# Supplementary Information

## Evidence aware multimodal auditing of museum image metadata consistency

Peng Yue; Jia Hou; Xiao Zhang

npj Heritage Science | Article

## Supplementary contents



### Supplementary Table S1 Corpus construction and annotation design

| Component | Unit | n | Description |
|---|---|---|---|
| P0-2A strict input | Legacy analysis-ready rows | 297 | Provenance-qualified strict rows entering physical-object identity resolution. |
| Canonical corpus | Physical objects | 295 | Two exact legacy alias rows collapsed; one physical object = one statistical unit. |
| Canonical images | Official assets | 295 | Exactly one unique official OpenData image asset per final object within the strict cohort; no manual multiview selection required. |
| Object-type Gold claims | Field-level claims | 295 | Two-author independent review; triggered cases adjudicated by the third author. |
| Motif Gold claims | Field-level claims | 248 | Created only where the motif ontology was assessable. |
| Technique Gold claims | Field-level claims | 291 | Secondary visual-inference task; |

| | | | |
|---|---|---|---|
| | | | two-author review plus third-author adjudication. |
| Combined Gold set | Field-level claims | 834 | 295 object type + 248 motif + 291 technique. |
| Objects with all three fields | Physical objects | 245 | Object type, motif and technique available for the same object. |
| Final de-leaked benchmark | Rows | 1,478 | 739 authentic + 739 controlled metadata perturbations across all fields. |

*P0-2C exact-path matching, JPEG decode, SHA256 hashing and canonical duplicate scanning passed for 295/295 objects. The strict cohort is an analytical subset, not an exhaustive census of the live NPM lacquerware universe.*

## Supplementary Table S2 Final Gold-label distribution

| Field | n | Consistent | Ambiguous | Unassessable | Potential inconsistency | Consistent % |
|---|---|---|---|---|---|---|
| Object type | 295 | 281 | 14 | 0 | 0 | 95.3% |
| Motif | 248 | 173 | 73 | 2 | 0 | 69.8% |
| Technique | 291 | 285 | 4 | 2 | 0 | 97.9% |
| Combined | 834 | 739 | 91 | 4 | 0 | 88.6% |

## Supplementary Table S3 Inter-rater reliability before adjudication

| Scope | n | Exact agreement | Raw agreement | Cohen κ | Adjudicated n | Note |
|---|---|---|---|---|---|---|
| Stage 1 overall | 543 | 474 | 87.3% | 0.587 | 69 | Object type + motif |
| Object type | 295 | 289 | 98.0% | 0.802 | 6 | Final |
| Motif | 248 | 185 | 74.6% | 0.458 | 63 | Final |
| Technique | 291 | 276 | 94.8% | 0.302 | 17 | κ prevalence-skewed; interpret with raw agreement |

## Supplementary Table S4 Fully de-leaked author-evaluator controlled-perturbation benchmark

| Field | Rows | Correct | Incorrect | Abstain | Decisive accuracy | Overall accuracy | Coverage |
|---|---|---|---|---|---|---|---|
| Object type | 562 | 478 | 52 | 32 | 90.2% | 85.1% | 94.3% |
| Motif | 346 | 191 | 85 | 70 | 69.2% | 55.2% | 79.8% |
| Technique | 570 | 298 | 88 | 184 | 77.2% | 52.3% | 67.7% |
| Combined | 1478 | 967 | 225 | 286 | 81.1% | 65.4% | 80.6% |

*Decisive accuracy is calculated among non-abstained judgments; coverage is the proportion receiving a decisive judgment. All judgments were made by study authors in de-leaked opaque batches.*

## Supplementary Table S5 Authentic acceptance and controlled-perturbation detection

| Field | Authentic n | Authentic acceptance | Authentic abstention | Perturbed n | Exact detection | False acceptance | Perturbation abstention |
|---|---|---|---|---|---|---|---|
| Object type | 281 | 87.5% | 5.0% | 281 | 82.6% | 11.0% | 6.4% |
| Motif | 173 | 82.7% | 12.1% | 173 | 27.7% | 43.9% | 28.3% |
| Technique | 285 | 68.1% | 28.4% | 285 | 36.5% | 27.4% | 36.1% |
| Combined | 739 | 78.9% | 15.7% | 739 | 52.0% | 25.0% | 23.0% |

## Supplementary Table S6 Physical-object-cluster bootstrap uncertainty intervals

| Metric | Estimate | 95% CI lower | 95% CI upper |
|---|---|---|---|
| Decisive accuracy | 81.1% | 79.0% | 83.3% |
| Overall accuracy | 65.4% | 63.1% | 67.9% |

| | | | |
|---|---|---|---|
| Coverage | 80.6% | 78.6% | 82.8% |
| Authentic acceptance | 78.9% | 76.0% | 81.8% |
| Perturbation exact detection | 52.0% | 48.6% | 55.5% |
| Perturbation false acceptance | 25.0% | 21.9% | 28.1% |
| Perturbation abstention | 23.0% | 20.0% | 25.9% |

*Percentile 95% intervals from 2,000 bootstrap resamples of physical_object_id (seed 20260908); all rows belonging to a sampled object are resampled together.*

## Supplementary Table S7 Frozen model-family comparison

| Model | Role | Macro AUROC | PR-AUC | Balanced acc. | F1 | MCC | Test rows |
|---|---|---|---|---|---|---|---|
| M1 structured image×claim interaction | Exploratory structured fusion | 0.759 | 0.755 | 0.703 | 0.699 | 0.407 | 282 |
| M2 nonlinear E0+claim fusion | Exploratory nonlinear fusion | 0.723 | 0.733 | 0.678 | 0.639 | 0.357 | 282 |
| CLIP-EN ViT-B/32 | Pre-specified primary VLM | 0.679 | 0.677 | 0.571 | 0.454 | 0.165 | 282 |
| CLIP-BI | Pre-declared bilingual sensitivity | 0.655 | 0.658 | 0.625 | 0.605 | 0.259 | 282 |
| E0 prototype | Image-conditioned baseline | 0.641 | 0.661 | 0.601 | 0.656 | 0.205 | 282 |
| Claim-only logistic | Metadata shortcut control | 0.547 | 0.570 | 0.604 | 0.529 | 0.197 | 282 |
| Image-only logistic | Negative control | 0.500 | 0.500 | 0.500 | 0.667 | 0.000 | 282 |

*M1 and M2 are descriptive exploratory models because their family ranking had already been inspected on the held-out partition. CLIP-EN remains the pre-specified primary pretrained VLM.*

## Supplementary Table S8 Field-specific CLIP-EN and M1 performance

| Field | Test rows | CLIP AUROC | M1 AUROC | Δ CLIP−M1 | CLIP pair-win | M1 pair-win | Protocol-resolved |
|---|---|---|---|---|---|---|---|
| Object type | 108 | 0.812 | 0.899 | −0.087 | 79.6% | 87.0% | 75.9% |
| Motif | 60 | 0.568 | 0.587 | −0.019 | 60.0% | 56.7% | 20.0% |
| Technique | 114 | 0.657 | 0.791 | −0.134 | 87.7% | 77.2% | 35.1% |

## Supplementary Table S9 Author-evaluator/model concordance

| Author-evaluator stratum | Rows | CLIP accuracy | CLIP AUROC | M1 accuracy | M1 AUROC |
|---|---|---|---|---|---|
| Correct | 184 | 0.614 | 0.801 | 0.766 | 0.865 |
| Abstain | 60 | 0.467 | 0.444 | 0.600 | 0.665 |
| Incorrect | 38 | 0.579 | 0.465 | 0.711 | 0.728 |

## Supplementary Table S10 Robustness and model-agreement analyses

| Analysis | N | Agreement | κ / Spearman ρ | Changed predictions | Interpretation |
|---|---|---|---|---|---|
| CLIP vs M1 binary predictions | 282 | 56.4% | 0.105 | — | Low thresholded agreement indicates different decision structures. |
| CLIP vs M1 score ranking | 282 | — | 0.244 | — | Only modest score correlation despite shared benchmark inputs. |
| CLIP-EN vs CLIP-BI | 282 | 75.5% | 0.927 | 69 | High score-rank correlation but lower threshold agreement |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | indicates score-scale / operating-point sensitivity rather than ranking sensitivity. |

## Supplementary Table S11 CLIP-M1 object-bootstrap AUROC differences

| Scope | M1 AUROC | CLIP AUROC | Δ CLIP−M1 | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| Macro | 0.759 | 0.679 | −0.080 | −0.148 | 0.006 |
| Object type | 0.899 | 0.812 | −0.087 | −0.192 | 0.040 |
| Motif | 0.587 | 0.568 | −0.019 | −0.173 | 0.152 |
| Technique | 0.791 | 0.657 | −0.134 | −0.219 | −0.020 |

The S11 paired difference intervals are retained from the frozen exploratory P3 analysis and used 1,000 physical-object resamples. Harmonized 5,000-resample individual-model figure intervals and the additional object×exact-claim dependence sensitivity are reported in Supplementary Table S21. Cross-family comparisons remain descriptive, not confirmatory model-selection tests.

## Supplementary Table S12 Field-, variant- and evaluator-batch-adjusted author-evaluator/model sensitivity analysis

| Model | Contrast | Odds ratio | 95% CI | Raw p / Holm-adjusted p | Adjustment | Inference |
|---|---|---|---|---|---|---|
| CLIP-EN | Abstain vs correct | 0.289 | 0.140–0.599 | <0.001 / 0.003 | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| CLIP-EN | Incorrect vs correct | 0.233 | 0.084–0.649 | 0.005 / 0.016 | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| CLIP-EN | Perturbed vs authentic | 12.322 | 5.151–29.478 | <0.001 / — | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| CLIP-EN | Evaluator batch Y vs X | 1.705 | 0.907–3.205 | 0.097 / — | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| M1 | Abstain vs correct | 0.510 | 0.272–0.955 | 0.035 / 0.071 | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| M1 | Incorrect vs correct | 0.867 | 0.355–2.119 | 0.755 / 0.755 | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| M1 | Perturbed vs authentic | 1.412 | 0.815–2.448 | 0.219 / — | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |
| M1 | Evaluator batch Y vs X | 0.975 | 0.595–1.598 | 0.921 / — | Field + variant + evaluator-batch side; cluster-robust SE by physical object | Post-hoc descriptive; frozen predictions |

Outcome is correctness of the already frozen thresholded prediction. Models adjust for metadata field, authentic-versus-perturbed status and pseudonymous evaluator-batch side (X/Y), with physical-object-clustered standard errors. Holm adjustment is applied across four author-outcome contrasts (CLIP abstain/incorrect and M1 abstain/incorrect versus correct); variant and batch coefficients are adjustment covariates and are not part of that family. M1 abstain versus correct therefore does not remain significant after multiplicity correction (Holm p = 0.071). No model, prompt, threshold, label or split was modified.

## Supplementary Table S13 Evaluator-role and blinding transparency

| Component | Evaluator structure | Blinding / independence | External expert | Interpretive boundary |
|---|---|---|---|---|
| Gold annotation | Two authors independently reviewed each claim; a third author adjudicated triggered cases. | The two annotating authors did not see one another's decisions before adjudication. | None | Internal protocol reproducibility, not external expert consensus. |
| Final de-leaked benchmark | Study authors judged complementary opaque batches; each final benchmark row has one author-evaluator judgment. | No author-evaluator received both variants of the same source field claim; record identity, donor/truth and variant status were hidden. Procedural blinding does not make investigators evaluator-naïve to images seen in earlier study stages. | None | Protocol resolution is reconstructed across complementary rows. No final-row blinded inter-rater reliability or external expert consensus is claimed. |
| Model analysis | Study authors developed and analyzed the computational workflow. | Human benchmark and model predictions were frozen before post-hoc sensitivity analyses. | None | Post-hoc analyses do not change test-set tuning or predictions. |
| Evaluator identity | Frozen benchmark files retain pseudonymous stage/batch IDs (B1R1-X/Y; R1-X/Y). | Named author-to-ID mapping was not preserved in a defensible release artifact and is not reconstructed retrospectively. | None | Sensitivity models control X/Y batch side rather than inventing named identity. |

## Supplementary Table S14 Locked CLIP prompt templates and ontology summary

| Field | EN prompt 1 | EN prompt 2 | EN prompt 3 | ZH sensitivity 1 | ZH sensitivity 2 |
|---|---|---|---|---|---|
| object_type | a museum photograph of a {label} | a lacquerware object that is a {label} | a Chinese lacquerware {label} | 一件{label}形制的漆器 | 博物中的{label}漆器 |
| motif | a lacquerware object decorated with {label} | a museum lacquer object with a {label} motif | decorative lacquerware showing {label} | 一件带有{label}的漆器 | 漆器上的{label}装饰 |
| technique | lacquerware made with {label} | a museum lacquer object showing {label} | a Chinese lacquerware object decorated using {label} | 一件采用{label}工的漆器 | 漆器的{label}工艺 |

*The prompt/ontology lock contained 116 exact claims: 31 object-type, 58 motif and 27 technique. All templates and claim translations were frozen before CLIP inference. The full 116-row bilingual ontology, including instantiated prompts, is supplied in Supplementary Data 1.*

## Supplementary Table S15 Controlled-perturbation donor selection and reuse audit

| Field | Perturbed pairs | Unique donor objects | Median object reuse | Max object reuse | Unique donor claims | Max claim reuse |
|---|---|---|---|---|---|---|
| object_type | 281 | 204 | 1.0 | 3 | 25 | 91 |
| motif | 173 | 173 | 1.0 | 1 | 38 | 17 |
| technique | 285 | 173 | 2.0 | 3 | 23 | 112 |

All source and donor claims were final Consistent claims; source and donor were different physical objects from the same period. Object-type labels differed and motif/technique label sets were disjoint. Frozen activation records describe a balanced deterministic assignment. The full Stage-1 object+motif assignment had maximum donor-object reuse of four; technique used 173 unique donor objects with maximum reuse of three. The exact source-to-donor map is released row-wise. Implementation-level tie-break ordering was not separately preserved, so reproducibility should use the frozen map rather than regenerate assignments from rules alone.

## Supplementary Table S16 Claim-frequency shortcut sensitivity

| Field | Learned claim-only AUROC | Empirical train-claim-frequency AUROC | Interpretation |
|---|---|---|---|
| object_type | 0.676 | 0.637 | Field-specific claim-distribution signal |
| motif | 0.267 | 0.256 | Field-specific claim-distribution signal |

| | | | |
|---|---|---|---|
| technique | 0.697 | 0.691 | Substantial field-specific claim signal; structured results are not pure visual-semantic measures |
| Macro | 0.547 | 0.528 | Overall shortcut is modest, but field heterogeneity is material. |

The empirical score is the Laplace-smoothed fraction of authentic rows for each exact displayed claim estimated from the training split and applied unchanged to held-out rows. It is a diagnostic, not a tuned model.

## Supplementary Table S17 Structured M1/M2 model specification

| Model/field | Image z | Claim c | Feature map | Selected settings |
|---|---|---|---|---|
| M1 object type | PCA d=16 | q=50 | [z,c,vec(z c^T)] = 866 | SGD logistic; alpha=0.001; threshold=0.866452 |
| M1 motif | PCA d=32 | q=65 | [z,c,vec(z c^T)] = 2,177 | SGD logistic; alpha=0.01; threshold=0.428524 |
| M1 technique | PCA d=16 | q=44 | [z,c,vec(z c^T)] = 764 | SGD logistic; alpha=0.001; threshold=0.713620 |
| M2 object type | PCA d=32 | q=50 | [z,c] = 82 | HGB; depth=3; lr=0.10; threshold=0.444041 |
| M2 motif | PCA d=32 | q=65 | [z,c] = 97 | HGB; depth=5; lr=0.05; threshold=0.605635 |
| M2 technique | PCA d=32 | q=44 | [z,c] = 76 | HGB; depth=3; lr=0.10; threshold=0.514915 |

c concatenates field-specific component indicators with an exact-claim indicator; the 116 exact claims are released in Supplementary Data 1. PCA was fitted on training objects only and the listed settings/thresholds were selected on development data. The source execution workbook does not separately preserve every unselected estimator default as a manuscript field; unrecorded settings are not reconstructed retrospectively.

## Supplementary Table S18 Upstream corpus-selection and provenance routing

| Frame | n | Routing / status | Interpretive boundary |
|---|---|---|---|
| Live NPM lacquerware category snapshot | 762 | Reference count checked 2026-08-26 | Not fully row-wise recovered; not the direct analysis frame |
| Inherited candidate registry | 583 | 388 prior DIRECT + 195 prior HOLD; 179 live-source records unrecovered | Incomplete provenance frame |
| P0-2A hard-identity eligible | 404 | 388 inherited DIRECT + 16 Legacy45 NEW-DIRECT | Routed to file-level audit |
| Exact-file strict PASS | 297 | Exact official image-file provenance; entered P0-2B | Actual analytical-entry frame |
| File-level HOLD | 107 | Object/page identity available but exact image-file class/binding unverified | Excluded from strict analysis |
| P0-2B final physical objects | 295 | Two exact legacy aliases collapsed | Statistical unit |
| P0-2C byte-locked files | 295 | 295/295 decoded, SHA256 locked; zero canonical duplicate groups | Final analytical corpus |

Selection was provenance-driven rather than probabilistic. Raw period composition differed markedly between the 297 strict rows and 107 file-HOLD records (Qing 89.6% versus 52.3%; Ming 10.4% versus 29.0%; descriptive Cramér's V = 0.464). Because these are deterministic routing frames and several residual period cells are sparse, no population-inference chi-square p value is used. Standardized object-type ontology coding was performed only after strict-file entry, so comparable object-type distributions for the file-HOLD frame were not reconstructed.

## Supplementary Table S19 Pseudonymous author-evaluator batch metrics

| Stage | Batch ID | N | Correct | Incorrect | Abstain | Overall accuracy | Coverage | Decisive accuracy |
|---|---|---|---|---|---|---|---|---|
| Stage 1 object type + motif | B1R1-X | 454 | 323 | 66 | 65 | 71.1% | 85.7% | 83.0% |
| Stage 1 object type + motif | B1R1-Y | 454 | 346 | 71 | 37 | 76.2% | 91.9% | 83.0% |
| Stage 2 technique | R1-X | 285 | 143 | 36 | 106 | 50.2% | 62.8% | 79.9% |
| Stage 2 technique | R1-Y | 285 | 155 | 52 | 78 | 54.4% | 72.6% | 74.9% |

Each final de-leaked benchmark row has one author-evaluator judgment. These batch summaries do not constitute blinded inter-rater reliability. Gold annotation reliability is reported separately in Supplementary Table S3.

## Supplementary Table S20 Claim-balanced held-out discrimination sensitivity

| Model | Field | Rows | Exact claims | Row AUROC | Claim-balanced AUROC | Δ balanced−row |
|---|---|---|---|---|---|---|
| CLIP-EN | Object type | 108 | 18 | 0.812 | 0.729 | −0.083 |
| CLIP-EN | Motif | 60 | 22 | 0.568 | 0.443 | −0.125 |
| CLIP-EN | Technique | 114 | 20 | 0.657 | 0.527 | −0.130 |
| CLIP-EN | Macro | 282 | — | 0.679 | 0.566 | −0.113 |
| M1 | Object type | 108 | 18 | 0.899 | 0.893 | −0.006 |
| M1 | Motif | 60 | 22 | 0.587 | 0.577 | −0.010 |
| M1 | Technique | 114 | 20 | 0.791 | 0.832 | +0.041 |
| M1 | Macro | 282 | — | 0.759 | 0.767 | +0.008 |

*Within each field, every exact displayed claim was assigned equal total weight; macro values are unweighted means of field AUROCs. Claim balancing is a sensitivity estimand, not a replacement for the frozen row-weighted primary benchmark. The larger CLIP shift indicates material dependence on displayed-claim prevalence.*

## Supplementary Table S21 Physical-object and object×claim bootstrap sensitivity

| Model | Field | Point | Object bootstrap 95% CI | Object×claim 95% CI | Valid two-way reps |
|---|---|---|---|---|---|
| CLIP-EN | Object type | 0.812 | 0.736–0.880 | 0.639–0.974 | 5,000 |
| CLIP-EN | Motif | 0.568 | 0.482–0.663 | 0.299–0.795 | 5,000 |
| CLIP-EN | Technique | 0.657 | 0.612–0.713 | 0.299–0.781 | 4,999 |
| CLIP-EN | Macro | 0.679 | 0.641–0.721 | 0.528–0.785 | 4,999 |
| M1 | Object type | 0.899 | 0.808–0.965 | 0.745–0.987 | 5,000 |
| M1 | Motif | 0.587 | 0.439–0.729 | 0.320–0.818 | 5,000 |
| M1 | Technique | 0.791 | 0.678–0.885 | 0.567–0.986 | 4,999 |
| M1 | Macro | 0.759 | 0.691–0.820 | 0.634–0.856 | 4,999 |

*Model sensitivity intervals use 5,000 percentile resamples with seed 20260908. Object bootstrap resamples physical objects and retains all associated rows; field AUROCs are computed within replicate and macro-averaged. The two-way pigeonhole sensitivity independently resamples source objects and exact displayed-claim clusters, with row weights equal to the product of cluster multiplicities. Replicates lacking both classes in a required field are omitted only for the affected interval. The deliberately wider two-way intervals quantify claim-cluster dependence and are not confirmatory model-selection tests.*

## Supplementary Table S22 Frozen CLIP-EN operating-point diagnostics

| Field | Threshold | TP | FN | Authentic sensitivity | TN | FP | Perturbation specificity | Balanced accuracy |
|---|---|---|---|---|---|---|---|---|
| Object type | 0.3370 | 15 | 39 | 27.8% | 49 | 5 | 90.7% | 59.3% |
| Motif | 0.3272 | 12 | 18 | 40.0% | 20 | 10 | 66.7% | 53.3% |
| Technique | 0.3323 | 23 | 34 | 40.4% | 44 | 13 | 77.2% | 58.8% |
| All fields | — | 50 | 91 | 35.5% | 113 | 28 | 80.1% | — |

*Authentic claims are treated as positive and controlled perturbations as negative. Thresholds were frozen from development data. The 1:1 controlled benchmark is not an estimate of naturally occurring museum inconsistency prevalence; therefore PPV, NPV and operational review burden are not inferred from these counts.*

## Supplementary Table S23 Threshold-free pair-margin sensitivity

| Model | Field | Protocol-resolved n | Mean margin resolved | Other n | Mean margin other | Resolved−other |
|---|---|---|---|---|---|---|
| CLIP-EN | Object type | 41 | 0.0255 | 13 | 0.0281 | −0.0026 |
| CLIP-EN | Motif | 6 | −0.0031 | 24 | 0.0079 | −0.0110 |
| CLIP-EN | Technique | 20 | 0.0201 | 37 | 0.0078 | +0.0123 |
| M1 | Object type | 41 | 0.6421 | 13 | 0.6800 | −0.0379 |
| M1 | Motif | 6 | −0.1781 | 24 | 0.1549 | −0.3330 |
| M1 | Technique | 20 | 0.6463 | 37 | 0.3177 | +0.3286 |

*Margin = authentic score minus matched perturbed score; positive values favor the authentic claim. The field-adjusted macro resolved-minus-other difference was −0.0004 for CLIP-EN (5,000-object-bootstrap 95% CI −0.0085 to 0.0072) and −0.014 for M1 (−0.207 to 0.191). This threshold-free sensitivity does not corroborate a general margin separation by protocol resolution; author-evaluator/model co-variation is therefore treated as operating-point-dependent and descriptive.*

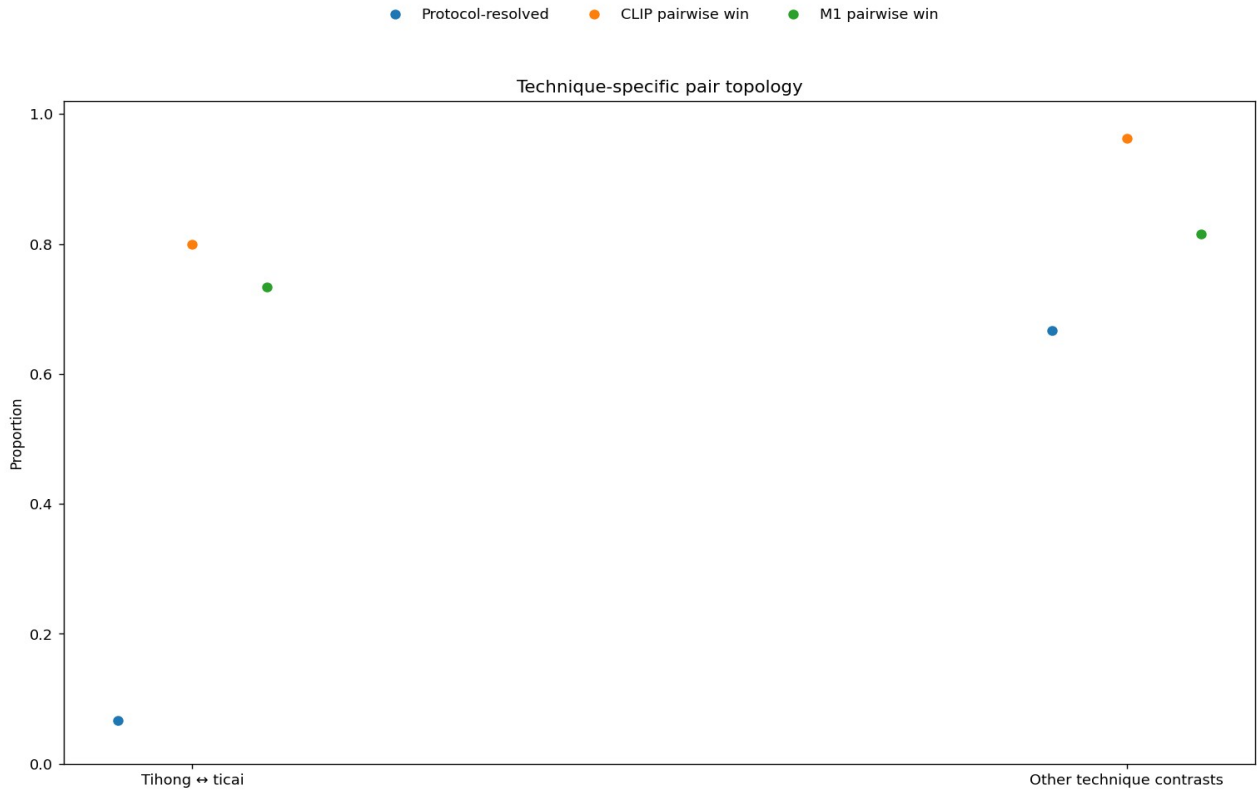
## Supplementary Table S24 Statistical analysis and reproducibility specification

| Item | Specification | Interpretive role |
|---|---|---|
| Statistical unit | Physical object for split and primary clustering; benchmark rows repeat within objects. | Avoid row-level pseudoreplication. |
| Sample size | No a priori power calculation; cohort size determined by provenance eligibility and frozen 177/59/59 object split. | Not a prevalence sample. |
| Primary metrics | Field AUROC/PR-AUC; macro = unweighted field mean. | Threshold-free ranking. |

| | | |
|---|---|---|
| Human uncertainty | 2,000 percentile object-cluster resamples; seed 20260908. | All rows of sampled object retained. |
| Model/figure uncertainty | 5,000 percentile object-cluster resamples; seed 20260908. | Harmonized figure intervals. |
| Claim dependence | Claim-balanced AUROC + 5,000 object×exact-claim two-way bootstrap. | Diagnoses repeated claim clusters. |
| Post-hoc GLM | Field + variant + batch adjusted; object-cluster robust SE. | Descriptive only. |
| Multiplicity | Holm across four author-outcome contrasts. | M1 abstain adjusted p=0.071. |
| Threshold-free check | Authentic−perturbed pair margin; field-stratified macro + 5,000 object bootstrap. | Operating-point sensitivity audit. |
| Missing data | No imputation; exact n reported; invalid single-class bootstrap replicates omitted only for affected interval. | Transparent denominators. |
| Reporting lock | No post-hoc retuning of model, prompt, threshold, Gold label or split. | Protects held-out test. |

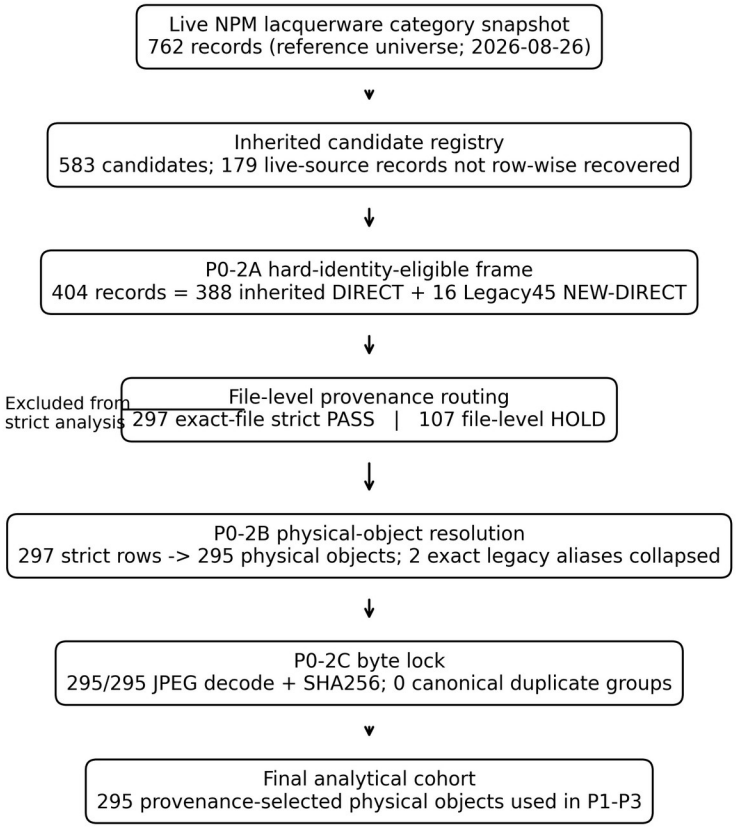
*The separately supplied reproducibility archive starts from the 282 frozen held-out prediction rows and reproduces Supplementary Tables S20–S23 without retraining. Historical model-training defaults not independently preserved in the frozen source artifacts are not reconstructed.*

## Supplementary Figure S1  Technique-specific error topology




Supplementary Figure S1. Technique-specific pair topology. Direct carved red lacquer, tihong (剔紅), versus carved polychrome lacquer, ticai (剔彩), contrasts constitute more than half of technique pairs and are rarely protocol-resolved under the internally blinded author-evaluator procedure despite relatively high model pairwise ranking accuracy.

## Supplementary Figure S2  Full upstream corpus-selection and provenance flow




Counts describe provenance routing, not probability sampling. The final cohort is not an exhaustive or representative census of NPM lacquerware.

Supplementary Figure S2. Upstream corpus-selection and provenance routing. The live NPM lacquerware category reported 762 records at the source-lock snapshot, while the inherited registry contained 583 candidates and did not row-wise recover 179 records. P0-2A routed 404 hard-identity-eligible records to file-level provenance audit: 297 exact-file strict PASS rows entered analysis and 107 remained file-level HOLD. P0-2B collapsed two exact legacy aliases, yielding 295 physical objects, and P0-2C byte-locked all 295 canonical images. These counts describe provenance routing rather than probability sampling; the final cohort is not an exhaustive or representative census.

## Supplementary Data 1

The separately supplied Supplementary Data 1 workbook contains sanitized object-level and row-level data supporting all reported analyses, including the 295-object manifest, 834 finalized Gold claims, the 1,478-row de-leaked controlled-perturbation benchmark, object-level split assignments, CLIP scores, held-out structured-model predictions, physical-object and object×claim bootstrap outputs, expanded field/variant/evaluator-batch analyses, pseudonymous evaluator assignments and batch metrics, upstream corpus-selection records, controlled-donor reuse diagnostics, claim-frequency and claim-balanced sensitivity, frozen CLIP operating-point diagnostics, threshold-free pair-margin sensitivity, exact prompt templates, the full 116-claim bilingual ontology and structured-model specifications. Source museum images are not redistributed in the workbook.