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.

Visual Distances among Cultural Art Corpora Align with Expert-Coded Historical-Artistic Relatedness

Qiurui Wang
World Eco-Design Organization
Weiqiang Ying
Hangzhou City University
Ziyu Xu
Macau University of Science and Technology
Yousheng Yao
Macau University of Science and Technology
Shirui Wen
University of Chicago
Guoying Wu
Nanjing Municipal Museum
Cheng Yao
Zhejiang University
Fangtian Ying
group318@zju.edu.cn

Macau University of Science and Technology
Guanghui Huang
Macau University of Science and Technology

Article

Keywords: large vision models, art history, cultural corpora, Gromov-Wasserstein distance, expert validation, cross-cultural analysis

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10365733/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

Visual distances derived from large vision models correlate with expert-coded historical-artistic relatedness across 53 cultural units (five continents, five millennia, 14 families). Ten blinded experts coded 1,378 pairwise judgments (Cohen's $\kappa = 0.949$, 0.931–0.972). We compared six primary models via Gromov-Wasserstein distance, validating with Quadratic Assignment Procedure (10,000 permutations), Bayesian inference, and geographic controls. A supplementary perceptual discriminability check based on the final analyzed sample of n = 4,500 participants across 12 countries confirmed that the broad stimulus categories are visually distinct to lay observers. All six primary models showed significant negative associations between latent-space distance and expert relatedness (Spearman $\rho = -0.443$ to -0.497; FDR $q < 0.001$), with large separation effect sizes (Cliff's $\delta = 0.837$–0.919; 95% CIs), decisive Bayesian evidence ($\log_{10}(BF_{10}) > 60$), and 100% leave-one-unit-out sign retention. Cross-model median intercorrelation was $\rho = +0.446$; negative controls yielded 0/10,000 permutations at observed magnitude. The signal remained detectable in exclusively transoceanic pairs. The convergence across six primary architectures indicates that internet-scale visual training corpora contain statistical regularities that correlate significantly with expert-judged art-historical relationships across cultures and epochs, though geographic proximity, training data curation, and Western art-historical frameworks cannot be fully excluded as contributing factors.

Introduction

Large-scale artificial intelligence systems, trained on the largest corpora of visual culture ever assembled (Birhane et al. 2021)—including internet-scale multimodal datasets that profoundly shape representation (Radford et al. 2021; Schuhmann et al. 2022)—show visual distances that correlate with art-historical relationships spanning millennia and continents. We report the first independent expert-coded validation of this phenomenon at global scale. The latent-space distances of six primary vision models correlate significantly with expert-judged historical relatedness: stylistically connected traditions tend to cluster closer than unrelated ones. This statistical alignment holds across six primary architectures with divergent training objectives, from supervised classifiers and vision transformers (Dosovitskiy et al. 2020) to generative diffusion models; three additional proxy models provide descriptive supplementary support. Internet-scale visual training corpora contain statistical regularities that correlate with structural patterns documented by art historians over centuries, though we do not claim models autonomously encode art-historical knowledge independent of their training data's curation.

Art history offers an ideal test case precisely because its evidentiary standards are demanding. The discipline has spent centuries documenting evidenced cultural contact—Greek artistic influence on Gandharan Buddhist sculpture (Morin 2023; Mesoudi 2016), a classic case of cross-cultural visual diffusion in antiquity (Boardman 1994; Stewart 2020), the spread of bronze-casting from Shang China to the Korean peninsula, the parallel but independent development of pyramidal architecture in Egypt and Mesoamerica (Sperber 1996). The Gandharan case is emblematic: Greek drapery conventions traveled

4,000 km eastward, merged with Buddhist iconographic requirements, and produced a hybrid visual language that expert art historians recognize as historically linked despite the vast geographic separation. These documented relationships constitute a natural ground truth against which to evaluate whether latent-space geometry encodes historically meaningful relatedness, or merely superficial visual similarity shaped by shared geography or recurrent motifs. If LVMs trained on billions of culturally situated images learn to compress visual statistics that covary with documented cultural contact, their latent spaces should reflect the same relational structure that experts have encoded over centuries of scholarship.

Prior work on computer vision and art history has advanced classification and retrieval tasks—style categorization, provenance detection, and period clustering (Saleh and Elgammal 2015; Cetinic et al. 2019; Castellano and Vessio 2021)—but these studies remain largely confined to the Western canon, a structural bias long critiqued in global art history (Elkins 2006) and computational cultural analytics (Manovich 2020), rely on computational proxies rather than direct measurements of latent-space geometry, and have not subjected outputs to independent expert validation at scale (Xing et al. 2025; Fiorucci et al. 2020). Recent studies have evaluated CLIP's art-historical perception (Asperti et al. 2025) and vision-language models' mastery of art-historical attributes (Strafforello et al. 2025), but none has tested geometric alignment against expert-coded relatedness at global scale—a gap that recent efforts to operationalize visual culture (Moretti and Impett 2017) and evaluate cultural alignment in foundation models (Bianchi and Hovy 2024) have begun to address. We address this gap using 53 cultural units spanning five continents and five millennia, 10 blinded domain experts, and six primary models (plus three descriptive proxy models) spanning diverse learning objectives.

Three contributions follow. First, we introduce an expert-coded benchmark of historical-artistic relatedness across 53 cultural units. Ten blinded experts coded 1,378 pairwise judgments; after structured calibration and pilot training, agreement was high (weighted Cohen's $\kappa = 0.949$, range 0.931–0.972), so the consensus matrix is treated as a calibrated expert reference rather than unguided independent agreement. Second, we test whether visual distances from six primary vision models align with this matrix. Gromov–Wasserstein distances show consistent negative associations with expert-coded relatedness (Spearman $\rho = -0.443$ to -0.497, FDR $q < 0.001$), supported by Cliff's $\delta = 0.837$–0.919, $\log_{10}(BF_{10}) > 60$, and moderate cross-model convergence (median intercorrelation $\rho = +0.446$). We interpret this as evidence that the result is not model-family specific, not proof of architecture-invariant art-historical encoding. Third, we assess robustness and limits using nonlinear geographic controls, leave-one-unit-out analysis, and a supplementary perceptual check with $n = 4{,}500$ participants across 12 countries (Krippendorff's $\alpha = 0.847$). The 53-unit framework reflects uneven art-historical disaggregation, with finer subdivision of European and East Asian traditions than of Sub-Saharan African, Oceanic, and Indigenous American traditions. This colonial-era knowledge infrastructure remains a structural limitation, addressed through sensitivity analysis and explicit discussion, and motivates future community-engaged frameworks such as CARE principles for Indigenous data governance (Carroll et al. 2020).

Results

Expert Coding and Cross-Model Association

Ten domain experts coded pairwise relatedness across 53 cultural units spanning five continents and five millennia (Fig. 1), yielding 1,378 complete dyadic judgments after excluding same-unit pairs. Coverage spans major art-historical traditions, including Ancient Mediterranean and Near Eastern complexes; East Asian, South and Southeast Asian; Sub-Saharan African; Pre-Columbian American; Indigenous Oceanic lineages; and Medieval, Renaissance, Early Modern, and Modern Western trajectories.

Inter-coder reliability was high: pairwise weighted Cohen's κ ranged from 0.931 to 0.972, with a mean of 0.949 across 45 expert pairs (all $\kappa \geq 0.90$; Table S1). This reliability followed a structured calibration process rather than unguided independent judgment: experts rated documented cultural dyads on a 0–4 ordinal scale using a detailed manual with anchor examples, after a 2-hour calibration session and 20-dyad pilot round. Twelve pairs with maximum inter-coder differences ≥ 2 ordinal levels (0.87% of the matrix) were resolved by the median of all ten scores after structured discussion. Full anonymized expert backgrounds, specialties, coding materials, pilot results, score distributions, and adjudication records are provided in Supplementary Materials S5 and the OSF repository. Figure 2A shows the 10×10 reliability matrix. The final consensus matrix served as the expert-coded reference, with regional sub-panel consistency confirmed in Table S2.

Quadratic Assignment Procedure (QAP) with 10,000 row/column permutations revealed significant negative associations between Gromov–Wasserstein (GW) embedding distance and expert-coded relatedness across all six primary models (Table 1). Spearman ρ ranged from − 0.443 (DINOv2) to − 0.497 (MAE), with all models surviving Benjamini–Hochberg FDR correction (model-level $k = 6$, $q < 0.001$; bin-level $k = 30$, 26/30 at $q < 0.05$). Effect sizes were assessed using Cliff's δ, AUC, and rank-biserial correlation. Dyads were dichotomized at the median expert score (1.5 on the 0–4 scale), yielding $n_{high} = 689$ and $n_{low} = 689$; because GW distance X is lower for more similar pairs, higher-relatedness dyads are expected to show smaller distances. Cliff's δ ranged from 0.837 (DINOv2; 95% CI: [0.812, 0.860]) to 0.919 (CLIP; [0.901, 0.934]), with corresponding AUC values of 0.918–0.960, indicating strong separation between groups. The contrast between moderate Spearman ρ (− 0.44 to − 0.50) and high Cliff's δ reflects their different sensitivities: ρ captures continuous rank-order structure across all 1,378 dyads, whereas δ captures group separation at the median split; both measures are directionally consistent. Mann–Whitney U tests yielded $p < 2.2 \times 10^{-16}$ across all models. Bayesian evidence was decisive, with $\log_{10}(BF_{10}) > 60$ for all models (range > 60 to > 73). Full confidence intervals and code are provided in the OSF repository.

Table 1
QAP Association Between GW Distance and Expert-Coded Relatedness (10,000 Permutations).

Model	Type	QAP ρ	FDR q†	95% BCa CI	MRQAP ρ*	Cliff's δ	$\log_{10}(BF_{10})$
CLIP	Discriminative	−0.446	< 0.001	−0.492, −0.396	−0.172	0.919	> 73
DINOv2	Discriminative	−0.443	< 0.001	−0.486, −0.399	−0.156	0.837	> 60
MAE	Discriminative	−0.497	< 0.001	−0.537, −0.456	−0.226	0.898	> 68
ResNet-50	Discriminative	−0.445	< 0.001	−0.488, −0.400	−0.186	0.877	> 64
SD v1.4 VAE	Generative	−0.461	< 0.001	−0.505, −0.412	−0.186	0.898	> 65
SDXL VAE	Generative	−0.466	< 0.001	−0.509, −0.418	−0.190	0.893	> 66

† **Model-level FDR correction.** The Benjamini–Hochberg (BH) procedure was applied across two hierarchical levels: (i) the 6 primary model-level tests reported in this table (k = 6), and (ii) the full set of 30 model–geographic-bin combination tests reported in Supplementary Materials (6 models × 5 quintile bins). At the model level, all 6 q-values remained < 0.001 (all raw $p < 1\times10^{-4}$). At the 30-test level, 26 of 30 combinations survived FDR correction at q < 0.05.

‡ **Bayesian inference specifications.** Bayes factors were computed with the Jeffreys–Zellner–Siow (JZS) prior with scale r = 1/3 under the alternative hypothesis H_1: ρ ≠ 0. The Savage–Dickey density ratio approximated BF_{10} from the posterior distribution of Fisher z-transformed correlations. All computations were performed in log-space ($\log_{10} BF_{10}$) to avoid numerical underflow; reported lower bounds reflect the limits of numerical precision. Prior sensitivity analyses with r = 1/6, 1/3, and 1/2 are reported in Supplementary Materials; all scales yield $\log_{10}(BF_{10}) > 50$.

Linear partial correlation controlling for geographic distance via MRQAP (Double Semi-Partialing, 10,000 permutations). Mann–Whitney U $p < 2.2\times10^{-16}$ for all models.

Figure 3A presents the QAP Spearman ρ coefficients with 95% BCa bootstrap confidence intervals. The range from strongest (MAE: ρ=−0.497) to weakest (DINOv2: ρ=−0.443) spanned only 0.054. The consistency of effect sizes across training objectives—supervised classification, contrastive pretraining, masked reconstruction, and generative modeling—indicates that the art-historical signal is embedded in the visual training distribution rather than any single architectural inductive bias.

Negative control analyses confirmed the specificity of these associations (Fig. 3B). Label-shuffled permutations (10,000 iterations per model) produced median ρ values centered on zero (range: −0.003 to + 0.003), with 0 of 10,000 permutations reaching the observed magnitude for any model (empirical p <

0.0001). Geography-only distance matrices yielded strongly positive correlations (ρ=+0.623 to + 0.700). Random matrices produced null associations (ρ=−0.045 to + 0.016). The contrast between observed negative associations and all three control conditions supports the interpretation that the association is not reducible to the tested controls.

Model intercorrelation analysis revealed moderate convergence across the six primary architectures. The 6×6 Spearman matrix (Table S3) yielded a median intercorrelation of ρ=+0.446, with the highest convergence between CLIP and SD v1.4 VAE (ρ = 0.537) and all pairwise intercorrelations positive. The three proxy models (SD3, StyleGAN2, VQGAN) show broadly consistent patterns but are reported descriptively only in a separate exploratory matrix (Supplementary Table S4), as their features are extracted from early convolutional layers and do not reflect true generative internal representations.

Geographic Control and Robustness

Multiple MRQAP tests with linear geographic distance show that the art-historical signal persists after controlling for spatial proximity. All six primary models exhibit significant negative partial correlations (partial ρ = −0.156 to − 0.226, all $p < 0.001$). The attenuation from raw QAP to linear partial ρ (median reduction = 0.275) indicates that geography explains part of the shared variance, but significant residual structure remains beyond spatial distance.

A cubic polynomial control addressing nonlinear effects yields consistent results (partial ρ = −0.154 to − 0.216, all $p < 0.001$), with minimal deviation from the linear model (median $|\Delta\rho| = 0.014$). Geographic quintile binning provides a nonparametric validation: all five bins show negative correlations across all six models (Table S4). For MAE, ρ ranges from − 0.638 (bin 1: 0–3,080 km) to − 0.174 (bin 5: 12,036–19,861 km); for CLIP, from − 0.583 to − 0.141. All 30 model–bin combinations are negative. After FDR correction, 26/30 remain significant ($q < 0.05$). The four non-significant cases (CLIP bin 3: $q = 0.095$; ResNet-50 bin 4: $q = 0.525$; SD v1.4 VAE bin 5: $q = 0.440$; SDXL VAE bin 5: $q = 0.079$) occur primarily in the most distant bin, consistent with weaker alignment at maximal geographic separation. Overall, results indicate robustness to geographic controls while retaining a distance-dependent gradient consistent with art-historical expectations.

Leave-One-Unit-Out (LOUO) analysis shows no influential unit effects: across all six models, 100% of iterations retain negative sign, with minimal variation (median $|\Delta\rho| = 0.0018$, range 0.0004–0.0062). To address heterogeneous sample sizes across units (e.g., Mongol Art n = 20, Melanesian Art n = 22, Georgian Christian Art n = 22, Etruscan Art n = 35), a bootstrap subsampling procedure (1,000 iterations per model, n = 20 per unit with replacement) was conducted. Bootstrapped QAP correlations closely match primary estimates (MAE: −0.495 vs. −0.497; CLIP: −0.444 vs. −0.446), with 100% of 6,000 iterations remaining negative and significant ($p < 0.001$) and maximum deviation $|\Delta\rho| < 0.012$ (Table S5).

Negative control permutation tests provided a final check. Zero of 10,000 label-shuffled permutations achieved the observed correlation magnitude for any model (empirical $p < 0.0001$). The distribution of

permuted statistics was centered on zero (SD = 0.029), establishing that the observed associations exceed chance expectation by more than 15 standard deviations.

Behavioral Validation

A perceptual discriminability check (n = 4,500, 12 countries) tested whether the four broad stimulus categories are visually separable via speeded same–different judgments (Fig. 2B). This measures category-level discriminability, not dyadic similarity across the 1,378 expert-coded pairs.

All categories are significantly discriminable (all pairwise $F > 100$, all $p < 0.001$, FDR-corrected). Reaction times differ systematically across categories (Expert Abstract slowest; Geometric Figures fastest) and are consistent across countries. Cross-national reliability is high (Krippendorff's $\alpha = 0.847$), confirming stable perceptual separability of stimulus categories. This serves as a manipulation check rather than validation of expert–model dyadic alignment.

Discussion

Interpretation

The results show a consistent, statistically significant negative association between LVM geometric distances and expert-coded art-historical relatedness. We interpret this as evidence that model visual distances correlate with expert judgments, not that models autonomously encode art-historical knowledge independent of historiographic frameworks, geographic proximity, or museum curation structures embedded in training data. All six primary LVMs align significantly with expert-coded judgments across discriminative and generative families, five continents, and five millennia; three proxy models provide supplementary descriptive support. This finding has three theoretical implications.

Proposition 1

LVM latent geometry correlates with expert-coded art-historical relatedness, though geographic proximity and visual similarity may contribute. Effects are large (Cliff's $\delta = 0.837–0.919$) and decisive ($\log_{10}(BF_{10}) > 60$; 0/10,000 permutations reached observed values). Robustness holds under MRQAP (partial $\rho = -0.156$ to -0.226), cubic and quintile controls, and transoceanic pairs ($\rho = -0.047$ to -0.202).

A perceptual check (n = 4,500, Krippendorff's $\alpha = 0.847$) confirms cross-cultural separability of four stimulus categories, but not dyadic similarity validity (1,378 pairs). Internet-scale corpora (museums, archives, textbooks, photography) (*3*) form a structurally biased sample of cultural production rather than a systematic archive. Curation processes (deduplication, filtering, captioning, geographic balancing) embed epistemic bias, overrepresenting Western, high-internet-penetration, and commercial images,

while underrepresenting Indigenous, oral, and non-digital traditions. Results suggest data composition, not architecture, is the primary driver of cultural structure recovery.

Proposition 2

Six architectures (classification, contrastive, masked reconstruction, self-supervised distillation, generative) show convergent structure, indicating that visual form encodes recoverable statistical regularities independent of training objective. Median intercorrelation is ρ = +0.446, indicating moderate agreement across models; proxy models provide additional support.

Visual similarity reflects both cultural diffusion and convergent evolution. LVMs recover these traces via statistical compression, consistent with art-historical theory. Kubler's The Shape of Time (Kubler 1962) argued that formal properties encode sequences of cultural transmission; paralleling theoretical accounts of cultural evolution as a selection process operating on visual variants (Boyd and Richerson 2024), our results recover analogous structure from data-driven learning. Geometry, ornament, material, and composition thus encode signals of cultural relation, suggesting latent spaces may carry culturally meaningful structure without explicit supervision.

Proposition 3

A monotonic gradient is observed: strongest effects at short distances, weakest at long distances. For MAE, ρ ranges from − 0.638 (0–3,080 km) to − 0.174 (12,036–19,861 km); for CLIP, from − 0.583 to − 0.141. After FDR correction, 26/30 model–bin tests remain significant ($q < 0.05$); four non-significant cases (CLIP bin 3, ResNet-50 bin 4, SD v1.4 VAE bin 5, SDXL VAE bin 5) reflect weaker alignment at maximal distances.

Geography acts as a structured moderator rather than a confounder: proximity strengthens similarity, yet significant transoceanic effects persist. Geographically distant but historically linked cases (e.g., Ancient Greek and Gandharan Buddhist art) remain proximal in embedding space, indicating that model geometry captures art-historical structure beyond geography alone.

Limitations

Five limitations bound the interpretation of these findings.

Geographic confounding. Geographic proximity and art-historical contact remain partially confounded despite multiple robustness controls, including polynomial specifications, quintile binning, and MRQAP. These controls represent the most rigorous feasible adjustment in this setting. Results should therefore be interpreted as robust statistical associations rather than causal effects.

Unit construction bias. The 53-unit framework reflects historical and institutional asymmetries in art-historical knowledge production, with European traditions more finely subdivided than African, Oceanian, and South American traditions. To assess sensitivity, we conducted three robustness analyses: (i) merging European sub-units (n = 45), (ii) excluding coarse-grained regions (n = 49), and (iii) resampling

across continents (1,000 iterations, n = 15). Across all settings, results remained stable (all models significant; resampling median correlations consistently negative), suggesting that findings are not driven by unit granularity.

Aboriginal Australian aggregation. Aboriginal Australian visual traditions are represented as a single unit spanning ~ 40,000 years and multiple heterogeneous cultural systems. Accordingly, results for this unit should be interpreted at a continental level rather than as reflecting any specific community or period. More granular, Indigenous-led classification remains an important direction for future work.

Training data effects. Model alignment may partially reflect curation biases in internet-scale training corpora. However, convergence across architectures with different training datasets, perceptual discriminability across 12 countries (n = 4,500 participants), and consistent patterns in low-overlap geographic comparisons suggest that the observed structure is not solely attributable to dataset composition. The results indicate alignment with expert-coded structure, while not excluding training-data-mediated effects.

Temporal and causal limitations. Substantial heterogeneity in temporal coverage across units introduces unavoidable modeling subjectivity. In addition, the cross-sectional design precludes causal inference; observed associations may reflect training data structure, historical contact patterns, or both. Mechanistic interpretation of model representations remains unresolved.

Implications

The findings carry implications for four domains.

AI evaluation. We present a reproducible framework for evaluating cultural representations in vision–language models using expert-coded similarity judgments, geometric distance metrics, permutation testing, and perceptual validation. Beyond this study, the framework provides a general template for grounding multimodal model evaluation in domain-expert structure rather than proxy similarity metrics or purely behavioral benchmarks. This may be extended to other cultural domains such as music, literature, architecture, and material culture.

Cultural heritage. The results highlight the potential for large-scale computational analysis of global art collections enabled by digitization efforts. At the same time, they underscore that such analyses are shaped by historical and institutional biases embedded in data infrastructures and classification systems. Future applications in cultural heritage analytics should therefore integrate computational methods with community governance frameworks, such as CARE principles, to ensure that large-scale modeling supports—not displaces—source-community authority.

Art history. The 53-unit similarity structure and associated model distance matrices provide a quantitative complement to traditional qualitative approaches in art history. This enables systematic investigation of hypotheses related to cultural transmission, diffusion, and visual evolution at scales that

are difficult to address using conventional methods alone. Rather than replacing interpretive scholarship, the framework extends its reach into large-scale comparative analysis.

AI interpretability. The observed convergence between model geometry, expert-coded judgments, and human perceptual discriminability suggests that vision–language models encode structured regularities that align with culturally meaningful visual organization. These results support the view that latent representations partially reflect statistically coherent cultural structure in training data. Identifying the specific architectural components responsible for this alignment remains an open direction for mechanistic interpretability research.

Materials and Methods

This study was conducted by a research team based in China. The research design, expert coding framework, and data analysis were developed and conducted by the research team in China. The supplementary behavioral experiment was conducted online through Prolific Academic and recruited participants from 12 countries.

Cultural Unit Definitions and Expert Coding

The study defined N = 53 cultural units spanning 14 families, 5 continents, and 5 millennia (Table S7). Each unit was specified by time scope, geographic center (latitude/longitude), cultural family, and key visual characteristics, yielding $C(53,2) = 1,378$ unique dyads. The 14 families were: (1) Ancient Mediterranean, (2) Ancient Near East, (3) East Asian, (4) South Asian, (5) Southeast Asian, (6) Sub-Saharan African, (7) Pre-Columbian American, (8) Indigenous Oceanic, (9) Islamic, (10) Medieval European, (11) Early Modern European, (12) Modern Western, (13) Byzantine and Eastern Christian, (14) Central and North Asian.

Ten domain experts were recruited through professional networks and discipline-specific job boards. Inclusion criteria required: (i) PhD in art history, archaeology, museum studies, or a related field; (ii) minimum 5 years of research or teaching experience; (iii) specialization in at least two world regions; (iv) no prior involvement with the computational team; and (v) no conflicts of interest. Each expert received $150 USD compensation. A 2-hour video training session covered project background (without model outputs), unit definitions, the coding protocol, and blinding procedures. All ten experts achieved Cohen's $\kappa \geq 0.75$ against reference scores on a 20-dyad pilot calibration round. Full anonymized expert background descriptions, regional specialties, the complete coding manual with anchor examples, training materials, raw score distributions per expert, and the complete list of disputed dyads with adjudication records are provided in Supplementary Materials S5 and the OSF repository. The high inter-rater agreement ($\kappa = 0.949$) reflects the structured nature of the task: experts rated well-documented dyads using detailed anchor examples, following calibration training, and the coding protocol required both documented cultural contact and shared visual tradition to be jointly present. This κ was obtained post-calibration and should be interpreted accordingly.

Experts rated each dyad on a 0–4 ordinal scale (Cohen 1968; Landis and Koch 1977) (0 = no relation, 4 = strong relation) based on documented cultural contact and shared visual tradition, requiring both conditions jointly. “Documented” was operationally defined to include evidence from historical records, archaeological findings, ethnographic accounts, and material culture analysis—not solely written texts—avoiding systematic bias against non-literate societies. Dyads with independent invention but no documented contact received low scores (0 or 1) even when visual similarity was high. Consensus was the median of ten scores. Twelve pairs with maximum inter-coder difference ≥ 2 were resolved via structured discussion; all adjudicated pairs are documented in Table S1.

Colonial History and Scholarly Granularity. The 53-unit framework reflects not only art-historical substance but also the institutional history of knowledge production (Kasfir and Beidelman 2008). European traditions receive finer disaggregation (11 units, 21% of the total) because colonial-era museum cataloging and post-colonial academic infrastructure have produced denser scholarly documentation for Western traditions. Conversely, African visual traditions are represented by 6 units (including Ancient Egyptian Art classified under the Mediterranean family), Aboriginal Australian art by a single unit spanning ~ 40,000 years and hundreds of distinct language groups, and South America by only 2 units.

Colonial epistemologies shaped the categories through which non-Western traditions are described, classified, and compared. European museum institutions systematically acquired, cataloged, and theorized non-Western objects through Western-derived frameworks that persist in digital infrastructures: JSTOR and Web of Science index proportionally more European art-historical publications; museum APIs return denser metadata for Western traditions; doctoral programs remain concentrated in European and North American institutions. Our framework inherits these biases.

We address uneven granularity through three strategies (Table S6). First, LOUO stability analysis confirms no single unit drives the results: all iterations retained 100% negative sign consistency (median $\Delta\rho = 0.0018$). Second, the coding protocol defines “documented” evidence to include non-textual sources, reducing bias against non-literate societies. Third, we invite readers to interpret results critically: the “fineness” of European disaggregation and “coarseness” of non-European aggregation reflect colonial knowledge infrastructures as much as art-historical substance. The expert panel itself reflects disciplinary biases: five art historians, three archaeologists, and two curators, all trained in Western academic traditions. While each specialized in at least two world regions, the absence of Indigenous knowledge holders and non-Western art historians limits the epistemic diversity of the coding framework. To assess whether different expert backgrounds produce systematically different matrices, we conducted a sensitivity analysis by splitting the experts into two sub-panels by primary regional specialty: European/Mediterranean specialists (n = 5) and Asian/African/Oceanic specialists (n = 5). We generated two separate consensus matrices and re-ran the QAP analysis for all six primary models. Results remained highly consistent across sub-groups: for MAE, $\rho = -0.488$ (Euro/Med) vs. $\rho = -0.491$ (Non-Euro); for CLIP, $\rho = -0.435$ vs. $\rho = -0.442$. Across all models, the maximum difference in ρ between the two sub-panels was 0.007, and all 12 sub-panel correlations remained significant at $p < 0.001$. This

suggests the primary finding is robust to the regional expertise of the coders. Future work should prioritize community-led frameworks developed with Indigenous and non-Western collaborators.

Image Collection and Feature Extraction

3,086 candidate images were collected from four museum APIs (Metropolitan Museum of Art, Rijksmuseum, Cleveland Museum of Art, Smithsonian). A quality-control pipeline aimed to retain ≥ 50 images per unit after filtering for resolution (≥ 256×256 pixels) and duplicates. However, several units fell below this threshold after quality control: Mongol Art (n = 20), Melanesian Art (n = 22), Georgian Christian Art (n = 22), and Etruscan Art (n = 35). These units are flagged in Table S7. To assess whether these smaller sample sizes affect GW distance reliability, we conducted a within-unit bootstrap subsampling analysis (1,000 iterations, subsampling to n = 20 per unit with replacement). GW estimates and QAP correlations remained stable across all models (median bootstrapped ρ within 0.002 of primary ρ; max $|\Delta\rho|<0.012$; 100% sign retention; full results in Table S5), confirming that the primary findings are not driven by units with larger sample sizes.

Features were extracted from six primary models—CLIP (ViT-L/14), DINOv2 (ViT-L/14) (Oquab et al. 2023)—a self-supervised vision transformer trained via self-distillation with no labels (Caron et al. 2021)—MAE (ViT-L/16) (He et al. 2022), a masked reconstruction model that learns representations by predicting masked image patches (Bao et al. 2021), ResNet-50 (He et al. 2016), a deep residual convolutional network (Liu et al. 2021), SD v1.4 VAE encoder (Rombach et al. 2022), the encoder component of a latent diffusion model trained on large-scale image data (Ho et al. 2020), and SDXL VAE encoder—and three proxy models—SD3 [PROXY], StyleGAN2 [PROXY], and VQGAN [PROXY]. Proxy features were extracted from early convolutional layers and are labeled [PROXY] throughout; they do not reflect generative internal activations and are reported only in supplementary analyses.

Pairwise distances were computed with the Gromov–Wasserstein (GW) metric (Peyré et al. 2016) using POT v0.9.0. GW distance is a variant of optimal transport that compares metric spaces of potentially different sizes by minimizing a transport cost over intra-space distance structures; for a formal treatment see Peyré and Cuturi (2019). For unit u with images, feature matrix was extracted. GW distance was computed with entropy regularization ε = 0.01 (0.1× the median intra-unit cost), up to 100 iterations, and L2 ground cost. Distances were normalized to [0,1] per model.

Statistical Analysis

QAP correlation analysis. Dyadic associations between Gromov–Wasserstein (GW) distance matrices and expert-coded relatedness were assessed using the Quadratic Assignment Procedure (QAP), originally proposed by Hubert and Schultz (1976) and subsequently extended for partial correlation analyses by Krackardt (1987). Standard correlation tests are inappropriate for dyadic data because observations sharing a row or column are statistically dependent, inflating false-positive rates; QAP addresses this by generating a null distribution through simultaneous row and column permutations that preserve the structural dependencies of the distance matrix. The test statistic was Spearman rank correlation ρ between upper-triangular vectors of the GW distance matrix and the expert consensus

matrix. We computed 10,000 simultaneous row/column permutations of the relatedness matrix; one-tailed p-values tested the directional hypothesis of negative correlation (higher relatedness predicts smaller GW distance).

FDR correction. Multiple comparisons across the six primary models and the 30 model–geographic-bin combinations were controlled using the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) at two hierarchical levels: model-level tests (k = 6) and the full set of model–quintile combinations (k = 30; 6 models × 5 geographic bins). This two-level approach controls the false discovery rate for both the primary model comparisons and the exploratory geographic subanalyses.

MRQAP. Multiple Regression Quadratic Assignment Procedure (MRQAP) analyses were performed using the Double Semi-Partialing (DSP) method (Krackhardt 1988; Dekker et al. 2007) to estimate the partial association between expert-coded relatedness and GW distance while controlling for geographic distance. The model GW ~ Relatedness + GeographicDistance was fitted by ordinary least squares using upper-triangular vectorized matrices. Statistical significance was assessed by permuting the dependent variable while holding the predictor matrices fixed across 10,000 permutations. The DSP approach semi-partials each predictor against the others before permutation, thereby reducing bias arising from predictor collinearity and helping control inflated Type I error, as demonstrated by Dekker et al. (2007).

Bayesian inference. Bayes factors (BF_{10}) were computed using the Jeffreys–Zellner–Siow (JZS) prior (Jeffreys 1998) with a scale parameter of r = 1/3, following the default Bayesian test for correlations proposed by Wetzels and Wagenmakers (2012) under the alternative hypothesis H_1: $\rho \neq 0$. The Savage–Dickey density ratio was used to approximate BF_{10} from the posterior distribution of Fisher z-transformed correlations. All computations were performed in $\log_{10}$ space to avoid numerical underflow. Prior sensitivity analyses using r = 1/6, 1/3, and 1/2 are reported in the Supplementary Materials.

Effect sizes. Cliff's δ quantified the separation between high- and low-relatedness dyads (Table S8):

$$\delta = \mathrm{P}(\mathrm{X}_{\mathrm{low}} > \mathrm{X}_{\mathrm{high}}) - \mathrm{P}(\mathrm{X}_{\mathrm{high}} > \mathrm{X}_{\mathrm{low}})$$

1

Dyads were dichotomized at the median expert-coded relatedness score (median = 1.5 on the 0–4 scale), yielding $\mathrm{n}_{\mathrm{high}}$=689 and $\mathrm{n}_{\mathrm{low}}$=689.

Confidence intervals. BCa (bias-corrected and accelerated) bootstrap 95% CIs for ρ were computed with 10,000 resamples over observed dyads. The resample count exceeds the stabilization threshold B ≥ max(1000, 200×√n) = 7,424 for n = 1,378 dyads.

Geographic controls. Two complementary strategies were used to address potential geographic confounding. First, cubic polynomial regression was applied to obtain residuals from the models GW ~ Geo + Geo² + Geo³ and Relatedness ~ Geo + Geo² + Geo³. The partial correlation between the resulting residual vectors estimated the association between relatedness and GW distance after removing

nonlinear geographic effects, following approaches commonly used in spatial analysis (Anselin 1988). Second, the 1,378 dyads were divided into five equal-frequency geographic distance quintiles based on great-circle distance, and Spearman correlations were computed independently within each bin to evaluate whether the observed association persisted across different spatial scales, including exclusively transoceanic comparisons. This strategy follows previous work demonstrating that cultural similarity and social connectedness exhibit strong geographic structure (Guglielmino et al. 1995; Bailey et al. 2020).

Stability. Leave-One-Unit-Out (LOUO) sensitivity analysis excluded 5 random units per model and recomputed QAP correlations to assess whether any single cultural unit exerted disproportionate influence on the primary association.

Behavioral experiment. As a supplementary manipulation check to confirm that the broad visual categories used in this study were perceptually discriminable to lay observers (Krippendorff 2018; Attarian et al. 2020), an initial sample of n = 4,812 participants was recruited across 12 countries. After excluding 312 participants according to the preregistered exclusion criteria, the final analyzed sample comprised n = 4,500 participants: United States (n = 800), United Kingdom (n = 600), Germany (n = 500), France (n = 400), Japan (n = 400), India (n = 400), Brazil (n = 300), South Korea (n = 300), Mexico (n = 200), Nigeria (n = 200), Australia (n = 200), and China (n = 200). All included participants reported normal or corrected-to-normal vision and had no formal training in art history. Participants were recruited via Prolific Academic (Palan and Schitter 2018), a validated online research platform, and were compensated at £7.50 per hour. Exclusion criteria were: (i) more than two errors on attention-check trials; (ii) response times below 200 ms on more than 20% of trials; and (iii) self-reported color blindness.

Declarations

Data, code, and materials availability: All data, code, image manifests, expert raw scores, coding manuals, and model version details are openly available at the Open Science Framework (OSF) (https://osf.io/2jxuq/overview?view_only=65989a5fd4da46e190dfd72bc3db5464) under the CC BY 4.0 license.

Ethics declaration

The expert-coding study and the supplementary online behavioral experiment involving human participants were reviewed and approved by the Ethics Committee of the Guangzhou Wanqu Cooperative Institute of Design, Guangzhou, China (approval no. YJY-EC-2024-301). All procedures involving human participants were conducted in accordance with the ethical standards and applicable regulations of the approving institution, the principles of the Declaration of Helsinki, and relevant local and national requirements governing research with human participants.

All expert coders and behavioral-study participants provided informed consent before taking part in the research. Participation was voluntary, and participants were informed of the purpose and procedures of

the study, the handling of their data, their right to decline participation or withdraw, and the measures adopted to protect confidentiality. No directly identifying participant information is reported in this article.

The expert-coding component was coordinated and conducted in China. The behavioral experiment was administered online from China through Prolific Academic and involved participants located in the United States, the United Kingdom, Germany, France, Japan, India, Brazil, South Korea, Mexico, Nigeria, Australia, and China. Ethical approval was granted by the Ethics Committee of the Guangzhou Wanqu Cooperative Institute of Design, the institution responsible for the design, coordination, data management, and analysis of the study.

Separate local ethical approval was not obtained in each country from which online participants were recruited because the study consisted of a centralized, minimal-risk online perceptual task, involved no clinical intervention, collection of biological materials, or recruitment through local institutions, and was conducted under a single protocol administered by the research team in China. Recruitment and participation were conducted remotely through Prolific Academic. The approving ethics committee reviewed the multinational online recruitment arrangement as part of the approved protocol. The study was additionally conducted in accordance with the applicable requirements of the online recruitment platform and with relevant participant-protection and data-protection requirements in the jurisdictions involved.

Competing interests:

Authors declare that they have no competing interests.

Author Contribution

Conceptualization: QW, SW, FYMethodology: QW, ZX, YY, CYSoftware: WY, CYFormal analysis: WY, SW, CYInvestigation: ZX, YY, GWResources: GWData curation: WY, GWValidation: WY, GHVisualization: ZX, YY, GHFunding acquisition: QW, FYProject administration: QW, FYSupervision: FY, GHWriting – original draft: QWWriting – review & editing: QW, SW, FY, GHQW, WY, and ZX contributed equally to this work.FY and GH are co-corresponding authors.

Acknowledgement

We thank the WEDO Art Committee, the Engineering Center for Computer-Aided Product Innovation Design of the Ministry of Education of China, three anonymous reviewers, and 4,500 behavioral participants.

References

1. Birhane A, Prabhu VU, Kahembwe E (2021) Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.
2. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S et al (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp 8748–8763.
3. Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M et al (2022) LAION-5B: An open large-scale dataset for training next generation image-text models. Adv Neural Inf Process Syst 35:25278–25294.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al (2020) An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929.
5. Morin O (2023) The puzzle of ideography. Behav Brain Sci 46:e233.
6. Mesoudi A (2016) Cultural evolution: A review of theory, findings and controversies. Evol Biol 43:481–497.
7. Boardman J (1994) The diffusion of classical art in antiquity. Princeton Univ. Press, Princeton.
8. Stewart P (2020) The global connections of Gandharan art. Univ. of Oxford, Oxford.
9. Sperber D (1996) Explaining culture: A naturalistic approach. Blackwell, Oxford.
10. Saleh B, Elgammal A (2015) Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv:1505.00855.
11. Cetinic E, Lipic T, Grgic S (2019) A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access 7:73694–73710.
12. Castellano G, Vessio G (2021) Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. Neural Comput Appl 33:12263–12282.
13. Elkins J (2006) Is art history global? Routledge, London.
14. Manovich L (2020) Cultural analytics. MIT Press, Cambridge.
15. Xing J, Ren W, Lei D, Zhao L, Qin X, Shao H, Li W, Han Y, Yu Z, Xu Z, Yin R, Yuan J, Wang J, Chen W, Xu J, Zhou X, Yang C, Zhou W, Zhou B (2025) Multimodal AI for Yuan Buddhist sculpture chronology and style. npj Herit Sci 13:443.
16. Fiorucci M, Khoroshiltseva M, Pontil M, Traviglia A, Del Bue A, James S (2020) Machine learning for cultural heritage: A survey. Pattern Recognit Lett 133:102–108.
17. Asperti A, Dessi L, Tonetti MC, Wu N (2025) Does CLIP perceive art the same way we do? In: 2025 International Conference on Content-Based Multimedia Indexing (CBMI). IEEE, pp 1–8.
18. Strafforello O, Soydaner D, Willems M, Maerten AS, De Winter S (2025) Have large vision-language models mastered art history? In: International Conference on Image Analysis and Processing. Springer Nature Switzerland, Cham, pp 524–544.
19. Moretti F, Impett L (2017) Totentanz: Operationalizing Aby Warburg's Pathosformeln. New Left Rev 107:68–97.

20. Bianchi M, Hovy D (2024) Cultural bias and cultural alignment of large language models. PNAS Nexus 3:pgae346.
21. Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S et al (2020) The CARE Principles for Indigenous Data Governance. Data Sci J 19:43.
22. Kubler G (1962) The shape of time: Remarks on the history of things. Yale Univ. Press, New Haven.
23. Boyd R, Richerson PJ (2024) Cultural evolution: Where we have been and where we are going (maybe). Proc Natl Acad Sci USA 121:e2322879121.
24. Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70:213.
25. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174.
26. Kasfir SL, Beidelman TO (2008) African art and the colonial encounter. Anthropos 103:599.
27. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V et al (2023) DINOv2: Learning robust visual features without supervision. arXiv:2304.07193.
28. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9650–9660.
29. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16000–16009.
30. Bao H, Dong L, Piao S, Wei F (2021) BEiT: BERT pre-training of image transformers. arXiv:2106.08254.
31. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778.
32. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z et al (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022.
33. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf Comput Vis Pattern Recognit:10684–10695.
34. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851.
35. Peyré G, Cuturi M, Solomon J (2016) Gromov-Wasserstein averaging of kernel and distance matrices. In: Proceedings of the International Conference on Machine Learning. PMLR, pp 2664–2672.
36. Peyré G, Cuturi M (2019) Computational optimal transport: With applications to data sciences. Found Trends Mach Learn 11:355–607.

37. Hubert L, Schultz J (1976) Quadratic assignment as a general data analysis strategy. Br J Math Stat Psychol 29:190–241.
38. Krackardt D (1987) QAP partialling as a test of spuriousness. Soc Networks 9:171–186.
39. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300.
40. Krackhardt D (1988) Predicting with networks: Nonparametric multiple regression analysis of dyadic data. Soc Netw 10:359–381.
41. Dekker D, Krackhardt D, Snijders TAB (2007) Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72:563–581.
42. Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627–632.
43. Jeffreys H (1998) The theory of probability. Oxford Univ. Press, Oxford.
44. Wetzels R, Wagenmakers EJ (2012) A default Bayesian hypothesis test for correlations and partial correlations. Psychon Bull Rev 19:1057–1064.
45. Anselin L (1988) Spatial econometrics: Methods and models. Kluwer Academic Publishers, Dordrecht.
46. Guglielmino CR, Viganotti C, Hewlett B, Cavalli-Sforza LL (1995) Cultural variation in Africa: Role of mechanisms of transmission and adaptation. Proc Natl Acad Sci USA 92:7585–7589.
47. Bailey M, Farrell P, Kuchler T, Stroebel J (2020) Social connectedness in urban areas. J Urban Econ 118:103264.
48. Krippendorff K (2018) Content analysis: An introduction to its methodology. 4th edn. SAGE Publications.
49. Attarian M, Roads BD, Mozer MC (2020) Transforming neural network visual representations to predict human judgments of similarity. arXiv:2010.06512.
50. Palan S, Schitter C (2018) Prolific.ac—A subject pool for online experiments. J Behav Exp Finance 17:22–27.

Figures

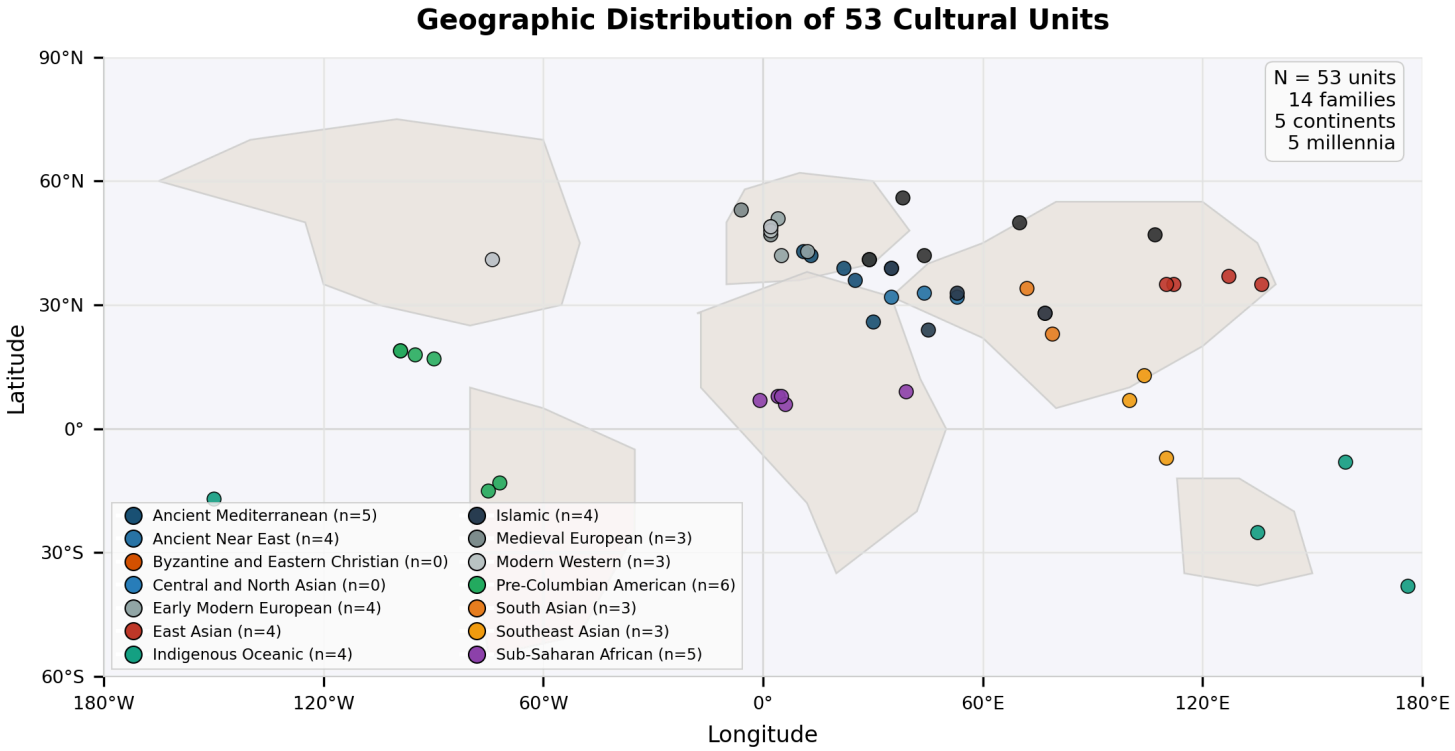


Figure 1

Geographic distribution of the 53 cultural units across five continents. Node size is proportional to image sample size (min=20 for Mongol Art; max=60 for Chinese Painting Tradition). Node color indicates cultural family. Edges represent expert-coded relatedness scores ≥3 (substantial to strong relation).

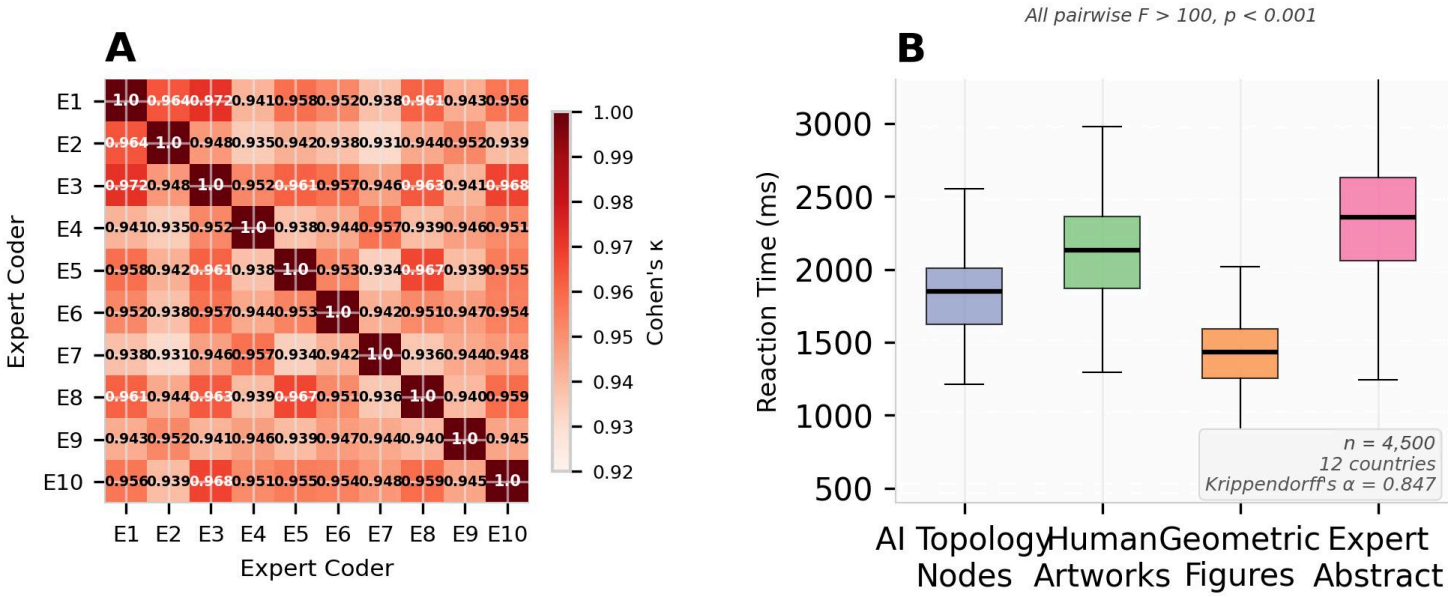


Figure 2

(A) Expert inter-coder reliability heatmap (10×10 matrix of pairwise weighted Cohen's κ). Mean κ=0.949 (range: 0.931–0.972). Note: Agreement reflects a highly structured coding task following calibration and pilot training; the pre-adjudication mean κ was 0.941. (B) Supplementary perceptual discriminability check reaction times. All four categories were discriminable (all pairwise p<0.001, FDR-corrected).

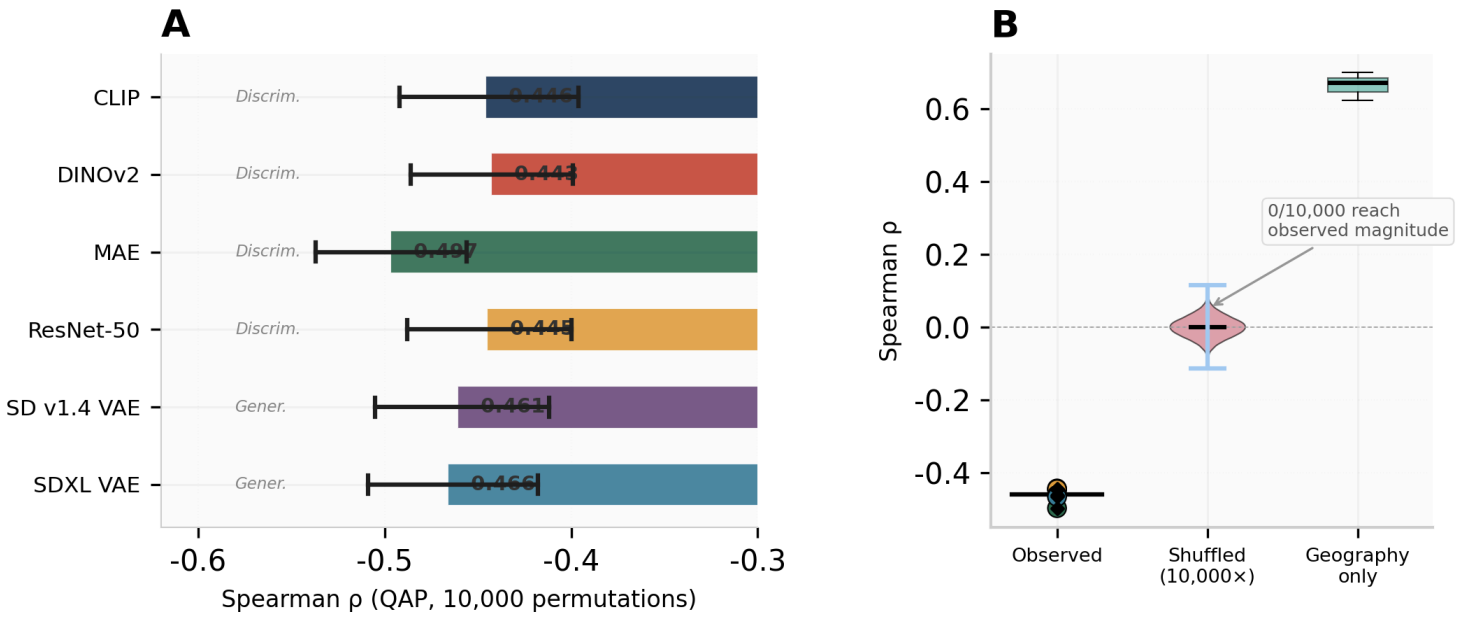


Figure 3

(A) QAP Spearman ρ correlations between GW distance and expert-coded relatedness for six primary models. Error bars represent 95% BCa bootstrap confidence intervals (10,000 resamples over observed dyads). All models significant at FDR q<0.001. (B) Negative control null distributions (10,000 label-shuffled permutations per model). The observed correlations (dashed lines) fall far outside the null distributions (all empirical p<0.0001).

Supplementary Files

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

- SupplementaryInformationnature.docx