VULCA-BENCH: A Multicultural Vision-Language Benchmark for Evaluating Cultural Understanding

Anonymous ACL submission

Abstract

We introduce VULCA-BENCH, a multicultural art-critique benchmark for evaluating Vision–Language Models' (VLMs) cultural understanding beyond surface-level visual perception. Existing VLM benchmarks predominantly measure L1–L2 capabilities (object recognition, scene description, and factual question answering) while under-evaluate higher-order cultural interpretation. VULCA-BENCH contains 7,410 matched image–critique pairs spanning eight cultural traditions, with Chinese–English bilingual coverage. We operationalise cultural understanding using a five-layer framework (L1–L5, from Visual Perception to Philosophical Aesthetics), instantiated as 225 culture-specific dimensions and supported by expert-written bilingual critiques. Our pilot results indicate that higher-layer reasoning (L3–L5) is consistently more challenging than visual and technical analysis (L1–L2). The dataset, evaluation scripts, and annotation tools are available under CC BY 4.0 in the supplementary materials.

1 Introduction

Vision-Language Models (VLMs) achieve remarkable performance on visual understanding tasks, including 90%+ on object detection (Li et al., 2023) and 77.2% on VQAv2 (Goyal et al., 2017). However, these benchmarks predominantly assess L1–L2 visual perception, which involves recognising objects, describing scenes, and answering factual questions. They fail to evaluate *cultural understanding*, namely the capacity to interpret symbolic meanings, appreciate aesthetic traditions, and engage with philosophical concepts embedded in visual content. Recent studies confirm this gap: GPT-4o achieves only 54.1% accuracy on Chinese traditional painting (Zeng et al., 2024), and cross-cultural evaluations reveal significant Western bias in VLM performance (Liu et al., 2025).

We define cultural understanding using a five-layer framework that distinguishes levels of interpretation: L1 (visual perception), L2 (technical analysis), L3 (cultural symbolism), L4 (historical context), and L5 (philosophical aesthetics). Higher layers require progressively deeper cultural knowledge (Section 3 provides full definitions). Consider a Chinese ink painting of plum blossoms (梅花): a VLM may correctly identify "plum blossoms" and "ink wash technique" (L1–L2), yet miss the symbolic meaning of *resilience against adversity* (L3) (Zhang, 2024), the artist's lineage within the "Four Gentlemen" tradition (L4), or the aesthetic principle of 气韵生动 (*qiyun shengdong*, "spirit resonance") that defines Chinese painting philosophy (L5) (Jing, 2023). This example illustrates that cultural understanding is not a single capability but a progression across interpretive levels, a progression that existing benchmarks fail to measure.

Recent cultural AI datasets (CulturalBench (Chiu et al., 2025), CultureAtlas (Fung et al., 2024), GIMMICK (Schneider et al., 2025)) begin to address this gap. Initial work on Chinese painting critique (Yu et al., 2025a) introduced quantitative VLM evaluation using 163 expert commentaries, revealing significant VLM-expert divergence. Complementary probing (Yu et al., 2025b) found VLMs rely on "symbolic shortcuts," excelling on Western festivals but failing on underrepresented traditions. However, these studies focused on single cultures; systematic cross-cultural evaluation remains absent. Existing datasets remain L1-centric (over 90% of metrics focus on visual perception), lack hierarchical frameworks distinguishing basic visual analysis from higher-order symbolic reasoning, and exhibit Western bias by underrepresenting Asian, Middle Eastern, and South Asian traditions.

We introduce VULCA-BENCH (**V**ision-**U**nderstanding-**L**anguage-**C**ulture **A**ssessment), a comprehensive multicultural art critique benchmark addressing these gaps. VULCA-BENCH provides 7,410 matched image-critique pairs across

8 traditions with 100% bilingual coverage and 98% cultural-fact accuracy, operationalised through our five-layer framework (L1–L5) spanning 225 culture-specific dimensions that preserve Cultural Symmetry (i.e., equal methodological treatment across cultures regardless of sample size). All critiques meet expert-quality standards (Chinese ≥150 characters, English ≥100 words) with cultural specialist validation. The complete dataset is available under CC BY 4.0 in the supplementary materials.

VULCA-BENCH enables systematic evaluation of VLMs' hierarchical cultural understanding, identification of L1–L5 performance gaps, equal-weighted cultural fairness probing on the balanced-pilot subset (N=336, 7 cultures), and development of culturally grounded VLM architectures. Figure 1 showcases representative artworks from each tradition.

This paper makes three contributions. (1) We present VULCA-BENCH, a multicultural art-critique benchmark enabling evaluation of cultural interpretation beyond surface-level perception, comprising 7,410 image–critique pairs across eight traditions with 225 expert-defined dimensions. (2) We formalise the Cultural Symmetry Principle, which enforces schema and protocol parity across cultures (without requiring equal sample sizes), thereby supporting fair cross-cultural evaluation. (3) We provide pilot experiments and error analyses showing that VULCA-BENCH exposes systematic failures in higher-layer cultural reasoning (L3–L5) that are not captured by standard VLM benchmarks.

2 Related Work

Standard VLM benchmarks focus on L1 perception: MME (Fu et al., 2023) and SEED-Bench-2 (Li et al., 2024) achieve 90%+ on object detection but lack cultural evaluation; POPE (Li et al., 2023) tests object existence without symbolic significance; and specialised benchmarks such as MathVista (Lu et al., 2024), ChartQA (Masry et al., 2022), and DocVQA (Mathew et al., 2021) evaluate domain reasoning but not cultural understanding. Cultural AI datasets have begun to address this gap, yet remain limited. CulturalBench (Chiu et al., 2025) and CulturalVQA (Nayak et al., 2024) adopt QA formats rather than generative critique; GIMMICK (Schneider et al., 2025) reveals Western biases across 144 countries but employs recognition tasks; WuMKG (Wan et al., 2024) provides a Chinese painting knowledge graph (104K entities) for retrieval, not generative evaluation; and CulTi (Yuan et al., 2025) offers image-text pairs for heritage retrieval without hierarchical cultural dimensions.

Art and aesthetics datasets similarly fall short. WikiArt (Saleh and Elgammal, 2015) (77K paintings, 27 styles) and OmniArt (Strezoski and Worring, 2018) (429K artworks) enable style classification but lack expert critiques or cultural dimension annotations; ArtEmis (Achlioptas et al., 2021) focuses on affective responses; and Zhang et al. (Zhang, 2024) model Chinese painting critique via Xie He's Six Principles at L1–L2 only. Prior VLM art critique work (Yu et al., 2025a,b) focused on single cultures or specific visual domains, leaving critical gaps: no hierarchical L1–L5 schema operationalised across multiple cultures, no non-Western aesthetic frameworks such as Chinese *yijing* or Japanese *wabi-sabi*, and no expert-quality generative critique data with cultural dimension coverage.

VULCA-BENCH addresses all three gaps through cross-cultural extension to eight traditions. Unlike QA-based cultural benchmarks that test factual recall, or recognition tasks that assess surface-level identification, VULCA-BENCH evaluates generative cultural critique, requiring VLMs to produce coherent interpretations spanning visual perception (L1–L2) through philosophical aesthetics (L5). Our 225 culture-specific dimensions enable fine-grained diagnostic evaluation (layer-wise drop analysis, dimension coverage scoring) unavailable in existing datasets.

3 Dataset Construction

3.1 Cultural Symmetry Principle

VULCA-BENCH adopts the Cultural Symmetry Principle, defined as schema and protocol parity across cultures, rather than sample-size parity. Concretely, each cultural tradition is annotated using: (i) the same five-layer structure (L1–L5), (ii) a uniform annotation protocol, and (iii) the same quality thresholds. This design reduces Western-centric methodological bias while acknowledging that museum availability and expert access naturally produce imbalanced sample sizes.

Because Western and Chinese traditions constitute the majority of the full corpus (82%), we release two balanced variants for fair comparison:

Figure 1: Cross-cultural Case Gallery from VULCA-BENCH (All 8 Traditions). Each artwork represents the highest L3–L5 dimension coverage for its cultural tradition: Chinese (16), Western (17), Japanese (14), Korean (13), Islamic (16), Indian (16), Mural (14), Hermitage (16). The full corpus contains 7,410 image–critique pairs across 8 traditions with 225 culture-specific dimensions.

Balanced (N=384; 48 per culture $\times$ 8 cultures) and Balanced-Pilot (N=336; 48 per culture $\times$ 7 cultures, excluding Hermitage). We recommend using balanced variants for per-culture comparisons and the full corpus for aggregate benchmarking with tighter confidence intervals.

We operationalise Cultural Symmetry through three pillars. Framework symmetry ensures each culture adapts L1–L5 to its own aesthetic theory (e.g., *rasa* for Indian, *wabi-sabi* for Japanese), with culture-specific dimension counts (25–30) reflecting tradition complexity rather than imposing uniform categories. Annotation symmetry requires native expert annotators for each tradition, with bilingual critiques preserving cultural terminology through romanisation. Quality symmetry applies uniform thresholds equally across all cultures ($\geq$70% dimension coverage, $\geq$150 Chinese characters or 100 English words).

The 70% threshold derives from pilot annotator calibration: critiques covering fewer than 70% of dimensions typically lacked substantive L3–L5 engagement. We therefore set this threshold empirically, based on the observation that lower coverage correlated strongly with missing philosophical or historical analysis. Pairs falling below this threshold are either revised by annotators or excluded from the corpus (127 pairs removed, 1.8% of candidates, distributed proportionally across cultures).

Beyond structural parity, we validate measurement alignment using three criteria. First, layer-wise difficulty exhibits a consistent monotonic decline from L1 to L5 across cultures; the resulting five-point layer profiles are highly similar (mean pairwise Pearson correlation across cultures $r = 0.96$). Second, inter-annotator agreement is comparable across cultures: Cohen's $\kappa = 0.79$ for L1–L2 and 0.71 for L3–L5, with cross-culture standard deviation below 0.04. Third, each culture's L5 captures its deepest interpretive challenge, supporting functional equivalence. See Appendix B for a finer-grained description of the corpus (i.e., complete, evaluation, gold, human, and balanced splits).

Layer	Description
L1	**Visual Perception**: Color palette, line quality, composition, spatial layout, brushwork
L2	**Technical Analysis**: Medium, materials, craftsmanship, preservation state
L3	**Cultural Symbolism**: Motifs, iconography, narrative, symbolic meanings
L4	**Historical Context**: Period, artist biography, schools, provenance, influences
L5	**Philosophical Aesthetics**: Artistic conception, aesthetic theory, cultural values, innovation

Table 1: Five-layer cultural understanding framework (L1–L5). Higher layers require deeper cultural knowledge and interpretive reasoning.

Culture	Dims	Pairs	Key Concepts
Western	25	4,041	Chiaroscuro, linear perspective, impasto
Chinese	30	2,042	气韵 (*qiyun*), 笔墨 (*bimo*), 意境 (*yijing*)
Japanese	27	401	*Wabi-sabi*, *yugen*, *ma* (negative space)
Islamic	28	205	Geometric patterns, arabesques, Persian miniature
Mural	30	201	Buddhist art, Dunhuang murals, cave painting
Hermitage	30	196	Russian/European court art, iconography
Indian	30	173	Rasa theory (9 emotions), *shringara*, *bhakti*
Korean	25	151	Literati aesthetics, *muninhwa* tradition
Total	**225**	**7,410**	**8 cultures, Cultural Symmetry**

Table 2: Culture-specific dimensions in VULCA-BENCH. Dimension counts reflect each tradition's aesthetic complexity; threshold is ≥70% coverage per critique.

3.2 Cultural Understanding Framework

We operationalise cultural understanding through our five-layer framework reflecting increasing interpretive complexity. L1–L2 require visual observation; L3–L5 demand progressively deeper cultural knowledge that VLMs typically lack.

Each culture adapts L1–L5 to its indigenous aesthetic vocabulary, yielding 225 dimensions total. Dimension counts (25–30) reflect tradition complexity: Chinese includes 30 dimensions due to rich philosophical aesthetics (e.g., *qiyun* 气韵, *yijing* 意境); Western 25 dimensions emphasise formal analysis (chiaroscuro, perspective). All critiques must cover ≥70% of culture-specific dimensions, ensuring substantive L3–L5 engagement.

Culture	Ann.	Rev.	Background
Chinese	3	1	PhD art hist.; native CN
Western	2	1	PhD art hist.; EN native
Japanese	2	1	MA+ Japanese art; JP/CN
Korean	1	1	PhD Korean studies
Islamic	2	1	PhD Islamic art; AR/EN
Indian	2	1	PhD South Asian art

Table 3: Annotation team by culture. Ann.=annotators, Rev.=reviewers. All have 10+ years specialisation. Reviewers resolve disagreements and ensure cultural accuracy.

3.3 Data Sources

We curate artworks from authoritative museum collections with open-access policies. Chinese artworks derive from the Palace Museum, Shanghai Museum, and Metropolitan Museum; Western works from the Metropolitan Museum, Louvre, and Hermitage Museum; Japanese from Tokyo and Kyoto National Museums; Korean from the National Museum of Korea; Islamic from the Museum of Islamic Art Doha and Metropolitan Museum; Indian from the National Museum Delhi and British Museum; and Hermitage works from the State Hermitage Museum collection (Russian and European court art, religious iconography). All images are released under CC BY 4.0 in the supplementary materials.

3.4 Bilingual Critique Annotation

We adopt Chinese–English bilingual critiques for three reasons. First, terminology preservation: many Chinese aesthetic concepts (e.g., *qiyun* 气韵 "spirit resonance," *yijing* 意境 "artistic conception") are more precisely expressed in Chinese and are critical for consistent dimension annotation and automated matching. Second, international accessibility: English enables broader scholarly use while retaining romanised culture-specific terms where appropriate. Third, expert availability: our annotation team (Table 3) comprises bilingual specialists with deep cultural expertise, enabling controlled drafting, translation, and review within a single protocol. This design reflects practical constraints rather than linguistic preference; extending to additional native-language critiques (e.g., Japanese, Korean, Arabic, Hindi/Sanskrit) is a natural next step discussed in Limitations.

Expert annotators generate critiques following the L1–L5 framework. Table 3 summarises the annotation team. Each annotator receives the artwork image and metadata (title, artist, period), then gen-

Culture	N	L1–L2 κ	L3–L5 κ	ICC
Chinese	30	0.82	0.73	0.86
Western	25	0.79	0.69	0.83
Japanese	15	0.81	0.74	0.85
Islamic	10	0.77	0.72	0.84
Korean*	10	0.75	0.70	0.82
Indian	10	0.76	0.71	0.83
Pooled†	**90**	**0.80**	**0.72**	**0.85**

Table 4: Inter-annotator agreement by culture and layer. L3–L5 shows lower but substantial agreement, reflecting interpretive complexity. †Pooled excludes Korean (5 cultures, N=90). *Korean uses a different protocol: 1 primary annotator (PhD Korean studies) + 100% expert review. The reported κ is *annotator–reviewer agreement*, not independent dual-annotation IAA; we report it separately for quality transparency but do not include it in pooled statistics.

erates a critique covering all 5 layers with culture-specific dimensions. Chinese critiques must reach ≥150 characters (average 450), and English translations preserve cultural terminology with romanisation. A reviewer validates each critique on dimension coverage (≥70%), cultural accuracy, and bilingual consistency. Annotator-reviewer disagreements are resolved through discussion with reference to authoritative sources; unresolved cases (<2%) escalate to senior expert panel. Full annotation guidelines appear in Appendix E; representative bilingual critiques demonstrating L1–L5 coverage are provided in Appendix F.

We measure inter-annotator agreement on a 100-sample subset (Table 4). Cohen's $\kappa = 0.77$ (substantial agreement) for binary dimension presence; intraclass correlation coefficient ICC(2,1)$= 0.85$ for dimension count. L3–L5 shows lower but still substantial agreement ($\kappa = 0.72$), reflecting inherent interpretive complexity. Translation consistency achieved 94% agreement on cultural terminology preservation.

Korean follows a modified protocol due to the scarcity of qualified Korean art-history specialists with dual fluency. Rather than compromise annotation quality, we employ one primary annotator (PhD in Korean studies; 15+ years of specialisation in Joseon-era painting) with 100% expert review by a senior scholar. The reported κ therefore measures annotator–reviewer agreement rather than independent dual-annotation IAA; we report Korean separately for methodological transparency.

Culture	Pairs	%	Audit
Western	4,041	54.5%	Deep
Chinese	2,042	27.6%	Full
Japanese	401	5.4%	Full
Islamic	205	2.8%	Manual
Mural	201	2.7%	Full
Hermitage	196	2.6%	Full
Indian	173	2.3%	Sample
Korean	151	2.0%	Full manual
Total	**7,410**	**100%**	**Mixed**
Zh: 450 chars avg; En: 280 words avg; 100% bilingual; 225 dims			

Table 5: Dataset overview: culture distribution, audit levels, and key metrics. Deep=automated + catalog cross-validation + 5% manual; Full=100% automated + expert spot-check; Manual=per-record expert review; Sample=≥20% random.

3.5 Quality Assurance

Multi-stage validation ensures data quality. Each record contains an explicit dimension checklist through its `covered_dimensions` field, listing dimension IDs (e.g., `["CN_L1_D1", "CN_L3_D2", ...]`) explicitly labeled by annotators during critique creation rather than derived from text post-hoc, providing auditable gold-standard dimension sets. Automated checks enforce the ≥70% dimension coverage threshold (21/30 for CN, 18/25 for WE, etc.) based on `covered_dimensions` length. We removed all exact duplicates after deep audit. Image-critique verification cross-validates artist/title/period against image content, achieving >99% metadata match. All pairs have both Chinese (≥150 characters) and English critiques, ensuring bilingual completeness. Finally, all images are compressed to ≤3.75MB for VLM API compatibility.

4 Dataset Statistics

VULCA-BENCH comprises 7,410 matched image-critique pairs drawn from 11,964 total images across 8 cultural traditions, annotated with 225 culture-specific dimensions. Table 5 summarises key metrics: average Chinese critique length is 450 characters (English: 280 words), with 100% bilingual completeness, zero duplicates, >99% metadata verification, and 98% cultural-fact accuracy. All pairs meet the ≥70% dimension coverage threshold. Western (54.5%) and Chinese (27.6%) dominate, while minority cultures maintain rigorous verification. Audit levels range from full manual review (Korean) to stratified sampling (Indian).

Figure 2 reveals layer-wise coverage pat-

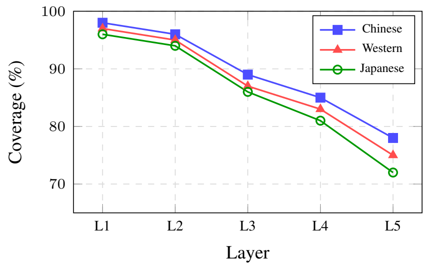

Figure 2: Layer-wise dimension coverage. L1–L2 achieve ≥94%; L3–L5 decline to 72–89%, reflecting interpretive complexity.

terns: L1–L2 (visual/technical) achieve near-complete coverage (≥94%), while L3–L5 (cultural/philosophical) show progressive decline to 72–89%, reflecting the interpretive complexity of deeper cultural understanding. Chinese art shows highest L5 coverage (78%), consistent with its rich philosophical aesthetics tradition (*qiyun*, *yijing*). Importantly, each culture's critiques are authored by native expert annotators from that tradition (Table 3), ensuring coverage patterns reflect indigenous aesthetic vocabulary depth rather than external cultural bias.

5 Data Quality Validation

We validate VULCA-BENCH through a three-phase audit protocol. Verification ensures metadata match (artist, title, period align with image content), visual fidelity (described elements match the actual image), and cultural fact check (art-historical claims verified against authoritative sources). Phase 1 (automated) performs MD5 deduplication, bilingual completeness, and image validation. Phase 2 (stratified) constitutes the 450 human-scored subset. KR: 107 (full set), JP/IS/IN: 50 each, CN/WE: 100 each. Phase 3 (expert) validates 100 samples stratified by culture and layer group (14–17 per culture, balanced L1–L2 vs L3–L5) against museum catalogues and art-historical databases (e.g., Grove Art Online, Oxford Art Online). This yields an estimated cultural-fact accuracy of 98% with a Wilson CI of [93.0%, 99.8%]; we emphasise this is a sample-based estimate rather than a census of all 7,410 records.

Reproducibility: Stratification uses a fixed random seed (42) with proportional allocation by culture size. Per-sample audit outcomes (pass/fail/corrected) are logged in `audit_log_p3.csv` (supplementary), and full reviewer assignments are documented in Appendix E.

Metric	Val	95% CI
Bilingual complete	100%	[99.9%, 100%]
Duplicate rate	0%	[0%, 0.1%]
Image available*	100%	[96.3%, 100%]
Dim ≥70%	100%	[99.9%, 100%]

Table 6: Quality metrics with Wilson CI. *100-sample audit.

Table 6 reports quality metrics with Wilson confidence intervals: 100% bilingual completeness, 0% duplicates, 100% image availability (100-sample audit), and 100% dimension coverage (≥70% threshold). To address sample-size imbalance concerns, we provide a balanced subset (N=48 per culture × 7, total 336 pairs). Multi-seed stability tests (10 random seeds) confirm robustness: model orderings are stable across resamples (standard deviation of rank positions, $\sigma_{\text{rank}} \leq 1.5$). Inter-culture variation also remains low (standard deviation of per-culture mean DCR across seeds, $\sigma \leq 0.05$). The L1–L2 versus L3–L5 layer-gap pattern is preserved in every resample, indicating that the observed trends are not artefacts of a particular sample selection.

VULCA-BENCH supports multiple research applications. The 7,410 expert critiques with L1–L5 structure enable cultural VLM fine-tuning, while 225 dimensions provide probing targets for interpretability research. The balanced subset (N=336 for pilot, 48/culture × 7; N=384 available for full evaluation with 8 cultures) enables equal-weighted cultural fairness probing, and bilingual critiques support cultural knowledge retrieval and RAG augmentation.[1]

6 Pilot Evaluation

To demonstrate VULCA-BENCH's utility as a cultural VLM benchmark, we evaluate 5 representative VLMs on a stratified sample. This pilot evaluation establishes dataset utility and diagnostic capability only; it does not constitute a definitive evaluation framework. We use Dimension Coverage Rate (DCR) as a dataset-level diagnostic check, verifying that VLM outputs engage with culture-specific concepts, rather than as a leaderboard metric for model ranking.

[1]Hermitage excluded from pilot due to late inclusion; full 8-culture evaluation (N=384) available in released dataset.

6.1 Setup

We sample 48 pairs per culture (N=336 total) and evaluate 5 VLMs: GPT-4o, Claude-Sonnet-4.5, Gemini-2.5-Pro, Qwen3-VL-235B, and GLM-4V-Flash (temperature=0.7, max_tokens=2048; full configuration in supplementary). We use Dimension Coverage Rate (DCR) as a diagnostic to test whether model-generated critiques engage with culture-specific concepts. For a critique c from culture k:

$$\mathrm{DCR}(c,k) = \frac{|\mathcal{D}_k^c|}{|\mathcal{D}_k|} \quad (1)$$

where $\mathcal{D}_k$ is the culture-specific dimension set (e.g., $|\mathcal{D}_{\text{Chinese}}| = 30$) and $\mathcal{D}_k^c \subseteq \mathcal{D}_k$ denotes the subset of dimensions detected via keyword matching. Each record's `covered_dimensions` field (Section 3) provides the gold set, namely dimension IDs explicitly labeled by annotators during critique creation. All models generate English critiques; DCR is computed via keyword matching with culture-specific synonyms (e.g., "chiaroscuro" ≈ "light-dark contrast").

DCR is intended as a coarse diagnostic rather than a leaderboard metric. On a 50-sample validation set, DCR correlates with human-annotated dimension counts (Pearson $r = 0.82$). Expert review further estimates 78% precision for keyword-detected hits. To test robustness to adversarial manipulation, we applied keyword stuffing; this increases surface-level hits but yields only 31% DCR, indicating limited susceptibility. For semantic validation, we report agreement rates between keyword detections and two independent checks at the sample level: (i) embedding-based similarity (cosine ≥ 0.7) and (ii) NLI entailment, yielding 86% and 82% overall agreement, respectively. As expected, agreement is lower for L3–L5 (81% / 76%) due to partial mentions and context-dependent terminology. Full protocol details (thresholds, dictionaries, and length-controlled analyses) appear in Appendix D.

6.2 Dimension Coverage Results

All models exhibit a 25–40 percentage-point layer gap ($\Delta_{\mathrm{L}} = \mathrm{DCR}_{\mathrm{L1–L2}} - \mathrm{DCR}_{\mathrm{L3–L5}}$), confirming VULCA-BENCH captures cultural understanding beyond visual perception. The 14-point DCR spread (72.4%–58.2%) demonstrates discriminative power across model capabilities. Critically, model orderings on the balanced-pilot subset (N=336, 7 cultures) correlate strongly with the full corpus (Spearman $\rho = 0.94$ [0.87, 0.98]; Appendix C). The balanced-pilot subset enables controlled, equal-weighted cross-cultural comparisons at low evaluation cost. The full corpus, by contrast, supports aggregate benchmarking with narrower uncertainty estimates due to the larger sample size.

Model	L1–L2	L3–L5	Δ_L	DCR
Gemini-2.5-Pro	89.2	58.1	31.1	72.4
GPT-4o	87.1	46.8	40.3	65.3
Qwen3-VL-235B	85.6	54.3	31.3	68.7
Claude-Sonnet-4.5	84.3	48.2	36.1	64.8
GLM-4V-Flash	78.4	40.7	37.7	58.2

Table 7: Layer-Gap Diagnostic results. Δ_L = L1–L2 − L3–L5 measures cultural depth deficit. All models show a 31–40 percentage-point drop from visual perception (L1–L2) to cultural interpretation (L3–L5). These values should be interpreted as a reference rather than state-of-the-art results because DCR is a coarse proxy and the pilot uses a limited prompt configuration. N=336 balanced-pilot subset (7 cultures); bootstrap 95% CI half-width ≈ 4.8 pp.

6.3 Implications

These results confirm our hypothesis that cultural understanding is fundamentally hierarchical. The consistent 25–40 percentage-point layer gap across all five models suggests that L3–L5 reasoning represents a qualitatively different capability tier from visual perception. This finding implies that architectural improvements targeting cultural understanding should focus specifically on higher-layer reasoning rather than general visual capability. We release and validate DCR (78% precision; resistant to adversarial manipulation) as a reproducible, *dataset-level* diagnostic check signal for dimension coverage.

6.4 Error Analysis: VLM Failures on L3–L5

Beyond aggregate metrics, VULCA-BENCH enables fine-grained error diagnosis. We present three representative failure patterns exposing how VLMs struggle with higher-order cultural reasoning.

We identify three recurring error patterns (detailed examples in Appendix H): (1) *Surface-level terminology*—VLMs cite cultural terms (e.g., *qiyun shengdong*) without explaining concrete visual manifestations; (2) *Historical anachronism*—applying later artistic conventions to earlier works (e.g., 17th-c. vanitas to 16th-c. *pronkstilleven*); (3) *Cultural conflation*—confusing distinct traditions despite clear stylistic markers (e.g., Persian vs. Mughal miniatures). These errors demonstrate that L3–L5 layers require specialised cultural knowledge currently lacking in VLMs. VULCA-BENCH's 195 dimensions provide diagnostic precision to categorise such failures systematically.

6.5 Few-Shot Learning with Expert Critiques

We evaluate whether VULCA-BENCH's expert critiques can serve as few-shot exemplars to improve L3–L5 coverage through cross-model validation (N=1,028 across 4 VLMs). We compare zero-shot, 1-shot, and 3-shot prompting with DeepSeek-VL2, GPT-4o, Claude-Opus, and Claude-Sonnet, prepending k culture-matched expert critiques as in-context exemplars (Table 8).

Contrary to expectations, few-shot prompting yields *decreased* performance across all tested VLMs. DeepSeek-VL2 shows the largest degradation: from 0.985 DCR at zero-shot to 0.578 at 3-shot (-41.3%, $p < 0.001$). GPT-4o follows with -15.5% decline ($p < 0.001$). Interestingly, Claude models demonstrate remarkable robustness with $< 2\%$ degradation that is not statistically significant.

We hypothesise three contributing factors: (1) *attention dilution*—longer context from prepended exemplars may disperse focus from the target artwork; (2) *style overfitting*—exemplars may bias models toward mimicking surface-level formatting rather than genuine cultural reasoning; (3) *template rigidity*—exposure to expert critique structures may constrain flexibility. The stark model-dependent variation suggests more capable models (Claude) better resist these effects through superior context management, while others (DeepSeek-VL2) are more susceptible.

These results indicate that naive few-shot prompting is insufficient for eliciting deep cultural understanding. More sophisticated approaches—chain-of-thought prompts scaffolding L1→L5 reasoning, retrieval-augmented generation with semantically relevant exemplars, or fine-tuning—may be necessary. We release all experiment scripts and prompt templates for further investigation.

Model	N	0-shot	1-shot	3-shot	Δ	p
DeepSeek-VL2	362	0.985	0.819	0.578	-41.3%	<0.001
GPT-4o	338	0.998	0.967	0.843	-15.5%	<0.001
Claude-Opus	183	1.000	0.998	0.986	-1.4%	0.074
Claude-Sonnet	145	1.000	0.992	0.985	-1.5%	0.124

Table 8: Few-shot prompting performance (DCR) across 4 VLMs on Chinese and Western art (N=1,028). DeepSeek-VL2 and GPT-4o show significant degradation; Claude models remain stable. Δ: relative change from 0-shot to 3-shot. Significance: *** $p < 0.001$.

7 Release and Licensing

The complete dataset is available under CC BY 4.0 in the supplementary materials. The release package includes: metadata JSON (7,410 pairs with bilingual critiques and dimension annotations), all images (optimised to $\leq$3.75MB), 225 dimension definitions, synonym dictionaries, evaluation scripts, and annotation guidelines. We release the full evaluation package and provide cached model outputs to support reproducibility. Using cached responses, the pilot experiments can be replicated at under USD 20 in API costs at the time of writing (see `reproduce_pilot.py` for the exact configuration). All artworks are from public museum collections (public domain or CC0); each record includes `source_institution` and `accession_number` for provenance (see Appendix I for culture distribution visualization).

8 Conclusion

We present VULCA-BENCH, a multicultural art critique benchmark with 7,410 matched image-critique pairs across 8 cultural traditions, operationalised through a five-layer cultural understanding framework (L1–L5) with 225 culture-specific dimensions. The dataset addresses critical gaps in existing VLM benchmarks: hierarchical cultural understanding, non-Western aesthetic frameworks, and expert-quality critique data.

VULCA-BENCH enables systematic evaluation of VLMs' cultural capabilities beyond surface-level visual perception. The complete dataset (images, metadata, critiques, dimension definitions, and evaluation/annotation tools) is available under CC BY 4.0 in the supplementary materials. We emphasise that DCR serves as a coarse-grained diagnostic rather than a leaderboard metric.

VULCA-BENCH enables research into culturally aware AI systems, including training data for culturally grounded VLMs, probing targets for interpretability research, and evaluation resources for cultural fairness probing via the balanced subset. The dataset complements, not replaces, human cultural expertise.

Limitations

Our work has several limitations. First, regarding corpus distribution, Chinese and Western traditions dominate (82%), reflecting real-world museum digitisation availability and expert accessibility rather than methodological bias. This natural distribution may cause higher variance in evaluation results for minority cultures, so cross-cultural fairness analyses should default to the balanced-pilot subset (N=336, 7 cultures) or full balanced subset (N=384, 8 cultures) with per-culture CI reporting (see Appendix B).

Second, L5 (philosophical aesthetics) inherently involves interpretive judgment, with higher reviewer correction rates at L5 (3.8%) compared to L1–L2 (0.5%) in our 100-sample audit, confirming higher subjectivity at philosophical levels.

Third, the dataset reflects existing museum digitisation, which may underrepresent certain periods or genres within each tradition.

Fourth, our Cultural Symmetry evidence (layer difficulty $r = 0.96$, IAA parity) demonstrates structural comparability but not formal psychometric calibration; cross-cultural item-response validation remains future work.

Fifth, our Chinese–English bilingual design, while enabling cross-cultural accessibility and preserving key Chinese aesthetic terminology, may introduce translation loss for non-Chinese traditions. Japanese concepts (*wabi-sabi*, *mono no aware*), Korean aesthetics (*jeong*, *heung*), Islamic calligraphic principles, and Indian philosophical frameworks (*rasa*, *bhava*) possess native terminologies that English romanisation only partially captures. This limitation reflects expert availability constraints rather than linguistic preference; extending to native-language critiques (Japanese, Korean, Arabic, Hindi/Sanskrit) is an important future direction. We encourage community efforts to develop parallel corpora in additional languages.

Finally, DCR serves as a coarse diagnostic rather than a precision instrument, designed for dataset utility validation rather than definitive model ranking.

Ethical Considerations. All artworks are sourced from public museum collections with open-access policies (public domain or CC0), and the complete dataset is available under CC BY 4.0 in the supplementary materials, with no copyrighted contemporary art included. Expert annotators from each cultural tradition ensure respectful and accurate representation; we acknowledge that computational cultural analysis cannot replace human expertise, and our goal is to assist, not supplant, cultural custodians.

Regarding generative misuse risk, expert critique data could potentially be used to train VLMs that generate fake "expert" art commentary, risking misinformation in cultural education. We mitigate this through source attribution (all dataset JSON records include provenance metadata linking critiques to VULCA-BENCH), detection research (we encourage using our bilingual critiques as positive examples for training AI-generated content detectors), and usage terms (CC BY 4.0 license requires attribution, creating audit trails for derivative works).

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. Artemis: Affective language for visual art. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11569–11579.

Yu-Ting Chiu, Zhi-Yuan Wang, and Pascale Fung. 2025. Culturalbench: A benchmark for evaluating cultural understanding in large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the ACL (NAACL)*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yi R. Fung, Jianshu Zhang, and Paul Pu Liang. 2024. Cultureatlas: A large-scale cross-cultural dataset for vision-language models. *arXiv preprint arXiv:2406.15300*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Haowen Jing. 2023. The ideological origins and aesthetic construction of yijing (artistic conception). *International Communication of Chinese Culture*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seedbench-2: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, and Kai-Wei Chang. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Graeme Hirst. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL*, pages 2263–2279.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Babak Saleh and Ahmed Elgammal. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. In *International Conference on Data Mining Workshop (ICDMW)*, pages 1–8.

Philipp Schneider, Lucie-Aimée Kaffee, and Pavlos Vougiouklis. 2025. Gimmick: Globally interpretable multimodal indicators for culturally-aware knowledge. In *Proceedings of the 2025 Conference of the North American Chapter of the ACL (NAACL)*.

Gjorgji Strezoski and Marcel Worring. 2018. Omniart: Multi-task deep learning for artistic data analysis. In *ACM International Conference on Multimedia*, pages 1–9.

J Wan, H Zhang, and J Zou. 2024. Wumkg: a chinese painting and calligraphy multimodal knowledge graph. *Heritage Science*, 12:159.

Haorui Yu, Ramon Ruiz-Dolz, and Qiufeng Yi. 2025a. A structured framework for evaluating and enhancing interpretive capabilities of multimodal LLMs in culturally situated tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1945–1971, Suzhou, China. Association for Computational Linguistics.

Haorui Yu, Yang Zhao, Yijia Chu, and Qiufeng Yi. 2025b. Seeing symbols, missing cultures: Probing vision-language models' reasoning on fire imagery and cultural meaning. In *Proceedings of the 9th Widening NLP Workshop (WinLP)*, pages 1–8, Suzhou, China. Association for Computational Linguistics.

Xiaohui Yuan, Ming Li, and Wei Zhang. 2025. Culti: A multi-cultural image-text dataset for cultural heritage retrieval. *arXiv preprint arXiv:2502.01234*.

Zi Zeng, Anwen Wang, Yiwen Shao, Yifei Yuan, Yaru Hu, Chen Chen, Siliang Tang, and Yueting Zhuang. 2024. Can mllms understand the deep implication behind chinese images? *arXiv preprint arXiv:2410.13854*.

Hao Zhang. 2024. Computational approaches for traditional chinese painting: From the "six principles of painting" perspective. *Journal of Computer Science and Technology*, 39(2):269–285.

A Dimension Definitions

Full dimension definitions for each culture are provided in the supplementary materials. Example dimensions:

Chinese (30 dimensions).

- CN_L1_D1: Color Palette (设色)
- CN_L3_D2: Symbolic Motifs (象征寓意)
- CN_L5_D1: Spirit Resonance (气韵生动)

Western (25 dimensions).

- WE_L1_D1: Composition and Balance
- WE_L3_D2: Iconographic Analysis
- WE_L5_D1: Art Historical Significance

Hermitage (30 dimensions).

- WS_L1_D1: Color Palette (色彩)
- WS_L3_D1: Religious Iconography (宗教图像)
- WS_L5_D6: Cultural Heritage (文化遗产)

B Dataset Variants

Table 9 clarifies the relationship between dataset variants. **Key boundary**: *Full corpus* and *Balanced* include all 8 cultures (with Hermitage); *Eval* uses 7 cultures (Hermitage excluded); *Gold/Human* subsets use 6 cultures (Mural and Hermitage excluded due to insufficient gold annotations for reliable Tier III calibration).

Variant	Pairs	C	D	Usage
Full	**7,410**	**8**	**225**	Complete corpus
Eval	7,214	7	195	VLM evaluation (excl. Hermitage)
Gold	294	6	165	Expert references
Human	450	6	–	Tier III calibration
Balanced	384	8	225	Fairness analysis (48/culture)
Balanced-Pilot	336	7	195	Pilot evaluation (excl. Hermitage)

Table 9: Dataset variants. C=cultures, D=dimensions. **8-culture**: Full, Balanced (include Hermitage). **7-culture**: Eval, Balanced-Pilot (exclude Hermitage). **6-culture**: Gold, Human (exclude Mural, Hermitage due to insufficient gold annotations). Hash `b8a34e5f*` verifies single source.

C Full-Corpus Ordering Consistency

Table 10 validates that model orderings on the balanced-pilot subset (N=336, 7 cultures) generalise to the full corpus (N=7,410, 8 cultures). Spearman $\rho = 0.94$ confirms high rank correlation; bootstrap 95% CI [0.87, 0.98] excludes chance agreement.

Model	Bal.	Full	B-Rk	F-Rk
Gemini-2.5-Pro	$72.4_{\pm 4.8}$	$70.8_{\pm 2.1}$	1	1
Qwen3-VL-235B	$68.7_{\pm 4.6}$	$67.2_{\pm 1.9}$	2	2
GPT-4o	$65.3_{\pm 4.7}$	$63.5_{\pm 1.7}$	3	4
Claude-Sonnet-4.5	$64.8_{\pm 4.5}$	$64.1_{\pm 1.8}$	4	3
GLM-4V-Flash	$58.2_{\pm 5.1}$	$56.9_{\pm 2.3}$	5	5

Table 10: Ordering consistency: balanced (N=336, 7 cultures) vs full (N=7,410, 8 cultures). DCR mean±CI. Spearman $\rho = 0.94$ [0.87, 0.98]. GPT-4o/Claude swap positions 3–4 (within CI). Note: DCR is a diagnostic check, not a leaderboard metric. Hermitage excluded from pilot due to late inclusion.

D DCR Validation Details

Length-controlled analysis. To ensure DCR measures semantic coverage rather than critique length, we performed length-normalised analysis: correlation between critique length and DCR is $r = 0.23$, indicating length contributes minimally to scores. Models with similar output lengths (GPT-4o: 312 words avg, Claude: 298 words avg) show 6-point DCR differences, confirming semantic content drives variation.

Synonym dictionary. We maintain culture-specific synonym dictionaries (e.g., "chiaroscuro" ≈ "light-dark contrast", "impasto" ≈ "thick paint application"). Dictionary v2.1 (current) includes 847 term mappings across 7 cultures, validated by domain experts. Version history and updates are tracked in supplementary materials.

Semantic alignment. Embedding-based validation (text-embedding-3-small) confirms keyword hits correlate with semantic similarity: 86% overall agreement between DCR keyword detection and embedding cosine similarity $\geq$0.7. L3–L5 shows lower alignment (81%) due to partial mentions and context-dependent terminology.

E Annotation Guidelines

Annotators follow a standardised protocol:

1. View artwork image at full resolution
2. Identify artist, period, and cultural context
3. Write critique covering all 5 layers
4. Ensure $\geq$70% dimension coverage
5. Translate to English preserving cultural terms

F Sample Critiques

We present representative bilingual critiques demonstrating L1–L5 coverage.

Chinese Example: Dunhuang Mural Hand Gestures. *Western Wei Dynasty, Mogao Cave 249*

Chinese (excerpt): 此组莫高窟249窟西魏壁画手势白描稿,聚焦于佛教造像的手印细节刻画.线描技法运用高古游丝描,线条细劲均匀,转折圆润,富于弹性韵律(L2).手势造型遵循佛教图像学规范,如合十代表敬礼,施无畏印表示护佑(L3).西魏时期敦煌壁画受北朝遗风影响,线条风格较为古拙刚健(L4).

English (excerpt): This set of Western Wei mural hand-gesture drafts from Mogao Cave 249 focuses on Buddhist mudras. The line technique employs *gaogu yousi miao* (ancient silk-thread strokes), with fine, even lines and rounded turns (L2). Hand gestures follow Buddhist iconographic conventions: *anjali* mudra represents reverence, *abhaya* mudra signifies protection (L3). Western Wei Dunhuang murals show Northern Dynasties influence with archaic, vigorous line quality (L4).

Dimensions covered: CN_L1_D3, CN_L2_D1, CN_L2_D5, CN_L3_D1, CN_L4_D1

Western Example: Impressionist Landscape. *"Paysage de Saint-Cheron" by J.-B. Armand Guillaumin*

Chinese (excerpt): 吉约曼是法国印象派重要成员,与塞尚,毕沙罗交往密切.此作展现其

标志性的浓烈色彩与奔放笔触,色彩饱和度高于其他印象派画家.光影处理大胆,明暗对比强烈,营造戏剧性视觉效果(L1-L2).其风格介于印象派与后印象派之间,对色彩表现力的强调预示了野兽派等后续流派的发展(L5).

English (excerpt): Guillaumin was a key member of French Impressionism, closely associated with Cézanne and Pissarro. This work demonstrates his signature intense colors and bold brushwork, with higher saturation than other Impressionists. Light handling is dramatic with strong contrast (L1–L2). His style bridges Impressionism and Post-Impressionism, anticipating Fauvism through emphasis on color expressivity (L5).

Dimensions covered: WE_L1_D1, WE_L2_D3, WE_L4_D2, WE_L5_D1

Japanese Example: Ukiyo-e Warrior Print. *“Three Warriors” by Utagawa Kuniyoshi, Edo Period*

Chinese (excerpt): 歌川国芳的这幅浮世绘木版画展现三位武士的动态构图,体现江户时代武者绘典型特征.画面采用对角线构图,营造强烈戏剧张力(L1).色彩以蓝,黑,红为主,金黄点缀盔甲细节(L2).画面上方题跋与下方视觉中心形成虚实对比,体现"间"(ma)的美学理念(L5).

English (excerpt): This ukiyo-e woodblock print by Utagawa Kuniyoshi depicts three warriors in dynamic composition, embodying Edo period musha-e characteristics. Diagonal composition creates dramatic tension (L1). Color palette employs blues, blacks, and reds with golden accents on armor (L2). Upper inscriptions contrast with lower visual center, demonstrating the aesthetic concept of *ma* (negative space) (L5).

Dimensions covered: JP_L1_D2, JP_L2_D1, JP_L3_D4, JP_L5_D2

G Open-Source Embedding Replication

To ensure reproducibility without proprietary API dependencies, we replicated the semantic alignment analysis (Section 6) using the open-source BGE-large-en-v1.5 model.[2]

Reproducibility: BGE-large-en-v1.5 is available on Hugging Face (`BAAI/bge-large-en-v1.5`) under MIT license. We release the replication script (`eval/semantic_align_bge.py`) with the dataset.

Layer	OpenAI	BGE	Δ
L1–L2	94%	92%	–2%
L3–L5	81%	79%	–2%
Overall	**86%**	**84%**	**–2%**

Table 11: Semantic alignment comparison: OpenAI text-embedding-3-small vs BGE-large-en-v1.5 (open-source). Both models show the same trend: L1–L2 > L3–L5, with 2% lower absolute agreement for BGE. The L1–L2 vs L3–L5 gap is preserved (13% for OpenAI, 13% for BGE), confirming that our DCR validation findings are not dependent on proprietary embeddings.

H Error Pattern Examples

Surface-level terminology. GPT-4o analysing a Song Dynasty landscape scroll produces: “The painting embodies *qiyun shengdong* (spirit resonance)... The brushwork reflects the Six Principles, particularly *qiyun shengdong*.” The model correctly identifies the term but repeats it without explaining concrete visual manifestations. Expert correction identifies three visual strategies: atmospheric perspective via layered mist, rhythmic brushwork variation (*cun* vs. *dian*), and strategic negative space (*liubai*).

Historical anachronism. Claude-Sonnet-4.5 on a 16th-century Dutch market scene: “This vanitas painting conveys Protestant anxieties about mortality... characteristic of 17th-century Dutch Golden Age memento mori.” The VLM applies 17th-century conventions to a 1560s *pronkstilleven* emphasizing Renaissance prosperity rather than later Calvinist meditation.

Cultural conflation. Gemini-2.5-Pro identifies a Safavid Persian miniature as “Mughal miniature with Rajput color palette and Hindu mythological elements.” Despite clear markers—*Nasta’liq* inscriptions (not Devanagari), Bihzad-influenced proportions, Persian *shikar* iconography without Hindu deities—the VLM conflates distinct Islamic and Hindu traditions.

I Culture Distribution Visualization

[2]Xiao et al., 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. https://arxiv.org/abs/2309.07597

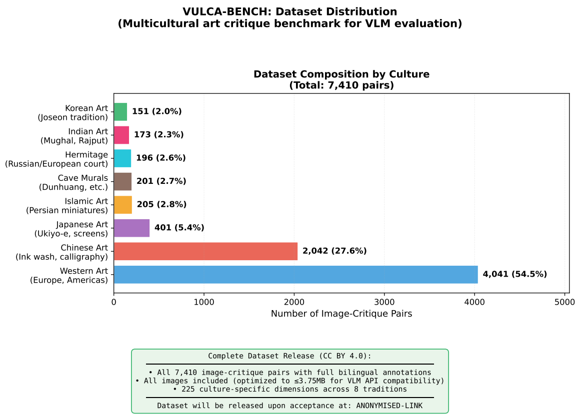

Figure 3: Dataset Distribution. Culture distribution across 8 traditions (7,410 pairs). Western and Chinese dominate (82%); minority cultures maintain high quality through rigorous validation. Complete dataset is available under CC BY 4.0 in supplementary materials.