SIGHT: A Structured Image-Grounded Hierarchical Taxonomy for Evaluating Multimodal Cultural Alignment

Anonymous ACL submission

Abstract

We introduce SIGHT, a culturally grounded multimodal benchmark of **5,086 expert-annotated instances** drawn from 5 million deduplicated image–caption pairs across **450+ authoritative sources**, with a structured human-in-the-loop annotation pipeline (Krippendorff's $\alpha = 0.680$). Evaluating state-of-the-art VLMs under zero-shot, few-shot, and supervised fine-tuning settings reveals persistent limitations in hierarchical cultural reasoning: the best model achieves only 56.0% category-level and 32.7% pair-level μF_1. Fine-tuning on $\sim$4,000 instances more than doubles pair-level μF_1, showing that the primary bottleneck is in-domain supervision rather than model scale. SIGHT is publicly available at this URL[1].

1 Introduction

Cultural context is a fundamental component of visual meaning, yet it remains weakly represented in multimodal evaluation (Yadav et al., 2025; Adilazuarda et al., 2024). Real-world images often encode socio-cultural, institutional, linguistic, and historical semantics that extend far beyond object recognition (Callow and Schiffman, 2002; Cabezas-García and Reimerink, 2022). Interpreting scenes such as religious gatherings, multilingual public signage, heritage architecture, or archival photography requires models not only to identify visible entities, but also to infer contextual meaning shaped by ritual, geography, governance, language, and history (Birhane and Prabhu, 2021; Liu et al., 2025).

Recent benchmarks including CultureVLM (Liu et al., 2025), IndicVisionBench (Faraz et al., 2025), CulturalVQA (Nayak et al., 2024), VULCA-Bench (Yu et al., 2026), and BLEnD-Vis (Tan et al., 2026)

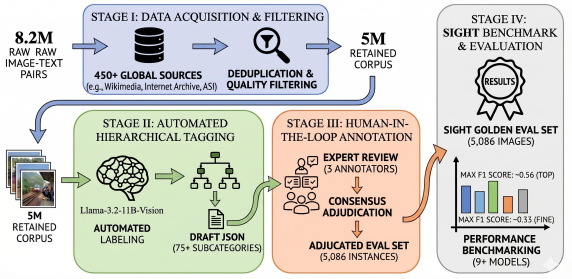

Figure 1: Overview of the SIGHT construction and evaluation pipeline, from large-scale web crawling and filtering to hierarchical annotation and benchmarking of VLMs under zero-shot, few-shot and SFT settings.

have begun to probe culturally grounded multimodal understanding, while broader benchmarks such as Winoground and CREPE expose weaknesses in compositional and contextual grounding (Thrush et al., 2022; Ma et al., 2023). However, existing evaluation paradigms remain limited in three key ways. First, many rely on constrained formats such as multiple-choice QA, enabling shortcut reasoning rather than robust semantic understanding (Zhang et al., 2023; Li et al., 2023). Second, culturally significant domains—including multilingual scene text, historical archives, indigenous scripts, civic institutions, and region-specific social practices remain underrepresented (Feng et al., 2025; Li et al., 2026; Kargaran et al., 2026). Third, most benchmarks treat cultural understanding as a flat prediction problem, despite the inherently overlapping and hierarchical nature of cultural semantics, where a single image may simultaneously encode religion, language, social activity, and institutional context (Nayak et al., 2024; Liu et al., 2025).

To address this gap, we introduce **SIGHT** (**S**tructured **I**mage-**G**rounded **H**ierarchical **T**axonomy), a benchmark for culturally grounded hierarchical visual understanding. SIGHT defines a multi-label hierarchical IMAGE CLASSIFICATION task, requiring models to jointly infer structured semantic labels across culturally

[1] https://anonymous.4open.science/r/SIGHT/. During the review phase, we release the manually annotated evaluation set; the complete 5M-scale corpus and additional resources will be released in future versions.

grounded categories and subcategories, reflecting the compositional nature of real-world visual understanding.

Constructing such a benchmark requires high-quality semantic annotation, as cultural interpretation often demands contextual expertise. Given the cost of large-scale manual curation (Deng et al., 2009; Krishna et al., 2017), we develop a human-in-the-loop pipeline combining multimodal model assistance with qualitative verification and multi-annotator workflow. We observe strong disagreement between pseudo-labels and ground truth, especially for ambiguous hierarchical cases, highlighting limits of automated labeling in culturally grounded settings.

We benchmark state-of-the-art VLMs on the SIGHT evaluation set, spanning both frontier closed-source and open-weight models. Even the strongest system achieves only 56.0% category-level μF_1 and 32.7% pair-level μF_1, revealing a persistent 22+ point gap between coarse category recognition and accurate subcategory grounding. Fine-tuning on only $\sim$4,000 SIGHT instances enables an 8B model to surpass all zero-shot baselines, including models more than $4\times$ larger, demonstrating that the primary bottleneck is not model scale, but access to high-quality human-annotated data with hierarchical structure.

Our main contributions are:

1. We introduce SIGHT, a large-scale multimodal benchmark comprising **5M+ image–caption pairs** curated from **450+** global archival, institutional, and cultural sources, with strong representation of South Asian visual domains alongside global diversity.
2. We propose a hierarchically structured multi-label taxonomy spanning **13 culturally grounded semantic categories** and **75+ fine-grained subcategories**, enabling systematic evaluation of hierarchical and contextual reasoning in VLMs.
3. We construct a high-quality expert-annotated evaluation set of **5,086 images**, each independently reviewed by **3 annotators** with quality control and adjudication, achieving a macro-average Krippendorff's $\alpha = 0.680$.
4. We benchmark **open- and closed-weight VLMs** under **zero-shot, few-shot, and SFT** settings, observing a persistent **22+ point gap** between category- and pair-level performance across all prompting regimes, and showing that fine-tuning a small **8B model** on $\sim$4,000 instances surpasses all zero-shot baselines.

2 Related Work

Multimodal benchmarking has advanced rapidly through foundational datasets such as *ImageNet* (Deng et al., 2009), *Visual Genome* (Krishna et al., 2017), and *WIT* (Srinivasan et al., 2021), which established large-scale evaluation protocols for object recognition, compositional grounding, and multilingual image–text alignment. Parallel efforts including *MMBench* (Zhang et al., 2023), and *SEED-Bench* (Li et al., 2023) further probed reasoning and visio-linguistic compositionality in general-domain settings. While these benchmarks have driven substantial progress in visual recognition, they treat images as conveyors of universal visual facts rather than as culturally situated representations (Birhane and Prabhu, 2021).

Culturally grounded evaluation. Recent work has increasingly focused on culturally grounded evaluation of VLMs. *MaRVL* (Jiang et al., 2024) first highlighted cross-lingual transfer failures in multilingual VLMs, while later benchmarks expanded geographic and linguistic coverage. *CVQA* (Romero et al., 2024) spans 30 countries and 31 languages, and *ALM-Bench* (Vayani et al., 2025) and *CultureVerse* (Liu et al., 2025) scale evaluation to 100+ languages and 188 countries respectively. Other efforts, including *CulturalVQA* (Nayak et al., 2024), *SEA-VQA* (Urailertprasert et al., 2024), *SCB* (Satar et al., 2025), *RICE-VL* (Pranav et al., 2025), *K-Viscuit* (Park et al., 2025), and *IndicVisionBench* (Faraz et al., 2025), further evaluate region-specific cultural understanding and multilingual reasoning. Retrieval-augmented approaches such as *RAVENEA* (Li et al., 2025) and structured cultural reasoning benchmarks such as *Appear2Meaning* (Jiang et al., 2026) demonstrate that current VLMs still capture only partial cultural understanding. A recent survey (Yadav et al., 2025) similarly concludes that Western-centric biases remain pervasive across multimodal evaluation. A detailed comparison of existing benchmarks is provided in Appendix §B (Table 7).

Underrepresented cultural domains. Existing benchmarks focus predominantly on food, attire, rituals, and landmarks, leaving multilingual scene text, archival imagery, civic infrastructure, and historical media underexplored (Feng et al., 2025; Kargaran et al., 2026; Li et al., 2026). *VULCA-*

Bench (Yu et al., 2026) and *GlotOCR-Bench* (Kargaran et al., 2026) partially address these gaps, but broader culturally contextual visual understanding remains insufficiently covered.

Annotation reliability and evaluation design. Most benchmarks rely on VQA or multiple-choice formats that reduce culturally complex scenes to single-answer predictions (Zhang et al., 2023), overlooking the layered nature of cultural meaning (Nayak et al., 2024). Annotation transparency and inter-annotator agreement are rarely reported (Suviranta and Hiippala, 2025; Xie et al., 2026). SIGHT addresses these gaps through a multi-label hierarchical classification taxonomy with a fully documented human-in-the-loop pipeline and agreement analysis across 5,086 annotated instances.

3 The SIGHT Benchmark

3.1 Raw Dataset Curation

The dataset is curated from a large-scale web crawl spanning over **450 publicly accessible portals**, covering cultural heritage, multilingual scene text, archival media, civic infrastructure, and diverse visual content. Representative sources are discussed in Appendix §A.1.

Only images accompanied by captions or descriptive metadata were retained during collection. The raw corpus contains approximately **8.2M image–caption pairs**; after filtering and deduplication, roughly **5M** pairs remain. From this corpus, a stratified quality-filtered subset of 6,351 image–caption pairs was sampled to ensure meaningful representation across all 13 taxonomy categories. Following human annotation and quality verification, 5,086 instances were retained as the final SIGHT evaluation benchmark. A detailed category-wise breakdown is provided in Appendix §A.1 (Table 6).

3.2 Dataset Processing

Starting from a large-scale corpus collected from 450+ globally distributed and regionally grounded sources, we constructed a high-quality benchmark subset using a five-stage preprocessing pipeline.

- ***Stage 1: Hard Filtering.*** Rule-based filtering removed invalid or low-information entries, including blank captions, captions outside a 4–200 word range, repetitive metadata (licensing notices, photographer credits), exact duplicates, and non-English captions, reducing the corpus from 35,694 to 15,639 pairs.

Cultural & Societal	Arts, Music & Dance; Cultural Symbols; Festivals & Rituals; Food & Cuisine; Occupations & Professions; Religious & Spiritual Imagery; Traditional Attire & Jewelry
Religion & Tradition	Devotional Objects; Pilgrimage Sites; Religious Ceremonies; Sacred Texts & Symbols; Temples & Places of Worship
Geographic & Environmental	Ecological Zones; Flora & Fauna; Landscapes & Terrain; River Systems & Water Bodies; Weather Patterns
Infrastructure & Public Spaces	Government Institutions; Private Institutions; Public Utility Visuals; Smart City Assets; Transport Systems
Education & Literacy	Classroom Photographs; Educational Institutions; Exam Papers; Learning Materials; Literacy Campaigns; Textbooks
Commerce & Economy	Agricultural Practices; Banking & Finance; Industrial Activity; Markets & Trade; Street Vendors
Digital & Print Media	Advertisements; Archival Print; Digital Screens; Memes & Visual Narratives; News Clippings; Posters & Banners; Social Media Posts
Historical & Archival Data	Archival Documents; Cultural Artifacts; Heritage Collections; Historical Maps; Historical Photographs; Museum Artifacts; Photographic Archives; Vintage Media
Sports & Competition	Athletic Competitions; Motor Racing; Professional Sport Events; Sports Infrastructure; Traditional Sports
Science & Engineering	Engineering Projects; Research Labs; Scientific Equipment; Space & Defense; Technology & Innovation
Scripts & Multilingual Text	Code-Switched Text; Cross-Script Content; Handwritten Text; Multilingual Captions; OCR-Ready Images; Printed Text; Script Identification
Community & Social Life	Community Gatherings; Daily Routines; Family & Household; Public Events; Street Life; Villages & Rural Life
Urban Dev. & Buildings	City Planning; Commercial Complexes; Construction Sites; Heritage Architecture; Residential Buildings

Figure 2: SIGHT Hierarchical Taxonomy: 13 top-level categories with associated subcategories.

- ***Stage 2: Text Quality Scoring.*** Each caption received a heuristic quality score based on caption length, descriptiveness, and metadata penalties. Near-duplicate captions within each category are identified via fuzzy string matching ($\tau_{\text{fuzzy}} = 0.75$) and assigned a 50% score reduction. The mean text quality score is 0.781 after deduplication.
- ***Stage 3: Visual Deduplication.*** Near-duplicate images were removed using perceptual hashing (pHash) and structural similarity heuristics, reducing repetitive archival scans, monument photographs, and visually redundant media content. This stage retained 13,527 image–caption pairs.
- ***Stage 4: Cross-Modal Alignment Scoring.*** Each

Category	Raw	Filtered	Eval Set
Sports & Competition	1,950	742	520
Community & Social Life	2,840	1,021	615
Science & Engineering	1,760	694	560
Education & Literacy	1,740	681	545
Historical & Archival Data	3,260	1,072	610
Cultural & Societal	4,180	1,534	575
Commerce & Economy	2,120	801	500
Infrastructure & Public Spaces	2,410	973	495
Geographic & Environmental	7,120	2,614	480
Religion & Tradition	1,230	402	450
Digital & Print Media	1,980	756	420
Urban Development & Buildings	2,140	803	340
Scripts & Multilingual Text	2,964	430	241
Total	**35,694**	**12,523**	**6,351**

Table 1: Category-wise corpus statistics for the final 13-category taxonomy. **Raw** denotes collected images, **Filtered** denotes samples retained after automated filtering and validation, and **Eval Set** denotes the final subset used for manual annotation in the benchmark.

image–caption pair was assigned a CLIP alignment score $s_{\text{clip}} \in [0, 1]$. Pairs below $\tau_{\text{clip}} = 0.18$ were discarded, resulting in 12,523 retained pairs with a mean CLIP score of 0.735.

- ***Stage 5: Composite Scoring and Selection.*** A composite quality score, $s_{\text{composite}} = 0.4 \times s_{\text{text}} + 0.6 \times s_{\text{clip}}$, was used to rank entries, with stronger weighting assigned to cross-modal alignment. After category-balanced filtering, 6,351 image–caption pairs remain for annotation (mean score: 0.815). The final evaluation set contains 5,086 instances derived from this subset (Table 1).

3.3 Taxonomy Design

Motivation and Taxonomy Structure. Existing multimodal datasets often rely on coarse or shallow visual labels, limiting their ability to capture culturally grounded and diverse visual understanding (Birhane and Prabhu, 2021). To address this limitation, we develop a hierarchical taxonomy spanning cultural, geographic, institutional, linguistic, historical, and societal dimensions of multimodal content, including both globally common and Indic-specific visual domains. The taxonomy consists of 13 top-level categories and more than 75 fine-grained subcategories. Figure 2 and Appendix Figure 12 summarize the top-level categories and their representative subcategories.

3.4 Annotation Methodology

Annotator Setup. The annotation was carried out by a team of **six** graduate-level summer interns with demonstrated knowledge of image-based annotation tasks, multimodal data processing, and cultural and linguistic contexts. Annotators were recruited from institutions with relevant domain expertise in South Asian cultural contexts, with proficiency in English and at least one regional Indian language. Annotators worked in **two pairs**, each pair jointly responsible for annotating approximately **2,043 images**, ensuring overlapping coverage and facilitating pair-level consistency checks. A dedicated **quality controller** supervised the process, resolving annotation conflicts, correcting systematic biases, and ensuring adherence to the annotation guidelines. Approximately **135** instances were directly reviewed and annotated, while additional flagged samples were returned for re-annotation.

Prior to annotation, a detailed annotation guideline document was prepared covering category definitions, category boundaries, ambiguous cases, and label assignment procedures (Detailed annotation guidelines are provided in Appendix §A.3.).

Annotation Pipeline. Initial category and subcategory predictions were generated using *Llama-3.2-11B-Vision-Instruct* (Meta AI, 2024) and presented via a dedicated annotation portal to accelerate annotation. These served only as suggestions: annotators manually reviewed, corrected, or rejected all labels, with decisions logged for subsequent human–model disagreement analysis.

Annotation Task. The IMAGE CLASSIFICATION task requires hierarchical semantic annotation of multimodal visual content. Given an image and its associated metadata, annotators assign one or more primary and optional secondary categories, each paired with at least one corresponding subcategory. Secondary categories are used only when additional semantic concepts are visually prominent or contextually important. While captions and metadata may provide auxiliary context or source verification, annotations are grounded primarily in the visual content of the image.

3.5 Annotation Efforts and Quality Analysis

Annotation Effort. The annotation process required substantial human effort to review and refine automated predictions. A team of six annotators, organized into two groups of three, worked over a one-month period (approximately **720 total annotation hours**). Each image was independently reviewed by three annotators following annotation guidelines. From 6,351 curated instances, 5,086

Category	Krippendorff's α
Sports & Competition	0.886
Education & Literacy	0.772
Science & Engineering	0.768
Community & Social Life	0.766
Historical & Archival Data	0.753
Commerce & Economy	0.681
Cultural & Societal	0.679
Infrastructure & Public Spaces	0.677
Digital & Print Media	0.658
Geographic & Environmental	0.654
Religion & Tradition	0.646
Urban Development & Buildings	0.493
Scripts & Multilingual Text	0.400
Macro-average	**0.680**

Table 2: Per-category Krippendorff's Alpha (α) scores for the IMAGE CLASSIFICATION annotation task. The macro-average Alpha of 0.680 indicate moderate-to-substantial inter-annotator consistency.

quality-verified instances were retained in the final benchmark. Across the final evaluation set, 454 model predictions (7.15%) were empty or unparseable and required manual annotation. Overall, the combined ground-truth and predicted annotations required an average of **3.12 changes per image**, indicating substantial divergence between automated predictions and consolidated human annotations. Individually, Annotator 1 made an average of **3.16** changes per image, Annotator 2 made **3.58** changes, and Annotator 3 made **3.13** changes.

Inter-Annotator Agreement. To assess annotation reliability, we computed inter-annotator agreement (IAA) using Krippendorff's Alpha (Hayes and Krippendorff, 2007), a standard metric for evaluating consistency among multiple annotators on categorical labels. The macro-average Krippendorff's Alpha across all categories is **0.680**, with an overall Average Agreement Score of **0.8442**, indicating moderate-to-substantial consistency and validating the benchmark annotations (Table 2).

Agreement varied notably across categories, reflecting differences in annotation difficulty. Categories with visually distinctive and semantically well-defined content, such as *Sports & Competition* ($\alpha = 0.886$) and *Science & Engineering* ($\alpha = 0.768$), achieved the highest agreement. In contrast, categories involving multilingual text, dense urban scenes, or overlapping contextual signals showed lower agreement. *Scripts & Multilingual Text* ($\alpha = 0.400$), *Urban Development & Buildings* ($\alpha = 0.493$), and *Digital & Print Media* ($\alpha = 0.658$) were comparatively more challenging, largely due to inherent semantic overlap and the need for deeper contextual interpretation.

Per-category Krippendorff's α scores for the IMAGE CLASSIFICATION task are sorted by agreement level in Table 2. We group categories into three difficulty tiers based on agreement: *Low* ($\alpha \geq 0.74$), *Moderate* ($0.60 \leq \alpha < 0.74$), and *High* ($\alpha < 0.60$). Lower agreement in categories such as *Scripts & Multilingual Text* and *Urban Development & Buildings* primarily reflects genuine semantic ambiguity rather than annotation noise, making them particularly valuable for evaluating hierarchical cultural reasoning in VLMs.

4 Experimental Setup

We evaluate state-of-the-art VLMs on the SIGHT evaluation set, spanning open-weight and proprietary systems. Open-weight models include `Qwen3-VL-8B/32B` (Bai et al., 2025), `InternVL3-78B` (Zhu et al., 2025), `Llama4-Scout-17B` (Adcock et al., 2026), `GLM-4.6V` (GLM et al., 2024), `Kimi-VL-A3B`, `Gemma3-27B` (Kamath et al., 2025), and `Phi-4-Multimodal-14B` (Abouelenin et al., 2025); proprietary systems include `Claude-3-Haiku` (Anthropic, 2024) and `Gemini-3.1-Flash-Lite` (Team et al., 2023).

5 Evaluation

Given an input image, models predict one or more category–subcategory pairs from a predefined hierarchy spanning 13 semantic domains. We adopt a constrained instruction-following setup with the full taxonomy provided in the prompt and require structured JSON outputs for reliable parsing.

Evaluation Metrics and Protocol. We evaluate performance using multi-label Precision, Recall, and F_1 at two levels: (i) **category-level**, considering only top-level categories, and (ii) **pair-level**, requiring exact category–subcategory matches. We additionally report Exact Match (EM), which measures whether the full predicted label set exactly matches the ground truth. All predictions are normalized via lowercasing, whitespace cleaning, and constrained taxonomy mapping prior to scoring.

Since each image may contain multiple cultural, textual, geographic, and institutional concepts, evaluation follows a multi-label hierarchical setting. Ground-truth labels are obtained via multi-annotator consensus with quality-checking.

Model	Category-level					Pair-level			
	P	R	EM	M-F$_1$	μF$_1$	P	R	EM	μF$_1$
Open-source Models									
`Llama-4-Scout-17B-16E`	0.4765	0.6783	0.1040	0.5122	**0.5597**	0.2602	0.4410	0.0142	0.3273
`InternVL3-78B`	0.4759	0.6463	0.1484	0.5069	0.5482	0.2951	0.3643	0.0552	0.3261
`Gemma3-27B`	0.4638	0.7022	0.0997	**0.5233**	0.5586	0.2472	0.4438	0.0212	0.3175
`GLM-4.6V`	0.5098	0.6059	**0.1791**	0.5165	0.5537	**0.3142**	0.2909	**0.0745**	0.3021
`Kimi-VL-A3B-Thinking`	**0.5902**	0.4505	**0.2302**	0.4791	0.5110	**0.3742**	0.2366	**0.0999**	0.2899
`Qwen3-VL-8B`	0.5267	0.4696	0.2155	0.4653	0.4965	0.2497	0.2745	0.0820	0.2615
`Qwen3-VL-32B`	0.3896	**0.7869**	0.0684	0.4860	0.5212	0.1528	**0.4876**	0.0195	0.2327
Closed-source Models									
`Claude-3-Haiku`	0.4635	0.5958	0.1015	0.4539	0.5214	0.2517	0.3724	0.0163	0.3004
`Gemini-3.1-Flash-Lite`	0.3992	0.7255	0.0979	0.5021	0.5150	0.2306	0.4343	0.0397	0.3013

Table 3: Zero-shot evaluation performance across category-level and pair-level metrics on the benchmark (n = 5,086). EM = Exact Match, M-F$_1$ = Macro-F$_1$, and μF$_1$ = Micro-F$_1$.

Model	Category-level								Pair-level							
	Precision		Recall		M-F$_1$		μF$_1$		Precision		Recall		EM		μF$_1$	
	FS	Δ	FS	Δ	FS	Δ	FS	Δ	FS	Δ	FS	Δ	FS	Δ	FS	Δ
Open-source Models																
Gemma3-27B ↑	0.527	+0.063	**0.696**	-0.007	**0.536**	+0.013	**0.599**	+0.041	0.333	+0.086	0.421	-0.023	0.066	+0.045	**0.372**	+0.055
GLM-4.6V ↑	**0.538**	+0.028	0.610	+0.005	0.535	+0.018	0.572	+0.018	**0.388**	+0.073	0.317	+0.026	**0.093**	+0.018	0.349	+0.046
InternVL3-78B →	0.486	+0.010	0.584	-0.062	0.497	-0.010	0.531	-0.018	0.353	+0.058	0.306	-0.059	0.074	+0.018	0.327	+0.001
Kimi-VL-A3B-Thinking ↑	0.529	-0.061	0.562	+0.111	0.500	+0.020	0.545	+0.034	0.362	-0.012	0.291	+0.055	0.080	-0.020	0.323	+0.033
Qwen3-VL-32B ↑	0.415	+0.026	**0.778**	-0.009	0.500	+0.014	0.542	+0.020	0.239	+0.086	**0.479**	-0.009	0.015	-0.004	0.319	+0.086
Qwen3-VL-8B ↑	0.464	-0.063	0.630	+0.160	0.500	+0.035	0.534	+0.038	0.302	+0.053	0.331	+0.056	0.060	-0.022	0.316	+0.054
Llama-4-Scout-17B-16E ↓	0.522	+0.045	0.556	-0.123	0.491	-0.021	0.538	-0.022	0.295	+0.035	0.326	-0.115	0.050	+0.036	0.310	-0.018
Closed-source Models																
Claude-3-Haiku	0.4409	-0.0226	0.7042	+0.1084	0.4706	+0.0167	0.5423	+0.0209	0.2656	+0.0139	0.4611	+0.0887	0.0354	+0.0191	0.3370	+0.0366
Gemini-3.1-Flash-Lite	0.4081	+0.0089	0.7854	+0.0599	0.5051	+0.0030	0.5371	+0.0221	0.2531	+0.0225	0.4973	+0.0630	0.0425	+0.0028	0.3355	+0.0342

Table 4: Few-shot (5-shot) performance and relative improvement (Δ) over zero-shot across category-level and pair-level metrics. Positive Δ indicates improvement. M-F$_1$ = Macro-F$_1$ and μF$_1$ = Micro-F$_1$. Bold denotes the best few-shot value per metric column among open-source models.

Prompting Setup and Implementation Details. We evaluate models under zero-shot (ZS) and few-shot (FS) ($n = 5$) settings. The zero-shot prompt includes the full taxonomy, structured output constraints, and JSON-only generation instructions. The few-shot setting additionally provides five representative annotated examples demonstrating valid category–subcategory mappings. No parameter updates or task-specific fine-tuning are performed.

Prompt templates and model-specific hyperparameters are provided in Appendix §B, Figures 10 and 11, and Appendix §B.1.

6 Results and Observations

Tables 3 and 4 summarize ZS and FS performance across all models on the eval set of 5,086 instances.

6.1 Zero-Shot Performance

A consistent gap between category-level and pair-level μF$_1$ is observed across all models: the best category μF$_1$ reaches 0.560 (`Llama4-Scout-17B`), while the highest pair μF$_1$ is only 0.327—a drop of over 22 points. The gap is largest for `Qwen3-VL-32B` (0.521 vs. 0.233, nearly 29 points), confirming that category recognition is substantially easier than fine-grained category–subcategory grounding.

Aggressive over-prediction. `Qwen3-VL-32B` predicts an average of 3.56 labels per image against a ground-truth mean of 1.76, yielding high recall (0.787) but low precision (0.390). It predicts more than two labels in 77.6% of single-label instances. Despite leading on recall, its pair-level μF$_1$ (0.233) is the lowest among open-weight models: hallucinated categories introduce incorrect subcategory pairs that cannot be aligned with the ground truth. `Gemini-3.1-Flash-Lite` ($\bar{k}$ = 3.20) exhibits nearly identical behaviour in the closed-weight setting, high recall (0.726), low precision (0.399), and its pair μF$_1$ (0.301) is the weakest among proprietary models. `Gemma3-27B` and `Llama4-Scout-17B` show milder over-prediction ($\bar{k}$ = 2.67 and 2.51, respectively), with more than half of single-label samples receiving three or more predicted categories.

Setting	Cat-P	Cat-R	Cat-EM	Cat-M	Cat-μF1	Pair-P	Pair-R	Pair-EM	Pair-μF1
Zero-shot	0.5385	0.4784	23.52	0.4899	0.5066	0.2675	0.2989	10.87	0.2823
Few-shot	0.4927	0.6538	18.38	0.5287	0.5619	0.3147	0.3383	8.70	0.3261
SFT (LoRA)	**0.7564**	0.5695	**39.13**	0.5112	**0.6498**	**0.5878**	0.4093	**23.32**	**0.4826**

Table 5: Comparison of zero-shot, few-shot, and SFT settings for Qwen3-VL-8B-Instruct on the test set ($N = 506$). EM values are reported as percentages.

Conservative under-prediction. `Qwen3-VL-8B` and `Kimi-VL-A3B-Thinking` adopt the opposite strategy: average prediction counts of 1.35 and 1.57 yield the two highest precision scores (0.590 and 0.527) but recall below 0.47. Both models predict exactly one category for more than 70% of instances, effectively collapsing the multi-label task into single-label classification and systematically missing secondary categories such as *Digital & Print Media* and *Scripts & Multilingual Text*.

Balanced mid-range models. `GLM-4.6V-9B` and `InternVL3-78B` exhibit more balanced prediction behaviour ($\bar{k} \approx 2.1$–2.4), with the smallest precision–recall gaps (9.6 and 17.0 points, respectively). `InternVL3-78B` and `Llama4-Scout-17B` achieve the strongest pair-level μF_1 among open-weight zero-shot systems (0.326 and 0.327), suggesting that moderate label diversity produces more accurate category–subcategory predictions. `Claude-3-Haiku` ($\bar{k} = 2.34$) occupies a similar balanced position among closed-weight models, with precision 0.464 and recall 0.596. Its pair μF_1 (0.300), however, is the lowest among proprietary models, marginally below `Gemini-3.1-Flash-Lite` (0.301), indicating that API-scale advantages do not translate automatically to stronger hierarchical prediction.

Open-source vs. closed-source comparison. Closed-weight models cluster tightly in the 0.300–0.301 pair-level μF_1 range, falling below the strongest open-weight systems, `InternVL3-78B` (0.326) and `Llama4-Scout-17B` (0.327). This pattern is notable given the presumed scale, data, and optimization advantages of API-based models. Additionally, `Claude-3-Haiku` generates 505 out-of-taxonomy labels under zero-shot prompting, including near-synonyms such as *History And Archival Data* and unsupported labels like *Food And Cuisine*, indicating partial non-compliance with the prescribed label vocabulary. In contrast, `Llama4-Scout-17B` becomes more conservative under FS prompting, $\bar{k}$ drops from 2.38 to 1.78 as null predictions increase from 9 to 22.

Overall, the consistent 22+ point gap between category and pair μF_1 across every evaluated system underscores the fundamental difficulty of fine-grained hierarchical reasoning in multimodal cultural classification.

6.2 Few-Shot Performance and Zero-Shot Comparison

Five in-context examples yield broadly positive but uneven effects. Open-weight models improve more at the pair level than category level (+0.037 vs. +0.016 μF_1), indicating stronger gains for sub-category grounding. In closed-weight models, `Claude-3-Haiku` rises from 0.300 to 0.337 and `Gemini-3.1-Flash-Lite` from 0.301 to 0.336.

Multi-label prediction recalibration. The most consistent effect of few-shot prompting is a redistribution of predicted label counts. Previously under-predictive models expand their outputs: `Qwen3-VL-8B` increases from $\bar{k} = 1.57$ to 2.39, and `Kimi-VL-A3B` from 1.30 to 1.79. `Claude-3-Haiku` similarly expands from $\bar{k} = 2.34$ to 2.80, with a substantial recall gain (+0.104 at category level) at only a modest precision cost (-0.024). In contrast, `Llama4-Scout-17B` becomes more conservative under FS prompting, $\bar{k}$ drops from 2.38 to 1.78 as null predictions increase from 9 to 22. `Qwen3-VL-32B` and `Gemini-3.1-Flash-Lite` remain the most over-predictive systems in both settings ($\bar{k} = 3.54 \rightarrow 3.29$ and $3.20 \rightarrow 3.39$, respectively).

Largest improvements. `Qwen3-VL-32B` yields the largest absolute gain among open-weight models (+0.086 pair μF_1, from 0.233 to 0.319), driven by a precision increase from 0.153 to 0.239 while recall remains high. `Gemma3-27B` emerges as the strongest FS model overall, reaching pair μF_1 of 0.372 and pair sample-F1 of 0.397 through better precision calibration with only a minor recall reduction. Among closed-weight systems, both `Claude-3-Haiku` (0.337) and `Gemini-3.1-Flash-Lite` (0.336) benefit meaningfully from FS prompting, though neither sur-

passes the best open-weight FS model.

Regressions under few-shot prompting. `Llama-4-Scout-17B-16E` declines by 2.2 points in category μF_1 and 1.8 points in pair μF_1, driven by a recall collapse (0.678 → 0.556). Few-shot exemplars amplify *Cultural & Societal→Religion & Tradition* confusion (166 → 270) and introduce a new *Historical & Archival Data→Geographic & Environmental* confusion (142 instances), indicating over-generalisation from historical landscape examples to archival content. While, `InternVL3-78B` shows near-neutral change (pair μF_1: +0.001), with its dominant error, *Digital And Print Media→Community & Social Life*, remaining essentially unchanged (314 → 317).

Takeaway. Few-shot prompting improves precision–recall calibration but not hierarchical reasoning. Over- and under-predictive models adjust label coverage, yet the category–subcategory gap remains largely unchanged: even the best FS system (`Gemma3-27B`) reaches only 0.372 pair μF_1, confirming that in-context examples do not resolve the underlying hierarchical grounding limitation.

Detailed Failure analysis. We present category- and subcategory-level confusion analysis and representative failure cases in Appendix §D.1–D.4. Errors primarily arise from incorrect hierarchical grounding, cultural ambiguity, and poor multi-label calibration rather than basic visual recognition failure.

6.3 SFT as an Upper-Bound Probe

Zero-shot and few-shot results indicate that even proprietary models struggle with the hierarchical IMAGE CLASSIFICATION task. We examine whether the manually annotated labels are sufficient to bridge this gap via fine-tuning `Qwen3-VL-8B-Instruct` with LoRA on a stratified evaluation split and comparing it to its ZS and FS baselines.

The 5,086-instance evaluation set is split 80/10/10 (train/val/test, stratified by primary category). Fine-tuning uses LoRA (rank 64, $\alpha = 128$) for 3 epochs on 8 GPUs, converging to a training loss of 0.138 and evaluation loss of 0.116. All results are reported on a test set of 506 samples using the same evaluation protocol as the main Eval set.

Supervision resolves the hierarchical grounding bottleneck. Despite having only 8B parameters, the SFT model raises category μF_1 from 0.507 (ZS) to 0.650—surpassing the best zero-shot result across all evaluated models (0.560, `Llama4-Scout-17B-16E`, 4× larger). More strikingly, pair μF_1 nearly doubles: 0.282 → 0.483. This gain demonstrates that the persistent category–subcategory alignment failure under ZS and FS is not an intrinsic VLM limitation, but a direct consequence of absent in-domain supervision. The SFT model learns the annotation structure, including secondary-category thresholds, when to suppress over-tagging, and how to prefer fine-grained subcategories over shallow surface-level cues.

Precision calibration reflects learned annotation boundaries. The sharpest gains are in precision: category precision rises from 0.539 to 0.756, and pair precision from 0.268 to 0.588. The SFT model predicts an average of 1.31 categories per image—far closer to the ground-truth mean (1.74) than 5-shot (2.30)—reflecting learned adherence to expert annotators' conservative tagging convention. Category and pair exact match improve from 23.5% → 39.1% and 10.9% → 23.3%, respectively.

Results. The fine-tuned model surpasses every ZS and FS settings on every F1 metric (Table 5), with the largest gains at pair level, precisely where ZS and FS models most consistently fail.

Takeaway. Fine-tuning an 8B model on only ∼4,000 SIGHT instances outperforms every ZS baseline, including models more than 4× larger, and more than doubles pair-level μF_1. The primary bottleneck is not model scale, but the absence of high-quality annotated resources.

7 Conclusion and Future Directions

We introduce SIGHT, a culturally grounded multimodal benchmark for hierarchical multi-label classification, comprising 5,086 expert-annotated instances from a 5M-scale corpus. Across zero-shot, few-shot, and SFT settings, a persistent 22+ point category-to-pair μF_1 gap exposes fundamental limitations in hierarchical cultural reasoning, with recurring failures in document understanding, cultural distinction, and subcategory grounding. Fine-tuning on ∼4,000 instances enables an 8B model to surpass all ZS baselines and more than double pair-level μF_1, confirming that the primary bottleneck is in-domain supervision rather than model scale. Future work will expand the benchmark with additional tasks and annotations toward broader culturally inclusive multimodal evaluation.

Limitations

Scale of released annotations. The full SIGHT corpus contains approximately 5M deduplicated image–caption pairs collected from 450+ sources. Due to storage and annotation constraints, the current public release includes 5,086 expert-annotated instances forming a stratified evaluation benchmark, while annotation of additional samples is ongoing.

Annotation coverage and scalability. Annotating culturally grounded multimodal data at large scale is highly resource-intensive. The current benchmark required approximately 720 annotation hours from six trained annotators. To improve scalability, future work will explore multimodal models trained on the expert-annotated subset for large-scale pseudo-annotation with human verification for uncertain predictions.

Multilingual coverage. The current annotated subset represents an initial large-scale annotation effort and primarily focuses on English-captioned image–text pairs. Future versions of SIGHT will expand toward multilingual annotation and evaluation across diverse languages and scripts.

Taxonomy overlap and annotation ambiguity. The hierarchical multi-label taxonomy captures overlapping cultural semantics, which can lead to boundary ambiguity in visually similar categories (e.g., *Digital & Print Media* and *Scripts & Multilingual Text*). While such overlap reflects inherent complexity in South Asian visual culture, alternative taxonomic designs with different granularity or boundary definitions could yield different agreement patterns and model performance. We view the current taxonomy as a principled but non-exhaustive starting point, and future work may refine category boundaries and annotation guidelines to further reduce ambiguity.

Limited closed-source model coverage. Evaluation of proprietary frontier VLMs was constrained by API cost and rate limits. As a result, the current setup includes only two closed-source systems.

Ethics Statement

Our benchmark was curated exclusively from publicly accessible and legally accessible online sources, including government repositories, institutional archives, licensed cultural collections, and journalistic platforms. We restricted data collection to images accompanied by associated captions or metadata and excluded content that was unsafe, visually corrupted, irrelevant, or ambiguous during quality control. No personally sensitive information was intentionally collected, and the benchmark is intended solely for research purposes in multimodal understanding and hierarchical semantic classification.

The annotation process involved trained annotators operating under detailed annotation guidelines to ensure consistency, fairness, and accuracy. We acknowledge that despite extensive filtering and quality control, the benchmark may still contain residual cultural, geographic, or source-level biases inherited from publicly available web data. We therefore encourage responsible use of the benchmark and caution against deploying models trained on it in high-stakes or safety-critical settings without further evaluation and auditing. SIGHT is released under the CC BY-NC-SA 4.0 license, with the claims that the original copyright remains to the original authors of the news articles, under the existing copyright laws. The models are either publicly available on Hugging Face or proprietary and purchased through APIs. Our usage of these resources complies with their intended use. Upon acceptance, we will release our entire 5M benchmark corpus for academic-only purpose, under the same license.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, and 1 others. 2026. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Anthropic. 2024. Claude 3.5 haiku. https://www.anthropic.com/claude/haiku. Accessed: 2026-05-23.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv e-prints*, pages arXiv–2511.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE.

Melania Cabezas-García and Arianne Reimerink. 2022. Cultural context and multimodal knowledge representation: Seeing the forest for the trees. *Frontiers in psychology*, 13:824932.

Michael Callow and Leon Schiffman. 2002. Implicit meaning in visual print advertisements: a cross-cultural examination of the contextual communication effect. *International journal of advertising*, 21(2):259–277.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Ali Faraz, Shaharukh Khan, Raja Kolla, Akshat Patidar, Suranjan Goswami, Abhinav Ravi, Chandra Khatri, Shubham Agarwal, and 1 others. 2025. Indicvisionbench: Benchmarking cultural and multilingual understanding in vlms. *arXiv preprint arXiv:2511.04727*.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. CulFiT: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22413–22430, Vienna, Austria. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592.

Yuechen Jiang, Enze Zhang, Md Mohsinul Kabir, Qianqian Xie, Stavroula Golfomitsou, Konstantinos Arvanitis, and Sophia Ananiadou. 2026. Appear2meaning: A cross-cultural benchmark for structured cultural metadata inference from images. *Preprint*, arXiv:2604.07338.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2026. Glotocr bench: Ocr models still struggle beyond a handful of unicode scripts. *arXiv preprint arXiv:2604.12978*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *eprint arXiv: 2307.16125*.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and 1 others. 2025. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462*.

Zhang Li, Zhibo Lin, Qiang Liu, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiajun Song, Jiarui Zhang, Xiang Bai, and Yuliang Liu. 2026. Mdpbench: A benchmark for multilingual document parsing in real-world scenarios. *arXiv preprint arXiv:2603.28130*.

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. *eprint arXiv: 2501.01282*.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921.

Meta AI. 2024. meta-llama/llama-3.2-11b-vision. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision. Hugging Face model card, accessed 2026-05-17.

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*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2025. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-VLM collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21960–21974, Vienna, Austria. Association for Computational Linguistics.

Tushar Pranav, Eshan Pandey, Austria Lyka Diane Bala, Aman Chadha, Indriyati Atmosukarto, and Donny Soh Cheng Lock. 2025. Rice-vl: Evaluating vision-language models for cultural understanding across asean countries. *arXiv preprint arXiv:2512.01419*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449. ACM.

Rosa Suviranta and Tuomo Hiippala. 2025. Can multimodal large language models annotate low-and high-level features of multimodal artefacts? *International Journal of Humanities and Arts Computing*, 19(2):127–142.

Bryan Chen Zhengyu Tan, Weihua Zheng, Zhengyuan Liu, Nancy F. Chen, Hwaran Lee, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2026. BLEnD-vis: Benchmarking multimodal cultural understanding in vision language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4669, Rabat, Morocco. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Ming-Kun Xie, Jia-Hao Xiao, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. 2026. Are multimodal large language models good annotators for image tagging? *arXiv preprint arXiv:2602.20972*.

Srishti Yadav, Lauren Tilton, Maria Antoniak, Taylor Arnold, Jiaang Li, Siddhesh Milind Pawar, Antonia Karamolegkou, Stella Frank, Zhaochong An, Negar Rostamzadeh, and 1 others. 2025. Evaluation of cultural competence of vision-language models. *arXiv preprint arXiv:2505.22793*.

Haorui Yu, Diji Yang, Hang He, Fengrui Zhang, and Qiufeng Yi. 2026. Vulca-bench: A multicultural vision-language benchmark for evaluating cultural understanding. *arXiv preprint arXiv:2601.07986*.

Yuanhan Zhang Bo Li Songyang Zhang, Wangbo Zhao Yike Yuan Jiaqi Wang, Conghui He Ziwei Liu Kai

Chen, Dahua Lin Yuan Liu, and Haodong Duan. 2023. Mmbench: Is your multi-modal model an all-around player. *arXiv preprint arXiv:2307.06281*, 4.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Dataset Sources and Extraction Methodology

Data Source Coverage. The SIGHT collection pipeline incorporates a diverse range of publicly accessible sources spanning educational repositories, museum archives, multilingual news media, biodiversity datasets, tourism platforms, institutional repositories, and web-scale crawls. Representative global sources include Wikimedia Commons, the Wikimedia WIT dataset (Srinivasan et al., 2021), Internet Archive, Common Crawl, Wikisource, museum repositories, academic archives, and multilingual news media. Regionally grounded sources include IGNCA, the Archaeological Survey of India (ASI), ISRO, NASA, India Biodiversity Portal, NCERT resources, Institutional textbooks, Arxiv papers, Indian Railways, and major news portals and Internet archives. Together, these sources improve coverage of culturally grounded visual domains including historical archives, multilingual scene text, civic infrastructure, commerce, environmental contexts, and geographically diverse societal practices across both Indian and global settings.

A.2 Examples of Noisy Image Instances in the Scraped Data

During raw corpus collection, several categories of images were identified as unsuitable for annotation and removed prior to the human annotation phase. The examples below illustrate the primary noise types encountered, along with a demonstration of the flagging mechanism used by annotators when such instances were occasionally encountered during the annotation process.

Category	Raw	Retained
Historical & Archival Data	∼700K	∼470K
Cultural & Societal	∼690K	∼455K
Community & Social Life	∼660K	∼430K
Geographic & Environmental	∼650K	∼405K
Urban Development & Buildings	∼620K	∼375K
Infrastructure & Public Spaces	∼645K	∼420K
Commerce & Economy	∼630K	∼395K
Science & Engineering	∼610K	∼390K
Sports & Competition	∼590K	∼375K
Education & Literacy	∼565K	∼365K
Digital & Print Media	∼540K	∼350K
Religion & Tradition	∼520K	∼315K
Scripts & Multilingual Text	∼470K	∼255K
Total	**8.2M**	**∼5.0M**

Table 6: Category-level corpus statistics for the full SIGHT raw corpus. **Raw**: total images per category available across all 450+ crawled sources at collection time. **Retained**: images passing caption-presence filtering and deduplication (no further quality filtering applied at this stage). The overall retention rate is 61.0%. Categories are sorted by raw count.

Figure 3: **Out-of-Domain Content.** An image depicting a law enforcement arrest operation, likely sourced from a globally distributed news crawl. This image has no relevance to any of the 13 SIGHT taxonomy categories and was removed during hard filtering as non-Indian and out-of-domain content.

Annotator Flagging Mechanism. Although all instances of the noise types shown above were removed from the corpus prior to annotation, a small number of borderline cases were occasionally encountered by annotators during the annotation phase. In such cases, annotators were instructed to use the portal's built-in flagging mechanism rather than submit a low-confidence annotation. Figure 9 illustrates this process: an image depicting ascetics near the Ghats of Varanasi was flagged due to culturally sensitive and visually ambiguous content that could not be annotated with confidence. The flagging interface records the reason, severity, and annotator ID, routing the instance to the quality

Figure 4: **Archival Image with Copyright Watermark.** A mid-20th century black-and-white photograph sourced from the LIFE magazine archive, depicting a young woman in traditional attire. Despite its cultural relevance, the prominent *LIFE* watermark renders this instance unsuitable for model training and it was removed during quality filtering.

Figure 5: **Entertainment Promotional Content.** A promotional poster for a Malayalam film, sourced from a news media crawl. Such images represent copyrighted entertainment media unrelated to the culturally grounded semantic categories and were discarded during source-level filtering.

controller for review and potential removal. All flagged instances that could not be resolved through adjudication were discarded from the final benchmark.

A.3 Annotation Guidelines

Quality Control & Training Comprehensive annotation guidelines were prepared and distributed to all annotators prior to the start of the annotation process. The guidelines defined category boundaries, subcategory selection criteria, thresholds for assigning secondary categories, procedures for handling ambiguous samples, and common annotation errors with representative examples. Annotators were instructed to prioritize relevance, accuracy,

Figure 6: **Digital Advertisement — Astrology Content.** A digitally produced advertisement from the Times of India platform promoting horoscope and astrology services. While categorically proximate to *Digital & Print Media*, such algorithmically generated promotional banners lack the information required for reliable annotation and were excluded.

Figure 7: **Product Advertisement — Technology.** A digitally rendered promotional image for a smartphone, scraped from a technology news portal. Product advertisements of this kind introduce commercial visual noise inconsistent with our taxonomy and were removed during filtering.

Figure 8: **Horoscope Graphic — Text-Dominant Media.** A text-dominant digital graphic for a daily Cancer horoscope from the Times of India, April 3, 2026. Although this instance contains legible printed text and could superficially map to *Scripts & Multilingual Text* or *Digital & Print Media*, it lacks substantive visual or cultural content and was excluded as low-utility media noise.

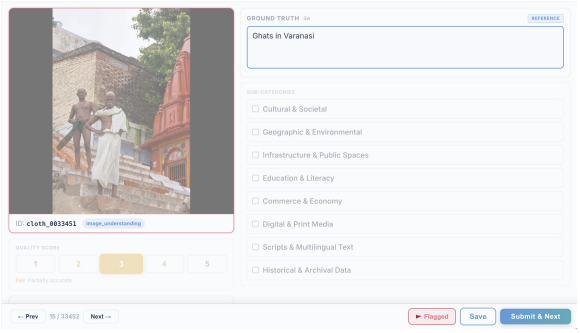

Figure 9: **Annotator Flagging Interface.** An image depicting ascetics near the Ghats of Varanasi, shown as it appeared in the annotation portal when flagged by an annotator. The instance was and flagged rather than submitted, as the visual content was deemed culturally sensitive and insufficiently unambiguous for confident hierarchical labelling. All such flagged instances were reviewed by the quality controller and discarded from the final benchmark if a reliable annotation could not be determined.

consistency, and completeness throughout the annotation workflow.

To further ensure dataset reliability, each image–annotation pair was assigned a quality score on a scale of 1–5 based on its usefulness for multimodal model training. The quality assessment considered factors such as subject clarity, alignment between the image and assigned labels, annotation specificity, and overall semantic usefulness. Images that were visually corrupted, taxonomically ambiguous, irrelevant to the defined taxonomy, or unsafe for annotation were flagged and removed during quality control. The final retained corpus therefore consists only of high-confidence annotations that passed both manual review and minimum quality thresholds.

Overview. The SIGHT annotation task requires hierarchical semantic classification of multimodal visual content. Given an image and its associated caption metadata, annotators assign one or more *primary categories* representing the dominant semantic theme, and optionally one or more *secondary categories* when additional semantic concepts occupy a substantial portion of the scene or are essential for contextual understanding. Every selected category must be accompanied by at least one subcategory. Annotations are grounded primarily in visual evidence; captions serve only as auxiliary context for source verification.

Key Definitions.

- **Category.** A top-level label placing an image within one of 13 broad semantic domains. Every image must be assigned at least one category.
- **Subcategory.** A fine-grained label nested within a category. Every confirmed category must have at least one subcategory; a category may never be submitted without one.
- **Primary Category.** The dominant category — the one that best represents what the image is fundamentally about. When uncertain, ask: *if I could attach only one label, which category would a researcher most want to retrieve this image under?*
- **Secondary Category.** An additional category assigned only when a secondary subject occupies approximately 20% or more of the visual frame, or is clearly central to understanding the full meaning of the image. Use sparingly.
- **Prediction.** A category or subcategory prefilled by the AI assistance system. Predictions are suggestions only — always review and confirm or correct before submitting.

Annotation Principles.

Principle	What It Means in Practice
Relevance	Select only categories reflecting clearly visible, meaningful content — no inferences.
Accuracy	Every label must match the actual, observable subject of the image.
Completeness	Include every category with meaningful visual presence; do not omit important subjects.
Consistency	Apply the same threshold uniformly across all images.

Annotation Workflow. Annotators follow six steps in order for every image:

1. **Examine the full image** (10–15 seconds) before selecting any label.
2. **Identify the primary category** that best describes the dominant subject.
3. **Select at least one subcategory** within the primary category.
4. **Check for secondary categories** add only if the subject meets the 20% threshold or is essential to understanding the scene.

5. **Assign a dataset quality score** from 1 to 5 based on training usefulness.

6. **Submit and monitor flags** if flagged by the quality controller, re-examine from scratch and resubmit.

Secondary Category Threshold.

A subject must occupy approximately **20% or more** of the image **or** be clearly central to understanding the scene, whichever applies first. If uncertain, do not add the secondary category.

Correct use. A classroom photograph where a cultural festival banner occupies roughly one-third of the frame. Primary: *Education & Literacy*. Secondary: *Cultural & Societal*.

Incorrect use. A market scene where a small temple is barely visible in the far background. The temple does not meet the 20% threshold; adding *Religion & Tradition* as a secondary category would be incorrect.

Working with Predictions.

Scenario	Required Action
Prediction is correct	Keep it — no changes needed.
Prediction is incorrect	Remove entirely; add the correct category and subcategory.
Prediction is partially correct	Keep the category; correct the subcategory.
Category is missing	Add it manually alongside valid predictions.
Too many predictions	Remove irrelevant categories; retain only those meeting the threshold.
No categories predicted	Annotate from scratch; record "No categories were predicted" in the Feedback field.

Dataset Quality Scoring.

Score	Label	Criteria
5	Excellent	Image clearly represents the category. Annotation is precise and unambiguous. A near-perfect training example.
4	Good	Mostly clear with minor noise or slight ambiguity that does not undermine the annotation.
3	Acceptable	Category is identifiable but representation is weak or subcategory requires judgment. Usable but not ideal.
2	Weak	Poor alignment or ambiguous ground truth that limits training value.
1	Unusable	Irrelevant, misleading, or too ambiguous to annotate reliably. Flag for removal.

Handling Poor or Irrelevant Images. Flag rather than submit with a low-confidence guess if any of the following apply: the subject is not identifiable; the image is severely corrupted; the content does not belong to any of the 13 categories; or the content is unsafe. Flagging is a quality safeguard, not a failure.

Conflict Resolution. Many disagreement cases arose from semantically entangled scenes where textual, architectural, institutional, and cultural cues co-occurred simultaneously. For example, public event photographs frequently contained multilingual signage, heritage infrastructure, and social gathering activities within the same frame, producing legitimate overlap between taxonomy branches rather than purely annotation noise.

When three annotators assigned completely different labels for the same image, a dedicated quality controller manually reviewed the instance, examining the image content, annotation rationale, and taxonomy consistency before determining the final label assignment. Instances were retained only when a semantically valid interpretation aligned with the taxonomy guidelines; otherwise, ambiguous or low-confidence instances were discarded.

Common Annotation Errors.

Error	Correct Approach
Accepting predictions without inspection	Always examine the image independently. The model makes systematic errors on culturally ambiguous content.
Over-tagging	Do not assign a category for minor or incidental elements.
Under-tagging	If a category is clearly visible and meaningful, it must not be omitted.
Assumption-based tagging	All selections must be grounded in visual evidence only.
Category without subcategory	Every confirmed category must have at least one subcategory before submission.
Misusing the quality score	Rate training usefulness, not photographic clarity.

Re-Annotation Process. When a quality controller flag is received, annotators must: (1) re-open the flagged image; (2) re-examine from scratch without relying on the previous submission; (3) remove incorrect categories; (4) add missing categories; (5) correct subcategories; (6) update the quality score if required; and (7) resubmit. Each flagged case is treated as a completely fresh annotation task.

Guiding Principle. Before submitting, annotators are asked: (1) Does the category accurately rep-

resent what is visible? (2) Is the annotation clear, specific, and defensible? (3) Would another trained annotator reasonably reach the same conclusion? If uncertain, the annotation must be reconsidered or the image flagged.

A.4 Examples Of Our Dataset on the Annotation Tool

Figure 12 presents the full SIGHT taxonomy with representative examples across all 13 categories. Figures 13–31 illustrate the annotation portal in detail, covering four aspects of the pipeline: (1) **core annotation** (Figures 13–22, 23–24), showing category and subcategory assignment, multi-label selection, model mis-prediction correction, and validation error handling; (2) **zoom and inspection** (Figures 15–16), demonstrating the full-screen overlay for fine-grained visual examination; (3) **quality control** (Figures 25–29), illustrating the admin flagging mechanism, annotator revision workflow, and flag resolution; and (4) **admin management** (Figures 30, 31), covering the flag manager dashboard and user management panel.

B Experimental Setup

Zero-shot Prompt

The zero-shot prompt template used for all evaluated VLMs is shown in Figure 10.

Few-shot Prompt (n=5)

The few-shot prompt template used for all evaluated VLMs is shown in Figure 11.

B.1 Model Configurations

The supervised fine-tuning (SFT) model is trained using LoRA with rank $r = 64$ and scaling factor $\alpha = 128$ for 3 epochs across 8× NVIDIA H200 GPUs. Training converges to a final train loss of 0.138 and evaluation loss of 0.116. All reported results are evaluated on a held-out test set of 506 samples using the same evaluation protocol as the main benchmark, ensuring direct comparability with ZS and FS experiments.

C Computation Requirement and Budget

The experiments were conducted using API-based access to state-of-the-art VLMs through OpenRouter. API credits were used for `Claude-3-Haiku` and `Gemini-3.1-Flash-Lite`, with a total estimated API expenditure of approximately 50. Computational experiments were carried out on an institutional high-performance computing cluster comprising nodes equipped with **8× NVIDIA H200 GPUs** and **224 CPU cores** per node.

D Results and Observations

D.1 Category-Level Failure Analysis

Table 8 reports per-category recall for all models, computed from instance-level prediction records using taxonomy-normalised category labels. Three structural failure modes emerge.

Universal confusion to *Digital And Print Media*. This category ($n = 970$, 19.1% of instances) is the single most difficult across all models, with ZS recall ranging from 0.182 (`Kimi-VL-A3B-Thinking`) to 0.511 (`Gemini-3.1-Flash-Lite`), and the majority of models below 0.40. The dominant confusion target is *Community And Social Life*: `InternVL3-78B` mislabels 314 *Digital And Print Media* images as *Community And Social Life*, `Llama-4-Scout-17B-16E` 373, `GLM-4.6V` 305, and `Kimi-VL-A3B` 173. A secondary confusion target is *Infrastructure And Public Spaces* (`Llama-4-Scout-17B-16E`: 241; `Qwen3-VL-32B`: 212). Among closed-source models, `Gemini-3.1-Flash-Lite` reports 558 false positives and 405 false negatives for this category under 5-shot evaluation, while `Claude-3-Haiku` accumulates 399 false positives and 426 false negatives under ZS. These errors reflect a consistent failure to recognise the *representational* or *mediated* nature of content such as news clippings, posters, screenshots, and digital displays: models attend to depicted subjects rather than the medium itself.

***Cultural And Societal↔Religion And Tradition* confusion.** This is the largest pairwise confusion in the benchmark (up to 395 instances per model under FS). Conservative models over-assign to *Religion And Tradition*, while over-predictive models hallucinate it as an additional label. Few-shot prompting worsens this error in two of seven open-source models (`Llama-4-Scout-17B-16E`: 166→270; `GLM-4.6V`: 175→273). Among closed-source models, `Claude-3-Haiku` produces the largest C&S→R&T confusion count: 245 instances ZS, reducing to 165 under FS prompting. `Gemini-3.1-Flash-Lite` is comparatively better calibrated in this direction (156 ZS, 124 FS) but exhibits a strong positive-class bias for *Religion And Tradition* overall (956 false positives under FS

evaluation).

Weak *Scripts And Multilingual Text* recognition. Despite being a key component of the benchmark, this category ($n = 126$) exhibits the lowest recall across all non-trivially-sized categories under ZS evaluation. `Kimi-VL-A3B-Thinking` achieves recall of only 0.095, `Qwen3-VL-8B` 0.270, and `Llama-4-Scout-17B-16E` 0.349. Only `Qwen3-VL-32B` (0.831) approaches reliable detection ZS. Among closed-source models, `Claude-3-Haiku` records the lowest recall at 0.270. In contrast, `Gemini-3.1-Flash-Lite` achieves near-perfect recall (0.968 ZS, 0.929 5-shot), though with very high false positives (1,802 under 5-shot and 1,457 ZS), indicating aggressive detection of script-like features in non-target images. Most models suppress this category in favour of *Cultural And Societal* or *Historical And Archival Data*, suggesting that script diversity is not treated as an independent visual signal.

D.2 Subcategory-Level Confusion

Table 9 summarizes the dominant subcategory-level confusion patterns across models, considering only instances where the parent category is correctly predicted.

Over-prediction of *Traditional Attire And Jewelry*. Within *Cultural And Societal*, virtually all models show a strong pull toward *Traditional Attire And Jewelry* as a substitute for more nuanced subcategories. `InternVL3-78B` mislabels *Arts, Music And Dance* as *Traditional Attire And Jewelry* 171 times and *Religious And Spiritual Imagery* 161 times. `Qwen3-VL-32B` makes this substitution for *Religious And Spiritual Imagery* 272 times. Among closed-source models, `Gemini-3.1-Flash-Lite` commits 118 such substitutions for *Arts, Music And Dance* and `Claude-3-Haiku` 87, indicating a shared bias toward salient surface features such as textiles and ornaments.

Collapse of *Heritage Collections* into *Historical Photographs*. Within *Historical And Archival Data*, `Qwen3-VL-32B` commits 314 such substitutions, `Qwen3-VL-8B` 302, `Gemma3-27B` 223, and `Kimi-VL-A3B` 152. *Archival Documents* are recurrently misclassified as *Historical Photographs* by `Gemma3-27B` (147), `Qwen3-VL-8B` (84), and `Claude-3-Haiku` (74 under 5-shot), and as *Cultural Artifacts* by `Gemma3-27B` (160) and `Kimi-VL-A3B` (53).

Hierarchy collapse: category name returned as subcategory. `GLM-4.6V` and `Qwen3-VL-32B` return the category-level label itself as a subcategory value rather than selecting a valid fine-grained subcategory. `GLM-4.6V` commits this in 540 instances for *Infrastructure And Public Spaces* and 318 for *Historical And Archival Data*; `Qwen3-VL-32B` in 893 and 1,074 respectively. `Gemini-3.1-Flash-Lite` exhibits the same pathology at larger scale: 1,624 instances of *Infrastructure And Public Spaces* hierarchy collapse in ZS evaluation, manifesting as 384, 271, and 189 substitutions for the *Public Utility Visuals*, *Government Institutions*, and *Transport Systems* subcategories respectively.

`Llama-4-Scout-17B-16E`'s bias toward *Cultural Symbols*. `Llama-4-Scout-17B-16E` shows a distinctive tendency to predict *Cultural Symbols* as the subcategory for *Cultural And Societal* images regardless of content, substituting it for *Arts, Music And Dance* (212 cases) and *Festivals And Rituals* (167 cases). No other model exhibits this pattern at comparable scale.

***Digital And Print Media* subcategory errors.** Even when the category is correctly identified, subcategory discrimination fails consistently. `Qwen3-VL-8B` substitutes *News Clippings* with *Posters And Banners* in 106 cases; `Gemini-3.1-Flash-Lite` makes the same substitution in 106 cases under ZS and 97 under 5-shot; `Claude-3-Haiku` does so in 83 ZS instances.

D.3 Subcategory Granularity Bottleneck

Conditional subcategory F_1 (computed only when the parent category is correctly predicted) reveals persistent difficulty in fine-grained discrimination. *Historical And Archival Data* yields conditional subcategory F_1 values of 0.20–0.54 across models, reflecting the difficulty of distinguishing among visually overlapping subtypes such as *Archival Documents*, *Cultural Artifacts*, and *Historical Photographs*—the same three types that dominate the subcategory confusion table. *Scripts And Multilingual Text* exhibits similarly low conditional subcategory F_1 (0.25–0.41), as does *Digital And Print Media* (0.34–0.46).

In contrast, *Geographic And Environmental* (0.71–0.82) and *Sports And Competition* are substantially easier, reflecting clearer perceptual sepa-

rability between subcategories such as *Landscapes And Terrain* and *Transport Systems*, and stronger low-level visual cues that contemporary VLMs exploit reliably. Notably, *Sports And Competition* achieves the highest category-level recall across all models (≥ 0.872), indicating near-ceiling performance on visually distinctive, culturally unambiguous content.

D.4 Summary of Failure Modes

We identify three dominant failure modes from instance-level prediction records:

1. **Document-context insensitivity.** Models classify visual content without recognising the representational medium, producing systematic errors on *Digital And Print Media* (recall 0.18–0.51 ZS). The primary confusion target is *Community And Social Life*: `InternVL3-78B` produces 314 such confusions in ZS and 317 in FS; `Gemini-3.1-Flash-Lite` 264 in ZS and 208 in FS; `Claude-3-Haiku` 196 in ZS and 96 in FS. FS prompting reduces some errors but performance remains unreliable (Gemini: 0.511→0.582; Claude-Haiku: 0.369→0.560). `Claude-3-Haiku` additionally confuses *Digital And Print Media* with *Cultural And Societal* (270 ZS) and *Infrastructure And Public Spaces* (245 ZS), suggesting model-family-specific representational biases. Across models, the common failure is attention to depicted subjects rather than the documentary frame itself.

2. **Hierarchical subcategory collapse.** Subcategory predictions systematically default to visually salient or high-frequency types such as *Historical Photographs*, *Traditional Attire And Jewelry*, and *Cultural Symbols*. More severely, `GLM-4.6V` and `Qwen3-VL-32B` return the category name itself as a subcategory value, producing 318–1,074 hierarchy-collapse errors per category. `Gemini-3.1-Flash-Lite` exhibits the same pathology at larger scale, with 1,624 such errors for *Infrastructure And Public Spaces* under ZS. A related failure appears in closed-source models: `Gemini-3.1-Flash-Lite` (198 instances) and `Claude-3-Haiku` (151 instances) predict truncated labels such as *Arts* instead of *Arts, Music And Dance*, suggesting instruction-following preferences for shorter outputs.

3. **Uncalibrated label-set cardinality.** Predicted label counts range from $\bar{k} = 1.30$ (`Kimi-VL-A3B`, ZS) to $\bar{k} = 3.54$ (`Qwen3-VL-32B`, ZS), against a ground-truth mean of 1.76. Closed-source models occupy the over-predictive regime: `Gemini-3.1-Flash-Lite` predicts $\bar{k} = 3.20$ categories per instance (ZS) and $\bar{k} = 3.39$ (FS), while `Claude-3-Haiku` predicts $\bar{k} = 2.34$ (ZS) and $\bar{k} = 2.80$ (FS). Both diverge further from the ground-truth mean under FS prompting. At the opposite extreme, `Claude-3-Haiku` produces 188 null predictions (3.7%) in ZS and 192 (3.8%) in FS, comparable to `Kimi-VL-A3B`'s 133 null predictions (2.6%). `Gemini-3.1-Flash-Lite`, by contrast, produces near-zero null predictions (<0.1%) in both regimes.

Overall, the results demonstrate that culturally grounded hierarchical visual classification remains substantially more challenging than standard multi-label recognition. Current VLM limitations stem less from coarse object or scene recognition failure and more from inadequate culturally grounded compositional understanding, script-aware reasoning, and hierarchical semantic calibration.

You are an expert image classifier. Your task is to classify a given image into one or more categories and their subcategories from the taxonomy below.

Taxonomy

1. **Commerce And Economy**: [Agricultural Practices, Banking And Finance, Industrial Activity, Markets And Trade, Street Vendors]
2. **Community And Social Life**: [Community Gatherings, Daily Routines, Family And Household, Public Events, Street Life, Villages And Rural Life]
3. **Cultural And Societal**: [Arts, Music And Dance, Cultural Symbols, Festivals And Rituals, Food And Cuisine, Occupations And Professions, Religious And Spiritual Imagery, Traditional Attire And Jewelry]
4. **Digital And Print Media**: [Advertisements, Archival Print, Digital Screens, Memes And Visual Narratives, News Clippings, Posters And Banners, Social Media Posts]
5. **Education And Literacy**: [Classroom Photographs, Education And Literacy, Educational Institutions, Exam Papers, Textbooks, Learning Materials, Literacy Campaigns]
6. **Geographic And Environmental**: [Ecological Zones, Flora And Fauna, Landscapes And Terrain, River Systems And Water Bodies, Weather Patterns]
7. **Historical And Archival Data**: [Archival Documents, Cultural Artifacts, Heritage Collections, Historical And Archival Data, Historical Maps, Historical Photographs, Museum Artifacts, Photographic Archives, Vintage Media]
8. **Infrastructure And Public Spaces**: [Government Institutions, Infrastructure And Public Spaces, Private Institutions, Public Utility Visuals, Smart City Assets, Transport Systems]
9. **Religion And Tradition**: [Devotional Objects, Pilgrimage Sites, Religious Ceremonies, Sacred Texts And Symbols, Temples And Places Of Worship]
10. **Science And Engineering**: [Engineering Projects, Research Labs, Scientific Equipment, Space And Defense, Technology And Innovation]
11. **Scripts And Multilingual Text**: [Code-Switched Text, Cross-Script Content, Handwritten Text, Multilingual Captions, Ocr-Ready Images, Printed Text, Script Identification]
12. **Sports And Competition**: [Athletic Competitions, Motor Racing, Traditional Sports, Professional Sport Events, Sports Infrastructure]
13. **Urban Development And Buildings**: [City Planning, Commercial Complexes, Construction Sites, Heritage Architecture, Residential Buildings]

Rules:

1. Select ONE or MORE categories that apply to the image.
2. For each selected category, select ONE or MORE of its listed subcategories.
3. Only use category and subcategory names EXACTLY as written above.
4. Return ONLY a valid JSON array with no additional text, explanation, or markdown.

Output Format

```
[
  {"category": "<exact category name>",
   "subcategories":
   ["<exact subcategory>", ...]},
  ...
]
```

Figure 10: Zero-shot prompt template used for all evaluated VLMs.

You are an expert image classifier. Your task is to classify a given image into one or more categories and their subcategories from the taxonomy below. You will be provided with five labeled examples before the test image.

Taxonomy

1. **Commerce And Economy**: [Agricultural Practices, Banking And Finance, Industrial Activity, Markets And Trade, Street Vendors]
2. **Community And Social Life**: [Community Gatherings, Daily Routines, Family And Household, Public Events, Street Life, Villages And Rural Life]
3. **Cultural And Societal**: [Arts, Music And Dance, Cultural Symbols, Festivals And Rituals, Food And Cuisine, Occupations And Professions, Religious And Spiritual Imagery, Traditional Attire And Jewelry]
4. **Digital And Print Media**: [Advertisements, Archival Print, Digital Screens, Memes And Visual Narratives, News Clippings, Posters And Banners, Social Media Posts]
5. **Education And Literacy**: [Classroom Photographs, Education And Literacy, Educational Institutions, Exam Papers, Textbooks, Learning Materials, Literacy Campaigns]
6. **Geographic And Environmental**: [Ecological Zones, Flora And Fauna, Landscapes And Terrain, River Systems And Water Bodies, Weather Patterns]
7. **Historical And Archival Data**: [Archival Documents, Cultural Artifacts, Heritage Collections, Historical And Archival Data, Historical Maps, Historical Photographs, Museum Artifacts, Photographic Archives, Vintage Media]
8. **Infrastructure And Public Spaces**: [Government Institutions, Infrastructure And Public Spaces, Private Institutions, Public Utility Visuals, Smart City Assets, Transport Systems]
9. **Religion And Tradition**: [Devotional Objects, Pilgrimage Sites, Religious Ceremonies, Sacred Texts And Symbols, Temples And Places Of Worship]
10. **Science And Engineering**: [Engineering Projects, Research Labs, Scientific Equipment, Space And Defense, Technology And Innovation]
11. **Scripts And Multilingual Text**: [Code-Switched Text, Cross-Script Content, Handwritten Text, Multilingual Captions, OCR-Ready Images, Printed Text, Script Identification]
12. **Sports And Competition**: [Athletic Competitions, Motor Racing, Traditional Sports, Professional Sport Events, Sports Infrastructure]
13. **Urban Development And Buildings**: [City Planning, Commercial Complexes, Construction Sites, Heritage Architecture, Residential Buildings]

Few-shot Examples **Example 1** Image: <IMAGE_1> Output:

```
[
  {"category": "Cultural And Societal",
   "subcategories": ["Festivals And Rituals", "Religious And Spiritual Imagery"]}
]
```

Example 2 Image: <IMAGE_2> Output:

```
[
  {"category": "Digital And Print Media",
   "subcategories": ["News Clippings", "Posters And Banners"]}
]
```

Example 3 Image: <IMAGE_3> Output:

```
[
  {"category": "Infrastructure And Public Spaces",
   "subcategories": ["Transport Systems", "Public Utility Visuals"]}
]
```

Example 4 Image: <IMAGE_4> Output:

```
[
  {"category": "Scripts And Multilingual Text",
   "subcategories": ["Printed Text", "Cross-Script Content"]}
]
```

Example 5 Image: <IMAGE_5> Output:

```
[
  {"category": "Historical And Archival Data",
   "subcategories": ["Historical Photographs", "Heritage Collections"]}
]
```

Test Instance Image: <IMAGE_TEST>

Task Classify the test image using the patterns learned from the examples and the taxonomy provided earlier.

Rules

1. Select one or more categories strictly from the taxonomy.
2. For each category, select one or more valid subcategories.
3. Do not introduce new labels or synonyms.
4. Ensure strict category–subcategory consistency.
5. Use exact string matching (case-sensitive).
6. Return only valid JSON with no explanation or markdown.

Output Format

```
[
  {
    "category": "<exact category name>",
    "subcategories": ["<exact subcategory>", ...]
  }
]
```

Figure 11: Few-shot prompt template used for all evaluated VLMs.

Dataset	Unique Img.	Ann. Inst.	Total Inst.	H.A.	Annotators	Topics / Coverage
CulturalVQA 2024	2,328	2,378 Q	2,378 Q 7,206 Ans.	✓	Community (11 countries)	Food, drink, clothing, rituals, traditions; 11 countries
SEA-VQA 2024	515	1,999 Q	1,999 Q	✓	Grad. students (8 countries)	Heritage, traditions, landmarks; 53 cultures from 8 South East Asia countries
K-Viscuit 2025	657	657 MCQ	657 MCQ	✓[v]	Native Korean speakers	Food, arts, architecture, celebrations; Korea only
RAVENEA 2025	1,868	2,986 total	2,331 Q 655 Cap.	✓	Experts (κ=0.83)	Cultural entities & captioning; 11 categories, 8 countries
SCB 2025	1,065	3,178 Q	3,178 Q (1,093 uniq.)	✓	Human annotators	Cultural artefacts; 138 cultural artifacts across 5 categories from 7 Southeast Asia countries
CVQA 2024	10,374	10,374 Q	10,374 Q	✓	75+ native speakers (30 countries)	Food, pop culture, landmarks, history, nature; 30 countries, 31 languages, at least 200 questions per Country-Language subset
BLEnD-Vis 2026	4,916	21K+ MCQ	21K+ MCQ	✓[v]	Human validation	Everyday cultural knowledge; 16 world regions
IndicVisionBench 2025	4,993	37,740 Q	37,740 Q	✓	Experts (11 languages)	OCR, MMT, and VQA on Indian content across 13 topics, with parallel annotations in 10 Indic languages.
ALM-Bench 2025	2,929	22,763 Q	22,763 Q	✓	Native speakers (800+ hrs)	Heritage, customs, architecture, music, sports; 100 languages, 73 countries, 19 cat., 24 scripts.
RICE-VL 2025	7,000	29K total	28K Q 1K VG	✓	Cultural experts (11 countries)	VQA & visual grounding; 14 cultural domains, 11 ASEAN countries
VULCA-Bench 2026	7,410	7,410 pairs	7,410 pairs	✓	Native bilingual critics	Art critique; 8 traditions, 5-layer framework, 225 dimensions (ZH–EN)
CultureVerse 2025	—	31K test[‡]	228K 31K test[‡]	✓[t]	Expert QC (test only)	15 cultural topics; 188 countries, 11,085 concepts
WorldCuisines 2025	6,045	1.15M[v]	1.15M	✓[v]	Human validation (semi-automatic)	Food & cuisine only; 2,414 dishes, 30 languages
SIGHT (ours)	**5M+**	**5,086**	**5,086 ann.**	✓	**6 annotators** + QC (κ=0.68)	**13-category taxonomy with 75+ fine-grained subcategories**, each annotated by 3 human annotators with additional quality verification.

[†] BLEnD-Vis images are AI-generated. [‡] CultureVerse training set is auto-constructed; only the 31,382-sample test set includes expert QC. [t] Test set only. [v] Human *verification* of semi-automatically generated content. ✗ Fully automated annotation. *Abbreviations:* H.A. = Human annotated; Cap. = captioning pairs; VG = visual grounding pairs; '—' = not independently reported; Unique Img. = number of unique images; Ann. Inst. = human-annotated instances; Total Inst. = total annotated instances.

Table 7: Comparison of culturally grounded multimodal benchmarks across scale, annotation effort, and coverage. The table highlights differences in dataset size, human annotation practices, and cultural diversity across existing resources. SIGHT is the only dataset with manually-annotated and verified hierarchical multi-label annotations spanning fine-grained cultural categories.

Category	n Instances	InternVL3-78B	Llama-4-Scout-17B-16E	GLM-4.6V-9B	Gemma3-27B	Qwen3-VL-32B	Qwen3-VL-8B	Kimi-VL-A3B	Gemini-3.1-Flash-Lite	Claude-3-Haiku
Sports & Comp.	520	.904	.872	.904	.904	.904	.894	.915	.947	.915
Religion & Trad.	450	.803	.732	.741	.700	.865	.672	.517	.881	.635
Hist. & Archival	610	.795	.762	.700	.902	.878	.625	.587	.624	.523
Geog. & Environ.	480	.779	.693	.614	.765	.804	.506	.570	.772	.773
Infra. & Pub.Sp.	495	.632	.801	.551	.507	.891	.413	.358	.727	.713
Cultural & Soc.	575	.632	.747	.682	.914	.862	.404	.424	.853	.677
Educ. & Literacy	545	.616	.631	.620	.537	.706	.495	.593	.632	.486
Community & Soc.	615	.589	.717	.616	.706	.773	.391	.389	.676	.449
Science & Eng.	560	.591	.655	.609	.555	.682	.636	.527	.741	.580
Commerce & Econ.	500	.527	.496	.629	.678	.572	.542	.485	.630	.268
Urban Dev.	340	.568	.410	.418	.637	.608	.326	.403	.868	.304
Digital & Print	420	.237	.252	.239	.391	.472	.374	**.182**	.511	.369
Scripts & Mult.	241	.669	.356	.432	.500	.831	.263	**.102**	.968	.270

Table 8: Per-category zero-shot recall across all evaluated models. n denotes the number of instances in the evaluation set. Cells at or below 0.25 are shaded red; cells at or above 0.80 are shaded green. Lowest recall per category is **bolded**. Closed-source models (`Gemini-3.1-Flash-Lite`, `Claude-3-Haiku`) use the same zero-shot evaluation protocol as the open-source models.

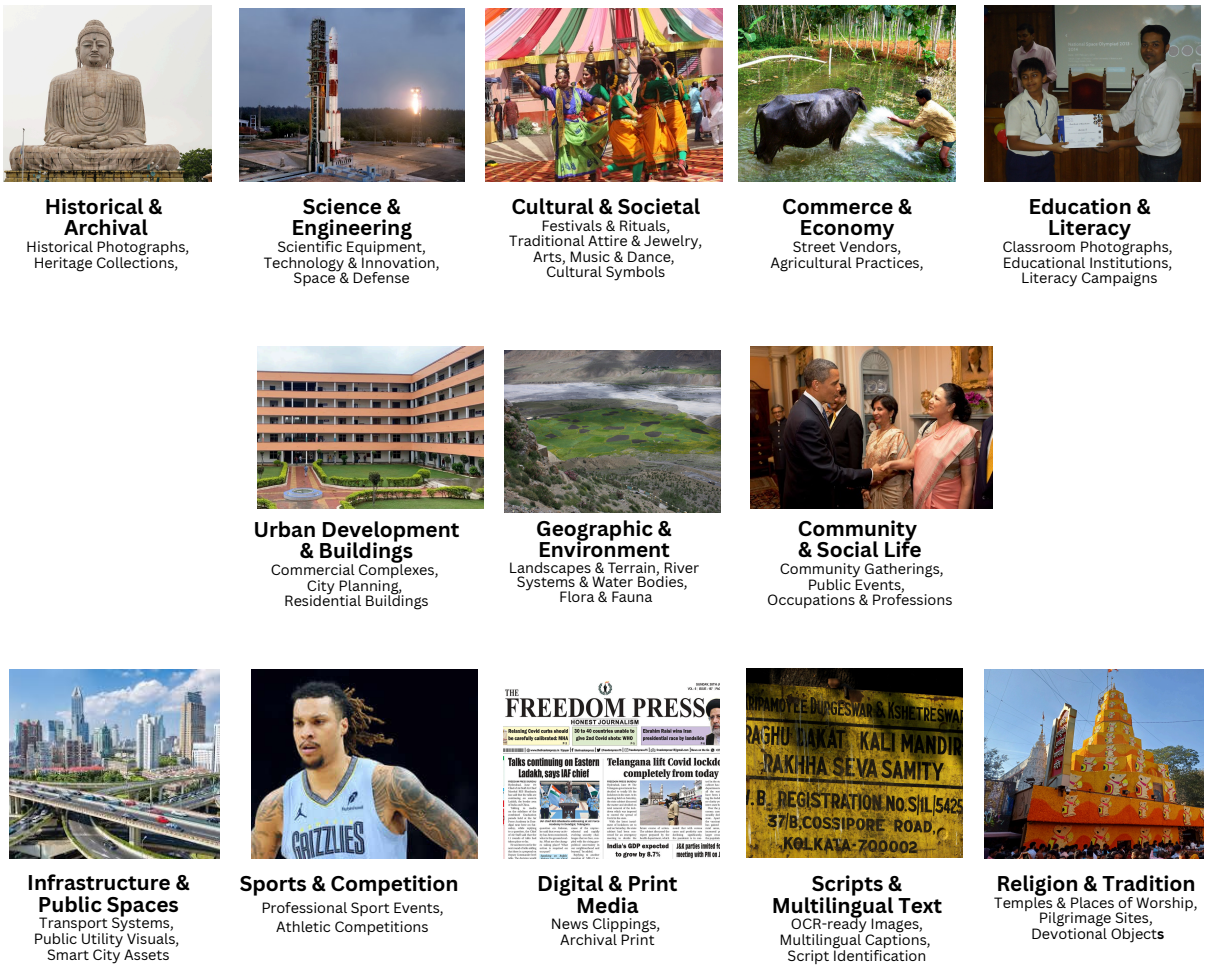

Figure 12: SIGHT taxonomy: 13 top-level categories and representative subcategories, illustrated with example images from the benchmark.

Model	GT Subcategory (missed)	Predicted Subcategory	n Instances
InternVL3-78B	*C&S* / Arts, Music & Dance	Traditional Attire & Jewelry	171
	C&S / Religious & Spiritual	Traditional Attire & Jewelry	161
	H&A / Heritage Collections	Historical Photographs	96
	H&A / Archival Documents	Historical Photographs	84
Llama-4-Scout-17B-16E	*C&S* / Arts, Music & Dance	Cultural Symbols	212
	C&S / Festivals & Rituals	Cultural Symbols	167
	C&S / Arts, Music & Dance	Traditional Attire & Jewelry	146
	I&P / Public Utility Visuals	Private Institutions	106
GLM-4.6V	*H&A* / Heritage Collections	Historical Photographs	181
	H&A / Heritage Collections	*Historical And Archival Data*†	168
	I&P / Public Utility Visuals	*Infrastructure And Public Spaces*†	162
	H&A / Heritage Collections	Cultural Artifacts	123
Gemma3-27B	*H&A* / Heritage Collections	Cultural Artifacts	224
	H&A / Heritage Collections	Historical Photographs	223
	H&A / Archival Documents	Cultural Artifacts	160
	H&A / Archival Documents	Historical Photographs	147
Qwen3-VL-32B	*H&A* / Heritage Collections	*Historical And Archival Data*†	314
	H&A / Heritage Collections	Historical Photographs	314
	C&S / Religious & Spiritual	Traditional Attire & Jewelry	272
	I&P / Government Institutions	Public Utility Visuals	153
Qwen3-VL-8B	*H&A* / Heritage Collections	Historical Photographs	302
	H&A / Heritage Collections	Photographic Archives	244
	H&A / Archival Documents	Historical Photographs	84
	D&P / News Clippings	Posters & Banners	106
Kimi-VL-A3B	*H&A* / Heritage Collections	Historical Photographs	152
	H&A / Heritage Collections	Cultural Artifacts	131
	H&A / Archival Documents	Cultural Artifacts	53
	C&S / Arts, Music & Dance	Traditional Attire & Jewelry	54
Gemini-3.1-Flash-Lite	*I&P* / Public Utility Visuals	*Infrastructure & Public Spaces*†	384
	I&P / Government Institutions	*Infrastructure & Public Spaces*†	271
	C&S / Arts, Music & Dance	*Arts*‡	198
	I&P / Transport Systems	*Infrastructure & Public Spaces*†	189
Claude-3-Haiku	*C&S* / Arts, Music & Dance	*Arts*‡	151
	C&S / Arts, Music & Dance	Cultural Symbols	106
	H&A / Photographic Archives	Historical Photographs	93
	C&S / Religious & Spiritual	Traditional Attire & Jewelry	92

† The predicted "subcategory" is identical to the parent category name, indicating a hierarchy collapse error where the model echoes the top-level category instead of resolving fine-grained labels. GLM-4.6V commits this error in 540 (*I&P*) and 318 (*H&A*) total instances; Qwen3-VL-32B in 893 and 1,074 respectively; Gemini-3.1-Flash-Lite in 1,624 (*I&P*) instances spanning all subcategories.

‡ The predicted subcategory is a truncated variant of the correct label (*Arts* instead of *Arts, Music And Dance*), indicating partial label reproduction rather than a genuine hierarchy collapse; nonetheless the prediction is scored as incorrect under exact-match evaluation.

Table 9: Top subcategory-level confusion patterns per model under zero-shot evaluation. Counts denote instances where the parent category is correctly predicted but the GT subcategory is absent from the prediction set. Larger confusion counts are lightly highlighted. Category abbreviations: *C&S* = Cultural & Societal; *H&A* = Historical & Archival Data; *I&P* = Infrastructure & Public Spaces; *D&P* = Digital & Print Media.

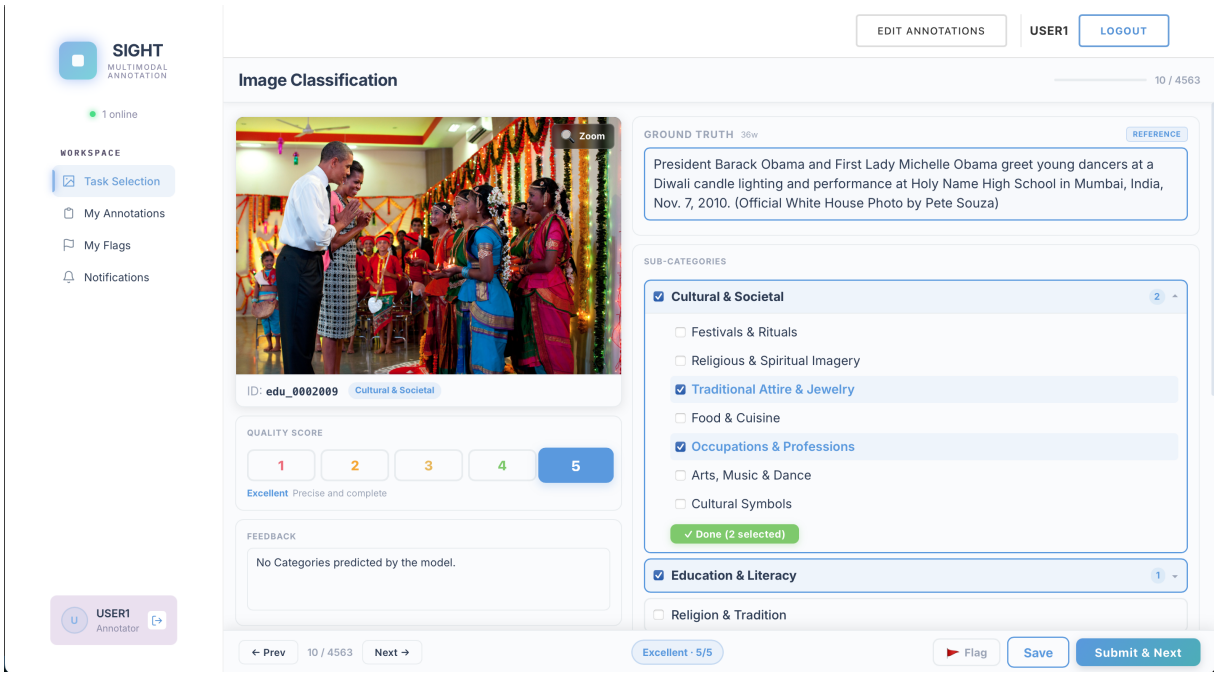

Figure 13: **SIGHT Annotation Interface.** An example of the Image Classification task UI. The ground truth describes President Obama at a Diwali event in Mumbai. No categories were predicted by the model, requiring the annotator to manually assign "Cultural & Societal" with subcategories.

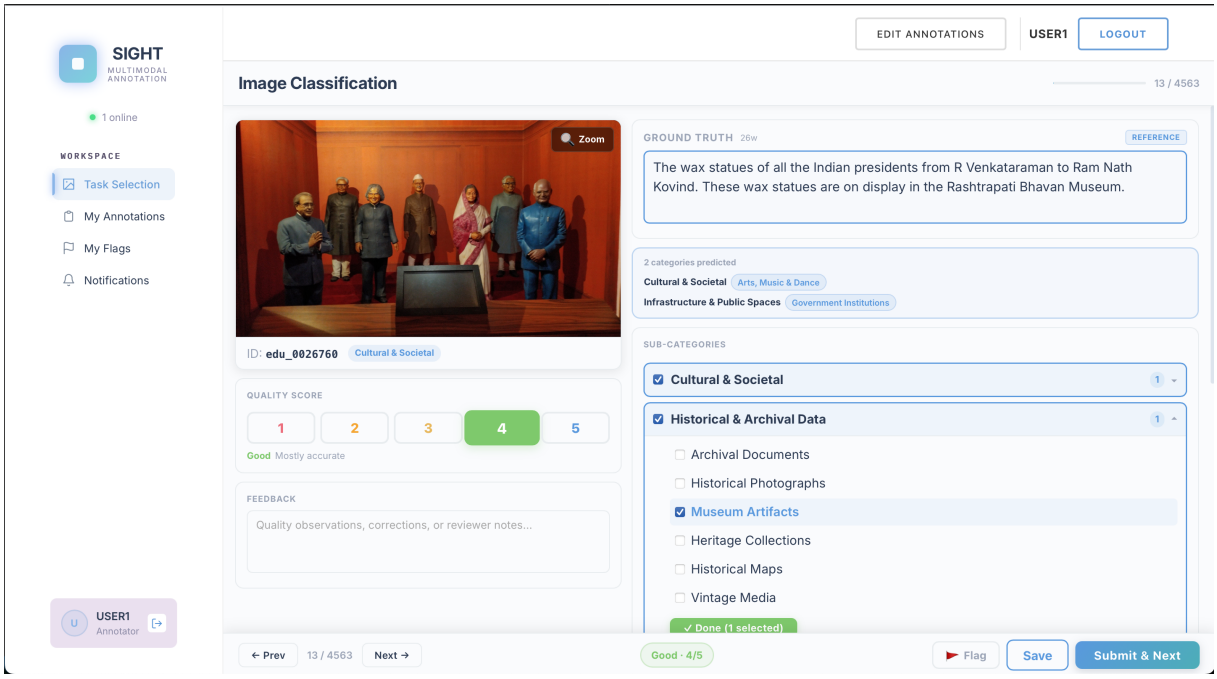

Figure 14: **Taxonomy Selection.** Annotation view for an image depicting wax statues of Indian presidents at the Rashtrapati Bhavan Museum. The model incorrectly predicted *Cultural & Societal* → *Arts, Music & Dance* and *Infrastructure & Public Spaces* → *Government Institutions*. The annotator overrides these with the correct label *Historical & Archival Data* → *Museum Artifacts*, illustrating the human correction step central to the pipeline.

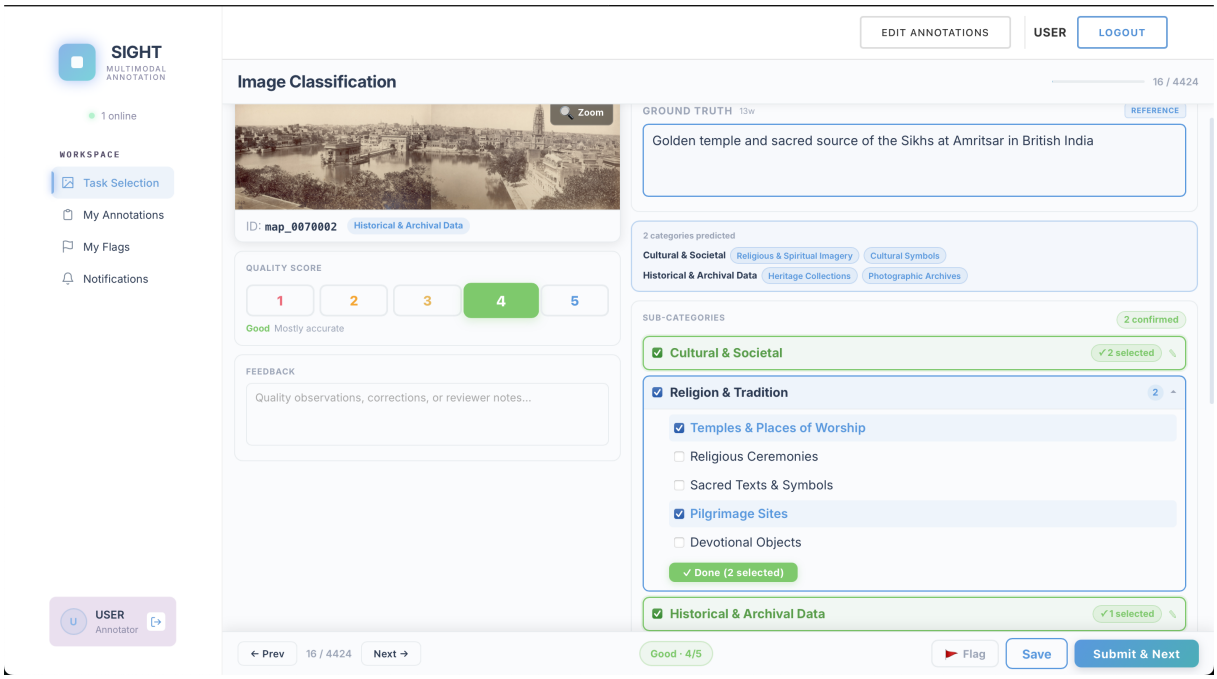

Figure 15: **Image related to Architecture and Religion.** An archival photograph of the Golden Temple at Amritsar, whose wide aspect ratio makes fine-grained inspection difficult in the default panel view. The annotator clicks "Zoom" to open the full-screen overlay before assigning the categories *Religion & Tradition* → *Temples & Places of Worship, Pilgrimage Sites* and *Historical & Archival Data*, ensuring accurate categorization of architecturally and religiously significant imagery.

Figure 16: **Full-Screen visual view.** The expanded view of the Golden Temple image, presented after the annotator clicks the "Zoom" button. This overlay allows for a detailed examination of fine-grained architectural and contextual features that might be obscured in the standard annotation task view.

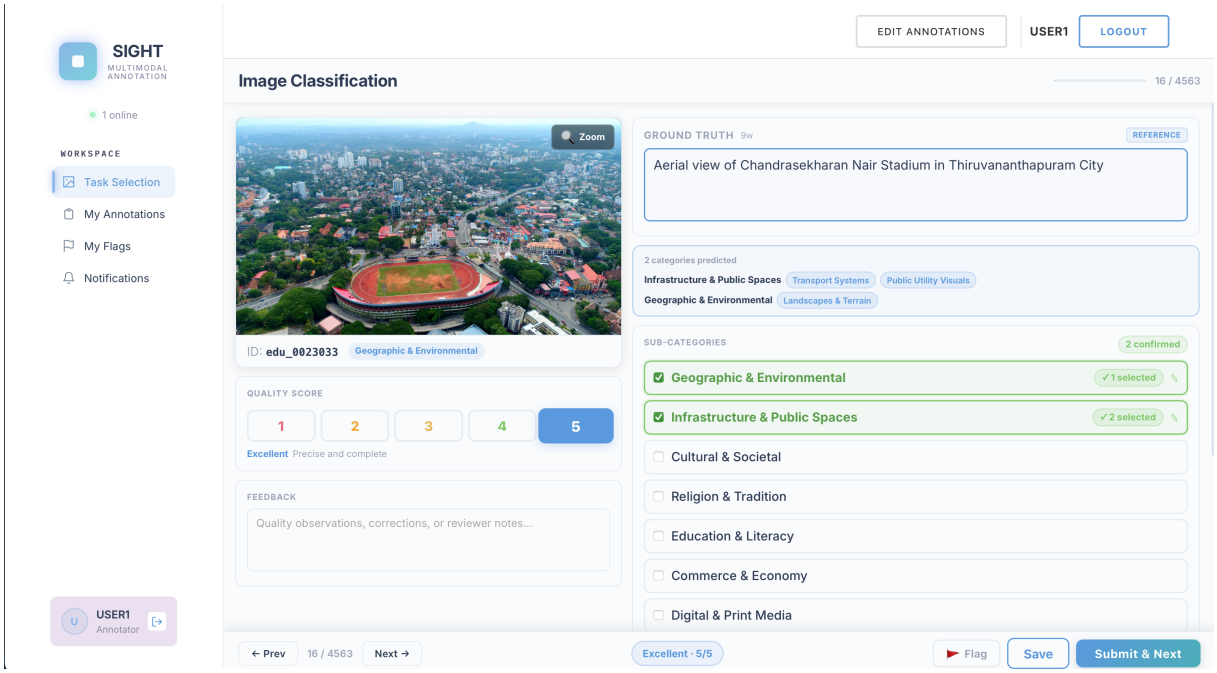

Figure 17: **Geographic & Infrastructure Annotation.** An aerial view of Chandrasekharan Nair Stadium in Thiruvananthapuram. The interface shows how an image is assigned multiple top-level categories, specifically "Geographic & Environmental" and "Infrastructure & Public Spaces."

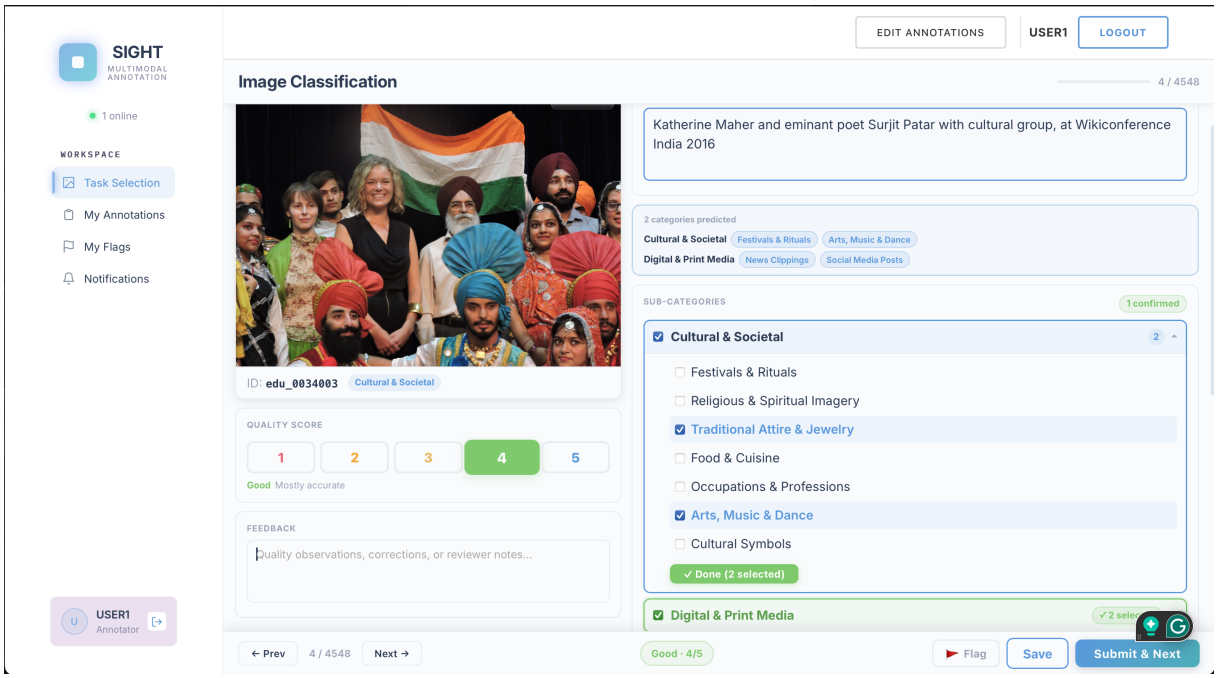

Figure 18: **Cultural and Event Annotation.** The interface capturing a cultural group at WikiConference India 2016, correctly categorized by the human annotator under "Cultural & Societal" (Arts, Music & Dance) and "Digital & Print Media."

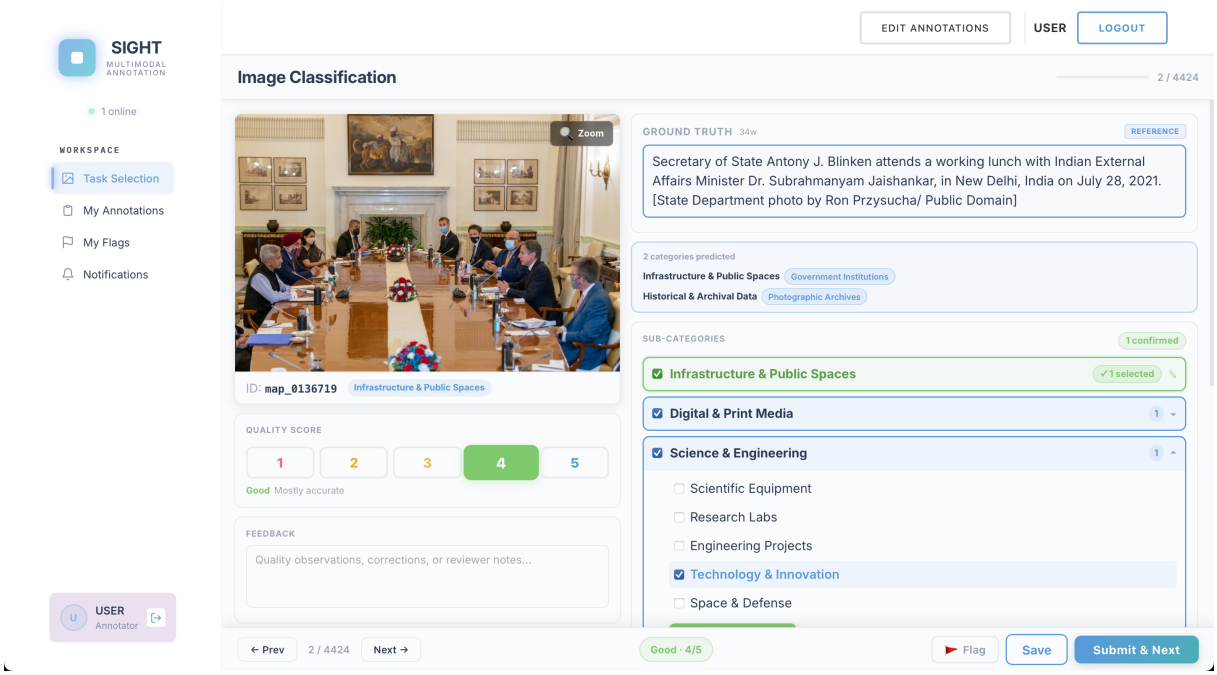

Figure 19: **Institutional and Engineering Annotation.** Annotation of a working lunch between US and Indian officials. The instance is classified under "Science & Engineering" alongside "Infrastructure & Public Spaces," demonstrating the taxonomy's coverage of civic and government events, accompanied by a 4/5 quality score.

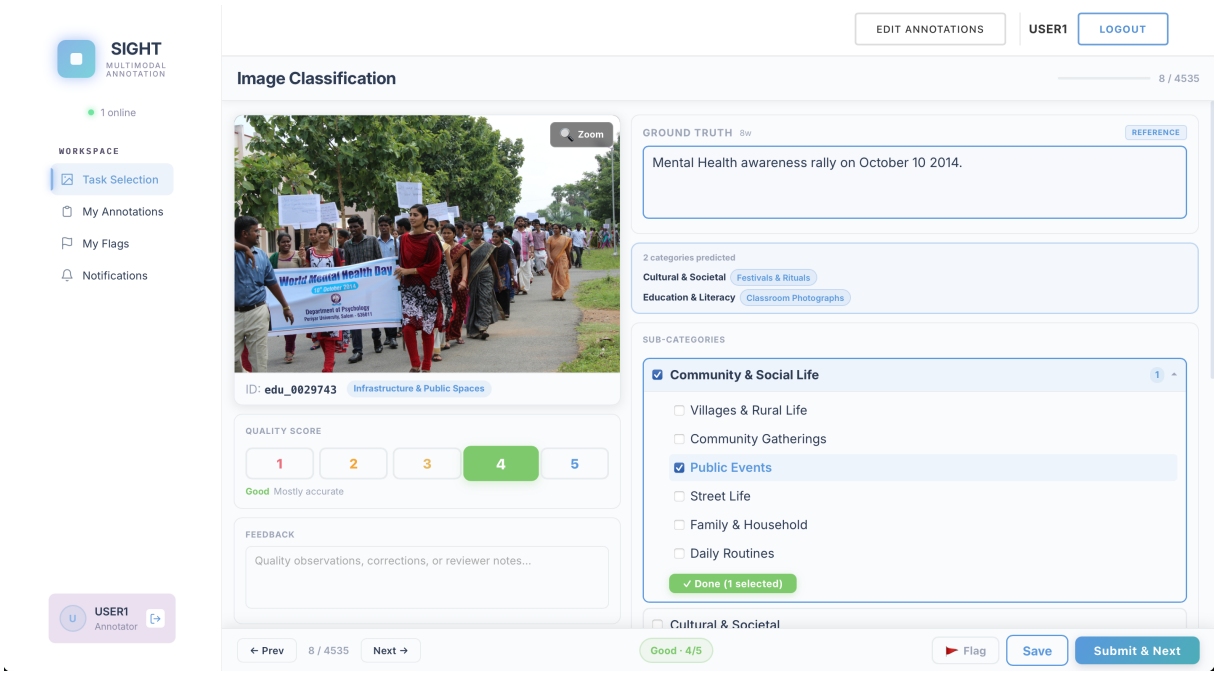

Figure 20: **Social Life and Public Gatherings.** A mental health awareness rally categorized under "Community & Social Life" with the specific "Public Events" subcategory selected, accompanied by a 4/5 quality score.

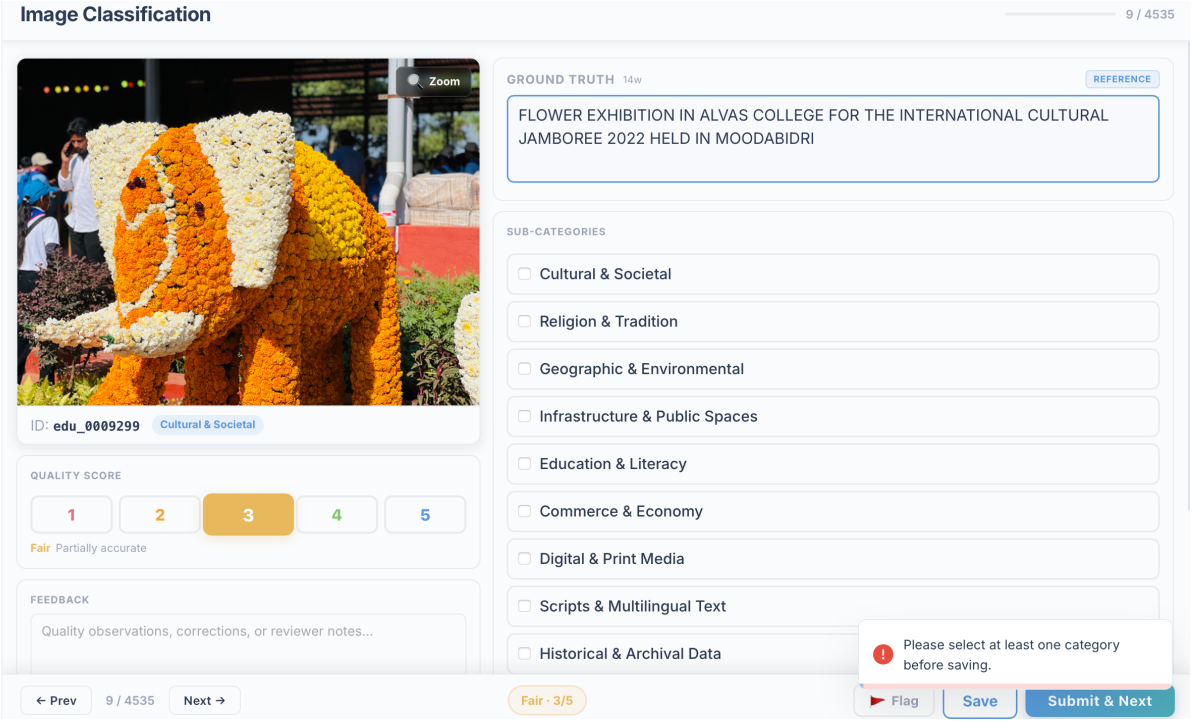

Figure 21: **Annotation Validation and Error Handling.** The interface displaying an pop-up for validation error ("Please select at least one category before saving") for an image of a flower exhibition, ensuring completeness in the human-annotation pipeline.

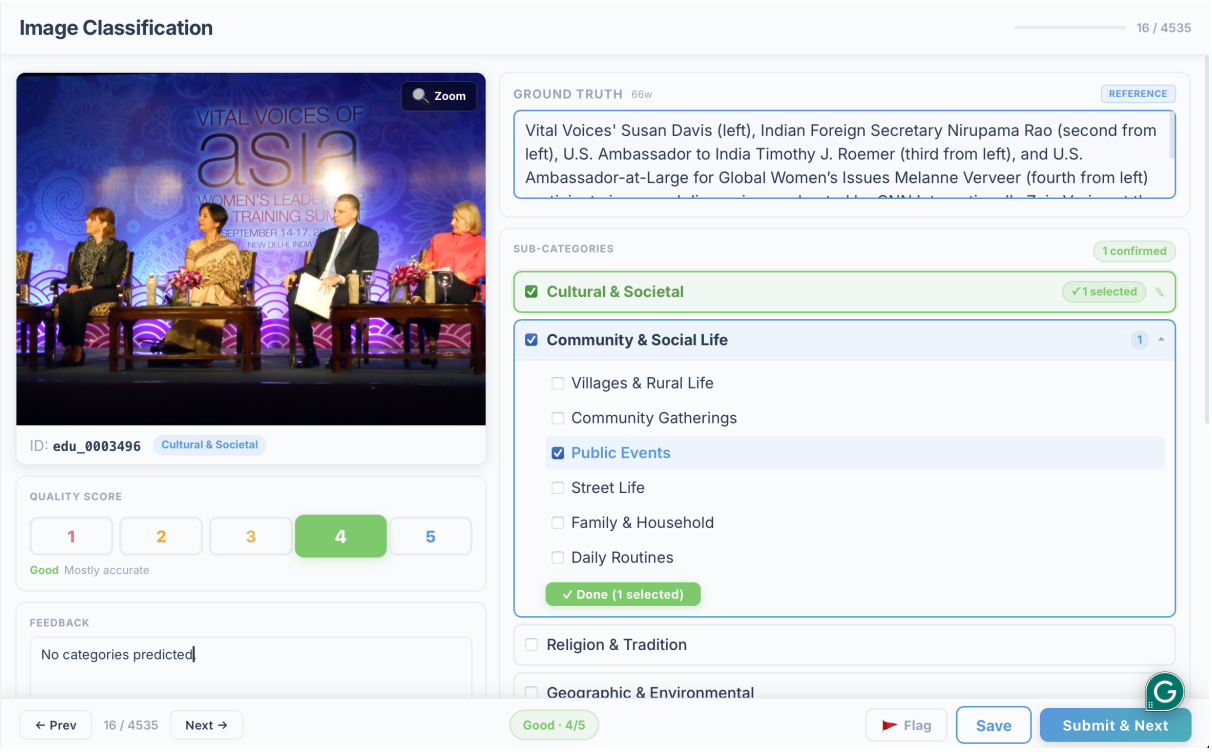

Figure 22: **Missing model-predicted labels.** A Vital Voices event is annotated under "Community & Social Life" (Public Events). The annotator has added in feedback panel explicitly that "No categories were predicted," highlighting the divergence between VLM capabilities and manual human annotation.

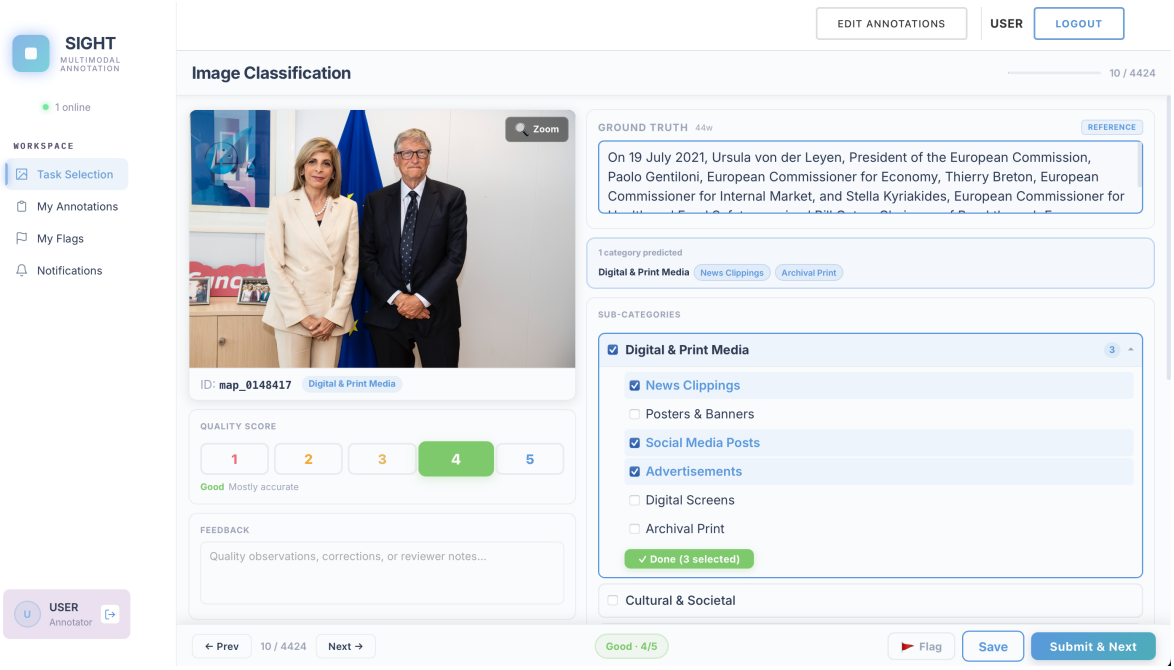

Figure 23: **Fine-grained Print & Media Classification.** Annotation of European Commission officials, categorized primarily under "Digital & Print Media" with multiple subcategories selected, including "News Clippings" and "Social Media Posts."

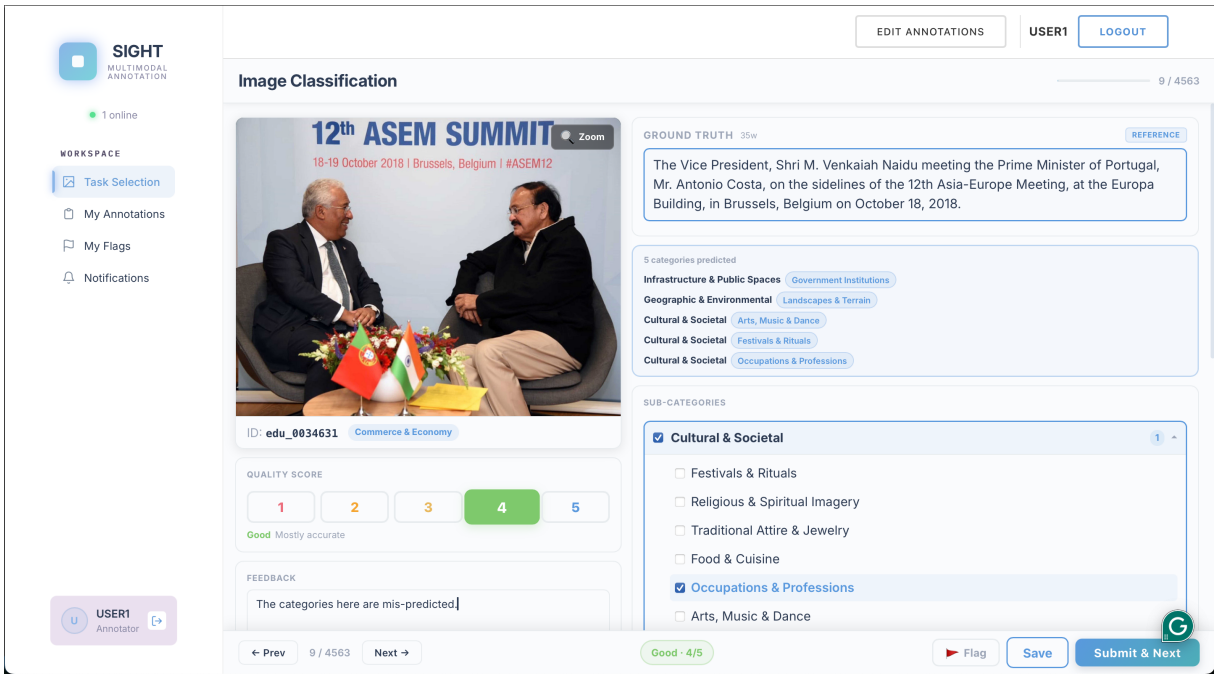

Figure 24: **Correcting Model Mis-predictions.** Annotation of the 12th ASEM Summit where the annotator has explicitly noted "The categories here are mis-predicted" by the model which annotator has added in the feedback textbox section, and manually selected the correct annotation.

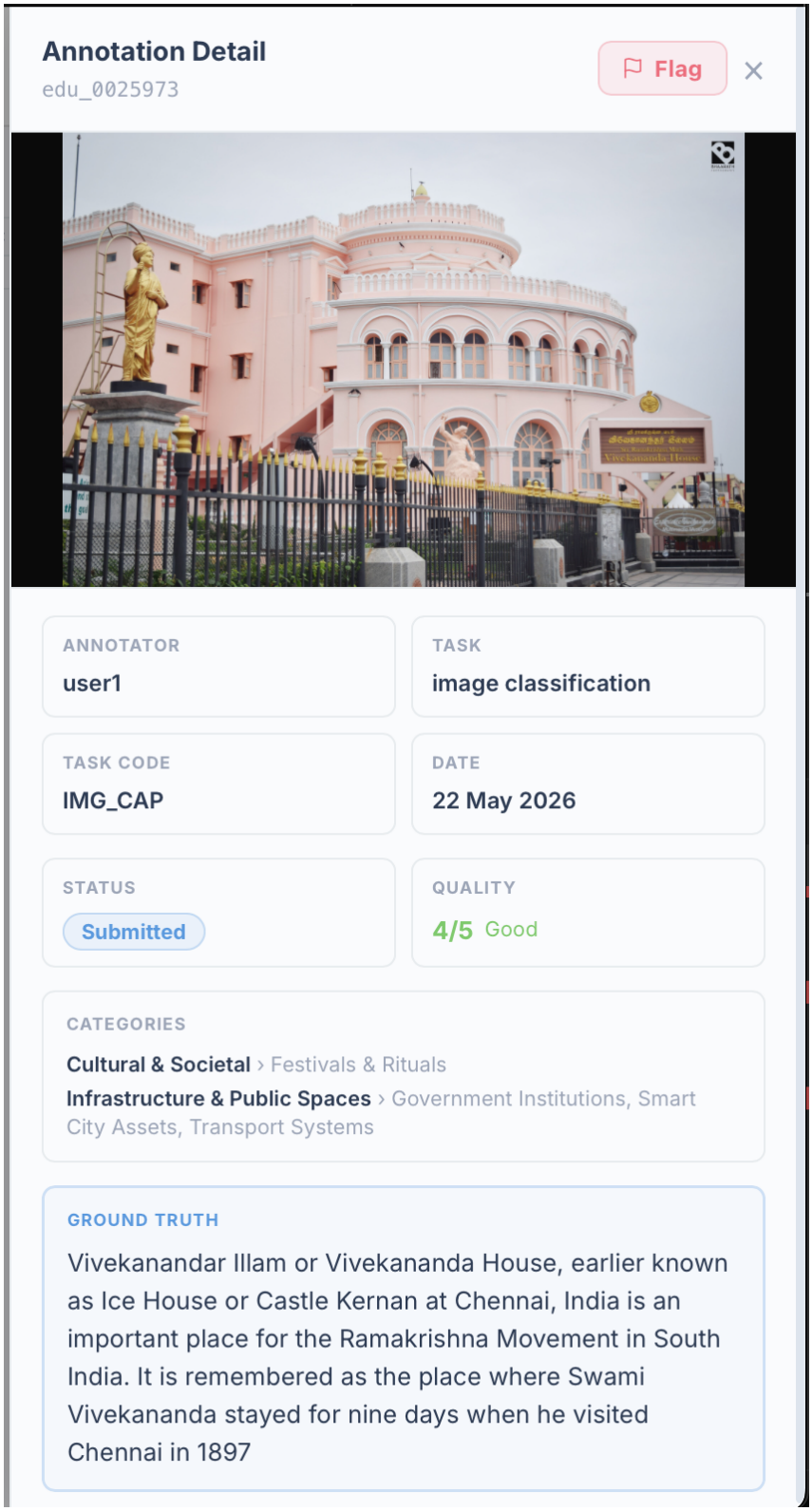

Figure 25: **Annotation Detail and Summary View (Admin Workflow)** A comprehensive summary panel for an annotated instance (Vivekananda House), displaying the annotator ID, task type, timestamp, quality score (4/5 Good), selected hierarchical categories, and the full ground truth text for quality review.

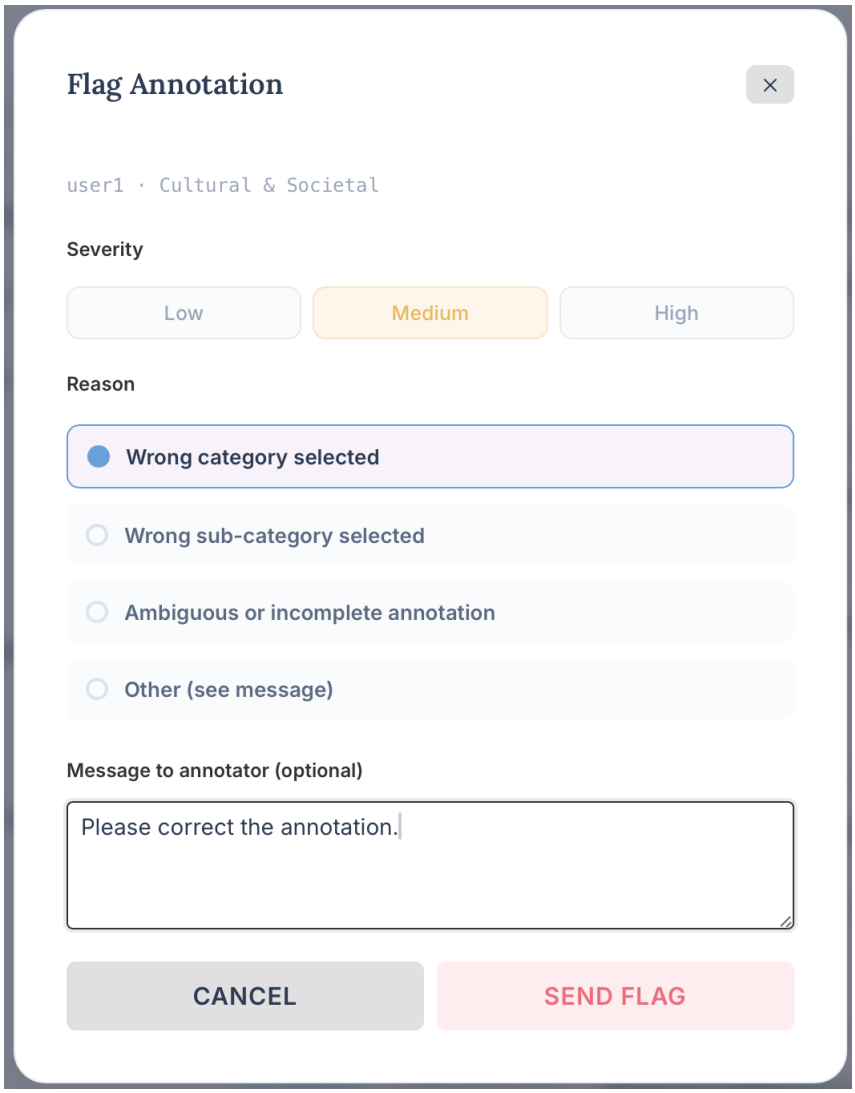

Figure 26: **Quality Control and Flagging Mechanism (Admin Workflow)** The "Flag Annotation" modal used by the admin to report annotation issues on a specific instance. The admin selects a severity level (Low/Medium/High), specifies a reason such as "Wrong category selected," and optionally includes corrective feedback directed to the annotator. Flagged instances are routed back into the annotator's revision queue.

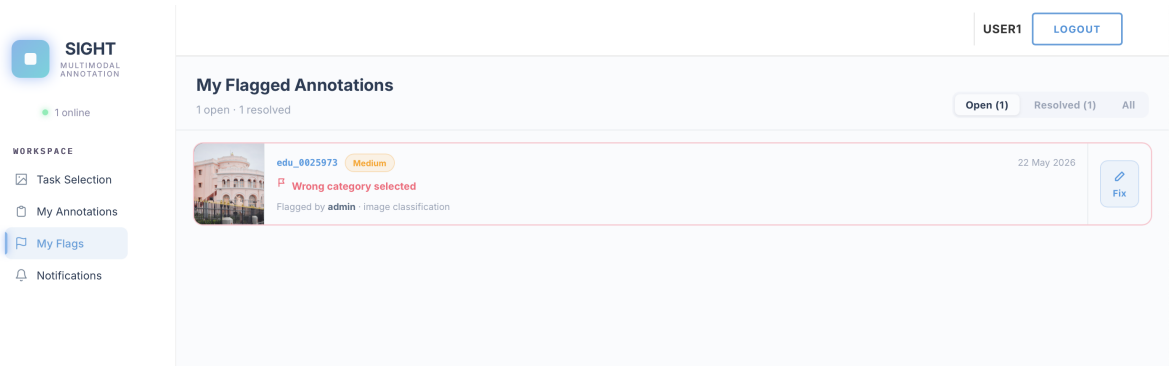

Figure 27: **Quality Control and Flagging Mechanism.(Annotator Workflow)** The "My Flagged Annotations" dashboard, where annotators track instances flagged by the admin. Each entry displays the image ID, flag severity, reason, and resolution status (Open/Resolved). Annotators revisit open flags, revise the annotation, and resubmit the corrected instance for review.

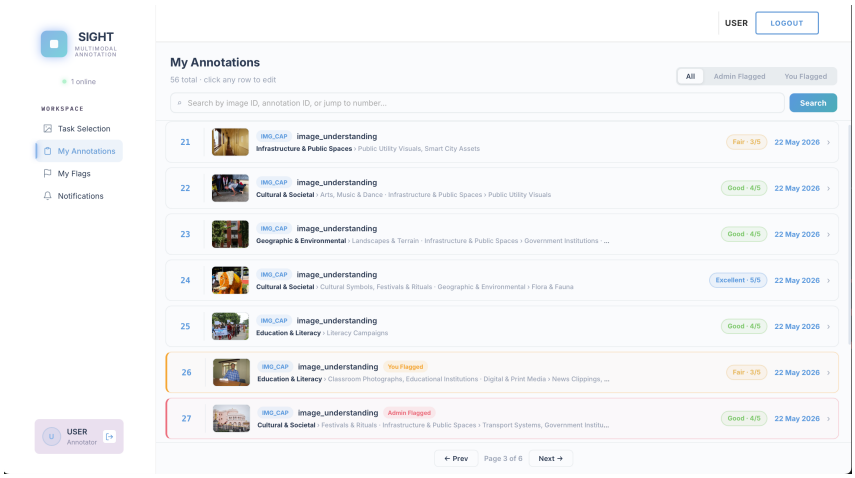

Figure 28: **Review and Editing Annotator Workflow.** The annotator's "My Annotations" dashboard also listing completed image-classification tasks together with assigned categories, quality scores, submission dates, and flag status indicators. This interface enables annotators to manage workload and quickly identify instances requiring re-annotation.

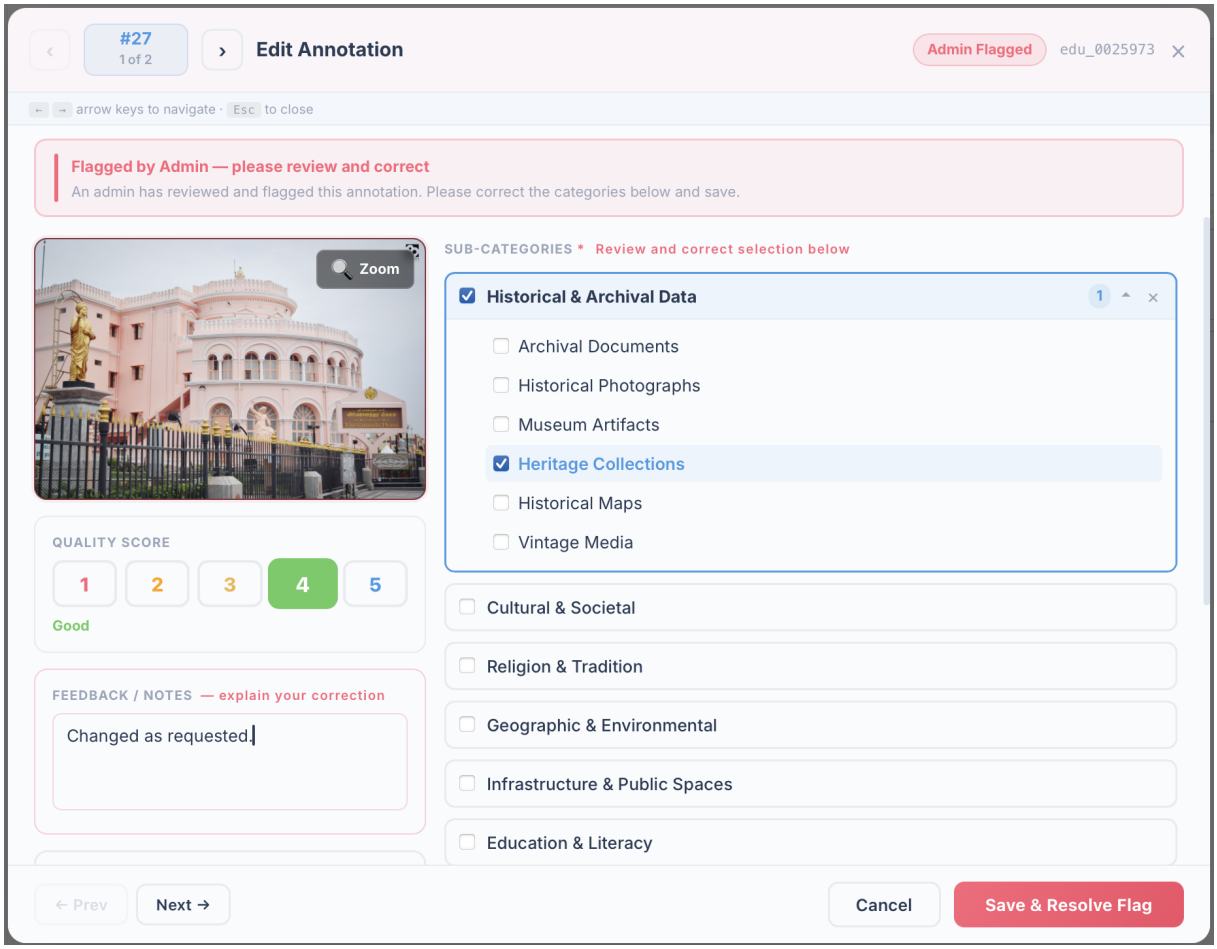

Figure 29: **Editing a Flagged Annotation (Annotator Workflow)** The "Edit Annotation" view for a Vivekananda House instance flagged by the admin with the message "Please review and correct." The annotator replaces the incorrect assignment with *Historical & Archival Data* → *Heritage Collections*, records "Changed as requested" in the feedback field, and resolves the flag through the "Save & Resolve Flag" action.

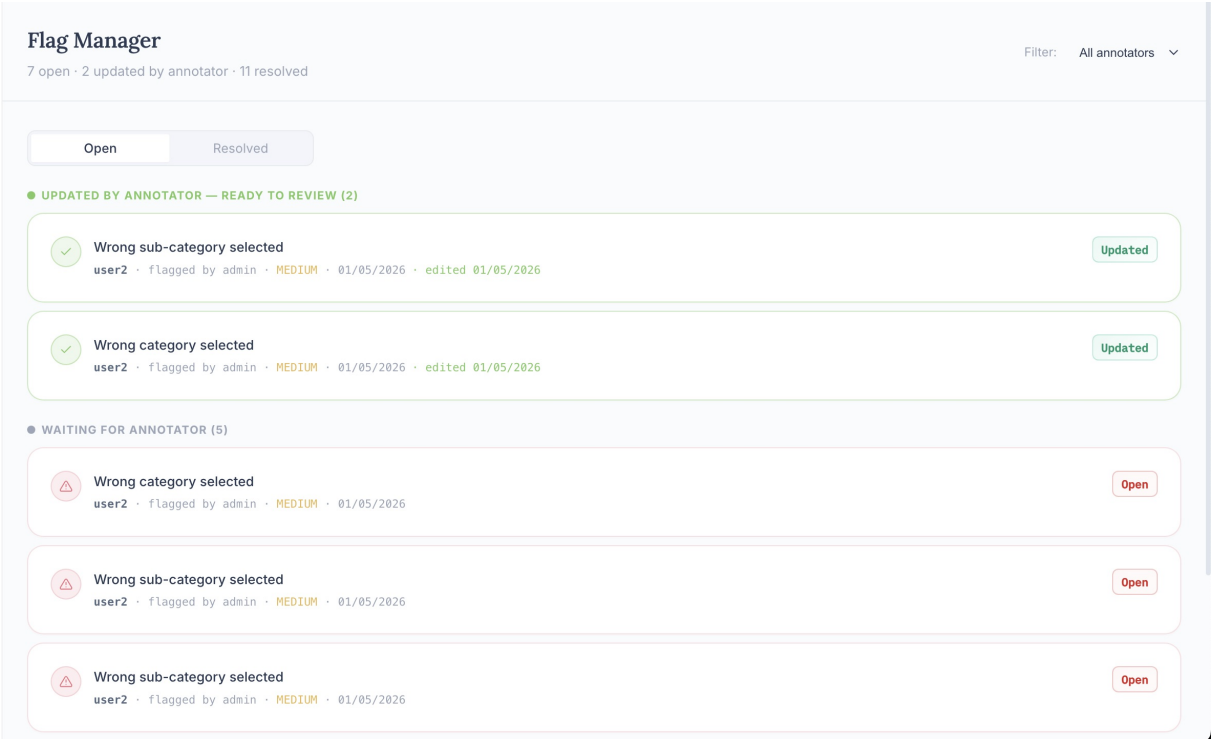

Figure 30: **Admin Flag Manager (Admin Workflow)** The Flag Manager dashboard used by the admin to monitor, filter, and resolve annotation disputes across the team. Flags are grouped by workflow status, including "Updated by Annotator — Ready to Review" and "Waiting for Annotator." Each entry records the flagging user, severity level, timestamp, and last-edited date, providing a complete audit trail for quality assurance.

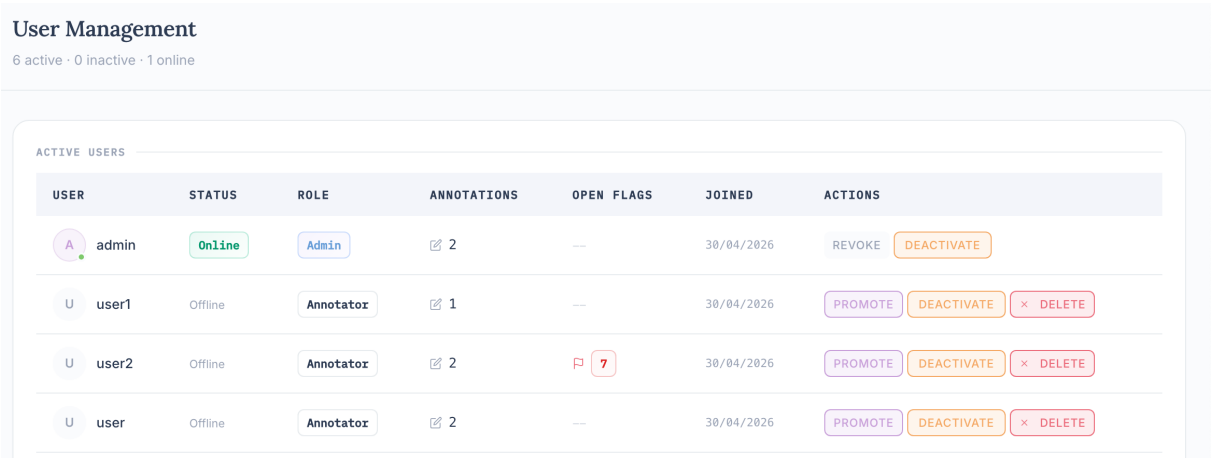

Figure 31: **Admin User Management Portal (Admin Workflow)** The User Management panel used by the admin to oversee the annotator team. Each row displays user status (Online/Offline), role (Admin/Annotator), total completed annotations, number of unresolved flags, join date, and management actions such as Promote, Deactivate, and Delete. This interface supports workload monitoring, team coordination, and access management throughout the annotation campaign.