Structured Annotation Framework for Culturally Grounded AI Evaluation in Historical Visual Culture

Recent advances in vision-language AI (e.g. GPT-4o, Gemma 3, Qwen 2.5 VL) offer new tools for analyzing visual materials, but these models often struggle with historical specificity and cultural nuance—frequently misreading visual cues, imposing present-day biases, or hallucinating details (Akbulut et al. 2024; Burda-Lassen et al. 2024). Addressing heterogeneity and contextual understanding, our project develops an annotation and evaluation framework grounded in archival photographs of Singapore's built environment and vernacular architecture. We will benchmark multimodal AI's ability to interpret cultural and temporal context accurately. Rooted in humanities expertise, our approach enables reciprocal impact: creating rich, structured annotations for humanistic inquiry while advancing the development of more context-sensitive, interpretable AI. These historical photographs are not just data points; they are cultural treasures. Our project treats them with respect, aiming to enhance their interpretive context through careful annotation.

Methodology and Timeline

The project will proceed over 24 months in five overlapping phases. We employ a human-in-the-loop design: an **external advisory panel of heritage archivists and historians** is engaged at key milestones to guide image selection, validate annotation quality, and interpret benchmark results. By coupling domain-expert feedback with our core team's interdisciplinary work, each phase of the project remains culturally informed and methodologically sound.

1. Dataset Curation (Month 1–6): We will assemble approximately 1,000 historical photographs and photo postcards (19th to mid-20th century) depicting Singapore's built environment and urban life, sourced from local archives and museum collections with appropriate permissions. Emphasis will be on images exemplifying vernacular architecture (shophouses, temples, street scenes) and urban transformations (colonial-era infrastructure to early modern developments), ensuring diversity across periods, neighborhoods, and formats (studio portraits, amateur snapshots, postcards). Throughout this phase we will work closely with archive curators to refine selection criteria and identify under-represented scenes. The annotation team will also establish a lightweight data management pipeline (CSV or a simple database) to systematically store images and annotations. We will ensure that enough images are acquired by Month 5 to keep annotators busy, and will finalize the complete image set by Month 6 in consultation with our archival partners. This ensures the collection is fixed before full-scale annotation begins.

2. Annotation Schema Design (Month 2–12): The Principal Investigator (PI) and one collaborator (trained in art/architectural history), together with three hired research assistants (RAs), will develop a multi-layer annotation schema and carry out the annotations. In the first two months, we will annotate a pilot set of 100–150 images to refine the schema and ensure annotation consistency, determining formats (e.g. a JSON or CSV template with distinct fields and free-text commentary). After confirming the annotation schema by Month 4 (incorporating lessons from the pilot), we will proceed to annotate the remaining ~850–900 images during Months 5–12. At the pilot milestone (Month 3), an expert advisor from our panel will review a sample of the preliminary annotations and provide feedback, helping to fine-tune the schema's

clarity and cultural appropriateness. Once full annotation is underway, the PI and collaborator will regularly review samples of annotations to ensure consistency and accuracy, and we will hold bi-monthly calibration meetings with the RAs to address any ambiguities or drift in standards. Additionally, we will conduct dedicated **cultural quality control checkpoints with external experts**: for example, at mid-point (around Month 8) and upon completion of the annotation phase (Month 12), heritage experts on the advisory panel will examine subsets of annotations. Their review will help catch interpretive oversights or biases and confirm that historical context is correctly and sensitively captured, before we lock the final dataset. The annotation schema is designed with three levels of detail:

- **Level 1 – Descriptive Overview:** Basic identification of visible elements (e.g. buildings, vehicles, people, natural features). At this stage, we list objects and structures ("two-story shophouses," "a trishaw on the street," "an arched bridge") and any text or signage visible. To streamline throughput, a first-pass annotation set will be produced by a state-of-the-art vision-language model (GPT-o3 / GPT-4o), whose current object-recognition accuracy is already competitive with human experts; these outputs will then be manually reviewed and adjusted by our annotators before the team advances to Levels 2 and 3.
- **Level 2 – Contextual and Historical Notes:** Deeper explanation of what those elements signify. This includes naming specific locations or structures if known (e.g. noting "Coleman Bridge" vs. just "a bridge") and describing architectural styles or functions (e.g. "ornamental balustrades and street lamps typical of colonial-era infrastructure"). We will incorporate archival knowledge such as construction dates, original names, or relevant historical events (for instance, identifying a market building and noting its role in colonial trade).
- **Level 3 – Interpretive Commentary:** High-level analysis from a humanistic perspective. Here the annotators provide insight into themes like urban transformation, colonial influence, or social life captured by the image. For example, an annotation might highlight that "the juxtaposition of low-rise shophouses and new high-rise flats marks a pivotal moment of transition in Singapore's urban planning history, as kampong-era structures were cleared for modern public housing." Such commentary connects what is seen in the photo to broader historical narratives (modernization, heritage loss, etc.), essentially grounding interpretation in the image evidence. Each annotation will remain grounded in visible clues (avoiding unfounded speculation), yet bring in contextual knowledge from historical research and cite sources when appropriate.

3. Benchmark Construction (Month 8–14, overlapping with annotation**):** The team will begin defining evaluation tasks once the pilot annotations are in place (by Month 8). As the full annotation set is completed by Month 12, we will finalize the benchmark by Month 14 using all curated annotations. This involves defining specific tasks (in consultation with our domain advisors) that AI systems can be tested on, using our expert-curated annotations as a gold standard or reference. We will ensure the benchmark tasks reflect challenges that are meaningful from a cultural heritage perspective. Example tasks include:

- **Image Description and Captioning:** Prompting a vision–language model to describe an image in detail, then comparing its output to our expert annotations. We will evaluate whether the model mentions critical context (does it recognize the scene as being in Singapore, or notice colonial architectural details?) and identify where it hallucinates (e.g. invents a building name or misidentifies the era). We can quantitatively score the outputs

using measures inspired by the Cultural Awareness Score (CAS) (Burda-Lassen et al. 2024)—checking if the model correctly includes cultural/historical details—and also qualitatively analyze errors.

- **Fact Verification and Hallucination Detection:** We will feed model-generated captions or explanations into an interpretability analysis: the team labels portions of the model's output as correct, incorrect, or unsupported. For instance, if a model describes an image and says "this scene is in Kuala Lumpur," that's a clear error (location misidentification). If it correctly describes a building's form but fails to note it is a specific heritage landmark, that indicates context insensitivity. Such mistakes will be recorded systematically.
- **Data Augmentation for Heritage Robustness**: To compensate for the sparse, uneven, and time-worn nature of archival photographs, we will create an *augmented* training subset that expands the effective sample size and simulates real-world degradations. Classic geometric and photometric transforms (e.g. rotation, mirroring, colour-jitter, Gaussian film grain) will be combined with heritage-specific artefacts such as scratches, torn edges, and high-compression noise that mimic newspaper scans. Where ethically appropriate, diffusion-based *in-painting* will insert contextually plausible period objects (for instance, 1930s trishaws or shop-sign typography) to oversample under-represented scenes. All augmented images will carry explicit provenance flags and will be used **only** for model training or for a dedicated "robustness sub-benchmark," leaving the core test set pristine. **Original images remain unaltered.** This step is expected to triple the usable training volume, balance rare categories (e.g. kampong neighbourhoods), and provide a controlled way to measure how well vision-language models cope with the visual noise typical of heritage imagery.

Crucially, we will develop an **interpretability protocol**—a structured method to assess model errors for current popular LLM tools such as **hallucinations, cultural misreadings, and context insensitivity** in outputs. This protocol will categorize error types (e.g. geographical misidentification, temporal anachronism, misinterpreted cultural symbols, unsupported hallucinations) and assign severity levels. In designing the protocol, we will incorporate feedback from our heritage advisors to ensure these error categories capture issues that domain experts consider important. By applying the protocol consistently across multiple model outputs, we can identify patterns of failure that inform how to improve both the AI models and our annotations.

4. Model Evaluation (Month 15–20): Using the benchmark tasks, we will evaluate at least three state-of-the-art vision–language models (for example, OpenAI's GPT-4o, Google's Gemma 3, Qwen 2.5 VL, and an open-source model like LLaMA 3.2 or DeepSeek Janus-Pro 7B). The evaluation will combine automated metrics (such as the Cultural Awareness Score for cultural content or BLEU/ROUGE-L for caption fidelity) with human judgment by our team and external expert advisors (applying the interpretability protocol). This mixed evaluation will yield detailed insights into **how each model handles these nuanced Singaporean historical images**. Do the models know Singapore's landmarks? Can they place the photos in the correct era? Do they confuse culturally specific artifacts or symbols? We anticipate uncovering many instances where models confidently output text that is factually wrong or culturally insensitive given the image—highlighting the need for better grounding. Notably, by involving domain experts in the evaluation process, we ensure that judgments of "correctness" or cultural

sensitivity are well-informed by deep historical knowledge, adding an extra layer of human-in-the-loop oversight to our analysis.

5. Dissemination and Project Conclusion (Month 18–24): As the evaluation concludes, we will transition into disseminating the results and finalizing project deliverables. Starting around Month 18, the team will begin drafting a research paper that synthesizes the benchmark findings and provides recommendations for improving vision–language models (e.g. suggestions for training data enrichment or model adjustments to reduce historical hallucinations). This writing continues through Month 24, aiming for a publication submission by the project's end. In parallel, we will compile the annotated dataset and benchmark documentation for public release. A final meeting with the expert advisory panel will take place around Month 22, where these external experts will review our draft findings and ensure the interpretations and recommendations are historically sound and relevant for cultural institutions. By Month 24, we will offer the completed annotated dataset (with rich metadata) to our partner archives for integration into their collections, and make the evaluation benchmark available to the research community. We also plan to present the project's outcomes at a suitable conference or public seminar, sharing insights with both AI researchers and heritage professionals.

Throughout the project, our interdisciplinary team of humanities scholars and AI specialists works in close collaboration, and we regularly consult external advisors for added domain insight. The **annotation process** itself is a form of humanities research (interpreting images via archival evidence and historical context), while the **evaluation phase** is an AI research exercise. By intertwining these modes of inquiry—and keeping human experts in the loop at critical points—we ensure the resulting dataset and findings speak meaningfully to both cultural heritage professionals and AI researchers alike.

Expected Outcomes

We anticipate impactful outcomes on two fronts—humanities scholarship and AI research—demonstrating the **bi-directional benefits** of this interdisciplinary effort:

- **For Humanities Research:** A new wealth of data and insights for scholars of Singapore and visual culture. The annotated image collection will enable **more nuanced analyses of visual history**. Researchers in urban studies or art history can, for example, trace patterns of architectural change over time or examine representations of colonial power in imagery with unprecedented granularity. Rich descriptions (with identified landmarks, architectural styles, signage, etc.) will help uncover narratives of spatial transformation—for instance, how traditional "horizontal" life of shophouses was gradually overshadowed by "vertical" modern housing projects. Previously, such insight required laborious individual study of each photo; our dataset makes it easily searchable and comparable. Moreover, the dataset can serve as a reference for **photo-historical studies of Singapore**. By identifying elements like colonial-era crests, inscriptions, or attire within images, the project helps humanities scholars detect the subtle presence of empire and indigenous culture in visual archives. The outcome is a foundation for historical inquiry that could spur new research questions and public history projects (such as museum exhibits or archives enhanced with our annotations). A copy of the annotated dataset will be offered to partner archives for integration into their collections to enhance metadata.

- **For AI Research: An evaluation benchmark for context-aware image understanding, along with an interpretability framework.** This will be a valuable resource for AI developers and researchers focused on multimodal systems. Concretely, we will deliver:
  - ◊ **Annotated Benchmark Dataset:** A set of few hundred images with ground-truth annotations and questions, usable for testing vision-language models. We envision other researchers using this dataset to benchmark their models' cultural/historical reasoning abilities. It can also be expanded to other regions including China, Japan, and other Southeast Asian countries or integrated into larger evaluation suites for multimodal AI.
  - ◊ **Assessment of Current Models:** We will highlight where even cutting-edge models fall short. For example, we expect to document instances of **hallucination** (the AI describing buildings that aren't actually present), **cultural misreading and biases** (e.g. mistaking a Buddhist temple for a Daoist one due to architectural ignorance, or misreading text on a shop sign in Malay), or **context insensitivity** (e.g. describing a 1920s street scene in modern terms). The findings will be synthesized into a set of **failure cases and analysis** that AI researchers can study. Identifying issues is the first step towards fixing them.
  - ◊ **Recommendations for Improvement:** Based on the evaluation, the team will formulate recommendations for AI model improvement. These might include suggestions on training data (e.g. incorporate local historical images or metadata), model architecture or prompting (to encourage asking clarifying questions), or post-processing techniques to verify facts. In **Month 18–24, we will draft a research paper** summarizing the benchmark results and recommendations, aiming to inform both AI developers and humanities scholars using such models.

Overall, we expect these outcomes to be seen as "major steps forward" in their respective fields. The dataset and benchmark fill a clear gap in AI resources, tackling the heterogeneity of cultural context problem head-on by operationalizing it in a concrete task. Meanwhile, the humanities outcomes push the envelope of digital humanities, showing how AI can unlock new perspectives on familiar historical materials. We request a **total budget of SGD $150,000 for the two-year project**. This budget will primarily support the personnel who will carry out the research, as well as minimal necessary infrastructure. In-kind support (access to university library archives, existing computational resources at our institutions, etc.) will further ensure we maximize output on this budget. We are confident that with this support, we can achieve the ambitious but feasible goals described above, delivering a high-impact dataset and analysis within two years.

Team Composition:

Lin Du
Principal Investigator
Incoming Assistant Professor (July 2025), Departments of Chinese Studies and Japanese Studies, National University of Singapore.
Dr. Du is a digital humanities scholar with expertise in media history, visual culture, and AI-driven archival research. She completed her Ph.D. at UCLA with a dissertation that employed advanced computer vision techniques to investigate wartime photojournalism and visual culture.

As PI, she leads the conceptual design and execution of the project, especially in dataset curation, annotation schema development, and the interpretive analysis of historical visual materials. Dr. Du's pioneering research on integrating computer vision with historical archives has been published in leading venues like the *ACM Journal on Computing and Cultural Heritage* (Q1, first author).

Fan Shi
Co-PI, Human-Centered AI Lead
Assistant Professor, Electrical and Computer Engineering, National University of Singapore; Director, Human-Centered Robotic Lab; Awardee of the Presidential Young Professorship.
Dr. Shi brings technical leadership in machine learning interpretability, human-centered design, and multimodal vision-language models. He oversees the development of the interpretability protocol and AI evaluation metrics, ensuring the alignment of cultural context with model diagnostics and performance assessment.

Zhuangzhuang Song
Collaborator, Urban Humanities and Design Lead
Urban historian and designer; Founder of Diduhui Studio (meaning "Drawings of the Imperial Capital"); Master of Architecture in Urban Design, Harvard Graduate School of Design.
Mr. Song contributes expertise in spatial history, vernacular architecture, and urban design. He supports the annotation of visual materials with insights into the evolving built environment of Singapore and co-develops use cases for the benchmark in heritage and urban humanities research.

Yingnian Wu
Collaborator, Statistical Modeling and Benchmarking Strategy
Professor, Department of Statistics, UCLA.
Professor Wu offers strategic guidance on benchmark design and evaluation, particularly in relation to statistical frameworks and interpretability. His experience in AI and probabilistic modeling ensures methodological rigor in the benchmarking protocols and dataset construction.

Reference

1. Akbulut, Canfer, Kevin Robinson, Maribeth Rauh, Isabela Albuquerque, Olivia Wiles, Laura Weidinger, Verena Rieser, Yana Hasson, Nahema Marchal, and Iason Gabriel. 2024. "Century: A Framework and Dataset for Evaluating Historical Contextualisation of Sensitive Images." In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=1KLBvrYz3V.
2. Amerini, Irene, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, et al. 2025. "Deepfake Media Forensics: Status and Future Challenges." *Journal of Imaging* 11 (3): 73. https://doi.org/10.3390/jimaging11030073.
3. Bhatia, Mehar, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. "From Local Concepts to Universals: Evaluating the Multicultural Understanding of Vision-Language Models." arXiv. https://doi.org/10.48550/arXiv.2407.00263.

4. Burda-Lassen, Olena, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. “How Culturally Aware Are Vision-Language Models?” arXiv. https://doi.org/10.48550/arXiv.2405.17475.
5. Chang, Jiat-Hwee. 2016. *A Genealogy of Tropical Architecture: Colonial Networks, Nature and Technoscience*. 1st ed. Book, Whole. London;New York; Routledge. https://doi.org/10.4324/9781315712680.
6. Cohere For AI [@CohereForAI]. 2025. “🚀 We Are Excited to Introduce Kaleidoscope, the Largest Culturally-Authentic Exam Benchmark. 📌 Most VLM Benchmarks Are English-Centric or Rely on Translations—Missing Linguistic & Cultural Nuance. Kaleidoscope Expands in-Language Multilingual 🌎 & Multimodal 👀 VLMs Evaluation Https://T.Co/OijYew5I2H.” Tweet. *Twitter*. https://x.com/CohereForAI/status/1910332931723194856.
7. Falconer, John, Gretchen Liu, and G. R. Lambert & Co. 1987. *A Vision of the Past: A History of Early Photography in Singapore and Malaya : The Photographs of G.R. Lambert & Co., 1880-1910*. Book, Whole. Singapore: Times Editions. https://go.exlibris.link/9NpTtd6C.
8. Gray, Anne, and George Lambert. 1996. *George Lambert, 1873-1930: Art and Artifice*. Book, Whole. Roseville East, N.S.W: Craftsman House. https://go.exlibris.link/TQGH54DV.
9. Klein, Lauren, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. 2025. “Provocations from the Humanities for Generative AI Research.” arXiv. https://doi.org/10.48550/arXiv.2502.19190.
10. Kleingrothe, C. J., Neil Jin Keong Khor, and Gretchen Liu. 2009. *Malay Peninsula: Straits Settlements & Federated Malay States*. Book, Whole. Kuala Lumpur: Jugra Publications. https://go.exlibris.link/glxpxmR3.
11. Lai, Chee Kien, Vikas Kailankaje, Hong Teng Koh, and Yeo Chuan. 2016. *Building Memories: People, Architecture, Independence*. Book, Whole. Singapore: Achates 360. https://go.exlibris.link/J4YdP024.
12. Mun, Chor Seng. 2020a. *Singapore Moments: A Book of Postcards*. Book, Whole. Singapore: Marshall Cavendish Editions. https://go.exlibris.link/Mjw8zS6l.
13. ———. 2020b. *Those Were the Days*. Book, Whole. Singapore: Marshall Cavendish Editions. https://go.exlibris.link/v79yH99C.
14. Nayak, Shravan, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. “Benchmarking Vision Language Models for Cultural Understanding.” arXiv. https://doi.org/10.48550/arXiv.2407.10920.
15. Pawar, Siddhesh, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. “Survey of Cultural Awareness in Language Models: Text and Beyond.” arXiv. https://doi.org/10.48550/arXiv.2411.00860.
16. Salazar, Israfel, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, et al. 2025. “Kaleidoscope: In-

Language Exams for Massively Multilingual Vision Evaluation." arXiv. https://doi.org/10.48550/arXiv.2504.07072.
17. Schneider, Florian, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. "GIMMICK -- Globally Inclusive Multimodal Multitask Cultural Knowledge Benchmarking." arXiv. https://doi.org/10.48550/arXiv.2502.13766.
18. Tim, Yap Fuan and National University of Singapore. Library. 1998. *A sense of history: a select bibliography on the history of Singapore*. Book, Whole. Singapore: National University of Singapore Library. https://go.exlibris.link/lnzb1g03.
19. Wang, Yuxuan, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. 2024. "CVLUE: A New Benchmark Dataset for Chinese Vision-Language Understanding Evaluation." arXiv. https://doi.org/10.48550/arXiv.2407.01081.
20. Wasielewski, Amanda. 2024. "Unnatural Images: On AI-Generated Photographs." *Critical Inquiry* 51 (1): 1–29. https://doi.org/10.1086/731729.
21. Yin, Da, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. "Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning." arXiv. https://doi.org/10.48550/arXiv.2109.06860.