Supplementary Information | Sampling variability and error ranking in lacquerware image inpainting

Additional protocols, numerical analyses, image provenance and two-expert foreground/material verification

Supplementary Note S1 — study lineage

Supplementary Table S1 | Study lineage and evidentiary scope.

Stage	Data products presently visible	Evidentiary limit
Pilot S1–S5	18 objects, 54 cases, 216 outputs reported in earlier local pilot; S5.1/S5.2 analyses	Distinct from the 64-object confirmation cohort; do not pool without documented design.
E3 confirmation	64 objects, plain/decorated classes, 192 cases, 4 seeds	Original E3 protocol lock received and hash-verified; internal timestamp precedes generation, without external preregistration.
E4 generation	768 files reported; validation status PASS in E5	768/768 image outputs, logs, and SHA-256 evidence independently inspected.
E5 confirmation	E5 report, 64-object metrics, seed rotations, bootstrap and selection summaries	Object-level measures independently recomputed from archived PNG outputs, references and masks; the reported E5 bootstrap intervals remain provenance-linked to the E5 analysis record.
E7 supplement	64-object post-hoc class-balanced and random comparators	Exploratory analyses designed after E5 results; not an external sample.

Supplementary Note S2 — units, score definitions and decision rule

The independent analysis unit was the lacquerware object (n=64). Each object contributed S/M/L occlusions and four seeded reconstructions. The main uncertainty score used the per-pixel RGB sample variance (ddof=1) across seeds 11002–11004, averaged within the binary occlusion mask; error was the masked RGB mean absolute error (MAE) for seed 11001 against the intact reference. Object scores aggregated across three cases. These definitions were checked against recovered E4/E5 source and original image outputs.

The original E3 protocol file has been independently checked against the E4/E5 archives. The prespecified primary success criterion was: object-level Pearson $r > 0$ with lower bound of the 10,000-replicate object-bootstrap 95% interval > 0.30; the nominal 40% selective-acceptance criterion required positive relative MAE reduction with bootstrap lower bound above zero. This was an internally timestamped protocol lock, not independently registered with an external preregistration service.

Supplementary Table S2 | Confirmatory estimates and verification status.

Analysis	Reported value	Status
Object Pearson r	0.7175933	Reproduced from supplied object CSV
Object Spearman ρ	0.9349359	Reproduced from supplied object CSV
Object bootstrap CI	0.558044–0.938146	Historical E5 interval retained; new independent bootstrap reported in Table S13
40% / 26-object accepted MAE	0.01810898	Recomputed from E5 object CSV
Full-cohort MAE	0.04654833	Recomputed from E5 object CSV
Relative reduction	61.096%	Recomputed from E5 object CSV
Class-matched random expected MAE	0.03258962	E7 simulation with 200,000 draws
Class-balanced 13+13 MAE	0.02639989	Recomputed from E5 object CSV
High-error objects (MAE >0.12)	3/64	Recomputed from object CSV

Supplementary Note S3 — rotation and class checks

Supplementary Table S3 | Seed-rotation sensitivity estimates.

Withheld seed	Pearson r (as available)	95% CI
11001	0.7176	[0.5580, 0.9381] (primary E5 CI)
11002	0.6852	[0.4651, 0.8994]
11003	0.7701	[0.6517, 0.8965]
11004	0.7760	[0.5420, 0.9340]

Statistical provenance note for Table S3: the archived rotation-only output listed [0.5602, 0.9380] for target 11001, whereas the prespecified primary E5 analysis lists [0.558044, 0.938146]. The table and revised Fig. 2 display the primary E5 CI for target 11001; both historical values are retained here for transparency. The reason for this small resampling discrepancy is unresolved. The other three rotation intervals remain as originally reported.

Four seed rotations share underlying generated images and should be interpreted as robustness of the same experiment. The seed-disjoint design addresses shared-output leakage in the primary orientation but does not establish external model or dataset transportability.

Supplementary Table S4 | Numerical object-level risk–coverage profile.

Nominal coverage	Accepted objects	Accepted MAE	Relative reduction
20%	13	0.01190	74.4%
40%	26	0.01811	61.1%
60%	39	0.02742	41.1%
80%	52	0.03611	22.4%
100%	64	0.04655	0.0%

Supplementary Note S4 — E7 exploratory composition controls

E7 compares four 26-object selection strategies: all-object random sampling, random sampling constrained to 22 plain and four decorated objects, the actual global uncertainty-ranked subset of 22 plain and four decorated, and class-balanced uncertainty ranking at 13 plain plus 13 decorated. Random baselines are simulation-derived expected means, not images newly restored by an alternative method. The comparison is post hoc and must remain explicitly labeled exploratory.

Supplementary Table S5 | Exploratory E7 object-selection strategies.

Strategy	Mean MAE	Class composition
Random 26 (expected)	0.04655	Unconstrained
Class-matched random (expected)	0.03259	22 plain / 4 decorated
Global uncertainty-ranked	0.01811	22 plain / 4 decorated
Balanced uncertainty-ranked	0.02640	13 plain / 13 decorated

E7 exploratory estimates: class-adjusted Pearson r=0.6613; stratified bootstrap relative-reduction 95% intervals 0.502–0.705 for global ranking and 0.324–0.510 for balanced ranking. The bootstrap reuses the same 64 objects and is not an independent verification of prospective deployment.

Supplementary Note S5 — interpretation of calibration

The risk-score endpoint used MAE >0.12 with only three positive events among 64 objects. Reported AUROC (0.9891) and Brier score (0.0236) cannot establish reliable class-wise calibration with so few events. The calibration slope 0.566 and intercept −1.784 should be presented as exploratory only. Neither the threshold nor the score constitutes a material conservation criterion.

Supplementary Note S6 — provenance and source documentation

All 64 source photographs match their registered SHA-256 hashes. Official accession and image URLs and frozen source metadata are retained for every object; 49 additional official rights-evidence JSON files are also archived. The responsible authors confirm the reuse conditions for the selected research images. The original object, image, and experiment identifiers remain unchanged.

The E3/E4 archive contains 768 separately named 512 × 512 PNG outputs, generation logs, protocol/model locks and a file-level SHA-256 manifest. The independent audit found 1,412 matched research files, 192 complete mask cases, and four seeded outputs per case. The E3 internal creation timestamp precedes E4 generation logs; this is evidence of the saved internal sequence, not third-party preregistration.

Supplementary Note S7 — data status

The E3/E4 original archive is retained intact and is a separate, large data asset. This supplementary document and the E8 writing package do not replace the original protocol files or generated images.

Supplementary Table S6 | Independent E3/E4 archive-integrity audit.

Audit measure	Independently verified result
Archive entries / manifest entries	1,415 / 1,412 (+ 3 archive administrative entries)
Manifest SHA-256 and length mismatches	0
Generated output files	768 / 768 unique SHA-256, all readable 512 × 512 PNG
Generation logs	768 success records, uniquely matching task IDs
Cases and seeds	192 / 192 cases, each with four locked seeds
Source images and source hashes	64/64 present and matching
Local rights-evidence JSON	49/64 present; 15 records have manifest/URL/CC0 status only
E5 object scores independently recomputed	64/64; max absolute Δ uncertainty $<2.5\times10^{-9}$; Δ MAE $<7.5\times10^{-9}$

Supplementary Note S8 — source image traceability

For 15 images without an additional local official-evidence JSON snapshot, the frozen object register retains the corresponding museum accession, object/image URL and recorded status. The study team has separately confirmed the reuse conditions; the archived accession and image URLs support traceability to the source records.

Supplementary Note S9 — Exploratory object-influence and retrospective oracle recheck

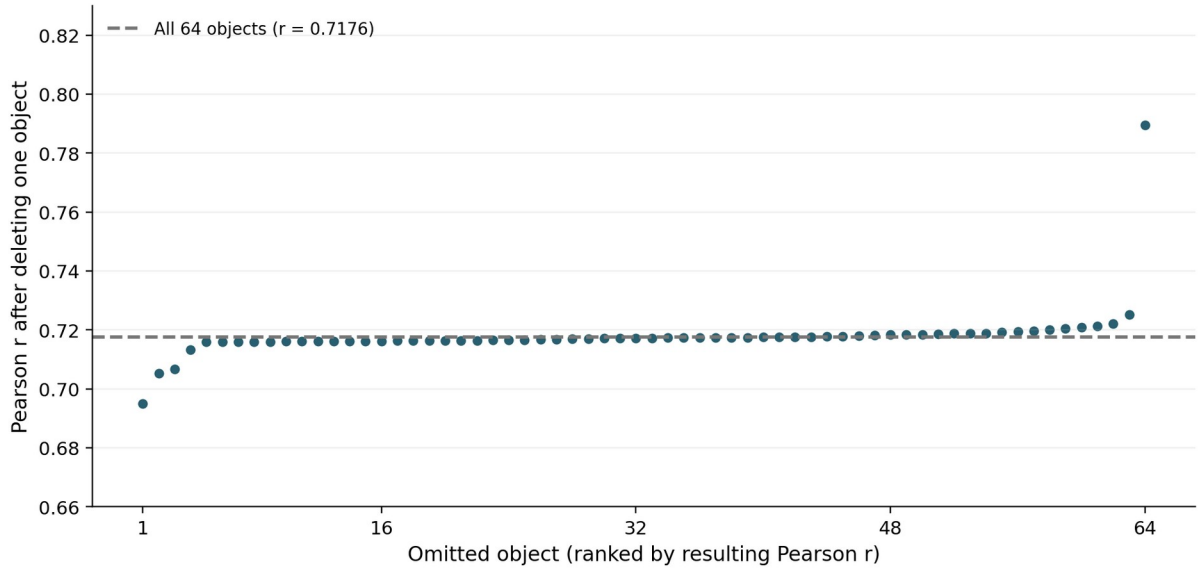
All analyses in this note were completed after the E5 confirmation-stage report. The input was the 64-row E7_2_RAW_IMAGE_RECOMPUTED_METRICS.csv from the archived E8 verification package, which directly reproduces object-level image-derived uncertainty and masked MAE. This is the original cohort, not an independent cohort. Methods and numerical results are explicitly exploratory.

Object influence was tested by dropping each entire object in turn and recomputing Pearson r and Spearman ρ across the remaining 63. A fresh 10,000-draw percentile bootstrap (NumPy generator seed 20261020) resampled independent objects, not their nested S/M/L cases. This new interval is shown separately from the original E5 interval and must not be substituted silently for the prespecified report.

Supplementary Table S7 | Exploratory robustness and object-influence sensitivity.

Analysis	Computed result
Independent original objects	64
Full-cohort Pearson r	0.7176
Full-cohort Spearman ρ	0.9349
Full-cohort Kendall τ	0.7778
Pearson r after excluding each single object	0.6950–0.7895
Spearman ρ after excluding each single object	0.9319–0.9400
Pearson r after omitting highest-error object	0.6950
Pearson r, new 10,000-resample percentile interval	[0.5590, 0.9366]
NumPy bootstrap generator seed	20261020
Interpretation	Post-hoc influence check; same 64-object cohort

Supplementary Fig. S1 | Leave-one-object-out Pearson association across 64 object deletions. Each point corresponds to a single omitted object; the dashed reference is the full-cohort Pearson r. The range quantifies single-object influence but does not constitute independent external validation.



Supplementary Table S8 | Object-level risk–coverage with retrospective oracle.

Selected objects	Realized coverage (%)	Uncertainty-ranked mean MAE	Oracle hindsight mean MAE
13	20.31	0.01190	0.00935
26	40.62	0.01811	0.01761
39	60.94	0.02742	0.02606
52	81.25	0.03611	0.03493
64	100.00	0.04655	0.04655

The oracle uses retrospective sorting by known masked MAE; the uncertainty-ranked strategy uses only the uncertainty score for sorting. Both are evaluated on identical objects and coverage counts. Neither the oracle nor the leave-one-out calculation uses new repaired images.

Supplementary Note S10 — Source-image complexity, museum subgroups, and structural quality

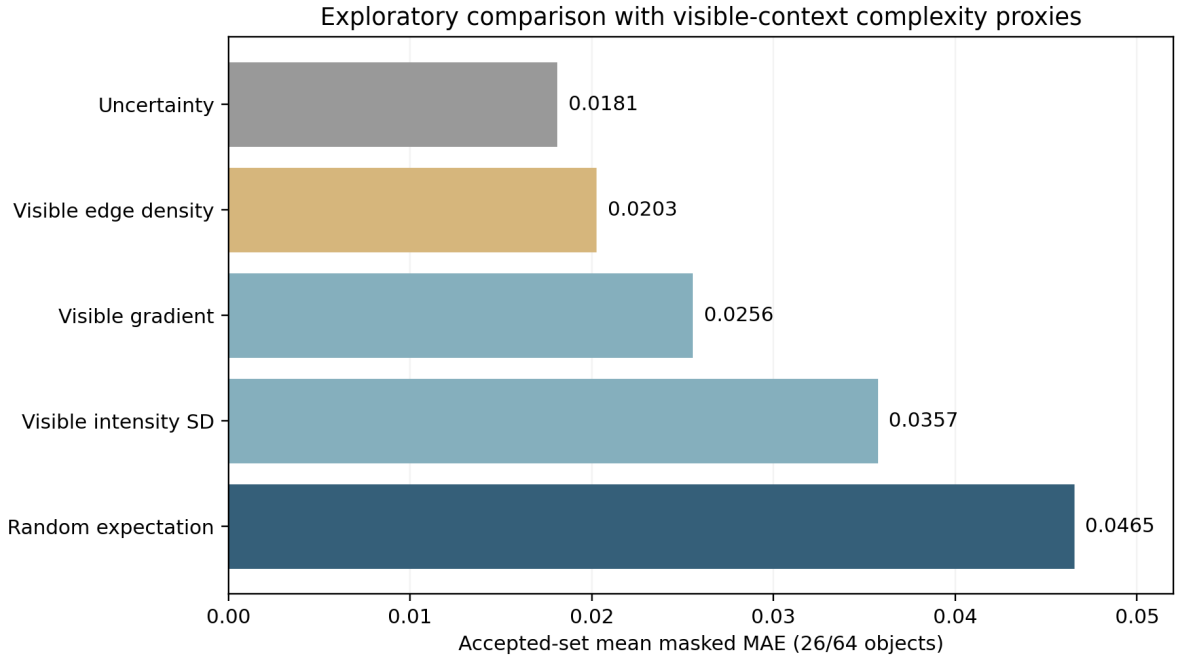
All analyses in this note are exploratory and use the previously supplied E3/E4 64-object original-image archive. Predictors for image complexity were extracted solely from visible pixels of each masked input. The withheld seed-11001 output was compared against the intact image only to compute the outcome. A single original object, not an occlusion case or seed, remains the inferential unit.

Supplementary Table S9 | Exploratory visible-context baseline performance at 26/64 object coverage.

Ranking score	Accepted MAE	Reduction vs full	Pearson r	Spearman ρ	Plain / 26
Three-seed uncertainty	0.01811	61.1%	0.7176	0.9349	22
Visible-context edge density	0.02026	56.5%	0.7830	0.7906	22
Visible-context Sobel gradient	0.02557	45.1%	0.7621	0.6532	18
Visible-context grayscale SD	0.03574	23.2%	0.2038	0.1674	16

For the 10,000-replicate paired object-bootstrap comparison of edge-density and uncertainty ranking, edge minus uncertainty accepted-set MAE = 0.00215, percentile 95% CI [−0.00005, 0.00651]. The interval contains zero. No ranking score was trained or tuned on an independent cohort. Full-cohort mean MAE = 0.04655.

Supplementary Fig. S2 | Exploratory 40.625%-coverage accepted-set mean MAE for three-seed uncertainty versus simple visible-context image descriptors. Random expectation equals full-cohort mean MAE; bars do not represent independent model test cohorts.



Supplementary Table S10 | Object-level uncertainty–MAE association by contributing museum.

Museum	n	Pearson r	Spearman ρ	Mean MAE
Walters Art Museum	40	0.6749	0.9435	0.04322
Cleveland Museum of Art	18	0.9147	0.9071	0.04330
Cooper Hewitt, Smithsonian Design Museum	6	0.8481	0.8286	0.07847

These associations reuse the full 64-object cohort split by accession provenance, not externally held-out or train/test museum groups. The Cooper Hewitt stratum has only six objects. Covariate-adjusted residual Pearson r = 0.3264 after simultaneous linear adjustment for museum, class, edge density, gradient and grayscale SD; exploratory and not a validated independent prediction coefficient.

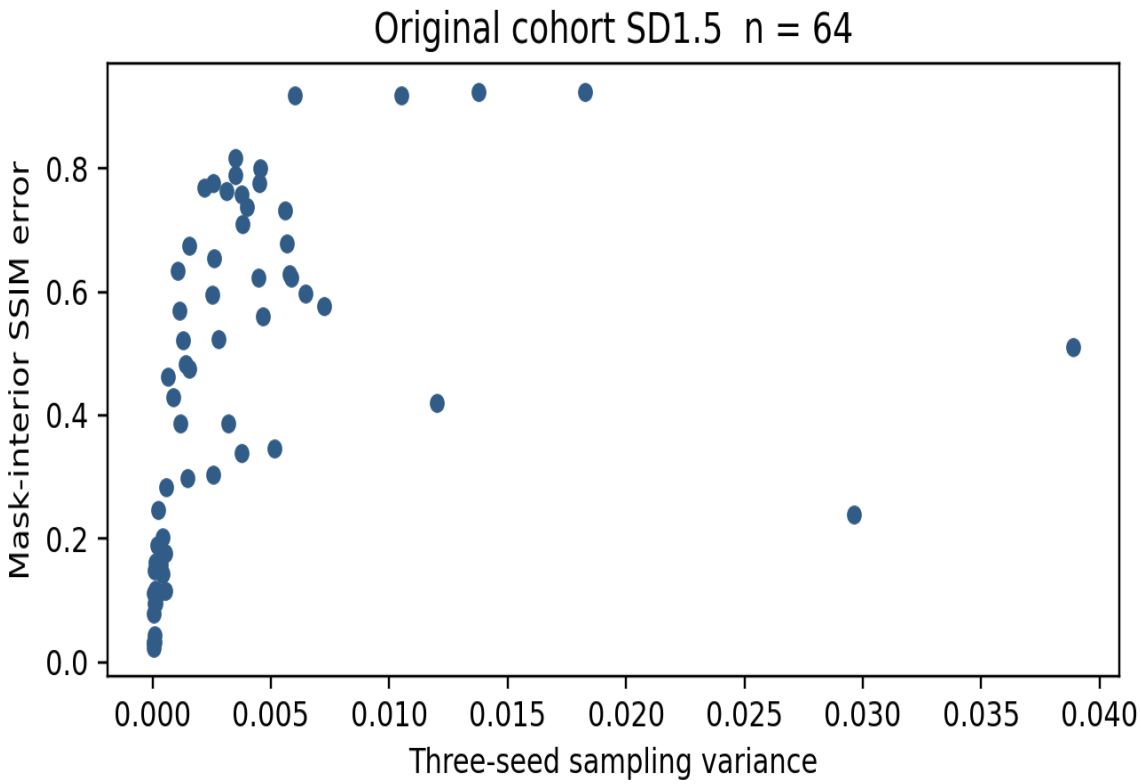
Supplementary Table S11 | Exploratory masked structural-similarity checks (S/M/L equal-weighted object means).

Statistic	Result
n objects / cases	64 / 192
Spearman (variance vs 1−SSIM)	0.7823
Pearson (variance vs 1−SSIM)	0.3323
Mean SSIM, all objects	0.5561
Mean SSIM, 26 lowest-uncertainty objects	0.8134
SSIM kernel / mask core	7 × 7 window / 7 × 7 square erosion

SSIM is now recalculated from the local RGB map (win_size 7, data_range 1), aggregated only over the mask after erosion by a 7 × 7 square. This explicit convention replaces the earlier erosion-dependent E10 summary (0.5582 overall). The new E10

mean is 0.5561, accepted mean 0.8134 and uncertainty versus SSIM-error Spearman correlation 0.7823. These secondary values do not alter primary U/MAE results.

Supplementary Fig. S3 | Object-level seed-disjoint uncertainty versus corrected mask-interior SSIM error across the 64 original photographed objects. Each point is one object; the trend is exploratory and is not an out-of-sample perceptual validation.



Reproducibility: the output artifact raw_192_case_visual_metrics.csv contains per-case baseline and SSIM metrics; raw_64_object_comparative_metrics.csv contains object-level aggregates, museum metadata and source-score comparisons. analyze_raw.py and evaluate.py encode the analysis, with no access to any new experiment beyond the archived E3/E4 images and metrics.

Supplementary Note S11 E11 design and procedural record

E11 adds 768 second-model outputs on the original 64 objects and 960 outputs on 40 external catalogue entries. The 72 preflight files were reused in the 768-output run; total new formal outputs are 1,728. Formal generation preceded completion of the subsequent two-expert external material/visible-class and ROI review. The archived generation-stage waivers are not retrospective registration. The original lock and full 40-entry results remain unchanged, while new blinded-to-method-status is NOT claimed for later expert classification. Archived object-ROI PNG files and overlap calculations are evaluated separately in Tables S25–S28.

The external circular masks have 489, 1,313 and 2,209 pixels (S/M/L): 0.1865%, 0.5009% and 0.8427% of a 512 × 512 canvas. These denominators differ from E10's independently segmented object-ROI fractions (approximately 7.5%, 20%, 35%). The 40-entry generation-stage lock retains its historical PROVISIONAL_PENDING_SECOND_REVIEW status. In a later documented two-expert review, 40 pairs of foreground ROIs and 40 material/ornament classifications were reconciled. Object-foreground pixel overlap was available for all 120 synthetic masks (S26). This does not establish a lacquer-layer-specific mask or equal damage area across experimental cohorts.

Supplementary Table S12 Image integrity and generation traceability

Run	Formal PNGs	Log records	Verification
E10 SD2	768	696 production + 72 preflight	All files decode 512 × 512 RGB; output/input/mask hashes match
External SD1.5	480	480	All expected combinations present; hashes match
External SD2	480	480	All expected combinations present; hashes match
Formal set total	1,728	1,728 unique tasks	1,728 unique file and decoded-pixel hashes
Preflight set	72	72	All hashes match formal copies; no additional sample size
Original E10 archive	768 outputs	1,412 manifest entries	All manifest hashes and lengths reproduced

Supplementary Table S13 Independent object level recalculation

Cohort model	n	Pearson r	Spearman rho	Full MAE	Accepted MAE
E10 SD1.5	64	0.7176	0.9349	0.04655	0.01811
E10 SD2	64	0.7664	0.9254	0.04679	0.01624
EXTERNAL SD1.5	40	0.9043	0.9698	0.07835	0.03934
EXTERNAL SD2	40	0.9162	0.9598	0.07680	0.03961

New percentile intervals use 10,000 entry bootstrap draws, seed 20261021, with selection repeated in each resample. These intervals are separate from archived E5 and E11 resampling outputs; small changes reflect the new Monte Carlo sequence. For 40 entries, physical-object independence is qualified by the ensemble/component pair.

Cohort model	Pearson 95% interval	Reduction	Reduction 95% interval
E10 SD1.5	[0.5605, 0.9367]	0.02844	[0.02247, 0.03453]
E10 SD2	[0.6293, 0.9125]	0.03055	[0.02371, 0.03649]
EXTERNAL SD1.5	[0.8802, 0.9653]	0.03901	[0.02779, 0.04835]
EXTERNAL SD2	[0.8781, 0.9570]	0.03720	[0.02784, 0.04674]

Supplementary Table S14 Baselines and permutation reanalysis

The archived Context_Edge feature is mean boundary-ring Sobel magnitude, not edge density. E10 original Canny baseline and E11 boundary Sobel are distinct calculations. Below, the new bootstrap seed is 20261021 and no family-wide secondary-baseline multiplicity adjustment is claimed.

External model	Visible descriptor	Accepted MAE	Difference	95% difference interval
SD1.5	Boundary Sobel	0.04908	0.00974	[0.00005, 0.01414]
SD1.5	Visible Sobel	0.05758	0.01824	[0.00285, 0.02919]
SD1.5	Visible gray SD	0.08422	0.04487	[0.02674, 0.06095]
SD2	Boundary Sobel	0.04981	0.01020	[0.00017, 0.01707]
SD2	Visible Sobel	0.05670	0.01709	[0.00286, 0.02920]
SD2	Visible gray SD	0.08163	0.04203	[0.02509, 0.05752]

Model	Permutation statistic	Unadjusted p	Holm p
SD1.5	Pearson two-sided permutation	0.000100	0.000400
SD1.5	Selection improvement one-sided permutation	0.000100	0.000400
SD2	Pearson two-sided permutation	0.000100	0.000400
SD2	Selection improvement one-sided permutation	0.000100	0.000400

Each permutation shuffles MAE across fixed entry-level predictors. Pearson uses an absolute two-sided statistic; selection uses the one-sided full-minus-accepted mean error. The plus-one p-value rule is (extreme+1)/(10000+1). The archived nonpositive fraction from ordinary effect bootstraps is not a null-calibrated test and is retained only as an archived calculation. The new tests are retrospective and assume exchangeable units; related-entry sensitivities are reported separately.

Supplementary Table S15 External entries and current official records

Official API records for all 40 catalogue entries were retrieved and locked accession/title fields matched; these support documentary identifiers but not the location of a lacquer-material boundary in the source photograph. The original cohort labels remain frozen as 20 plain and 20 decorated, and the lock still records provisional status. A later two-expert review classified the same 40 entries as 35 lacquer-led, four lacquer-as-component and one material-role unknown; visual labels were 35 decorated and five plain. Seventeen visible-class labels changed compared with the original frozen labels. Both expert classifications are reported in S25 and should not be substituted into prespecified E11 analyses without an explicit post-hoc designation. Official-source ambiguity and the ensemble/component pair still constrain material-specific generalization.

Entry	Accession	Official title	Origin
MET-39646	29.100.713	Box with pommel scroll	China
MET-39684	2015.500.1.74	Guri lacquer cup	China

Entry	Accession	Official title	Origin
MET-40189	2015.500.1.15a, b	Box with pommel-scroll design	China
MET-40192	2015.500.1.17a, b	Box with pommel scrolls	China
MET-40312	2015.500.2.18a, b	Pair of Negoro Lacquer Candlesticks	Japan
MET-40527	2015.500.3.5a, b	Stationery box decorated with peony scrolls	Korea
MET-448442	30.95.174.1a, b	Lacquer Covers of the Davis Album	Not specified
MET-452948	1979.460.2a, b	Lacquer Mirror Case	Not specified
MET-455286	2006.543.1	Qur'an Manuscript with Lacquer Binding	Not specified
MET-57797	91.1.658	Lacquer Comb with Flower-Shaped Roundels	Japan
MET-58436	1982.244.5	Lacquer Comb with Geese and Autumn Grasses	Japan
MET-58437	1982.244.6	Lacquer Comb with Maple Leaves	Japan
MET-58438	1982.244.7a, b	Lacquer Hairpin with Geese	Japan
MET-58439	1982.244.8a, b	Lacquer Hairpin with Flowers	Japan
MET-62027	2001.535	Lozenge-Shaped Tray	China
MET-648051	2014.257a, b	Lacquer Pen Box	Not specified
MET-648056	2014.258a, b	Lacquer Pen Box with Royal Audience Scenes	Not specified
MET-900380	2023.502a, b	Lacquer Pen Box with Royal Audience Scenes	Not specified
MET-910551	2024.336a–c	Black lacquer box with phoenix designs	Korea
MET-910552	2024.337a, b	Red lacquer box for an inkstone	Korea
AIC-25658	1966.331	Hot Water Pot	Japan
AIC-111380	1962.356	Seated Bodhisattva	Japan
AIC-226206	1961.60.4	Dish with Images of Ancient Bronzes	China
AIC-36691	1971.510	Covered Box with Butterflies, Gourds, and Scrolling Vines	China
AIC-45749	1973.639	Ritual Wash Basin	Japan
AIC-134	1888.165	Lacquer Tea Bowl 乾隆禦制雕漆盞	China
AIC-48617	1943.975.3	Inro Decorated with the Animals of the Zodiac	Japan
AIC-88258	1938.476	Book Binding	Iran
AIC-147048	1997.403.5	Fitted Box	China

Entry	Accession	Official title	Origin
AIC-12922	1961.56	Box with Boating Scene	China
AIC-13064	1961.105	Carved Lacquer Box	China
AIC-25700	1937.975	Carved Lacquer Boxes	China
AIC-147050	1997.406	Cinnabar Lacquer 'Scholar in Landscape' Box and Cover	China
AIC-49631	1975.657	Lacquer Saddle	China
AIC-12945	1961.67	Stem Cup with Leechee and Vines	China
AIC-88255	1938.475	Book with lacquer covers	Iran
AIC-262429	2021.447	Writing Brush and Cover	China
AIC-57791	1900.1170.3	Inro Case Depicting the Seven Sages of the Bamboo Grove	Japan
AIC-86912	1961.625-627	Enthroned Buddha with Elephant and Monkey Paying Homage	Thailand
AIC-86913	1961.626	Kneeling Elephant	Thailand

AIC-86912 is catalogued as Enthroned Buddha with Elephant and Monkey Paying Homage (1961.625–627), while AIC-86913 is Kneeling Elephant (1961.626). This is an ensemble/component relationship. Six Met records leave the culture field empty; that field is reported as unspecified rather than inferred from titles. The cohort includes lacquered bindings, sculpture, accessories and containers; it is not a homogeneous Chinese-lacquerware collection.

Supplementary Table S16 External sensitivity analyses

Model	Subset	n	Pearson r	Spearman rho
SD1.5	All locked	40	0.9043	0.9698
SD1.5	Exclude AIC-86913	39	0.9032	0.9688
SD1.5	Exclude AIC-86912,AIC-86913	38	0.9105	0.9779
SD1.5	Museum: Art Institute of Chicago	20	0.9163	0.9414
SD1.5	Museum: Met	20	0.9511	0.9714
SD1.5	Visual_Class: decorated	20	0.9215	0.9669
SD1.5	Visual_Class: plain	20	0.8872	0.9504
SD2	All locked	40	0.9162	0.9598
SD2	Exclude AIC-86913	39	0.9144	0.9585
SD2	Exclude AIC-86912,AIC-86913	38	0.9163	0.9591
SD2	Museum: Art Institute of Chicago	20	0.9393	0.9489
SD2	Museum: Met	20	0.9080	0.9624
SD2	Visual_Class: decorated	20	0.9110	0.9173
SD2	Visual_Class: plain	20	0.8974	0.9534

Excluding only the component leaves n=39 with round-half-up coverage 16/39; excluding both linked entries leaves n=38 with coverage 15/38. The locked-label subgroup analyses remain descriptive and must not be presented as expert-verified class analyses. A separate retrospective two-expert visual-class/material categorization and object-foreground annotation is reported in Tables S25–S28. Museum strata remain subsets of the same evaluation cohort.

Supplementary Table S17 Structural perceptual and boundary metrics

Cohort model	Interior SSIM	Composite SSIM	LPIPS crop	Boundary color	Boundary gradient
E10 SD1.5	0.5561	0.9968	0.1452	0.0431	0.1326
E10 SD2	0.5492	0.9968	0.1432	0.0430	0.1319
EXTERNAL SD1.5	0.4245	0.9969	0.1653	0.0706	0.2190
EXTERNAL SD2	0.4167	0.9969	0.1681	0.0691	0.2163

Corrected interior SSIM averages the local RGB map only in a mask eroded by a 7 × 7 square. The archived Masked_SSIM column instead computed a whole-image scalar after copying intact reference pixels into the visible region; values around 0.997 therefore cannot describe missing-region quality. Both quantities are shown to preserve the audit trail. LPIPS-Alex v0.1 uses a fixed 32-pixel-padded crop resized to 256 × 256. All 624 LPIPS cases were independently rerun on CPU; maximum difference from archived values was 1.79×10^{-7}. AlexNet checkpoint SHA-256: 7be5be791159472b1fbf3c69796f7cb30dca7ad8466c2df70058c37116cdee02.

Supplementary Table S18 Anonymous rating table audit

Two anonymized worksheets each contain 80 complete ratings for the same 80 model-output items (two model versions for each of 40 museum entries). The recovered Blind_Key.csv and all 160 original Stage 1/Stage 2 review panels agree with their indexed IDs and SHA-256 digests. The research team affirmed reviewer qualifications and ethics applicability; image-to-score traceability is not evidence of professional consensus. Four visual judgments yielded ordinal α = 0.230–0.287, and the further-human-review suitability item was identical for both raters, with 63/80 responses in category 3 for each. Original marginals are in Table S24; cluster-aware α and quadratic-weighted κ intervals, signed differences, and within-rater item redundancy are reported in the new Supplementary Note S20 and Tables S32–S34. Original ratings and primary image-model outcomes were not changed.

Rating dimension	Ordinal alpha	Exact agreement	Pairs
pattern_continuity	0.2299	56.25%	80
texture_consistency	0.2712	56.25%	80
suspicious_artifact	0.2872	57.50%	80
further_review	1.0000	100.00%	80
reference_consistency	0.2714	57.50%	80

Composite	SD2 minus SD1.5	95% interval	Wilcoxon p
Stage1	-0.0563	[-0.1750, 0.0563]	0.5638
Stage2	-0.0250	[-0.1625, 0.1250]	0.5431

Two-rater scores were averaged within each anonymous image and paired across SD1.5 and SD2 for the same 40 entries; two-sided zsplit Wilcoxon tests and the original 10,000 object-pair bootstrap draws reproduce the archived results. Neither stage yielded a significant between-model difference (Stage 1 p = 0.5638; Stage 2 p = 0.5431), which is not evidence of model equivalence. The Stage 1 four-item exploratory mean includes suitability for further human review, whereas Stage 2 asks reference consistency; no identical item was rerated across stages. Low ordinal α on visual judgments, systematic rater-score differences and apparent dimensional redundancy limit the evidentiary strength of these exploratory assessments (Table S24; Supplementary Tables S32–S34).

Supplementary Note S12 Recalculation and evidence availability

The E10/E11 computational verification bundle includes SHA-256 and decoded-pixel checks, expected-file and generation-log matching, 624 image-derived case rows, 208 model-by-entry aggregates, independent LPIPS, 40 official record snapshots, permutation statistics, ensemble-exclusion sensitivities and 160 rating-text linkage rows. Primary U and MAE agree with archived case values within 1×10^{-15}; additional grayscale/Sobel differences are below 2×10^{-7}. The independently retrieved blind-review evidence adds the original 80-entry administrative key, 80 Stage 1 panels, 80 Stage 2 panels, two reviewer worksheets and an image index; 160 panel hashes, 160 PNG decodes, 80 index/key entries and 80 mapping comparisons passed. Image-level cases and repeated seeds are nested measurements, not 624 independent heritage objects. The original 64 entries and external 40 entries are distinct cohorts and were not pooled for a single effect estimate.

Supplementary Note S13 — Matched low-cost selectors and retrospective coverage diagnostics

The following analyses were conducted after the E10/E11 confirmatory endpoints were fixed. The four model–cohort groups each contain independent object/catalogue-entry aggregates, not independent masked patches or seeds. All selectors were ranked ascending, without tuning their direction on held-out data. The E10 original Canny baseline differs from the shared boundary-Sobel descriptor used below. The AURC value here is the unweighted mean of accepted-set MAE over integer acceptance counts from ceil(0.25 × N) through N; smaller values indicate lower retrospective error, not certified deployment safety. Exact file-level source values are retained in the accompanying CSV package.

Supplementary Table S19 | Unified accepted-set error and average risk across coverages

Cohort	Model	Selector	n accepted	MAE at ~40%	Mean risk 25–100%
E10	SD1.5	U	26	0.01811	0.02798
E10	SD1.5	Context_Edge	26	0.01889	0.02909
E10	SD1.5	Visible_Mean_Sobel	26	0.03706	0.04023
E10	SD1.5	Visible_Gray_SD	26	0.06107	0.05571
E10	SD2	U	26	0.01624	0.02726
E10	SD2	Context_Edge	26	0.01669	0.02798
E10	SD2	Visible_Mean_Sobel	26	0.03695	0.04058
E10	SD2	Visible_Gray_SD	26	0.06256	0.05677
EXTERNAL	SD1.5	U	16	0.03934	0.05272
EXTERNAL	SD1.5	Context_Edge	16	0.04908	0.05605
EXTERNAL	SD1.5	Visible_Mean_Sobel	16	0.05758	0.06270
EXTERNAL	SD1.5	Visible_Gray_SD	16	0.08422	0.08266
EXTERNAL	SD2	U	16	0.03961	0.05213
EXTERNAL	SD2	Context_Edge	16	0.04981	0.05620
EXTERNAL	SD2	Visible_Mean_Sobel	16	0.05670	0.06191
EXTERNAL	SD2	Visible_Gray_SD	16	0.08163	0.07988

Differences are reported as baseline minus uncertainty. All confidence limits were derived by reselecting objects within each paired object-bootstrap replicate, not by independently resampling outcomes for each method. Only the best matched low-cost descriptor (boundary Sobel) is highlighted in the main text; the full table below covers all three descriptors. The 2,000-draw intervals reflect a new exploratory analysis and should not be substituted for the E5 predeclared primary intervals.

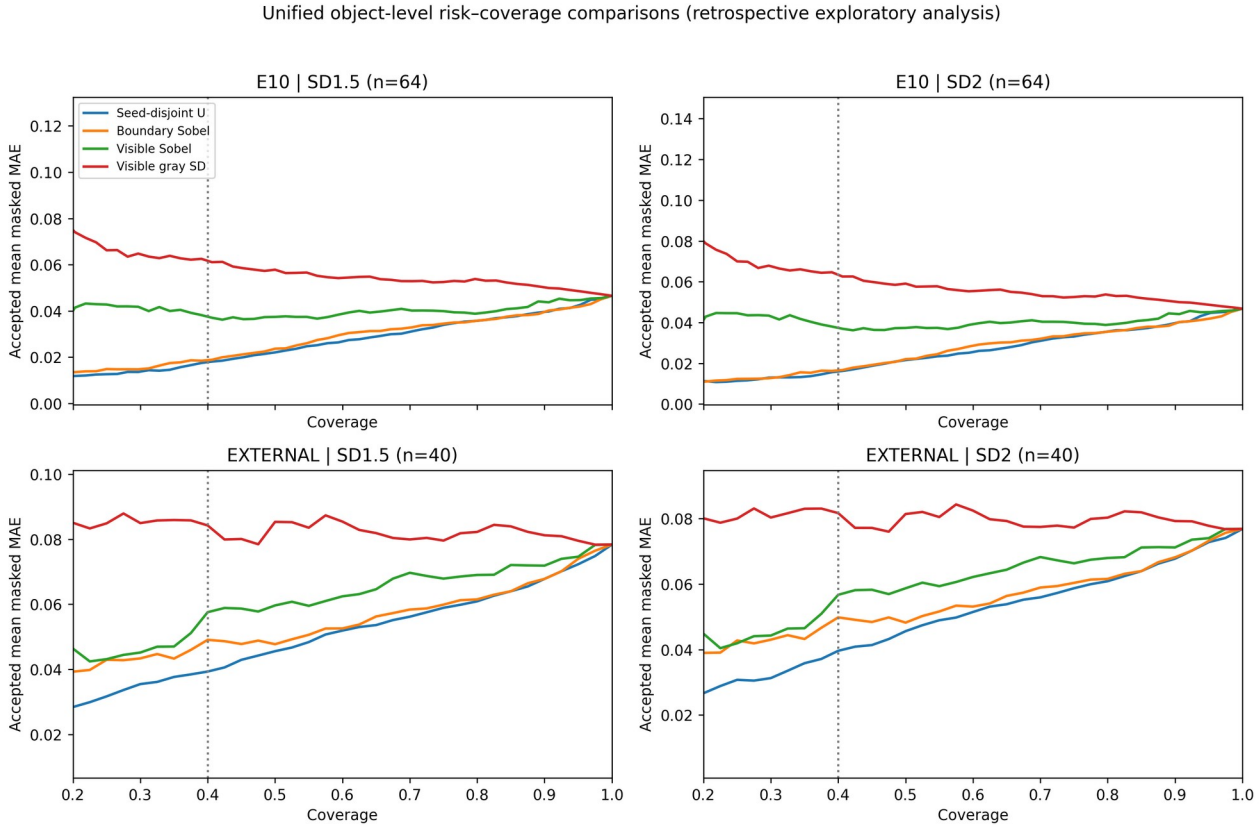
Supplementary Table S20 | Paired object-bootstrap differences for matched selectors

Cohort	Model	Baseline minus U	ΔMAE	95% lower	95% upper
E10	SD1.5	Context_Edge minus U	0.00078	-0.00013	0.00386
E10	SD1.5	Visible_Mean_Sobel minus U	0.01895	0.01070	0.03038

Cohort	Model	Baseline minus U	ΔMAE	95% lower	95% upper
E10	SD1.5	Visible_Gray_SD minus U	0.04296	0.02795	0.05874
E10	SD2	Context_Edge minus U	0.00045	-0.00061	0.00302
E10	SD2	Visible_Mean_Sobel minus U	0.02071	0.01127	0.03330
E10	SD2	Visible_Gray_SD minus U	0.04632	0.02976	0.06328
EXTERNAL	SD1.5	Context_Edge minus U	0.00974	0.00005	0.01444
EXTERNAL	SD1.5	Visible_Mean_Sobel minus U	0.01824	0.00291	0.02977
EXTERNAL	SD1.5	Visible_Gray_SD minus U	0.04487	0.02676	0.06096
EXTERNAL	SD2	Context_Edge minus U	0.01020	-0.00000	0.01780
EXTERNAL	SD2	Visible_Mean_Sobel minus U	0.01709	0.00244	0.02851
EXTERNAL	SD2	Visible_Gray_SD minus U	0.04203	0.02508	0.05736

Post-hoc rank residualization against four simultaneously fitted descriptors (boundary Sobel, visible Sobel, gray SD and mask-pixel count) yielded partial rank correlations of 0.7936 and 0.6818 in E10, and 0.9004 and 0.8913 externally (SD1.5 and SD2 respectively). These are rank-based partial correlations, NOT the earlier adjusted residual Pearson coefficient of 0.3264 (which additionally controlled museum and original visual class). The estimands and covariate sets differ and should not be directly compared or interpreted as causal attribution; see the CSV audit for 500-resample sensitivity intervals.

Supplementary Fig. S4 | Retrospective object-level risk–coverage profiles for U, five-iteration boundary Sobel, visible-region Sobel and grayscale SD across the original and external cohorts in both diffusion models. Each point at k/N coverage is the mean masked MAE among the k lowest-ranked entries by that descriptor. Dotted vertical line indicates nominal 40% coverage; no threshold is trained for unseen images.



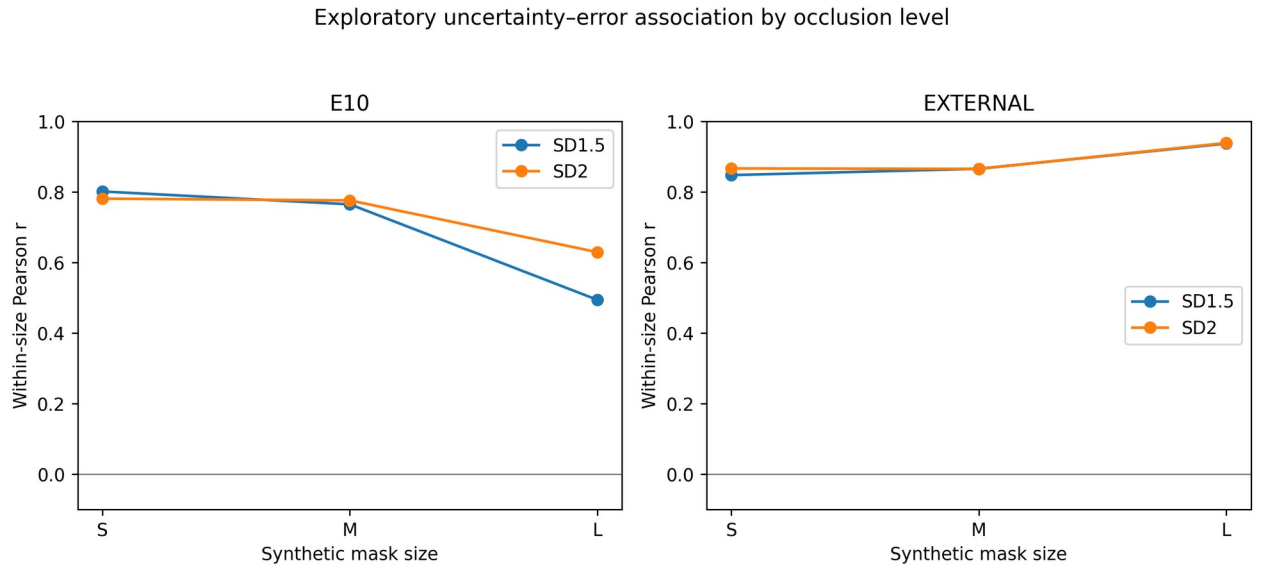
Supplementary Note S14 — Mask-area non-equivalence, visual-screen sensitivity and heritage material scope

E11 circular S/M/L masks contain 489, 1,313 and 2,209 pixels per 262,144-pixel canvas. The original images and masks are preserved in the external archive. Following the frozen generation and analyses, two experts reviewed the visible object foreground and category labels for 40 catalogue entries. Their two ROI passes generated 80 binary PNGs, all 512 × 512 and strictly binary; 120 pixel intersections with the original synthetic masks matched the submitted workbook without error. Using the second-pass ROI chosen in the final adjudication register, 101 masks had ≥80% foreground overlap, 11 had ≥50% but <80%, and eight <50%. Thirty-three catalogue entries passed the ≥80% criterion for all S/M/L masks (34 for M alone). Foreground and lacquer layers cannot be conflated. The old five visually flagged cases remain a historical qualitative screen, not an alternative primary inclusion rule. Figure S6 and Tables S25–S28 present the verification and exploratory sensitivities. A dated joint boundary-adjudication record subsequently confirmed the second-pass masks for MET-62027 and MET-910552 on the grounds that the first pass had included extensive non-object background; all published overlap categories and sensitivity subsets remain unchanged.

Supplementary Table S21 | S/M/L mask-specific uncertainty–error associations

Cohort	Model	Mask	n	Pearson r	Spearman ρ	Mean pixels
E10	SD1.5	L	64	0.49425	0.94341	2650.75
E10	SD1.5	M	64	0.76475	0.92491	1514.78125
E10	SD1.5	S	64	0.80107	0.92532	568.3125
E10	SD2	L	64	0.62929	0.91323	2650.75
E10	SD2	M	64	0.77575	0.94272	1514.78125
E10	SD2	S	64	0.78083	0.93539	568.3125
EXTERNAL	SD1.5	L	40	0.93681	0.92852	2209.0
EXTERNAL	SD1.5	M	40	0.86540	0.96041	1313.0
EXTERNAL	SD1.5	S	40	0.84743	0.95478	489.0
EXTERNAL	SD2	L	40	0.93852	0.93902	2209.0
EXTERNAL	SD2	M	40	0.86503	0.92833	1313.0
EXTERNAL	SD2	S	40	0.86636	0.92477	489.0

Supplementary Fig. S5 | Exploratory masked error association by S/M/L size, using one observation per object within each mask category. The E10 and E11 masks differ in their foreground-ROI denominators and must not be interpreted as the same physical damage size.



Supplementary Table S22 | External contact-sheet flag exclusion sensitivity (post hoc)

Model	Scenario	n	Selected	Selector	Selected MAE	Overall r
SD1.5	All40	40	16	U	0.03934	0.90433
SD1.5	All40	40	16	Context_Edge	0.04908	0.90433
SD1.5	Exclude5†	35	14	U	0.04730	0.90506
SD1.5	Exclude5†	35	14	Context_Edge	0.05060	0.90506
SD2	All40	40	16	U	0.03961	0.91616
SD2	All40	40	16	Context_Edge	0.04981	0.91616
SD2	Exclude5†	35	14	U	0.04724	0.91496
SD2	Exclude5†	35	14	Context_Edge	0.05215	0.91496

† Exclude5 removes five objects identified by earlier retrospective contact-sheet screening, not a prespecified eligibility criterion. The full 40 entries remain the primary E11 cohort. The subsequent two-expert ROI assessment now provides pixel-accurate *object-foreground* intersection, but it does not establish lacquer-layer-specific segmentation. For the prespecified endpoints no observations were removed retrospectively; the 33-entry mask-overlap and 35-entry lacquer-led subsets are exploratory sensitivity analyses shown in Table S27.

Supplementary Table S23 | Keyword-only material tags for external audit (not verified)

Provisional tag	n
lacquer-led or catalogued lacquer object (provisional)	21
wood/hard-substrate lacquer combination	10
book/cover/binding — mixed media	4
metal sculpture — lacquered/finished	3
papier-mâché / mixed media	2

The material tags in S23 were generated earlier from accession text by keyword rules; they are NOT the later two-expert material-role assessment in S25. The latter yielded 35 lacquer-led, four lacquer-as-component, one unknown. Retain all archived frozen labels and their original provisional statuses for reproducibility. The five plain-looking post-hoc labels are too few for reliable inferential plain-versus-decorated comparisons.

Supplementary Note S15 — Marginal rating distributions and agreement interpretation

Two raters applied ordinal 0–3 scores to each of 80 anonymous items. The following table exposes category marginals and exact matches rather than using pooled summary α as a substitute for detailed measurement. The unusually high exact agreement in the further-review item is partly influenced by skew: both raters gave score 3 to the same 63 items. The remaining visual items show different category usage by rater and do not warrant a claim of expert equivalence between generated models.

Supplementary Table S24 | Raw expert-rating marginals and exact agreement (n=80 pairs)

Criterion	Agreement	A=1	A=2	A=3	B=1	B=2	B=3
Pattern continuity	0.562	3	14	63	3	47	30
Texture consistency	0.562	5	12	63	3	46	30
Suspicious artifact	0.575	5	12	63	4	45	30
Further review	1.000	1	16	63	1	16	63

Criterion	Agreement	A=1	A=2	A=3	B=1	B=2	B=3
Reference consistency	0.575	3	13	63	4	45	30

Supplementary Note S16 — Post-hoc two-expert object-foreground and material adjudication

The research team confirmed that two relevant specialists performed the 40-object material, ornament and ROI reviews. The uploaded source mask archive contains 40 expert-A masks in roi_masks/ and 40 expert-B masks in roi_masks_second_pass/, each 512 × 512 and taking only pixel intensities 0/255. The retained final object ROI file recorded in the supplied adjudication workbook is the second-pass mask. Both recorded sets of visual class/material judgments agreed on 40/40 entries, as did all S/M/L 80%/50% placement categories. Nevertheless, continuous binary-boundary agreement differed for three objects (Table S28); no independent signature/audit of expert recruitment is claimed here. The expert-reviewed foreground is an anatomical object/image-silhouette annotation, NOT an analytically verified lacquer coating boundary. Archive original generated-mask and outcome data unchanged.

For mask M and object foreground R, on-object overlap was $|M \cap R| / |M| \times 100\%$. The original external masks each contain 489, 1,313 or 2,209 active pixels for S/M/L. The second-pass ROI cross-tabulation is included in Table S26, and all 120 numeric intersections, both reviewers' overlap proportions and the retained masks are recorded in the accompanying CSV archives. Two empirical thresholds (≥80% and ≥50% foreground overlap) were applied in a post-hoc sensitivity analysis. The nominal 40% accepted-set cut-off was implemented as floor(0.40 n + 0.5) entries in each subset; ties sorted by object identifier. Percentile 95% intervals in Table S27 come from 3,500 object-entry bootstrap resamples with within-resample reranking, seed 20261021, and describe selection uncertainty within these observational subsets. They do not imply that the post-hoc data filtering was prospective or blinded.

Supplementary Table S25 | Two-expert object material and ornament reclassification (post hoc)

Object ID	Frozen class	Expert A material	Expert B material	Consensus visual
AIC-111380	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-12922	plain	Lacquer-led	Lacquer-led	Decorated
AIC-12945	plain	Lacquer-led	Lacquer-led	Decorated
AIC-13064	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-134	plain	Lacquer-led	Lacquer-led	Decorated
AIC-147048	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-147050	plain	Lacquer-led	Lacquer-led	Decorated
AIC-226206	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-25658	decorated	Lacquer-led	Lacquer-led	Plain
AIC-25700	plain	Lacquer-led	Lacquer-led	Decorated
AIC-262429	plain	Lacquer-led	Lacquer-led	Decorated
AIC-36691	plain	Lacquer-led	Lacquer-led	Decorated
AIC-45749	plain	Lacquer-led	Lacquer-led	Plain
AIC-48617	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-49631	plain	Lacquer-led	Lacquer-led	Decorated
AIC-57791	decorated	Lacquer-led	Lacquer-led	Decorated
AIC-86912	plain	Lacquer as component	Lacquer as component	Decorated
AIC-86913	decorated	Lacquer as component	Lacquer as component	Decorated
AIC-88255	decorated	Lacquer as component	Lacquer as component	Decorated
AIC-88258	decorated	Lacquer-led	Lacquer-led	Decorated
MET-39646	decorated	Lacquer-led	Lacquer-led	Decorated
MET-39684	plain	Lacquer-led	Lacquer-led	Decorated
MET-40189	plain	Lacquer-led	Lacquer-led	Decorated
MET-40192	plain	Lacquer-led	Lacquer-led	Decorated
MET-40312	plain	Lacquer-led	Lacquer-led	Plain
MET-40527	decorated	Lacquer-led	Lacquer-led	Decorated
MET-448442	decorated	Lacquer-led	Lacquer-led	Decorated
MET-452948	decorated	Lacquer-led	Lacquer-led	Decorated
MET-455286	decorated	Lacquer as component	Lacquer as component	Decorated
MET-57797	decorated	Lacquer-led	Lacquer-led	Decorated
MET-58436	decorated	Lacquer-led	Lacquer-led	Decorated
MET-58437	plain	Lacquer-led	Lacquer-led	Decorated
MET-58438	plain	Lacquer-led	Lacquer-led	Decorated
MET-58439	plain	Lacquer-led	Lacquer-led	Decorated
MET-62027	plain	Lacquer-led	Lacquer-led	Plain
MET-648051	plain	Lacquer-led	Lacquer-led	Decorated
MET-648056	decorated	Lacquer-led	Lacquer-led	Decorated
MET-900380	decorated	Unknown	Unknown	Decorated
MET-910551	decorated	Lacquer-led	Lacquer-led	Decorated
MET-910552	plain	Lacquer-led	Lacquer-led	Plain

Expert A/B material and visible-ornament categories each agree in all 40 entries according to the submitted worksheet. The object-role counts are 35 lacquer-led, four lacquer-as-component and one unknown; 17 visual classifications changed relative to the frozen external lock. Original labels are retained and are not rewritten retrospectively.

Supplementary Table S26 | Expert foreground ROI intersection for S/M/L masks (second-pass retained)

Object ID	S %	M %	L %	Expert A/B IoU %	All S/M/L ≥80%
AIC-111380	100.0	100.0	100.0	100.00	Yes
AIC-12922	100.0	100.0	100.0	100.00	Yes
AIC-12945	100.0	100.0	99.8	100.00	Yes
AIC-13064	100.0	100.0	100.0	100.00	Yes
AIC-134	100.0	100.0	100.0	100.00	Yes
AIC-147048	100.0	100.0	100.0	96.83	Yes
AIC-147050	100.0	100.0	100.0	100.00	Yes
AIC-226206	100.0	100.0	100.0	100.00	Yes
AIC-25658	100.0	100.0	100.0	100.00	Yes
AIC-25700	79.1	66.3	61.2	100.00	No
AIC-262429	100.0	100.0	97.7	100.00	Yes
AIC-36691	100.0	100.0	100.0	100.00	Yes
AIC-45749	100.0	100.0	100.0	100.00	Yes
AIC-48617	100.0	100.0	100.0	100.00	Yes
AIC-49631	100.0	100.0	100.0	100.00	Yes
AIC-57791	100.0	100.0	100.0	100.00	Yes
AIC-86912	100.0	100.0	100.0	100.00	Yes
AIC-86913	100.0	100.0	100.0	100.00	Yes
AIC-88255	100.0	100.0	100.0	100.00	Yes
AIC-88258	100.0	100.0	100.0	100.00	Yes
MET-39646	100.0	100.0	100.0	100.00	Yes
MET-39684	100.0	100.0	100.0	100.00	Yes
MET-40189	100.0	100.0	100.0	100.00	Yes
MET-40192	100.0	100.0	100.0	100.00	Yes
MET-40312	0.0	0.0	0.0	100.00	No
MET-40527	100.0	100.0	100.0	100.00	Yes
MET-448442	100.0	100.0	100.0	100.00	Yes
MET-452948	100.0	100.0	100.0	100.00	Yes
MET-455286	100.0	100.0	100.0	100.00	Yes
MET-57797	100.0	100.0	100.0	100.00	Yes
MET-58436	100.0	100.0	100.0	100.00	Yes
MET-58437	96.9	81.5	74.1	100.00	No
MET-58438	27.4	35.9	39.2	100.00	No
MET-58439	62.8	57.8	56.0	100.00	No
MET-62027	100.0	100.0	99.3	46.13	Yes
MET-648051	100.0	100.0	100.0	100.00	Yes
MET-648056	72.6	64.1	60.8	100.00	No
MET-900380	12.1	42.7	52.7	100.00	No
MET-910551	100.0	100.0	100.0	100.00	Yes
MET-910552	100.0	100.0	100.0	57.10	Yes

All 120 mask areas and second-pass overlap pixel counts were independently reproduced from the original PNGs, without missing files, nonbinary ROI pixels or inconsistent spreadsheet entries. Foreground overlap is not lacquer-layer overlap.

Supplementary Table S27 | Retrospective sensitivity of external error ranking after expert ROI/material review

Scenario	Model	N	Accepted	Pearson r	Selected MAE U	Edge–U (95% bootstrap CI)
All 40	SD1.5	40	16	0.9043	0.03934	0.0097 [0.0002, 0.0142]
All 40	SD2	40	16	0.9162	0.03961	0.0102 [0.0002, 0.0176]
M≥80%	SD1.5	34	14	0.9050	0.04730	0.0033 [-0.0008, 0.0105]
M≥80%	SD2	34	14	0.9154	0.04724	0.0049 [-0.0010, 0.0123]
S/M/L≥80%	SD1.5	33	13	0.9035	0.04572	0.0062 [-0.0008, 0.0114]
S/M/L≥80%	SD2	33	13	0.9147	0.04502	0.0083 [-0.0011, 0.0130]
Lacquer-led	SD1.5	35	14	0.9107	0.03903	0.0091 [0.0003, 0.0151]
Lacquer-led	SD2	35	14	0.9199	0.04010	0.0089 [0.0000, 0.0190]
S/M/L≥80% & lacquer-led	SD1.5	29	12	0.9110	0.04608	0.0036 [-0.0002, 0.0120]
S/M/L≥80% & lacquer-led	SD2	29	12	0.9190	0.04549	0.0057 [-0.0012, 0.0140]

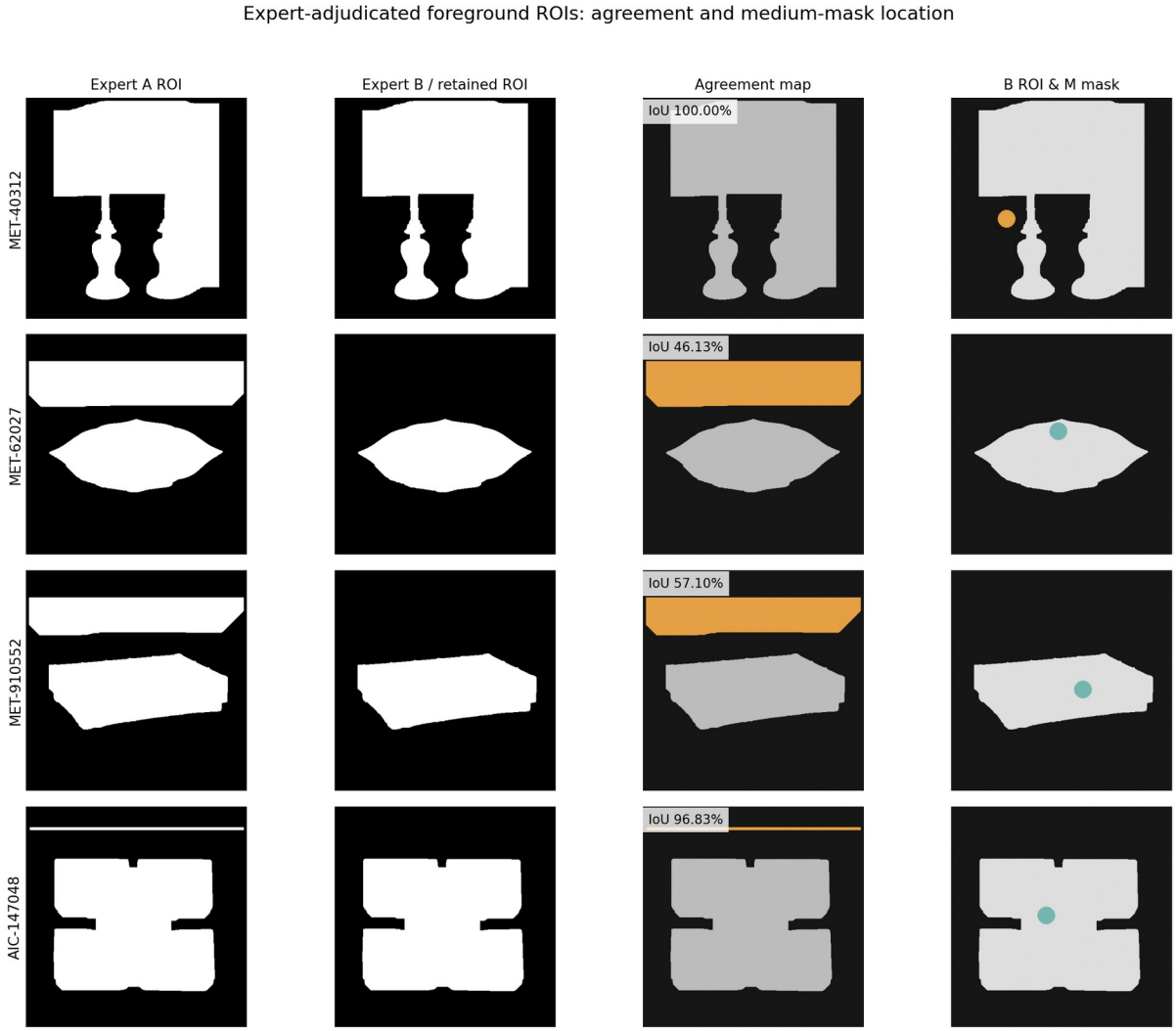
Post-hoc subgroups are not interchangeable with the original n=40 primary experiment. The external threshold was not learned or calibrated prospectively. Confidence intervals for differences frequently cover zero in reviewed foreground subsets, restricting claims of superiority over a cheap boundary-Sobel baseline. Mask-size effects and class proportions can be confounded; p values are descriptive unless multiplicity and selection issues are addressed.

Supplementary Table S28 | Three objects with non-identical binary foreground ROIs

Object ID	Expert IoU (%)	A ROI px	B ROI px	M overlap A/B (%)	Retained ROI file
MET-62027	46.13	94658	43669	100.0/100.0	roi_masks_second_pass/MET-62027_object_roi_second_pass.png
MET-910552	57.10	93043	53131	100.0/100.0	roi_masks_second_pass/MET-910552_object_roi_second_pass.png
AIC-147048	96.83	109523	106051	100.0/100.0	roi_masks_second_pass/AIC-147048_object_roi_second_pass.png

Two foreground ROIs had substantial first-to-second-pass boundary changes, and a third had a smaller difference. The dated E11 ROI Boundary Adjudication Record (21 October 2026) selected the second-pass mask for MET-62027 (IoU 46.1334%; 50,989 first-pass-only pixels removed) and MET-910552 (IoU 57.1037%; 39,912 first-pass-only pixels removed) because the first-pass masks included disconnected upper-image background. The second-pass masks are strict subsets of their first-pass counterparts, not intersections or averaged masks. For these two objects, second-pass overlap across S/M/L remained 99.321–100% and the three mask-placement categories were unchanged. The other non-identical pair, AIC-147048 (IoU 96.8299%), differs by 3,472 first-pass-only pixels; it retains the second-pass ROI already specified in the analysis register. Both original ROI versions and the decision record are archived. Agreement on mask-placement categories should not be misinterpreted as pixelwise agreement, and none of these object ROIs proves lacquer-layer segmentation.

Supplementary Fig. S6 | Object-foreground ROI agreement and position of the synthetic medium mask. The white silhouettes show expert-delineated foreground regions; the difference panel distinguishes shared versus expert-specific pixels. The fourth panel compares the retained second-pass foreground ROI with the original M-size synthetic mask. These are derived binary annotations, not original museum photographs or AI-inpainted restorations; no lacquer-layer boundary is established.



Supplementary Note S17 — Single-mask ranking and heritage selection coverage (post hoc)

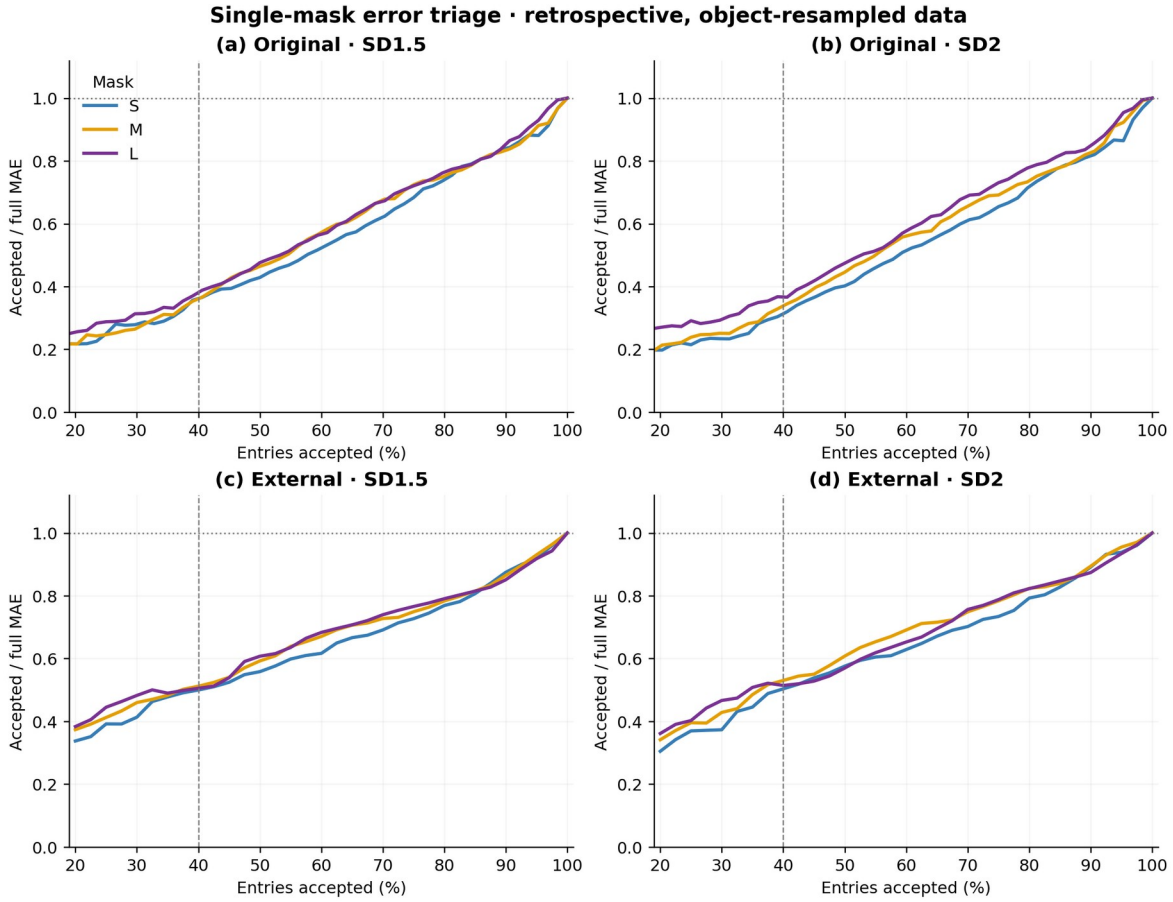
This secondary analysis uses the frozen 624 case-level records: original E10 SD1.5 and SD2 (64 objects × S/M/L each), and external E11 SD1.5 and SD2 (40 entries × S/M/L each). The original prespecified 64-object pooled S/M/L averages and full-40 external findings are unchanged. The unit is one object or catalogue entry within a given mask level, not individual generated images. Separately ranking a single mask condition avoids pooling masks, but does not create an independent prospective restoration task. All 12 mask-level uncertainty rankings gave lower selected-set mean MAE than their respective full sets; improvement over the inexpensive Context_Edge boundary-Sobel ranker is not consistent across every condition. The table displays descriptive 4,500-replicate object/entry bootstrap intervals for the paired Sobel-minus-uncertainty selected-MAE difference; these are not adjusted hypothesis tests. Reproducibility CSV includes all per-mask risk–coverage points and object resampling code.

Supplementary Table S29 | Single-mask 40% acceptance: uncertainty and low-cost boundary Sobel

Cohort/model	Mask	n/k	Full MAE	U-selected MAE	Sobel-selected MAE	Pearson r	Sobel–U CI
E10 SD1.5	L	64/26	0.0522	0.0203	0.0258	0.494	[0.0001, 0.0147]
E10 SD1.5	M	64/26	0.0460	0.0168	0.0188	0.765	[-0.0006, 0.0045]
E10 SD1.5	S	64/26	0.0414	0.0151	0.0155	0.801	[-0.0009, 0.0028]
E10 SD2	L	64/26	0.0511	0.0187	0.0211	0.629	[-0.0006, 0.0075]
E10 SD2	M	64/26	0.0464	0.0160	0.0177	0.776	[-0.0004, 0.0046]
E10 SD2	S	64/26	0.0429	0.0137	0.0153	0.781	[-0.0001, 0.0041]
E11 SD1.5	L	40/16	0.0836	0.0422	0.0481	0.937	[-0.0055, 0.0097]
E11 SD1.5	M	40/16	0.0785	0.0402	0.0471	0.865	[-0.0003, 0.0138]
E11 SD1.5	S	40/16	0.0730	0.0365	0.0460	0.847	[0.0020, 0.0195]
E11 SD2	L	40/16	0.0823	0.0423	0.0484	0.939	[-0.0020, 0.0117]
E11 SD2	M	40/16	0.0767	0.0407	0.0469	0.865	[-0.0016, 0.0160]
E11 SD2	S	40/16	0.0715	0.0360	0.0458	0.866	[0.0016, 0.0208]

U is three-seed masked sample variance; outcome MAE is from the disjoint fourth seed. CIs are percentile resampling intervals for the paired boundary-Sobel-selected minus U-selected MAE difference, not p values; a positive difference favors uncertainty ranking. Across mask-specific tests and models, sample membership and generated images are shared; inferential multiplicity is unadjusted. E10 and E11 mask geometry differ. Object-level aggregation in the main text is not replaced by these secondary results.

Supplementary Fig. S7 | Case-specific synthetic-mask risk–coverage curves in four dataset–model combinations. S, M and L are analyzed separately, with object/entry (not pixel or seed) as the unit. Curves show selected-set masked MAE divided by that mask-level full-set MAE; horizontal dotted line denotes 100% full-set risk and vertical dashed line nominal 40% acceptance. The figure is exploratory and does not imply general performance on natural damage.



Supplementary Note S18 — Class-coverage imbalance and operating-policy trade-off

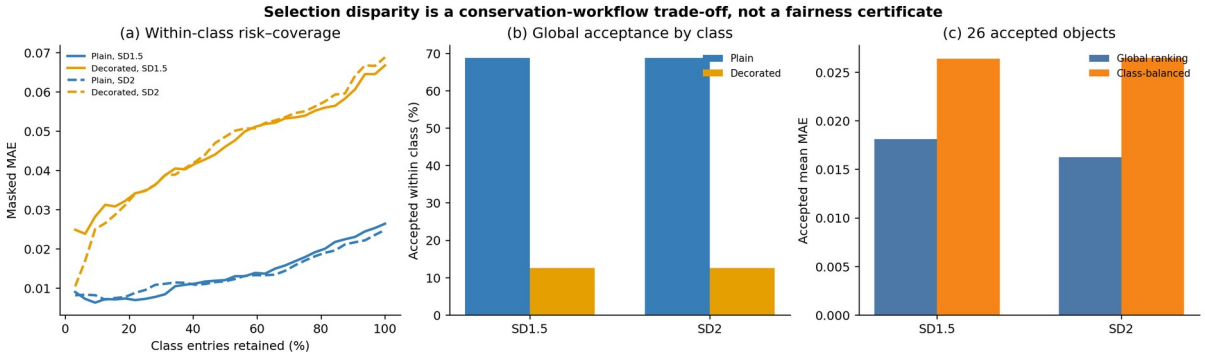
The original E10 cohort has 32 frozen visually plain and 32 decorated objects. Both model variants selected the same 22 plain and four decorated objects under their respective global uncertainty rankings (26/64 capacity). The class-balanced 13/13 selection increases decorated representation, but its accepted mean error is higher; this does not establish an optimal ethical or conservation policy. Group labels are image appearance classes, not certified art-historical or lacquer materials. Within-class curves are descriptive and share the same objects as the full-cohort primary study.

Supplementary Table S30 | Acceptance share and error by visual class and model

Model	Visual class	Available	Global accepted	Global accepted (%)	Full class MAE	Accepted within class MAE
SD1.5	Plain	32	22	68.75	0.02638	0.01573
SD1.5	Decorated	32	4	12.50	0.06672	0.03119
SD2	Plain	32	22	68.75	0.02479	0.01437
SD2	Decorated	32	4	12.50	0.06879	0.02650

At 26 accepted entries, the global U selection has mean MAE 0.01811 (SD1.5) / 0.01624 (SD2); the class-balanced 13/13 selection gives 0.02640 / 0.02647, respectively. Rankings are retrospectively selected from the same E10 cohort and are not prospective policy evaluations.

Supplementary Fig. S8 | Visual-class risk–coverage and representation–error trade-offs. (a) Separate within-class ranking curves, with solid SD1.5 and dashed SD2, represent accepted mean error conditional on each class. (b) Global 26-object policy accepts different proportions of plain and decorated objects. (c) Mean error for global and equal-class 13/13 selection at the same total capacity. Both original models reuse the same 64 objects; numbers are post hoc.



Supplementary Note S19 — Mask geometry and two-expert annotation provenance

E10 mask areas were defined relative to delineated object ROIs (nominal S/M/L = 7.5%, 20%, 35% of the object ROI). E11 used fixed-centre circular masks on 512 × 512 padded museum photographs. The absolute pixel areas are of similar order, but shape, position, content and denominator differ, so cross-cohort loss severity is not matched. Mean canvas percentages in Table S31 are descriptive geometric values, not E10 object-ROI area fractions.

Supplementary Table S31 | Original and external mask geometry, pixel counts and denominators

Cohort	Mask	Mean px	Min–max px	Mean canvas (%)	Protocol denominator
E10	S	568.3	247–1810	0.217	Object ROI
E10	M	1514.8	658–4821	0.578	Object ROI
E10	L	2650.8	1152–8438	1.011	Object ROI
EXTERNAL	S	489.0	489–489	0.187	Full 512x512 canvas
EXTERNAL	M	1313.0	1313–1313	0.501	Full 512x512 canvas
EXTERNAL	L	2209.0	2209–2209	0.843	Full 512x512 canvas

Two relevant specialists verified material/ornament categories and produced two recorded foreground-ROI mask sets for 40 external catalogue entries. Separate ROI versions, their agreement statistics and the dated adjudication are traceable. However, the surviving artifacts do not independently document whether both masks were drawn de novo or whether previous masks were visible during second review. Accordingly, we describe a two-expert verification/adjudication workflow and do not claim independently blinded inter-annotator segmentation reliability. There were 37/40 pixelwise-identical pairs; the other three retain their IoU values and adjudicated second-pass ROIs (Table S28). Material-role/visual-class categories agreed for all 40 entries as recorded, but category agreement should not be equated with ROI boundary identity or a validated lacquer-layer boundary. The two external visual-quality raters who reviewed 80 anonymized model outputs are a separate assessment function and should not be conflated with these object-ROI/material expert roles.

Supplementary Note S20 — Cluster-aware diagnosis of anonymous visual ratings

This post-hoc analysis reused the original, unchanged ordinal responses for 80 anonymized image items representing two model outputs for each of 40 external catalogue entries. Original primary endpoint calculations and all image-reconstruction outputs were unaffected. The preserved review form rated three visual attributes and suitability for further human review at Stage 1; the additional photographic-reference comparison at Stage 2 assessed reference consistency. These different questions do not define matched pre/post measurements. A score of 3 on the further-review suitability item indicates suitability for human examination, not an absence of concern.

Ordinal Krippendorff α and quadratic-weighted Cohen κ are complementary chance-adjusted reliability statistics with different weightings and opportunity-agreement definitions. Exact agreement, marginal distributions and signed differences are therefore reported alongside both coefficients. Percentile 95% intervals were taken from 4,000 paired-object bootstrap draws from the archived diagnostic workbook, retaining the two SD model outputs together for each sampled object. Scores from 80 images were not resampled as 80 independent heritage objects. No multiplicity-adjusted significance claim is made for exploratory reliability comparisons.

Supplementary Table S32 | Cluster-aware inter-rater reliability across five original rating items

Original rating item	Ordinal α [95% CI]	Quadratic κ [95% CI]	Exact / 80	R1−R2 mean	Difference ≥2
Pattern continuity	0.230 [−0.057, 0.471]	0.402 [0.174, 0.573]	45/80	+0.4125	0
Texture consistency	0.271 [−0.042, 0.530]	0.491 [0.176, 0.680]	45/80	+0.4125	0
Suspicious artifacts	0.287 [−0.037, 0.549]	0.522 [0.199, 0.701]	46/80	+0.4250	0
Reference consistency (Stage 2)	0.271 [−0.041, 0.527]	0.536 [0.199, 0.716]	46/80	+0.4250	0
Suitable for further review	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	80/80	0.0000	0

Table S32: intervals are post-hoc percentile intervals from 4,000 bootstrap resamples of the 40 catalogue-entry clusters (both model outputs retained). Negative lower limits for α reflect sampling uncertainty, not negative observed agreement. Quadratic κ does not correct away systematic differences in rater thresholds. The further-review item is almost ceiling-concentrated and is not a substitute for the four visual quality judgments.

Supplementary Table S33 | Observed 0–3 score marginals and rater-severity differences

Original rating item	Rater 01 (0/1/2/3)	Rater 02 (0/1/2/3)	R1=3 / 80	R2=3 / 80
Pattern continuity	0 / 3 / 14 / 63	0 / 3 / 47 / 30	63	30
Texture consistency	0 / 5 / 12 / 63	1 / 3 / 46 / 30	63	30
Suspicious artifacts	0 / 5 / 12 / 63	1 / 4 / 45 / 30	63	30
Reference consistency	1 / 3 / 13 / 63	1 / 4 / 45 / 30	63	30
Suitable for further review	0 / 1 / 16 / 63	0 / 1 / 16 / 63	63	63

Across the four visual judgments, all nonidentical pairs differed by only one category. Rater 01 assigned category 3 to 63/80 items in each visual criterion whereas rater 02 did so on 30/80. Each rater gave identical numerical scores across the four visual criteria for 77/80 images, and 33 images showed the same 3-versus-2 discrepancy on all four criteria. The mean signed rater-01 minus rater-02 difference was 0.4125–0.4250; cluster-bootstrap 95% intervals were 0.287–0.537 for the first two criteria and 0.312–0.550 for the next two, excluding zero. These shifts describe rating thresholds rather than verified inter-rater bias mechanisms or image-level invalidity.

Supplementary Table S34 | Model-stratified descriptive reliability of visual-rating items

Model	Original item	n images	Ordinal α	Quadratic κ	Exact (%)
SD1.5	Pattern continuity	40	0.259	0.480	60.0
SD1.5	Texture consistency	40	0.275	0.514	55.0
SD1.5	Suspicious artifacts	40	0.304	0.570	57.5
SD1.5	Reference consistency	40	0.304	0.629	60.0
SD2	Pattern continuity	40	0.193	0.312	52.5
SD2	Texture consistency	40	0.266	0.462	57.5
SD2	Suspicious artifacts	40	0.266	0.462	57.5
SD2	Reference consistency	40	0.230	0.401	55.0

The model-stratified coefficients are descriptive and reuse the same 40 underlying objects; they should not be interpreted as two independent reliability experiments. All reviewed items represent judgments of synthetic-loss image completions, not evaluations of genuine material conservation outcomes. The original two-expert object-foreground/material classification and ROI verification are a distinct assessment task from these two anonymous image-quality rating worksheets.