Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sampling variability ranks reconstruction error in lacquerware image inpainting across models and collections

Peng Yue[1]

yp6253344@163.com

Kangwon National University

Jia Hou[2]

Zhengzhou Business University

Xiao Zhang[3]

Christian-Albrechts-Universität zu Kiel

Article

Keywords: lacquerware, generative AI, image inpainting, sampling uncertainty, selective prediction, risk–coverage, heritage imaging

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11459219/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Generative inpainting can produce visually persuasive but inaccurate completions of lacquerware photographs. We examined whether variability across three seeded reconstructions predicts the masked error of a fourth, held-out reconstruction under synthetic occlusion. The analysis included 64 original objects, 40 external museum catalogue entries, and two diffusion models. Object-level Pearson correlations between variability and error were 0.7176 and 0.7664 in the original cohort, and 0.9043 and 0.9162 externally. Retaining approximately 40% of the lowest-variability entries reduced the mean masked error of selected outputs in each analysis, although no reconstruction was improved. The incremental advantage over inexpensive image descriptors was uncertain in the original cohort. Subsequent two-expert foreground and material assessment supported the association in restricted external subsets, while anonymous visual ratings showed limited agreement despite adjacent-grade differences. Sampling variability may help prioritize synthetic-loss reconstructions for human review; it does not establish historical authenticity or conservation suitability.

Introduction

Museum photographs document the visible appearance of lacquerware, including surface decoration and condition, but cannot establish its underlying material composition. Analytical studies of historic lacquer demonstrate layered substrates, pigments and decorative materials, while non-contact thermographic inspection reveals a separate domain of subsurface defects [1,2]. Computational approaches have also been applied to lacquerware design and augmented-reality reconstruction [3,4]. A recent diffusion-based study examined lacquerware image completion and texture quality [5]. Although these approaches expand the possibilities for visual reconstruction, convincing synthetic detail should not be mistaken for evidence of a lost historical appearance.

Controlled occlusion of an intact photograph permits a direct comparison between the generated completion and the concealed pixels. Such a reference is rarely available when historical material has actually been lost. Digital-heritage scholarship emphasizes evidence-aware workflows, transparency and the limits of reconstruction claims [6,7,8]. Reflective coatings, painted motifs and densely decorated surfaces pose different visual challenges, making it important to distinguish a photographic proxy task from restoration of a verified historical state. Material analyses and three-dimensional defect inspection likewise address information that a photographic reconstruction cannot recover [1,2].

Research on heritage imagery has examined mural enhancement, damage detection and diffusion-based completion [9,10,11,12]. Recent methods include heterogeneous diffusion networks, prompt-guided restoration, image-guided synthesis, damage segmentation and structural guidance [13,14,15,16,17]. Comparable approaches have been studied for artifacts, paintings and three-dimensional cultural objects [18,19,20,21]. Work on bronzeware has additionally considered the risk of over-restoration in spatially controlled reconstruction [22]. Taken together, these studies make clear that generating a visually restored heritage image is not in itself a methodological contribution.

Image-inpainting reviews cover convolutional, adversarial, transformer and diffusion architectures and show how mask geometry and evaluation criteria influence reported performance [23,24,25,26]. Approaches to large missing regions include Fourier convolutions and mask-aware transformers, alongside models designed to recover global structure or produce multiple plausible completions [27,28,29,30]. GAN inversion, high-frequency refinement and text-guided editing further illustrate the distinction between perceptual quality and fidelity to the missing content [31,32,33]. Diffusion-based inpainting also introduces choices concerning consistency between observed and generated regions [34,35,36]. In none of these settings can visual plausibility alone establish agreement with an unseen reference.

We therefore focus on a measurable question: whether disagreement among generated candidates can rank the reconstruction error of another candidate not used to estimate that disagreement. Prior uncertainty-aware restoration and semantic candidate filtering show that uncertainty can inform inpainting workflows, but do not validate the present seed-disjoint error-ranking score [37,38]. Measures of uncertainty during diffusion sampling and image synthesis also vary with the task and the quantity being estimated [39,40]. Deep ensembles demonstrate the practical appeal of sample-based disagreement, while formal analyses distinguish epistemic and data-related components that simple output variance cannot automatically separate [41,42]. Low variability does not rule out a shared systematic error. We therefore estimate variability from three seeds and evaluate error on a fourth seed excluded from that calculation.

Error ranking should also be distinguished from probability calibration. Selective classification and uncertainty ranking use coverage–risk trade-offs to prioritize lower-risk predictions [43,44]. In contrast, converting a score to a credible probability of exceeding an error threshold requires separate calibration evidence and adequate outcome prevalence [45,46]. Formal conformal risk control provides stronger guarantees only under its prescribed assumptions and calibration procedure [47]. We tested whether variability across separate seeds was associated with held-out masked error and whether it could identify lower-error objects at fixed coverage. We then examined sensitivity to image complexity, model, museum source and expert-reviewed foreground regions. The correlation and fixed-coverage analyses in the original 64-object cohort followed the internal protocol; subsequent class adjustments, baseline comparisons and material/ROI reviews were exploratory.

Results

Seed-disjoint uncertainty was associated with reconstruction error

The original cohort contained 64 lacquerware photographs (32 visually plain; 32 decorated), each evaluated under small, medium and large synthetic occlusions of a locked object ROI. Three mask conditions and four generation seeds yielded 192 cases and 768 outputs. Mask-level observations were averaged within objects before the primary statistical analysis. Generation settings, images and masked-error measurements were checked against the archived records (Supplementary Tables S1, S2 and S6).

At the object level, higher seed-disjoint uncertainty was associated with higher masked-region MAE (Pearson r = 0.7176, two-sided p = 2.52×10^{-11}; 10,000-replicate object-bootstrap 95% CI 0.5580–0.9381; Fig. 1). Spearman rank correlation was 0.9349. The result met the internally prespecified criterion (r > 0 and bootstrap lower bound > 0.30), but the observed association does not calibrate the probability of reconstruction error. The primary estimates and their verification status are detailed in Supplementary Table S2.

Leaving out each original object in turn yielded Pearson r values of 0.6950–0.7895 and Spearman ρ of 0.9319–0.9400; an independent 10,000-resample object bootstrap gave a Pearson 95% interval of 0.5590–0.9366 (Supplementary Fig. S1; Table S7). These checks assess sensitivity within the same cohort and do not constitute external validation.

Robustness across seed rotations and visual classes

All four leave-one-seed-out target rotations yielded positive object-level Pearson correlations (r = 0.6852–0.7760; Fig. 2). Within the two visual classes, the association was positive in plain (r = 0.7380) and decorated (r = 0.6715) objects. The association persisted across target-seed choices, although the rotations share generated outputs and cannot be regarded as independent replications. Rotation-specific numerical results are provided in Supplementary Table S3.

The prespecified seed-11001 interval is reported as primary; intervals from the other rotations are shown separately (Fig. 2; Supplementary Table S3).

Uncertainty-based acceptance reduced accepted-set mean error

Objects were ranked in ascending order of sampling uncertainty and the lowest-uncertainty 26 of 64 were accepted (nominal 40% coverage, realized 40.625%). Their mean masked MAE was 0.01811 versus 0.04655 over all 64 objects (relative reduction 61.10%, object-bootstrap 95% CI 49.14–70.85%; Fig. 3). Across increasing coverage, accepted-set mean MAE rose from 0.01190 at 20% to 0.03611 at 80% before converging to the full-cohort value at 100%. These differences result from the selective inclusion of objects; the process did not alter or improve any individual reconstruction. The complete numerical risk–coverage profile is reported in Supplementary Table S4.

The retrospective oracle reference in Fig. 3 orders objects by their observed masked MAE rather than estimated uncertainty. It is a hindsight lower bound on accepted-set error for these 64 objects, rather than a predictor available at the time of reconstruction. At 26/64 objects, the oracle mean was 0.01761 compared with 0.01811 for uncertainty ranking; the other plotted acceptance levels were independently recomputed from the image-derived object metrics (Supplementary Table S8). This comparison evaluates ordering within the analyzed cohort only.

Post-hoc controls for visual-class composition

The globally accepted 26-object set contained 22 plain objects and four decorated objects, even though the total dataset was balanced (32 per class). Mean MAE for all plain objects was 0.02638 versus 0.06672 for decorated objects. In a post hoc Monte Carlo comparison using 200,000 draws per relevant sampling strategy, random selection of 26 objects yielded expected mean MAE 0.04655; random selection maintaining 22 plain and four decorated objects yielded expected mean MAE 0.03259. Both were higher than the 0.01811 mean observed under uncertainty ranking.

An exploratory class-balanced strategy accepting the 13 lowest-uncertainty plain and 13 lowest-uncertainty decorated objects yielded MAE 0.02640 (43.3% below the full-cohort mean). Class-centered uncertainty and error retained Pearson r = 0.6613. Class-stratified object-bootstrap estimates for the relative reduction were 50.2–70.5% under global ranking and 32.4–51.0% under balanced ranking. These checks indicate that visual-class selection contributes to, but does not fully explain, the observed error ranking. Because strategies were selected after examining E5 data, these numbers are descriptive robustness evidence, not independent confirmation (Fig. 4). The numerical comparison of exploratory selection strategies is provided in Supplementary Table S5.

The plain/decorated distinction is only a rough measure of image difficulty. Although the class-based checks address part of the selection imbalance, they cannot rule out an association between low variability and intrinsically easier image content.

Exploratory tests using the archived reference images

For the original cohort, we compared uncertainty ranking with Canny edge density, annular Sobel magnitude and grayscale SD. At 26/64 acceptance, their masked MAEs were 0.02026, 0.02557 and 0.03574 versus 0.01811 under uncertainty ranking (Fig. S2; Supplementary Table S9). The paired Canny-minus-uncertainty bootstrap interval included zero (95% CI − 0.00005 to 0.00651). This Canny feature is not the later matched boundary-Sobel descriptor, which is compared across all four cohort–model groups (Fig. S4; Supplementary Tables S19–S20).

In exploratory analyses of the original cohort, within-museum Pearson associations were positive (r = 0.6749, 0.9147 and 0.8481 for Walters, Cleveland and Cooper Hewitt, respectively), but these groups were not independent transfer cohorts. After adjusting uncertainty and MAE for museum, visual class and three context-complexity descriptors, the residual Pearson association was 0.3264, consistent with partly overlapping information from image difficulty and generation variability (Supplementary Table S10).

Mask-core structural similarity provided a complementary, exploratory endpoint: uncertainty correlated with corrected mask-interior SSIM error (Spearman ρ = 0.7823; Pearson r = 0.3323). Mean corrected SSIM was 0.5561 across the original cohort and 0.8134 in its 26 lowest-uncertainty objects (Fig. S3; Supplementary Table S11). This measures resemblance to a photographic reference, not historical authenticity.

The displayed source-image crops include low-variability/low-error, ranking-discordant decorated and high-variability/high-error examples (Fig. 5). The discordant example shows that sample dispersion does not identify every error, notably in complex decoration. The displayed unmasked reference is available only because the losses are synthetic.

Transfer across the second model and external museum entries

The second-model experiment generated 768 additional outputs for the original 64 objects, while the external 40-entry cohort contributed 960 outputs across the two models (1,728 new outputs in total). Analyses were conducted at object or catalogue-entry level, rather than treating masks, seeds or individual output files as independent samples. Generation provenance and file-integrity evidence are detailed in Supplementary Table S12.

In the original cohort, the second model retained a positive object-level uncertainty–MAE association (Pearson r = 0.7664; Spearman ρ = 0.9254). Full-cohort mean MAE was 0.04679, compared with 0.01624 among the 26 lowest-uncertainty objects. The original-model estimates were also reproduced from the archived images (Fig. 6; Supplementary Table S13).

The external cohort comprised 20 Metropolitan Museum of Art and 20 Art Institute of Chicago catalogue entries, with distinct source-image hashes and accession identifiers from the original cohort. One entry, AIC-86912 (1961.625–627), includes AIC-86913 (1961.626) as a component; the 40 entries therefore cannot all be treated as unrelated physical objects. The cohort spans heterogeneous periods, materials and object types (Supplementary Table S15).

Across the locked external entries, Pearson r was 0.9043 for Stable Diffusion 1.5 and 0.9162 for Stable Diffusion 2; Spearman ρ was 0.9698 and 0.9598. At 16/40 coverage, accepted mean MAE was 0.03934 versus 0.07835 overall for Stable Diffusion 1.5, and 0.03961 versus 0.07680 for Stable Diffusion 2 (Fig. 7). A new 10,000-permutation reanalysis gave Holm-adjusted p = 0.0004 for each of the four association/selection tests. Unlike a bootstrap tail fraction, this permutation analysis tests a specified null hypothesis. Its interpretation remains subject to dependence among entries and differences in mask design (Supplementary Tables S13 and S14).

At the external 16/40 operating point, boundary-Sobel ordering retained mean MAE 0.04908 for SD1.5 and 0.04981 for SD2, compared with 0.03934 and 0.03961 under uncertainty ordering. Across 25–100% coverage, uncertainty also showed lower average risk than this descriptor in both external model runs (0.05272 versus 0.05605; 0.05213 versus 0.05620). Unlike the image-context descriptors, the variability score requires three additional seeded reconstructions. The comparisons are retrospective. No runtime or energy benchmark was available to determine whether this improvement offsets the additional generation cost (Supplementary Tables S19–S20; Fig. S4).

Corrected mask-interior SSIM averaged 0.4245 for SD1.5 and 0.4167 for SD2; fixed-crop LPIPS-Alex v0.1 averaged 0.1653 and 0.1681. Earlier near-unity SSIM values were based on whole-image reference substitution rather than mask-restricted SSIM; the distinction is documented in Supplementary Table S17. Illustrative cropped results display the reconstructed region and its masked error (Fig. 8); neither image metric establishes conservation suitability.

The external association remained positive after excluding the nested component entry (n = 39; r = 0.9032 and 0.9144) or both linked entries (n = 38; r = 0.9105 and 0.9163). A subsequent two-expert review classified the 40 catalogue entries as 35 lacquer-led, four lacquer-as-component and one material-role unknown; visible ornament was assessed as 35 decorated and five plain, revising 17 original labels. These post-hoc classifications are reported separately from the original frozen 20/20 labels (Supplementary Tables S15–S16, S23, S25).

Matched risk–coverage and exploratory mask-placement checks

Matched retrospective selectors were defined consistently across the two cohorts and two models. Uncertainty ranking had lower accepted-set MAE and average risk over 25–100% coverage than boundary Sobel, visible-region Sobel or grayscale SD (Fig. S4; Supplementary Table S19). At 40% acceptance, the boundary-Sobel-minus-uncertainty difference in the original cohort was small: 0.00078 (95% bootstrap CI − 0.00013 to 0.00386) for SD1.5 and 0.00045 (−0.00061 to 0.00302) for SD2. External differences were larger (0.00974 and 0.01020), but exploratory intervals and unmatched mask designs do not establish a general superiority claim (Supplementary Table S20).

Mask-level uncertainty–error correlations remained positive for S/M/L conditions in each cohort–model combination (Fig. S5; Supplementary Table S21). In a later two-expert review, foreground masks were produced for all 40 external entries and checked against the 120 original synthetic S/M/L masks. Of these, 101 masks placed at least 80% of their pixels on the visible object foreground, 11 placed 50% to less than 80%, and eight placed less than 50%. The two ROI passes agreed categorically on mask placement, although three entries had nonidentical outlines; a dated joint adjudication retained the second-pass ROIs (Fig. S6; Supplementary Tables S26, S28). Thirty-three entries satisfied the ≥ 80% overlap criterion for all three masks; their uncertainty–error Pearson correlations remained 0.9035 (SD1.5) and 0.9147 (SD2). Among 35 lacquer-led entries, the corresponding values were 0.9107 and 0.9199. The positive associations persisted in these exploratory subsets, but neither lacquer-layer placement nor a consistent advantage over boundary Sobel was established. The original 40-entry analysis remains primary; subgroup statistics are in Supplementary Table S27, the earlier five-entry screening sensitivity in Table S22, and supporting material/ROI records in Tables S25, S26 and S28.

Single-mask error triage and selection coverage by visual class

The primary analysis averages S/M/L conditions within objects. To examine ranking for individual synthetic-occlusion tasks, we also ranked S, M and L cases separately across the same 64 original objects or 40 external entries, accepting 26/64 or 16/40 cases at each level. All 12 mask–cohort–model analyses reduced accepted-set mean masked MAE relative to the corresponding full mask-level set. For SD1.5 in the original cohort, the small-, medium- and large-

mask Pearson associations were 0.8011, 0.7648 and 0.4943, respectively; the corresponding SD2 values were 0.7808, 0.7758 and 0.6293. In the external cohort, mask-specific correlations ranged from 0.8474 to 0.9368 under SD1.5 and from 0.8650 to 0.9385 under SD2 (Supplementary Table S29; Fig. S7). The larger-mask attenuation in E10 is not evidence of comparable physical damage levels in E11, where the masks are fixed circles. These post-hoc rankings show stability of an association within single synthetic occlusion levels; they do not establish prospective case-specific risk calibration.

Using a common 40% accepted fraction within each mask type, 4,500 object/entry-bootstrap resamples compared uncertainty selection with the same case-level boundary-Sobel descriptor. Several confidence intervals for the Sobel-minus-uncertainty difference included zero, indicating that correlation with error does not itself demonstrate a consistent advantage over this inexpensive baseline (Supplementary Table S29). Masks and model results share objects, and the 12 secondary comparisons were not subject to a new confirmatory multiplicity protocol.

The original cohort additionally displayed pronounced visual-class selection disparity. In both diffusion models, global uncertainty ranking accepted 22/32 plain and 4/32 decorated objects (68.75% and 12.50%, respectively). At the same 26-object capacity, balancing selection to 13/32 per class increased accepted mean masked MAE from 0.01811 to 0.02640 for SD1.5 and from 0.01624 to 0.02647 for SD2. Within-class risk–coverage curves and exact accepted entries are provided in Fig. S8 and Supplementary Table S30. These retrospective results quantify the trade-off between retained error and representation of decorated objects; they do not identify an optimal policy for museum use.

Audit of anonymous visual ratings

We linked the original anonymous allocation key, 160 image panels and two 80-item rating sheets without changing any recorded score. In a cluster-aware reanalysis, four visual criteria showed low ordinal Krippendorff agreement ($\alpha = 0.230–0.287$) and modest quadratic-weighted Cohen agreement ($\kappa = 0.402–0.536$), with exact agreement of 56.25–57.50%. All observed disagreements on these four criteria were between adjacent 0–3 categories; none differed by two or three grades. Rater 01 consistently assigned higher scores than rater 02 (mean paired difference 0.413–0.425 points); object-pair bootstrap intervals for these differences excluded zero. Both raters gave identical scores on the separate item assessing suitability for further human review, but 63/80 responses from each rater were in category 3, creating a highly concentrated distribution. Each rater assigned the same rating across the four visual items for 77/80 images, suggesting that these judgments may not represent independent measured dimensions. Stage 1 and reference-revealed Stage 2 rated different constructs, so no within-item before/after change is estimable. Paired model comparisons remained nonsignificant (Stage 1 Wilcoxon $p = 0.5638$; Stage 2 $p = 0.5431$); this is not evidence of model equivalence. These ratings describe disagreement in visual judgment but do not establish expert consensus or conservation validity (Supplementary Tables S18, S24 and S32–S34).

High-error classification and calibration remained underpowered

Only three of 64 original objects exceeded the exploratory high-error threshold (masked MAE > 0.12). With so few positive cases, classification statistics cannot establish reliable probability calibration (Supplementary Note S5). The threshold is not a validated conservation decision limit.

Discussion

Across the original and external datasets, disagreement among three generated reconstructions was associated with error in a fourth, held-out image. This separation avoids using the evaluated image to construct its own variability score, although the images share input and mask conditions. Ranking by variability selected lower-error outputs without altering their pixels; the association also persisted across the second diffusion model and subsets defined by subsequent expert foreground and material review. Sampling-based uncertainty is useful only insofar as its relationship to actual error is demonstrated [39,40,41]. The original 64-object estimate ($r = 0.7176$) was lower than that observed in the pilot, underscoring the limited predictive value of small preliminary samples.

When S, M and L conditions were ranked separately, variability remained associated with error and selected sets had lower average MAE. However, the primary analysis concerns three-condition object means, and neither analysis directly establishes performance for real historical damage. The mask-specific exploratory estimates should not be used as calibrated probabilities for one historically damaged object. Acceptance coverage is also uneven by visual class. A global ranker retained 68.75% of plain versus 12.50% of decorated originals, while forcing 13 accepted entries per class incurred higher mean error in both models. This trade-off matters for heritage applications: selection based only on estimated image error may disproportionately withhold complex decorated objects from the lower-risk subset (Fig. S8; Supplementary Table S30).

The apparent transfer across cohorts should nevertheless be interpreted cautiously. The model architectures remain related, museum photographs share visual conventions, and the external experiment uses small fixed circular masks rather than the original object-relative occlusions. The two cohorts therefore cannot be pooled as equivalent damaged-surface fractions. Positive mask-specific associations and foreground-overlap sensitivities test robustness within these defined synthetic tasks, not authenticity of reconstructed historical details (Fig. S5; Supplementary Table S21).

Selective acceptance lowers the average error of retained objects by changing which outputs are inspected; it does not improve the reconstructions themselves. This distinction follows the selective-prediction framework [43,44]. In museum imaging, such a ranking could help prioritize visual inspection, but accepted and rejected completions alike remain unverified proposals. This interpretation accords with evidence-aware approaches to virtual reconstruction [6,7].

The variability score requires three predictor generations, whereas simple descriptors are obtained directly from the masked input. It may be a useful by-product when multiple candidates are already generated, but it adds computation to a single-output workflow. Without runtime, energy or downstream review measurements, no efficiency claim is justified. Human inspection remains essential when several seeds reproduce the same plausible error.

Where only one proposal would otherwise be generated, the additional sampling cost must be considered alongside any improvement in review prioritization.

Image difficulty partly explains selection: the lowest-uncertainty 26 originals contained 22 plain-looking objects and four decorated ones. Matched-class random selection and class-balanced uncertainty ranking improved on unconstrained random selection but did not match the lowest global accepted-set error. Coverage by visual class should therefore accompany overall error statistics whenever the retained sample may underrepresent decorated objects.

An incremental advantage over inexpensive visual-complexity baselines was not consistently established. In the original cohort, accepted-set MAE under matched boundary Sobel was 0.01889 versus 0.01811 under uncertainty for SD1.5 and 0.01669 versus 0.01624 for SD2, with bootstrap differences spanning zero. Larger external differences (0.04908 versus 0.03934 and 0.04981 versus 0.03961) were retrospective and depend on distinct image and mask characteristics. Likewise, the positive residual association after adjusting for image context cannot exclude unmeasured difficulty. Any extra ranking utility must be weighed against the cost of repeated diffusion sampling (Supplementary Tables S19–S20).

A variability score could help organize review of generated proposals, but cannot authenticate reconstructed ornament. Because decorated items were less often accepted, practical deployment would also need to monitor unequal coverage across culturally important visual classes.

Existing conservation-imaging research provides an important point of comparison. Lacquer-specific image restoration already targets appearance and structural coherence [5], while studies of mural imagery have investigated guided diffusion, restoration of damaged regions and preservation of visual features [12,13,14,15,16]. Painting-conservation reviews stress the gulf between simulated supervision and the complexity of aged originals [8]. Similarly, work on bronzeware restoration and traditional-painting reconstruction raises distinct questions about intervention scope and visual fidelity [20,22]. Our work does not introduce a new inpainting architecture or establish greater historical fidelity. Rather, it tests whether candidate variability predicts held-out photographic error and examines the consequences of selection and image complexity. The established perception–distortion trade-off further cautions against treating lower pixelwise error as synonymous with higher perceptual quality[49].

All reference-based outcomes concern synthetic occlusions, and agreement with photographic pixels says nothing directly about pigments, lacquer technology or historical iconography. Moreover, the external circular masks differ geometrically from the object-relative masks in the original cohort. Expert review confirmed their overlap with visible object foregrounds, not with analytically identified lacquer layers. The external records also cover heterogeneous materials, include one linked ensemble/component pair and may overlap with model pretraining data. The 40-entry results and restricted-subset analyses should be interpreted within those boundaries, not as fully independent heritage-restoration validation.

The within-museum estimates offer descriptive comparisons, not a leave-one-museum-out test of transfer, particularly for the six-object Cooper Hewitt subset. Visual examples show both concordant and discordant cases without measuring their prevalence. The sparse high-error event count also precludes defensible probability calibration.

The verified image-to-rating mapping makes the anonymous visual assessments traceable, yet their reliability remains limited. Differences across the main visual criteria were confined to adjacent scores, yet the consistent positive rater-01 versus rater-02 shift (about 0.4 of a 0–3 grade) and low ordinal α prevent a claim of shared professional judgment. Quadratic-weighted κ gives a complementary account of grade-distance agreement, not a replacement for the low α values. Highly correlated patterns across the four ratings for each rater raise a concern about redundant scoring constructs; suitability for further human review is a separate workflow item, not a measure of freedom from visual defects. The Stage 1 and Stage 2 questions measure different constructs and cannot be treated as before/after changes. These rating diagnostics do not alter the independently calculated uncertainty–error associations, masked MAE or risk–coverage results. Separately, two-expert material and ROI review improves interpretability of external masking; because it followed outcome analysis, its restricted-cohort results remain retrospective sensitivities (Supplementary Tables S24–S28 and S32–S34).

Prospective studies on additional collections could evaluate runtime, review decisions and more representative geometries of loss. Real damage would require independently justified reference standards and material evidence, not synthetic photographic ground truth alone. Expert studies should also examine whether triage improves documented review decisions rather than treating agreement between two raters as conservation endorsement.

Seed-disjoint sampling variability ranked masked reconstruction error across two diffusion models and original and external museum-photo cohorts under synthetic loss. At fixed coverage, selective acceptance reduced the mean error of retained entries without improving individual images; gains over inexpensive boundary-Sobel ranking were small and uncertain in the original cohort. Subsequent two-expert foreground/material review supported positive associations in restricted external subsets, but did not establish lacquer-layer placement, historical authenticity or professional consensus. The results support a reproducible signal for prioritizing photographic reconstruction checks, subject to human review and the material context of each object.

Methods

Design and study units

The confirmation-stage dataset included 64 independently identified lacquerware objects (32 visually plain and 32 decorated) from the Walters Art Museum (n = 40), Cleveland Museum of Art (n = 18) and Cooper Hewitt, Smithsonian Design Museum (n = 6). The National Palace Museum Taipei was not represented in this cohort. Two blinded ROI assessments with subsequent adjudication defined the analyzed surface and visual class, as recorded in the frozen E4 object/ROI manifest. Eligibility screening used accession identifiers, original-image SHA-256 and perceptual-image checks; within each class the pre-recorded SHA-256 selection key determined the 32 final objects. Each adjudicated object surface was masked at approximately 7.5% (S), 20% (M) and 35% (L) of ROI area, yielding 192 image-mask cases nested within the 64 original objects. All masks lay within the locked ROI, and the intact image was withheld

as a pixel reference. Four fixed seeds generated 768 outputs in total. Inferential independence was defined at object level, not by treating nested masks, patches or repeated seeds as independent replicates; the logic of hierarchical resampling is documented in the wider statistics literature [50].

Image generation, environment and audit trail

A fixed pretrained Stable Diffusion v1.5 inpainting model was used throughout (versioned model repository: https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-inpainting; commit 8a4288a76071f7280aedbdb3253bdb9e9d5d84bb), with 30 diffusion steps, guidance scale 7.5 and 512 × 512 RGB outputs. Latent diffusion provides the broader model family, while diffusion-based inpainting studies describe masked conditional reconstruction [34,35,36]. Seeds were fixed at 11001, 11002, 11003 and 11004. The run lock records Windows, NVIDIA GTX 1080, Python 3.12.10, PyTorch 2.6.0 + cu124, diffusers 0.35.2 and transformers 4.57.6. The E3 protocol JSON creation timestamp (17 October 2026) precedes the logged E4 GPU experiment (18 October 2026), which establishes internal file chronology but not independently authenticated public preregistration. Direct SHA-256 auditing matched 768 image outputs against their manifest and generation logs, and E5 object-level uncertainty and error were independently reconstructed from output PNGs, masks and references.

Seed-disjoint uncertainty and masked error

For the prespecified primary seed orientation, per-case uncertainty was calculated from the three predictor-seed reconstructions (11002–11004) as their per-pixel RGB sample variance (ddof = 1), averaged over the masked region. The target score was the masked RGB mean absolute error for the excluded seed 11001 relative to the intact image, with pixel values scaled to [0,1]. Each object's S/M/L case values were weighted equally to obtain object-level scores. This evaluates a reproducible sample-dispersion statistic, not a Bayesian decomposition of uncertainty or a calibrated event probability [40,41,42]. Disjointness applies to the target and predictor images at this scoring step, while their image and experimental conditions remain shared. Reconstruction of the scores from the archived original PNGs agreed with the stored E5 metrics within floating-point tolerance.

Image intensities were normalized to [0, 1], and the three S/M/L observations were equally averaged within each object. For the exploratory leave-one-object-out check, one complete object (all nested conditions) was removed at a time before recomputing Pearson r and Spearman ρ on the other 63 objects. A separate check used 10,000 independent-object bootstrap resamples with replacement, NumPy random-generator seed 20261020 and percentile 2.5%/97.5% quantiles. These new diagnostic results were calculated from the archived E7.2 image-recomputed 64-object table; they were not part of the E3 prespecified analysis. Exact input identifier and output values are in Supplementary Table S7 and Supplementary Fig. S1.

Prespecified confirmation-stage inference

The E3_CONFIRMATORY_PROTOCOL_LOCK_V1.json file specifies object-level Pearson correlation of seed-disjoint uncertainty with masked MAE as the primary endpoint, two-sided alpha = 0.05 and 10,000 object-level bootstrap replicates. The prespecified success condition was positive Pearson r and a bootstrap 95% lower bound above 0.30. A second operating criterion required a reduction in accepted-set error at nominal 40% object coverage and a bootstrap lower bound above zero. Resampling the independent original objects rather than all nested masks or patches is consistent with methods for hierarchical measurements [50]. The internally recorded file chronology does not constitute external preregistration. Seed rotation and visual-category breakdowns are secondary descriptive checks.

Selective acceptance and confirmatory risk scoring

The 26 lowest-uncertainty objects were retained at nominal 40% acceptance (26/64 = 40.625%). Accepted-set risk is the mean masked MAE of the selected objects; relative reduction compares this value with mean masked MAE across the whole cohort. The E5 analysis used 10,000 object-level bootstrap replicates to summarize uncertainty in the relative reduction. Separately, a previously frozen model estimated the probability that an object's error exceeded the methodological MAE threshold 0.12. Discrimination and Brier-score summaries are reported, but only three objects were positive, so probability calibration cannot be established robustly [45,46]. The accepted-set ranking is not formal conformal risk control [47] and should not be interpreted as prospective safety certification.

The oracle comparator in Fig. 3 is defined, for any selected count k, as the mean observed masked error among the k lowest-error objects in the evaluated cohort. Thus oracle sorting assumes access to the very ground-truth errors the uncertainty score is intended to estimate; it is solely a hindsight lower-bound diagnostic. Coverage counts are k = 13, 26, 39, 52 and 64 (20.31%, 40.63%, 60.94%, 81.25% and 100%). The uncertainty-based selection uses the same counts after sorting by the variance score (Supplementary Table S8).

Post-hoc composition sensitivity (E7)

After examining the initial results, we conducted 200,000 independent uniform random selections of 26 objects and 200,000 class-matched selections retaining 22 plain and four decorated objects. A class-balanced uncertainty ranking accepted the 13 lowest-scoring objects in each visual class. Class-centered Pearson correlation and 10,000 class-stratified object-bootstrap iterations were also calculated, keeping the independently photographed object as the resampling unit [50]. These comparisons were exploratory and reused the original objects rather than providing an independent validation sample.

Further exploratory analyses used the original image archive, including the 64-object accession/museum register, 192 ground-truth/mask/input cases and all 768 generated PNG outputs. The recomputed 192 masked MAEs agreed with the original 64-object metric table within 1×10^{-8} after averaging three occlusion levels. Museum-stratified descriptions, visible-context image baselines and masked SSIM were calculated post hoc and are neither independent observations nor newly prespecified hypotheses (Supplementary Tables S9–S11).

For the complexity baselines, each masked 512 × 512 input was converted to uint8 grayscale. An annular context was defined as pixels within a 49 × 49 square dilation of the binary mask but outside its 7 × 7 square dilation, corresponding to an approximately 4–24-pixel Chebyshev-distance band of visible context. Pixels inside or immediately bordering the occlusion were excluded from all baseline summaries. We calculated standard deviation of grayscale intensities divided by 255; Canny edge density at thresholds 50 and 150 (OpenCV uint8 implementation); and mean Euclidean 3 × 3 Sobel gradient magnitude divided by 1443. Case values were equally averaged within each object. These image-based descriptors require only the masked input and are intentionally simple diagnostic baselines, not optimized computer-vision models.

At nominal 40% coverage, the 26 lowest-score objects were separately selected for each descriptor and their held-out target masked MAE was averaged. To contrast edge density with sampling variance, 10,000 bootstrap resamples of 64 objects with replacement were drawn using NumPy default_rng seed 20261020; the lowest 26 scores were selected anew within each resample and the difference in accepted mean MAE (edge minus uncertainty) recorded. The 2.5th and 97.5th percentiles summarize within-cohort uncertainty but not future collection performance. We additionally computed within-museum Pearson and Spearman associations, deletion-of-one-museum sensitivity, and the Pearson correlation of residuals from separate ordinary least-squares regressions of the two primary scores on museum indicators, plain/decorated class and all three complexity proxies. These regressions are post-hoc descriptive adjustments, not prospective predictive models.

For SSIM, the target seed-11001 RGB output was compared with the intact 512 × 512 reference using the local RGB SSIM map (data_range = 1, channel_axis = 2, win_size = 7; full=True). Only pixels inside a mask eroded by a 7 × 7 square contributed to the corrected summary. Scores were averaged equally over S/M/L within each object. Supplementary Table S11 reports this mask-interior definition; the primary uncertainty and MAE outcomes are unaffected.

Cross model and external collection extension

E11 used Stable Diffusion 2 inpainting (sd2-community/stable-diffusion-2-inpainting, revision 5f74973cbb64c8568780732c17f43eb269d63a0d) and the frozen Stable Diffusion 1.5 configuration. Generation logs specify 30 steps, guidance scale 7.5, four seeds and 512 × 512 RGB images. The external cohort comprised 40 catalogue entries (20 each from the Metropolitan Museum of Art and Art Institute of Chicago); official images were square-padded then resized to 512 × 512. Three fixed circular masks contained 489, 1,313 and 2,209 pixels: 0.1865%, 0.5009% and 0.8427% of the full image canvas. These canvas fractions are not directly comparable to E10 mask percentages of approximately 7.5%, 20% and 35% of an adjudicated object ROI. The external object list was frozen before generation, while specialist review of material role, ornament and visible foreground was completed afterward. Two experts assessed the original 40 entries without modifying their outcome data. Two sets of 40 512 × 512 binary ROI PNGs and the original S/M/L masks were independently intersected pixelwise, with foreground overlap defined as $|\text{mask} \cap \text{ROI}| / |\text{mask}|$; ≥80% on-object defined an exploratory foreground sensitivity subgroup. Both specialists gave identical category decisions for all 40 objects, while mask agreement was evaluated continuously with ROI IoU. The final archived second-pass mask is used for reported numerical overlap (Supplementary Tables S25, S26, S28; Fig. S6).

For every case, uncertainty is the masked RGB sample variance across seeds 11002–11004 and error is seed-11001 masked MAE on [0,1] values. The three mask-level values are equally averaged by entry. Original-cohort coverage retains 26/64; external coverage retains 16/40. Percentile intervals from 10,000 resamples use NumPy default_rng seed 20261021, resampling entries and reselecting within each draw. Ties resolve by ascending object ID. Original prespecified E5 intervals are retained separately. New external permutation tests shuffle entry-level MAE relative to fixed uncertainty scores, using a two-sided Pearson statistic and a one-sided full-minus-selected mean-MAE statistic. The four new p values use the plus-one Monte Carlo rule and Holm correction. They are retrospective reanalyses, conditional on entry exchangeability; the ensemble/component exclusions are sensitivity checks.

E11 boundary context is the five-iteration binary-dilation ring outside the mask; mean Sobel magnitude is extracted from reference pixels that are visible in the masked input. Additional descriptors are whole-visible-region Sobel magnitude and grayscale SD. For the unified retrospective baseline comparison, object-level values were computed by equal averaging over S/M/L for each of the four cohort–model groups, then separately ranked in ascending order of their respective scores. Realized 40% coverage was 26/64 or 16/40; an additional area-under-risk–coverage summary averages the accepted-set masked MAE over integer accepted counts corresponding to 25–100% coverage (S19). Exploratory paired object-resampling uses 2,000 bootstrap replicates (NumPy seed 20261021), reselecting after resampling, and returns percentile 95% intervals. Differences are baseline MAE minus uncertainty MAE; bootstrap tail fractions are descriptive, not permutation-test p values or external guarantees. A control-only rank-residual association was also calculated using boundary Sobel, visible Sobel, grayscale SD and mean mask-pixel count; 500 object-resamples provide exploratory intervals (S20). No runtime comparison was available. Within the inner five-iteration boundary band, color MAE and Sobel error were archived. The corrected SSIM summary uses an RGB local map with win_size 7, data_range 1 and a 7 × 7 square-eroded mask. Fixed-crop LPIPS-Alex v0.1 uses the mask bounding box padded by 32 pixels, bilinear resizing to 256 × 256 with align_corners=False and scaling to [− 1,1]. LPIPS is used here as a deep-feature perceptual-distance measure, not as a criterion of historical accuracy[48].

Anonymous visual review used 80 medium-mask image items (40 external museum entries × SD1.5/SD2, seed 11001). Stage 1 presented anonymized masked-context and reconstruction panels and assessed pattern continuity, texture consistency, suspicious visual artifacts and suitability for further human review; Stage 2 revealed the intact photographic reference and assessed reference consistency. A score of 3 on the further-review item denoted suitability to proceed to human inspection, not absence of a need for further review. Model and object names were withheld. Two worksheets of 80 ratings were aligned by anonymous ID and the 40 object-pair IDs, and the original Blind_Key.csv and 160 indexed panels were hash-matched. Original ordinal responses were not

edited. Agreement was summarized for each of five 0–3 items using pooled-frequency ordinal Krippendorff α; quadratic-weighted Cohen κ, exact agreement, score marginals and signed/absolute rater differences were supplementary diagnostics. A retrospective 4,000-draw percentile bootstrap resampled the 40 objects with both model outputs retained per draw, rather than treating 80 related images as independent units; paired intervals were calculated for α, κ and mean rater-score differences. We also inspected whether within-rater scores were identical across the four visual items, without assuming that they constitute a validated common scale. Stage 1 and Stage 2 did not repeat the same rating items, so a direct same-item stage-change analysis was not attempted. The original two-sided paired Wilcoxon model comparisons used per-item two-rater averages, zero_method=zsplit and 10,000 paired object-bootstrap draws; their values were retained. The first-stage four-item average is only an exploratory composite. Reviewer qualifications were confirmed by the study team, but matching images and worksheets alone does not document recruitment chronology or independent consent procedures (Supplementary Note S20; Tables S32–S34).

Single-mask risk ranking and visual-class selection sensitivity

For the single-mask analysis, each S/M/L case was considered separately, using one case per original object or external catalogue entry at that mask level (n = 64 or n = 40); the levels were analyzed separately and never treated as three independent replicates of a single object. At each level, objects were ranked by the same masked RGB variance computed from seeds 11002–11004, and outcome risk was the masked MAE of excluded seed 11001. The lowest floor(0.40n + 0.5) ranks were accepted (26/64 or 16/40). Comparator sorting used the original Context_Edge boundary-Sobel descriptor at the same mask level. A retrospective 4,500-replicate paired object/entry bootstrap sampled entries with replacement and reranked each selector within each draw, giving percentile 95% intervals for accepted MAE and boundary-Sobel-minus-uncertainty risk; model and mask effects were not considered independent replications and no adjustment for 12 post-hoc comparisons was made. The curves show the accepted-set risk from 20% through 100% of entries (Fig. S7; Supplementary Table S29).

For the visual-class audit, frozen original-cohort labels (32 plain and 32 decorated) were used separately for SD1.5 and SD2, without incorporating later E11 external labels. Global selection chose the 26 lowest-uncertainty objects; the class-balanced alternative retained the 13 lowest in each class. Acceptance proportions were computed with the original denominators of 32 objects per class. For within-class risk–coverage each class was ranked independently at integer counts 1–32. These are retrospective, descriptive selection-policy comparisons, not new prespecified tests (Fig. S8; Supplementary Table S30).

The two-expert foreground/material review is separate from the two anonymous inpainting-quality raters. Two archived ROI versions and two sets of category judgments were inspected and a dated boundary-adjudication record retained the second-pass foreground outlines. The records establish review by two experts but do not establish that each outline was drawn from scratch without access to the other; agreement in categorical judgments should not be interpreted as independently blinded segmentation reliability. Object foregrounds, not chemically verified lacquer coatings, define the reported intersection analysis. A compact mask-geometry comparison reports absolute pixels and original denominator conventions (Supplementary Table S31).

Data provenance and reuse conditions

The original image experiment used photographed objects and synthetic occlusions, not measurements of material conservation outcomes. E11 includes anonymous visual-rating records whose image-to-score mapping was independently hash-verified. The research team confirms that the two original reviewers had relevant subject-matter qualifications and formal ethics approval was not required under the applicable requirements; these author statements are not third-party certifications. Original and external source image IDs and file hashes do not overlap, although pretrained model exposure is unknown. One ensemble–component relationship remains documented. A later two-expert assessment confirmed object material/visible-ornament classes and 80 object-ROI mask files. Across all 120 external synthetic masks, the foreground overlap was independently recomputed from original PNGs; classifications of objects versus background are not evidence of lacquer-layer-specific mask placement. Full-40 results remain primary, and reviewed foreground/material subsets are retrospective sensitivity analyses (Supplementary Tables S25–S28).

Declarations

Competing interests

The authors declare no known competing financial or non-financial interests.

Ethics approval and consent to participate

Not applicable. The study analyzed museum object images and incorporated anonymous professional visual assessments of reconstruction quality. No identifiable personal or sensitive participant data were collected in the reported assessment. Under the applicable requirements, formal research ethics approval was not required, as confirmed by the authors.

AI-assisted work declaration

Stable Diffusion models were used for the image-inpainting experiments as described in the Methods section. OpenAI ChatGPT was used to assist with language revision, manuscript organization, literature-search support, and preliminary code and document drafting. The authors are responsible for verifying all data, analyses, sources, figures, and conclusions.

Author Contribution

P.Y. conceived the study, developed the methodology, performed validation and formal analysis, conducted the investigation, and contributed to manuscript drafting and revision. J.H. contributed to methodology, provided research resources, and participated in manuscript drafting and revision. X.Z. developed the software, contributed to validation and visualization, and participated in manuscript drafting and revision. All authors reviewed and approved the final manuscript.

Acknowledgments

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data supporting the findings of this study, including derived analytical tables, experimental records, and reproducibility materials, are available through the Open Science Framework (OSF) reviewer-access repository: https://osf.io/s8umb/overview?view_only=97f28d86cff3471691dcfcb2854d3508. Access to original museum photographs and certain administrative review records remains subject to source-specific usage conditions and applicable restrictions.

Code availability

Image-generation and analysis scripts, including the original experiments and exploratory checks, are archived with model specifications, environment files and checksums. The original study code includes fixed local Windows paths. Reviewer access is provided through the OSF project: https://osf.io/s8umb/overview?view_only=97f28d86cff3471691dcfcb2854d3508. Separate E11 generation and analysis scripts are retained in the supplied archives; the archive includes image-integrity checks, metric calculations, permutation tests and rating-table diagnostics.

References

1. Shen W, et al. Multi-analytical investigation of the Western Han Dynasty Lacquer Lian from Hunan, China. npj Herit Sci. 2026;14:456. https://doi.org/10.1038/s40494-026-02691-5.
2. Deng B, et al. 3D thermography for the surface and internal defects detection of lacquerware artifacts. npj Herit Sci. 2026. https://doi.org/10.1038/s40494-026-02821-z. Published online 11 July 2026.
3. Liang J, Li Y, Xiong Z, Huang Q. Advancing lacquerware design through human-AI collaboration with controllable diffusion models. Sci Rep. 2026;16:421. https://doi.org/10.1038/s41598-025-33119-y.
4. Tang X, Zhan C, Tang C, Kim S. AI and AR based digital reconstruction of Liangshan Yi lacquerware. Sci Rep. 2026;16:5550. https://doi.org/10.1038/s41598-026-35772-3.
5. Xu Q, Iwasaki M. Image restoration model of wood-structured raw lacquer cultural relics based on diffusion generation network. J Asian Archit Build Eng. 2025. https://doi.org/10.1080/13467581.2025.2603753. Published online 19 December 2025.
6. Paolanti M, Frontoni E, Pierdicca R. Towards trustworthy AI in cultural heritage. npj Herit Sci. 2026;14:131. https://doi.org/10.1038/s40494-026-02403-z.
7. Park SW, et al. A systematic and procedural framework for digital reconstruction of architectural cultural heritage using generative AI. npj Herit Sci. 2026;14:549. https://doi.org/10.1038/s40494-026-02761-8.
8. Van Vijle A, Hacıgüzeller P, Van der Snickt G. Machine learning for painting conservation: a state-of-the-art review. npj Herit Sci. 2025;13:437. https://doi.org/10.1038/s40494-025-01924-3.
9. Chu E, Chu X. Yongle Palace murals' image restoration using deep learning optimization algorithms. npj Herit Sci. 2025;13:683. https://doi.org/10.1038/s40494-025-02199-4.
10. Wei X, et al. Progressive enhancement and restoration for mural images under low-light and defective conditions based on multi-receptive field strategy. npj Herit Sci. 2025;13:63. https://doi.org/10.1038/s40494-025-01635-9.
11. Chen L, Wu L, Wan J. Damage detection and digital reconstruction method for grotto murals based on YOLOv10. npj Herit Sci. 2025;13:91. https://doi.org/10.1038/s40494-025-01607-z.
12. Li Y, et al. An improved mural image restoration method based on diffusion model. npj Herit Sci. 2025;13:347. https://doi.org/10.1038/s40494-025-01914-5.
13. Zhao F, et al. Diffusion-based heterogeneous network for ancient mural restoration. npj Herit Sci. 2025;13:206. https://doi.org/10.1038/s40494-025-01719-6.
14. Jiang C, Ren T, Cheng Z. All-in-one mural restoration with prompt-guided residual diffusion. npj Herit Sci. 2025;13:667. https://doi.org/10.1038/s40494-025-02232-6.
15. Hu J, Yu Y, Zhou Q. GuidePaint: lossless image-guided diffusion model for ancient mural image restoration. npj Herit Sci. 2025;13:118. https://doi.org/10.1038/s40494-025-01693-z.
16. Xu Z, et al. Automated mural restoration via semi supervised segmentation and prompt guided diffusion inpainting. npj Herit Sci. 2025;13:504. https://doi.org/10.1038/s40494-025-02047-5.

17. Lei T, et al. Low-rank structure guided diffusion for Shaanxi temple mural restoration. npj Herit Sci. 2025;13:295. https://doi.org/10.1038/s40494-025-01827-3.
18. Swathi B, Rao DB. J. Automated image inpainting for historical artifact restoration using hybridisation of transfer learning with deep generative models. Sci Rep. 2026;16:4810. https://doi.org/10.1038/s41598-026-35056-w.
19. Zhang C, Li X, Jean M. Restoration of ancient Chinese paintings using GANs with multi-scale defect fusion architecture and multi-pooling enhanced spatial-channel attention. Sci Rep. 2026. https://doi.org/10.1038/s41598-026-64962-2. Published online 17 August 2026.
20. Yang C. A coarse-to-fine restoration framework integrating deep semantic inpainting and texture-aware refinement for Jinling school landscapes. Sci Rep. 2026;16:19267. https://doi.org/10.1038/s41598-026-49966-2.
21. Yu J, Li H. Digital restoration of cultural relics based on 3D modeling and machine vision. Sci Rep. 2026;16:15124. https://doi.org/10.1038/s41598-026-45060-9.
22. Sui Q, Shou W, Yang H. Diffusion-based restoration of corroded bronzeware images under a minimal-intervention framework: spatial-compliance indices and over-restoration risk diagnostics. npj Herit Sci. 2026;14:259. https://doi.org/10.1038/s40494-026-02539-y.
23. Qin Z, Zeng Q, Zong Y, Xu F. Image inpainting based on deep learning: a review. Displays. 2021;69:102028. https://doi.org/10.1016/j.displa.2021.102028.
24. Zhang X, et al. Image inpainting based on deep learning: a review. Inform Fusion. 2023;90:74–94. https://doi.org/10.1016/j.inffus.2022.08.033.
25. Xiang H, et al. Deep learning for image inpainting: a survey. Pattern Recogn. 2023;134:109046. https://doi.org/10.1016/j.patcog.2022.109046.
26. Yang J, Ruhaiyem NIR. Review of deep learning-based image inpainting techniques. IEEE Access. 2024;12:138441–82. https://doi.org/10.1109/ACCESS.2024.3461782.
27. Suvorov R et al. Resolution-robust large mask inpainting with Fourier convolutions. Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, 2149–2159 (2022). https://openaccess.thecvf.com/content/WACV2022/html/Suvorov_Resolution-Robust_Large_Mask_Inpainting_With_Fourier_Convolutions_WACV_2022_paper.html
28. Li W et al. MAT: Mask-aware transformer for large hole image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10758–10768 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Li_MAT_Mask-Aware_Transformer_for_Large_Hole_Image_Inpainting_CVPR_2022_paper.html
29. Dong Q, Cao C, Fu Y. Incremental transformer structure enhanced image inpainting with masking positional encoding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11358–11368 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Dong_Incremental_Transformer_Structure_Enhanced_Image_Inpainting_With_Masking_Positio
30. Liu Q et al. Reduce information loss in transformers for pluralistic image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11347–11357 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Reduce_Information_Loss_in_Transformers_for_Pluralistic_Image_Inpainting_CVPR_2022_
31. Wang W, Niu L, Zhang J, Yang X, Zhang L. Dual-path image inpainting with auxiliary GAN inversion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11421–11430 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Dual-Path_Image_Inpainting_With_Auxiliary_GAN_Inversion_CVPR_2022_paper.html
32. Kim SY et al. Zoom-to-Inpaint: Image inpainting with high-frequency details. Proceedings of IEEE/CVF CVPR Workshops, 477–487 (2022). https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Kim_Zoom-to-Inpaint_Image_Inpainting_With_High-Frequency_Details_CVPRW_2022_paper.html
33. Wang S et al. Imagen Editor and EditBench: Advancing and evaluating text-guided image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18359–18369 (2023). https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Imagen_Editor_and_EditBench_Advancing_and_Evaluating_Text-Guided_Image_Inpainting_CVPR_2023_paper.html
34. Liu H et al. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8038–8047 (2024). https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Structure_Matters_Tackling_the_Semantic_Discrepancy_in_Diffusion_Models_for_CVPR_2
35. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
36. Lugmayr A et al. RePaint: Inpainting using denoising diffusion probabilistic models. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 11461–11471 (2022). https://doi.org/10.1109/CVPR52688.2022.01117
37. Ma X, et al. Uncertainty-aware image inpainting with adaptive feedback network. Expert Syst Appl. 2024;235:121148. https://doi.org/10.1016/j.eswa.2023.121148.
38. Giakoumoglou P, Karageorgiou D, Papadopoulos S, Petrantonakis PC. SAGI: Semantically aligned and uncertainty guided AI image inpainting. In Proc. IEEE/CVF Int. Conf. Comput. Vis. 16090–16101 (2025). https://doi.org/10.1109/ICCV51701.2025.01493
39. De Vita M, Belagiannis V. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, 3844–3854 (2025). https://openaccess.thecvf.com/content/WACV2025/html/De_Vita_Diffusion_Model_Guided_Sampling_with_Pixel-Wise_Aleatoric_Uncertainty_Estimation_WACV_2025_paper.html
40. Franchi G, Belkhir N, Trong DN, Xia G, Pilzer A. Towards understanding and quantifying uncertainty for text-to-image generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8062–8072 (2025). https://doi.org/10.1109/CVPR52734.2025.00755

41. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413 (2017). https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html
42. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Adv. Neural Inf. Process Syst. 2017;30:5574–84. https://papers.neurips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.
43. Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In Adv. Neural Inf. Process Syst. 2017;30:4878–87. https://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.
44. Galil I, Dabbah M, El-Yaniv R. What can we learn from the selective prediction and uncertainty estimation performance of 523 ImageNet classifiers? In Int. Conf. Learn. Represent. (2023). https://mlanthology.org/iclr/2023/galil2023iclr-we/
45. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In Proc. 34th Int. Conf. Mach. Learn., PMLR 70, 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html
46. Minderer M et al. Revisiting the calibration of modern neural networks. In Adv. Neural Inf. Process Syst 34 (2021). https://mlanthology.org/neurips/2021/minderer2021neurips-revisiting/
47. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. In Int. Conf. Learn. Represent. (2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html
48. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
49. Blau Y, Michaeli T. The perception-distortion tradeoff. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6228–6237 (2018). https://doi.org/10.1109/CVPR.2018.00652
50. Saravanan V, Berman GJ, Sober SJ. Application of the hierarchical bootstrap to multi-level data in neuroscience. Neurons, Behavior, Data Analysis, and Theory 3(5), 1–25 (2020). https://nbdt.scholasticahq.com/article/13927-application-of-the-hierarchical-bootstrap-to-multi-level-data-in-neuroscience

Figures

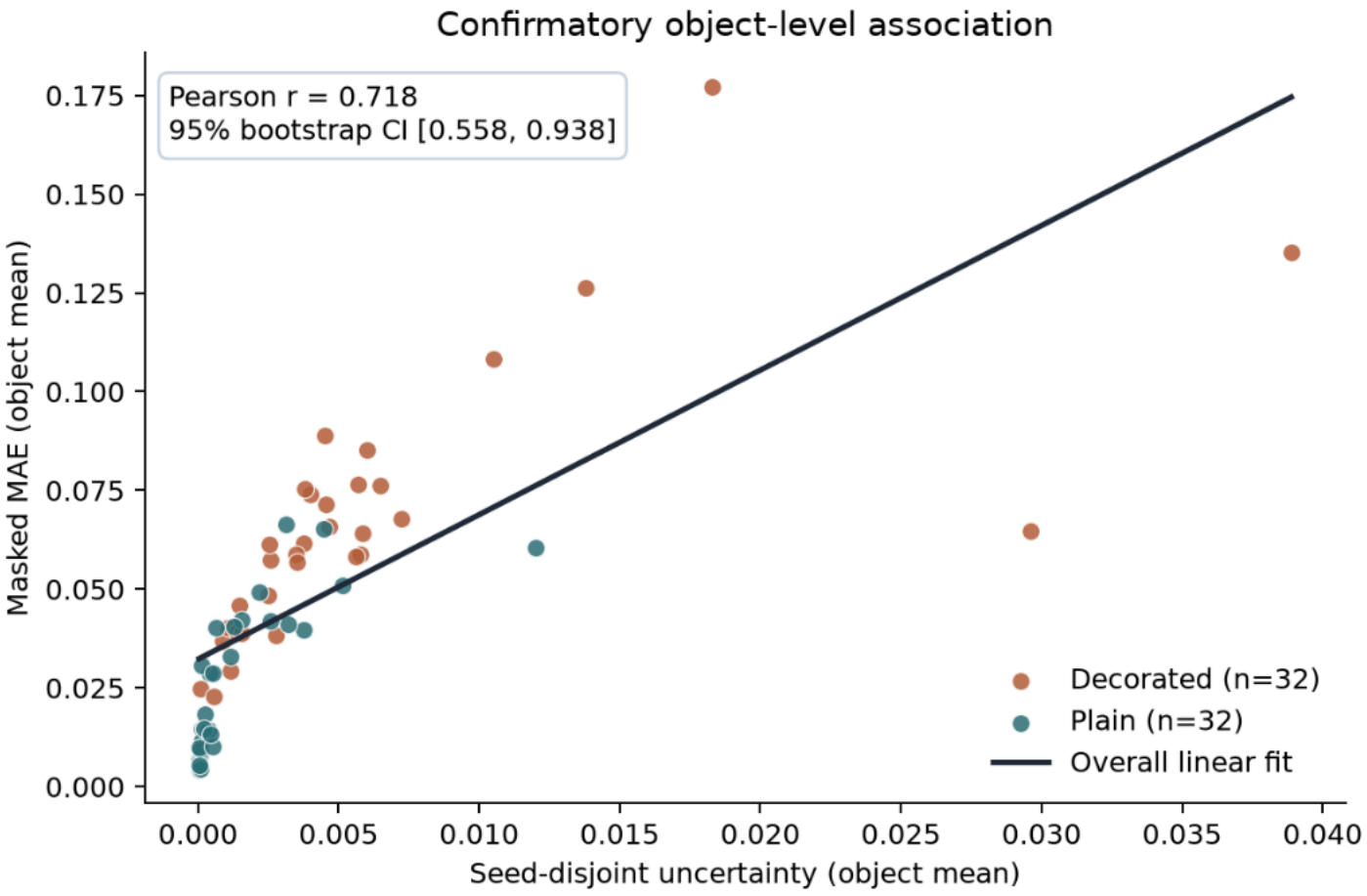


Figure 1

Relationship between seed-disjoint sampling uncertainty and masked reconstruction error across 64 independent lacquerware objects. Each point represents one object aggregated across three controlled occlusion cases; plain and decorated are visual categories, not laboratory-verified material types.

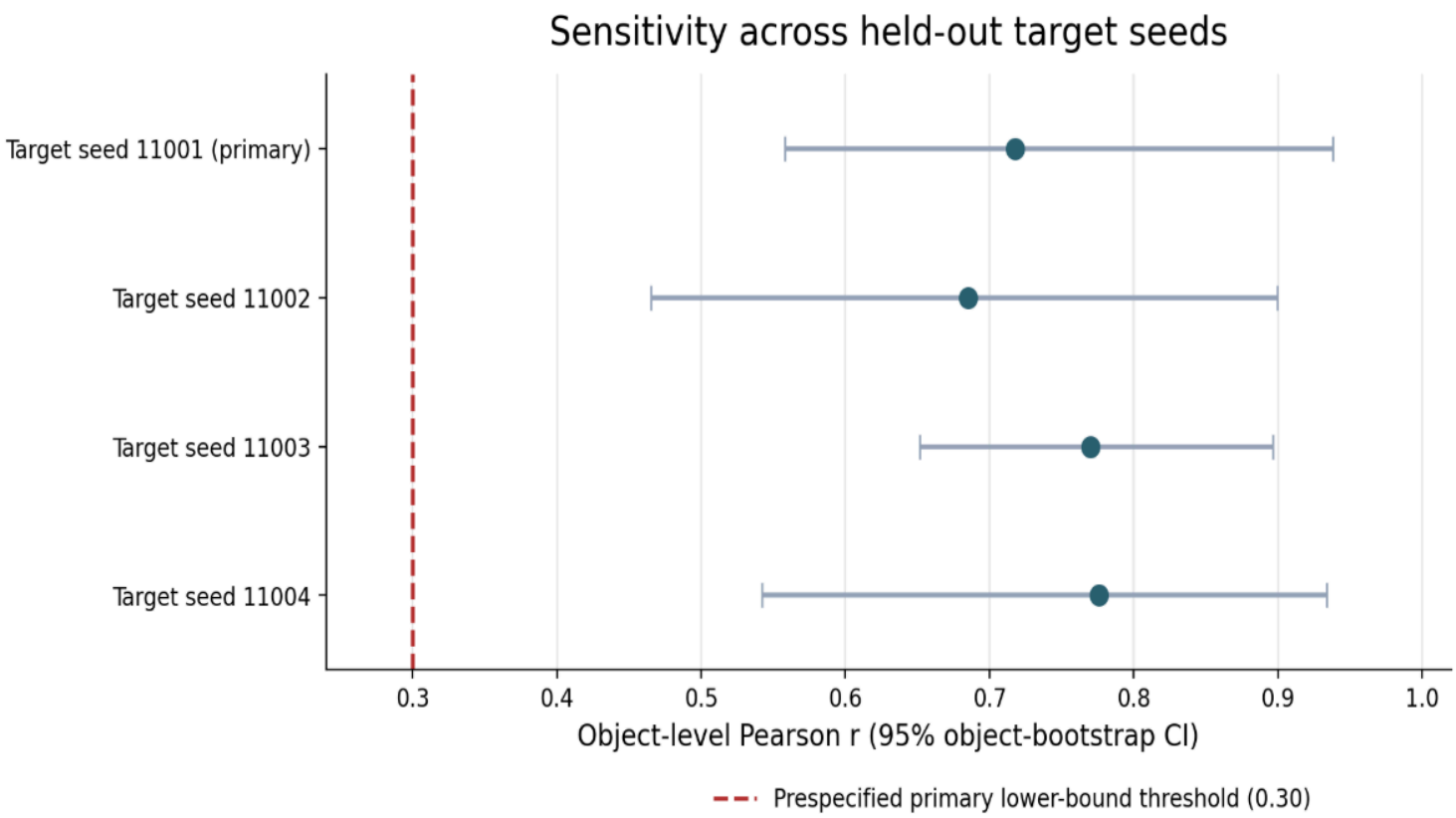


Figure 2

Object-level sensitivity to the held-out target seed. Target 11001 is shown with the prespecified E5 primary 95% object-bootstrap interval; 11002–11004 use the stored secondary rotation intervals. The rotations share underlying images and are not independent validation sets. The earlier 11001 rotation-only interval differs slightly from the primary interval and remains documented in Supplementary Table S3.

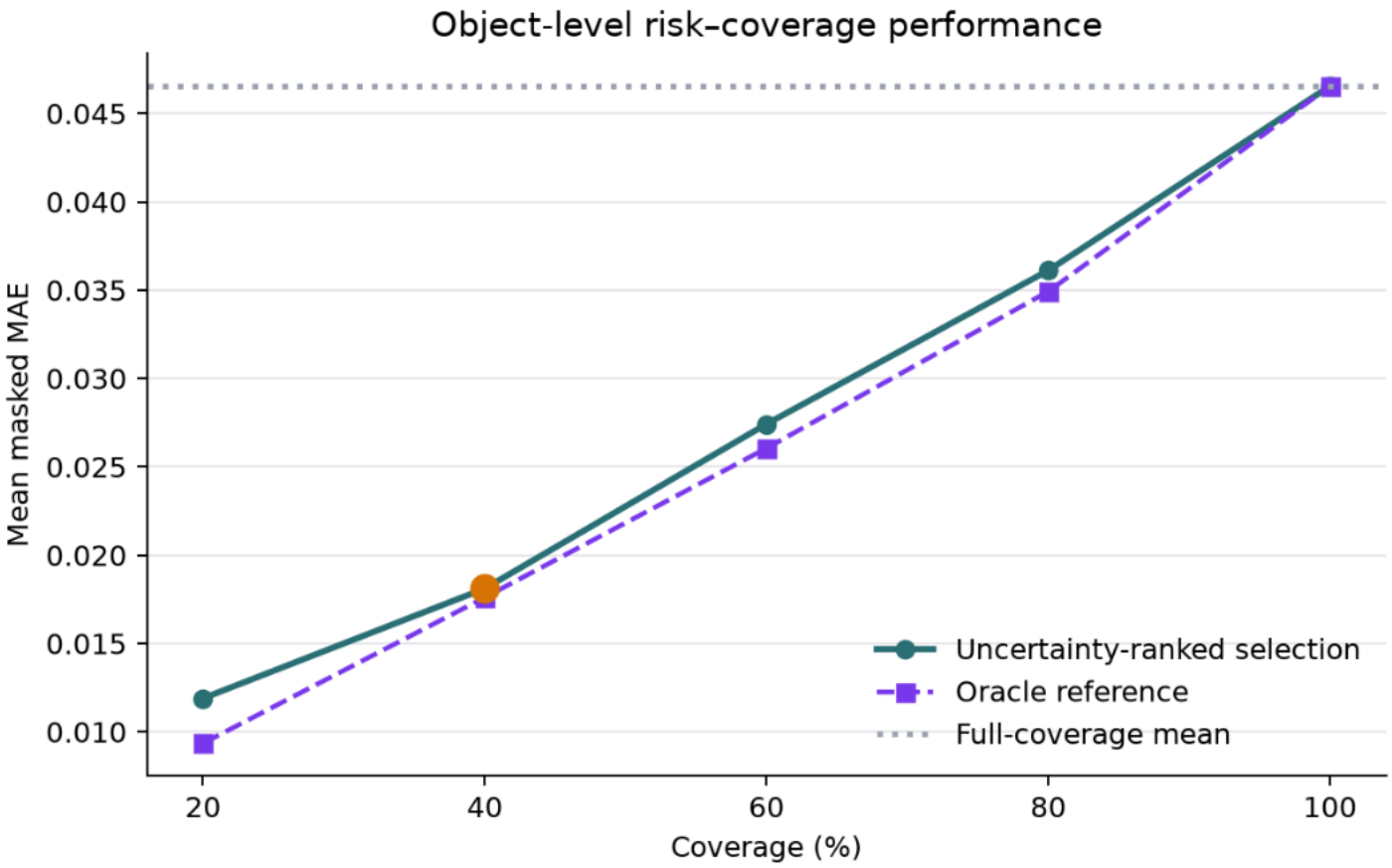
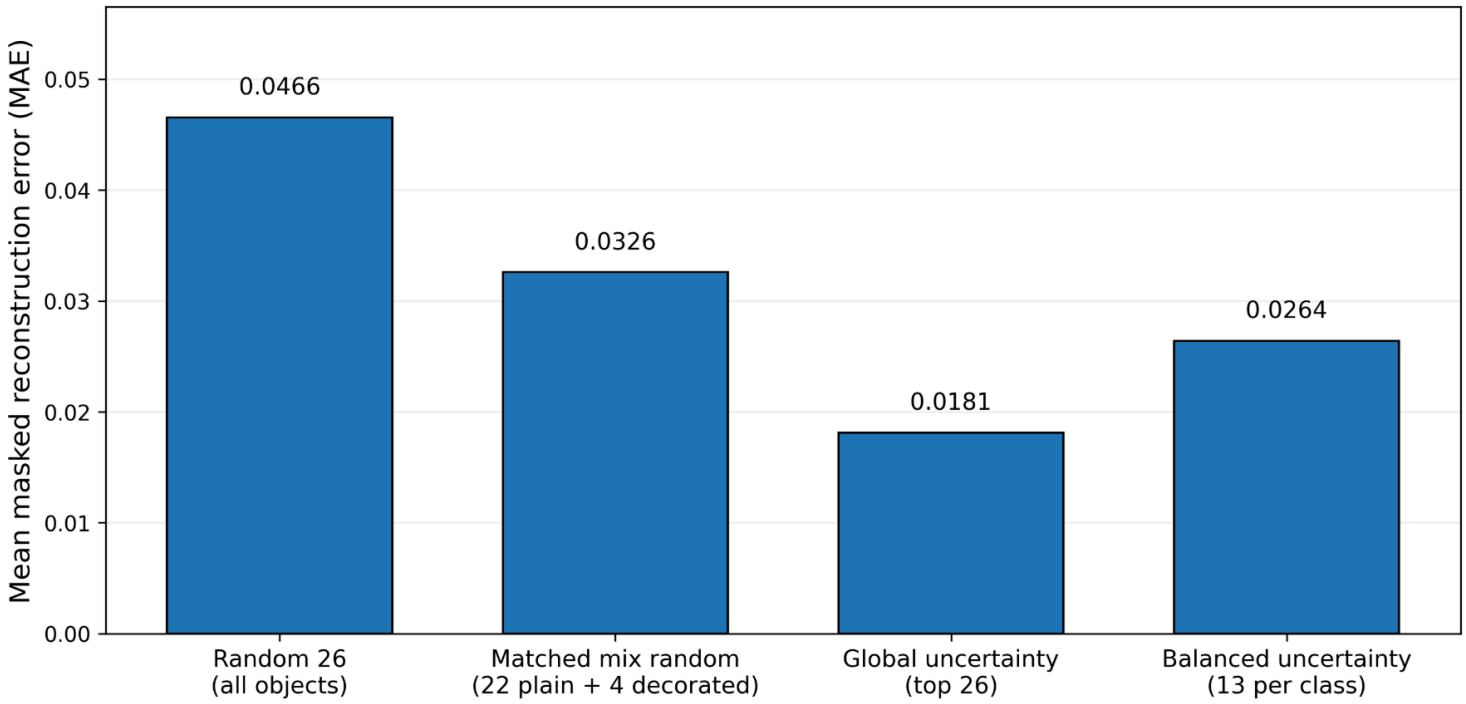


Figure 3

Object-level error–coverage curve. The lowest-uncertainty 26/64 objects were accepted at the nominal 40% operating point. The oracle reference sorts by observed masked MAE and is an unattainable retrospective within-cohort lower bound, not a prospective decision tool. The dotted line indicates full-cohort mean error; the ranking does not improve any individual reconstruction.



Exploratory E7 comparison; random bars show Monte Carlo expected values (200,000 draws).
Observed ranking bars use the same 64-object E5 cohort; no external test cohort.

Figure 4

Exploratory E7 strategies at 26 accepted objects (realized coverage 40.625%): uncertainty-ranked global selection, random selection, class-matched random selection and class-balanced uncertainty selection. The comparison was designed after E5 results were visible; simulated random selections are not independent test data.

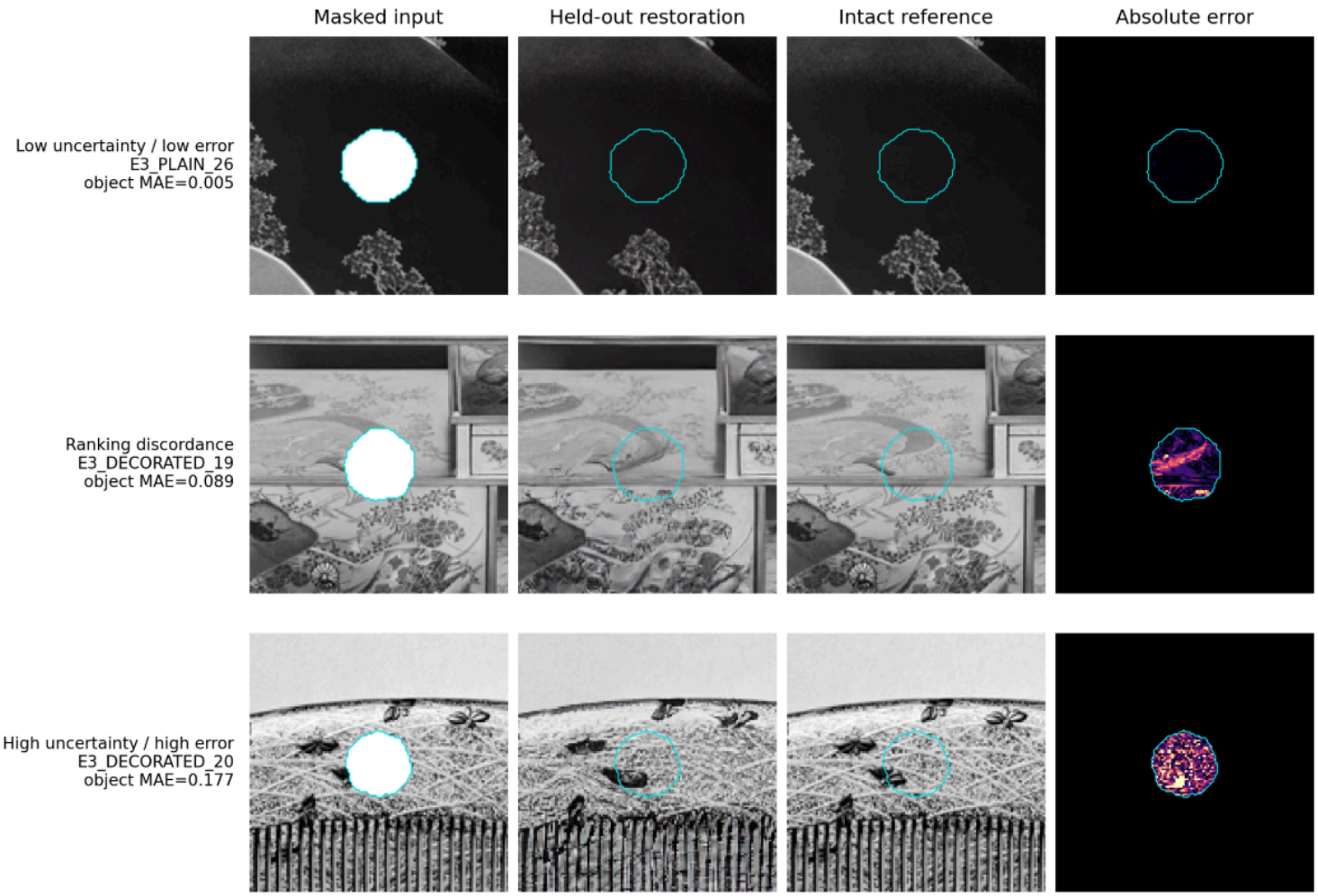


Figure 5

Illustrative medium-occlusion cases from the original cohort. Each panel shows the masked photographic input, an AI-generated held-out seed-11001 inpainting result, the intact photographic reference and mask-restricted absolute RGB error (fixed 0–0.4 display scale). The cyan contour indicates the synthetic missing region. Crops were chosen after analysis, and do not constitute new validation samples or documented historical restorations.

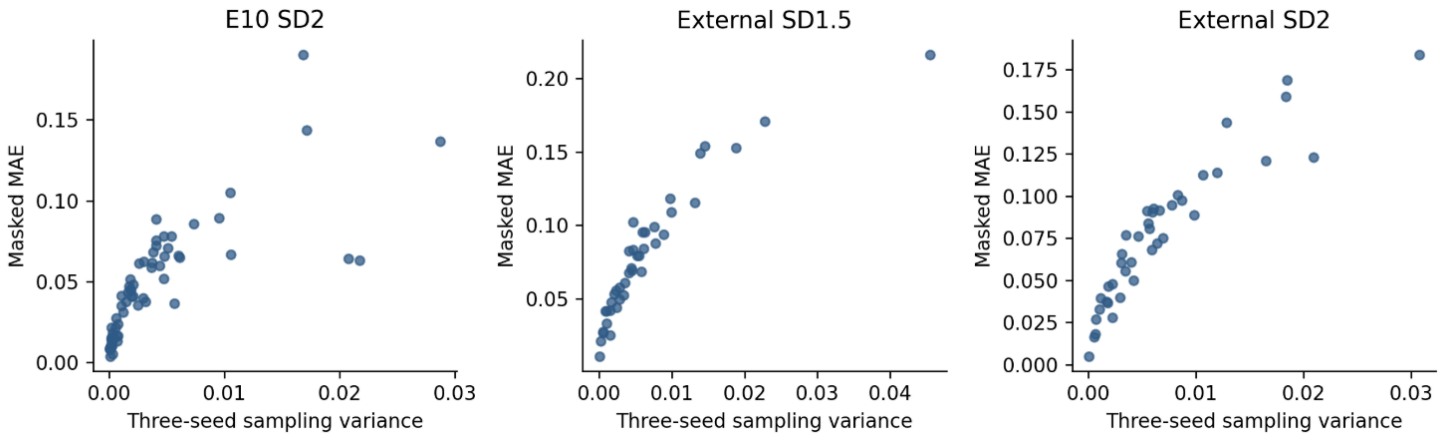


Figure 6

Seed-disjoint uncertainty and masked MAE for the second model on the original cohort and for both models on the external cohort. Each point averages three masks for one catalogue entry. External entries include one documented ensemble-component relationship; omission sensitivity is reported in Supplementary Table S16.

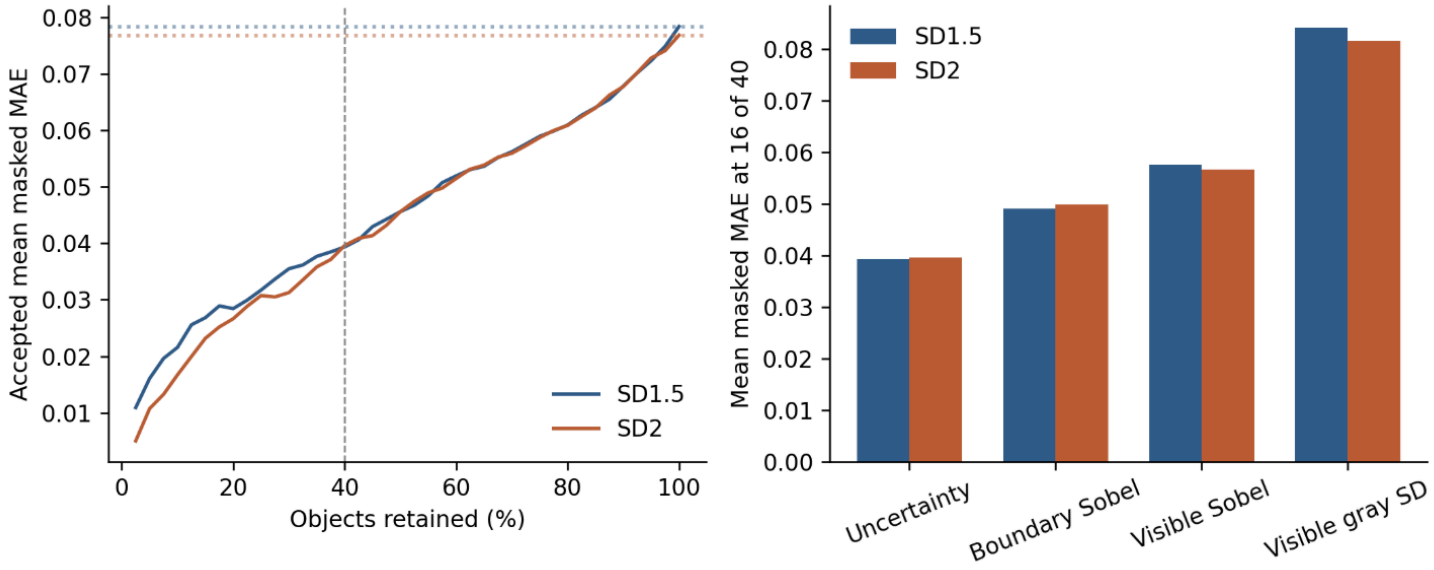


Figure 7

External-cohort selection results. Left, mean masked MAE as the lowest-uncertainty entries are retained; dotted lines show each full-cohort mean. Right, 16/40 accepted-set MAE under uncertainty and three visible-context descriptors. Boundary Sobel is the mean magnitude in a five-iteration dilation ring; it is not Canny edge density. These selectors change the accepted sample, not the individual generated images.

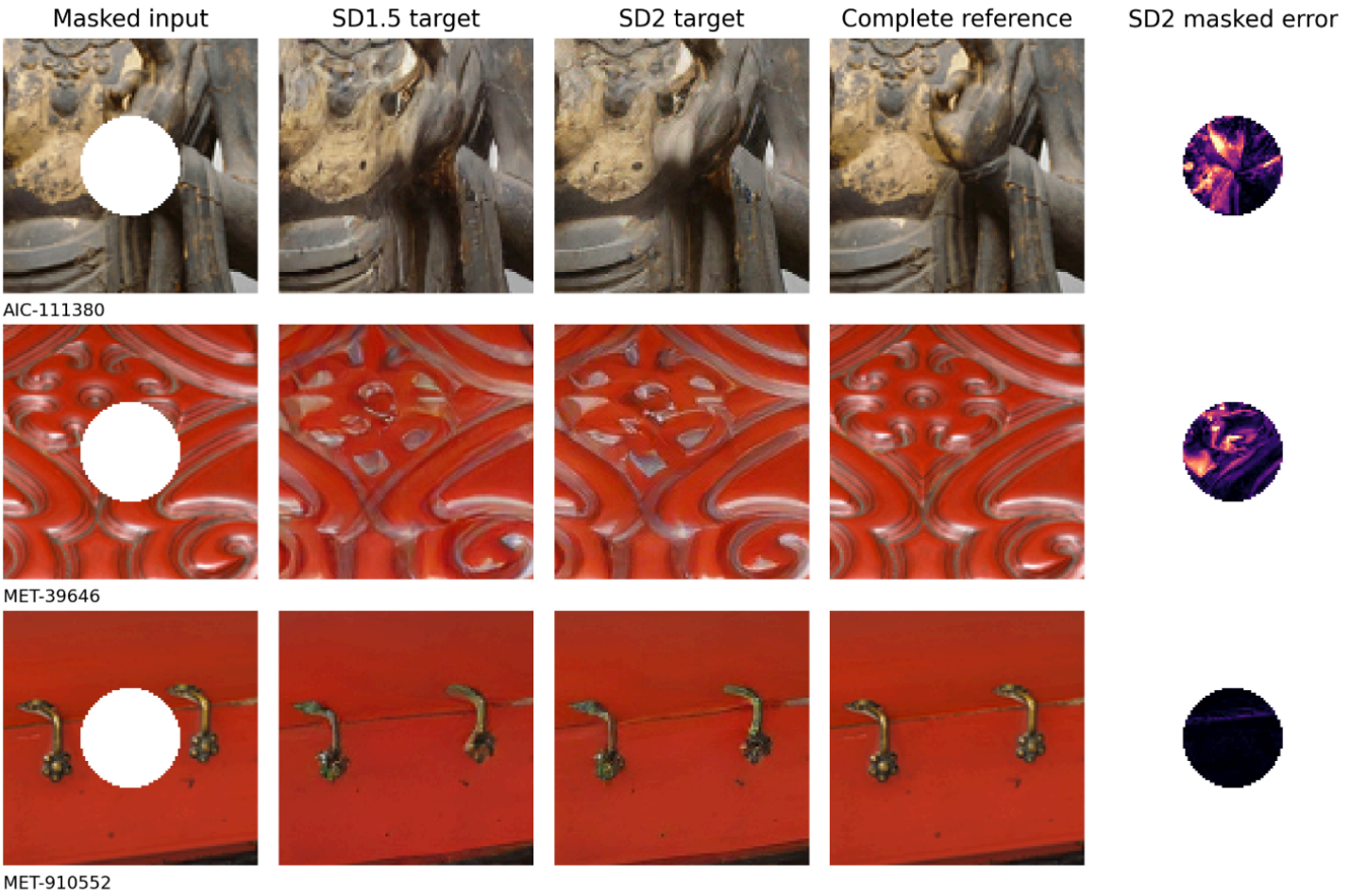


Figure 8

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf