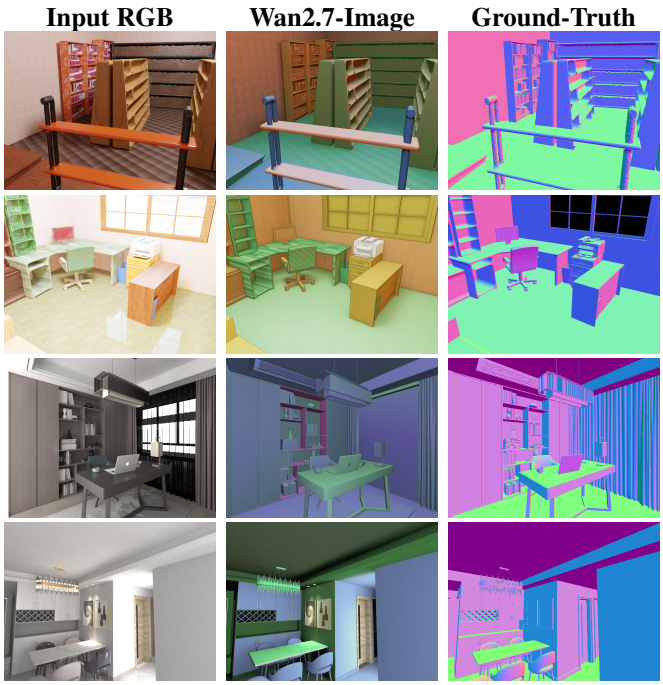

Figure 12: Illustrative Wan2.7-Image normal-representation failures under the evaluated normal access setting. Each row shows the input RGB image, the Wan2.7-Image output, and the corresponding ground-truth normal map. The corresponding run is recorded as a representation failure rather than an official Mean Angular or Acc@22.5 entry. These panels illustrate that run-level status and are not used to select individual outputs for exclusion.

H Reproducibility and Experiment Status

This appendix records manifests, system versions, run status, and release artifacts. Complete official runs, incomplete probes, and representation failures are distinguished so that unavailable results are not treated as evaluated entries. Scene and illumination metadata support the sampling and stress definitions in Appendix B; they are not model inputs or scoring targets.

H.1 Benchmark Metadata and Scene-Level Auditing

Scene-level metadata are used only for sampling and split auditing, not as model inputs or scoring targets. The retained records include source-native scene category, illumination-related sampling fields, and scene-quality audit fields such as layout plausibility, functional coherence, object-arrangement coherence, and overall plausibility. Initial annotations and reviewer arbitration follow the fixed schema used during benchmark construction; the released records preserve the resulting metadata and audit decisions.

H.2 Evaluation Records

Official results are matched by stable sample identifiers and released manifests listing every retained scene and image. Run records preserve the model/backend, access setting, prompt, input order, seed, relevant exemplar identifiers, and completion or failure status; all invalid or incomplete outcomes are logged. Official and diagnostic records remain separate, and proprietary outputs are omitted when provider policy prohibits redistribution.

H.3 Model Versioning and Experiment Status

Table 31: Model identifiers, backends, and target-wise main-axis run status.

System	Model identifier / backend	D	N	A	R	M
Doubao Seedream 5.0	`doubao-seedream-5-0-260128` Volcengine Ark	C	C	C	C	C
FireRed-Image-Edit	`FireRed-Image-Edit-1.0` Local Diffusers	C	C	C	C	C
GPT-Image-1.5	`gpt-image-1.5` Azure OpenAI	C	C	C	C	C
GPT-Image-2.0	`gpt-image-2.0` Azure OpenAI	C	Inc.	C	C	C
Gemini 3 Pro Image	`gemini-3-pro-image-preview` AIHubMix	C	Inc.	C	C	C
FLUX.1-Kontext-dev	`FLUX.1-Kontext-dev` Local Diffusers	C	Fail	C	C	C
LongCat-Image-Edit	`LongCat-Image-Edit` Local Diffusers	C	Fail	C	C	C
Qwen-Image-2.0	`qwen-image-2.0` DashScope	C	Fail	C	C	C
Wan2.7-Image	`wan2.7-image` DashScope	C	Fail	C	–	C
Doubao Seed 2.0 Pro	`doubao-seed-2-0-pro-260215` Volcengine Ark	–	–	Aux	–	–

Note. D/N/A/R/M denote depth, normal, albedo, roughness, and metallic. C: complete official run; Inc.: incomplete and excluded; Fail: representation failure without assigned score; Aux: diagnostic-only use; –: not included in official comparison.

Gemini 3 Pro Image was accessed through the third-party AIHubMix endpoint using the preview identifier shown in Table 31. The intermediary may update routing or the provider-side preview behind a stable alias; the reported results therefore apply to the request records and timestamps preserved in the frozen run manifests, not to every deployment carrying the same public-facing model name.

Table 32: Exceptions to complete target-wise main-axis coverage.

System	Target	Status	Explanation
GPT-Image-2.0	Normal	INC.	The recorded official run is incomplete under the three-exemplar protocol; per-pair coverage and the cause are not documented here, so no official aggregate is reported.
Gemini 3 Pro Image	Normal	INC.	The recorded official run is incomplete under the three-exemplar protocol; per-pair coverage and the cause are not documented here, so no official aggregate is reported.
FLUX.1-Kontext-dev	Normal	FAIL	Outputs do not instantiate the required dense RGB-encoded normal-map representation.
LongCat-Image-Edit	Normal	FAIL	Outputs do not instantiate the required dense RGB-encoded normal-map representation.
Qwen-Image-2.0	Normal	FAIL	Outputs do not instantiate the required dense RGB-encoded normal-map representation.
Wan2.7-Image	Normal / Roughness	FAIL / –	Normal outputs do not instantiate the required dense RGB-encoded normal-map representation. No Wan roughness generation was recorded before the experiment manifest was frozen, so roughness is unmeasured rather than a model failure.
Doubao Seed 2.0 Pro	Albedo A1	AUX	Used only to generate shared query-image-only illumination-analysis records for the diagnostic variant, not as an evaluated image editor.

H.4 RELEASE PACKAGE AND EVALUATOR REPRODUCIBILITY

The accompanying release provides source manifests, evaluation and aggregation utilities, prompt templates, access-setting documentation, and scene-audit utilities. Full proprietary-model outputs are not redistributed when provider policies prohibit it; their run records retain the protocol and status information described above.

For a fixed target, every method receives the same resizing, masking, validity checks, deterministic post-processing, and metric computation. The sole representation conversion is the shared, frozen Lotus/StableNormal normal-convention adapter in Appendix D.5.2; the release retains its calibration manifest. Normal prediction-valid and ground-truth-valid pixel counts remain available for the coverage audit in Table 13.