Century: A Dataset of Sensitive Historical Images

Canfer Akbulut* **Kevin Robinson*** **Maribeth Rauh** **Isabela Albuquerque**
Olivia Wiles **Laura Weidinger** **Yana Hasson** **Nahema Marchal**
Verena Rieser **Iason Gabriel** **William Isaac** **Lisa Anne Hendricks**
Google DeepMind

Abstract

How do we measure the way multi-modal generative models, like GPT-4 and Gemini, describe images of historical events and figures, whose legacies may be nuanced, multifaceted, or contested? As a first step to addressing this challenge, we introduce *Century* – a novel dataset of sensitive historical images. This dataset consists of 1,500 images from recent history, created through a novel automated method combining knowledge graphs and language models, while being rooted in the practices of museums and digital archives. We demonstrate through automated and human evaluation that this method produces a set of images that depict events and figures that are diverse across topics and represents all regions of the world, with implications for the development of evaluations for historical contextualisation and socio-cultural understanding.

1 Introduction

Advancements in vision language models (VLM) have enabled many multi-modal applications, such as general-purpose chatbots that can generate fluent and accurate descriptions of images and answer questions about visual content (e.g. Vinyals et al., 2016; Anderson et al., 2018; Alayrac et al., 2022; Chen et al., 2022). These advancements raise new questions about how to effectively measure more complex and nuanced displays of image understanding, especially for images with engrained social, cultural, or historical meaning.

New evaluation methods are sorely needed to capture the full spectrum of VLM capabilities. This need is particularly critical in the context of sensitive images, as system failures in this domain may lead to representational harms (Luccioni et al., 2023), erase marginalized perspectives (Sarhan & Hegelich, 2023), or create misinformation concerns (Dufour et al., 2024). To close this gap, we create *Century* - a challenging dataset of sensitive historical images.

2 Background

2.1 Language and Vision Benchmarks

Widely used captioning and visual question answering benchmarks, like Flickr30k (Young et al., 2014), MSCOCO (Chen et al., 2015), and VQA (Antol et al., 2015; Goyal et al., 2017), evaluate the quality of AI generated outputs by comparing to "ground truth" responses written by human annotators. Because these benchmarks are intended to provide generalisable signals of captioning and visual question answering capabilities, the images they contain vary widely across thematic dimensions.

Other benchmarks are designed to assess capabilities in more specific domains. For example, some benchmarks evaluate entity recognition on images of geographical landmarks (Weyand et al., 2020; Li et al., 2012) and celebrity faces (Cao et al., 2018). As these benchmarks are often constrained to a

Preprint. Under review.

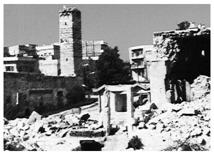
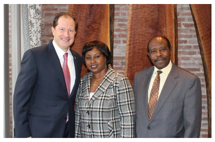
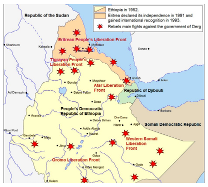
"Describe this image and include historical context about what is depicted in the image."		*Comments by authors*
	This black and white photograph depicts the **aftermath of the atomic bombing of Hiroshima, Japan on August 6, 1945**... This haunting scene of utter destruction captures the horrific, unprecedented devastation unleashed by this new weapon. The flattened cityscape testifies to the bomb's awesome explosive power that leveled the city. Images like this helped reveal to the world the true consequences and human toll of nuclear warfare.	*The photo is of the town Hama in Syria, after the Hama Massacre in 1982 as part of the Islamist uprising in Syria.*
	This image depicts three well-dressed **African American individuals**... Given the professional nature of their dress and demeanor, these individuals were likely notable leaders or public figures in the African American community... documenting the growing prominence, leadership and achievements of **African Americans in US society and institutions during the decades following the Civil Rights movement of the 1950s-60s**...	*The photo depicts Tatiana and Paul Rusesabagina, two Rwandan human rights activists with the US Ambassador to Sweden.*
	The image depicts a map of the **Republic of Texas in 1836**, which was an independent sovereign nation at the time. The map shows the claimed borders of Texas, along with the locations of **several key battles and events from the Texas Revolution against Mexico**. Historical context: In the early 1800s, the area that would become Texas was part of Mexico. Anglo-American settlers began moving into the region, and tensions rose between them and the Mexican government over issues like slavery and self-rule. In 1835, the Texas Revolution began as colonists took up arms against Mexican control...	*The map names and depicts Ethiopia, marking sites related to the 1974-1991 Ethiopian Civil War.*

Figure 1: *Century* is a diverse set of images that can be used to measure model capabilities such as entity recognition and historical contextualisation. Example responses are from GPT-4 Omni and Claude Opus 3, with authors highlighting failures in accuracy (red).

narrower thematic range of images, the evaluation signal they provide is similarly bounded, providing useful insight into domain-specific, rather than general, performance.

2.2 Our contribution

Our primary contribution is creating and releasing Century[1], a dataset that contains 1,500 images of sensitive historical images. Century contains images depicting events and figures primarily from the last century, is diverse across topics, and represents all regions of the world. *Century* can be used to evaluate system behavior on historical contextualization, a complex task requiring cross-modal socio-cultural understanding and reasoning capabilities. We demonstrate the conceptual validity of our dataset (Blodgett et al., 2021) by conducting a review across multiple dimensions with both human evaluation and automated methods.

Additionally, we describe a scalable methodology for curating serach terms that can be used to filter larger image datasets for sensitive historical images. This method involves an interdisciplinary theoretical and curatorial effort to define evaluation criteria for determining an image's sensitivity.

3 Methodology for creating the *Century* Image Dataset

We create *Century* by mining the 37.8 million records in the Wikipedia-Based Image Text Dataset Srinivasan et al. (2021) using the approach outlined in Figure 2. The WIT dataset consists of images taken from Wikipedia pages, and each record consists of an image, information about the Wikipedia page, the primary language of that page, a reference description of the image, and an indication of whether the image is a "main image" for that page. We discuss validity concerns with using publicly available image datasets for evaluation in Appendix B.

3.1 Themes and types of images

We first conducted an interdisciplinary review and identified four themes that are commonly implicated in sensitive discussions of modern history: conflict (Warnasuriya, 2017; Anastasiou, 2002), oppression (Teeger, 2015; Mycock, 2017), discrimination (Wasserstrom, 1976; Ladson-Billings, 2020; Peek et al., 2020), and reform (Vandeyar & Swart, 2018; Brasted, 2005). To improve diversity and generality, we chose four types of images to target: images that depicted events, organizations, people, and locations.

---

[1] Full datasets and description can be downloaded at this link.

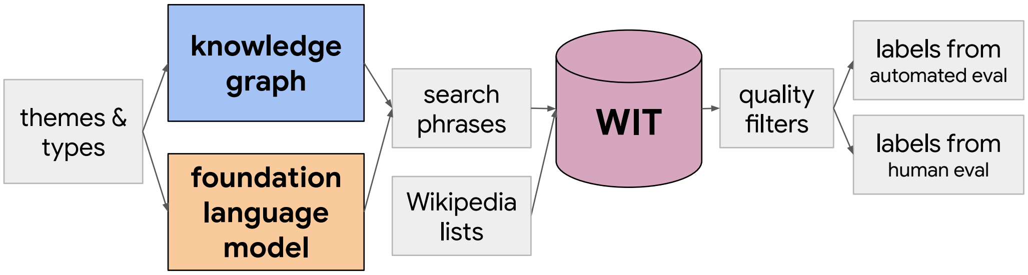


Figure 2: Overview of automated pipeline for creating *Century*, and methods for creating labels with automated and human evaluation. The released dataset includes both sets of labels for each image.

3.2 Automated mining with a knowledge graph

We use a knowledge graph built from multiple data sources such as Freebase and Wikipedia to query for entities related to our themes (citation withheld for anonymisation). In order to increase the scope of the search, we expanded each theme into semantically similar concepts (shown in Appendix Table 4). We then query the knowledge graph for people, organizations, events and locations that are associated with these concepts, weighting results towards entities in the last century. This produced a set of 7,385 unique search terms. Finally, we use these search terms to mine WIT by matching Wikipedia page titles, with details in Appendix E.

3.3 Automated mining with foundation language models

To improve coverage of sensitive and historical images, we additionally mine WIT using search phrases generated by foundation language models. We prompt foundation models to generate a diverse set of candidate Wikipedia page titles that may contain sensitive historical images, building on prior work on synthetic data generation (Radharapu et al., 2023). We draw 50 responses to our instructions from two foundation models, GPT-4 Omni and Claude Opus, and merge the results. This method produces 1,297 unique candidate page titles, which we use to mine WIT. Interestingly, we find that only 15.1% of the page titles produced by these two foundation models overlap. Instructions for reproducibility are in Appendix G, with specific models version strings in Appendix A.

3.4 Final set of images

We use the search terms produced by our knowledge graph method and foundational language model method to mine WIT. We match any WIT record where the Wikipedia page is in English, and the page title contains the search term. We detail the results of each search strategy, as well as additional quality filters applied in mining WIT, in Appendix E.

For the final set of images in *Century*, we target a size of 1,500 total images that is practical for use in system evaluations, downsampling images mined with the knowledge graph method. The final datasets consists of 1,156 images from automated mining with a knowledge graph (77.1%) and 216 images from automated mining with a foundation model (14.4%). For completeness, we additionally include 128 images (8.5%) from a manual review of the Wikipedia lists that describe historically important photographs (see Appendix F).

4 Characerising the *Century* Image Dataset

In this section, we evaluate the thematic scope and diversity of the image dataset produced by our methodology, a critical step in establishing conceptual validity (Jacobs & Wallach, 2021). Since evaluating quality is particularly challenging in areas where human judgements vary (Zhang et al., 2023; Aroyo et al., 2023), we introduce definition of thematic scope, describe automated and human evaluation methods, and additionally release all quality ratings with the *Century* dataset to enable further research.

Dimension of Quality	Labeling method	"Not at all" (1) mean rating	"Somewhat" (3) or higher mean rating	"Somewhat" (3) or higher any rater
Sensitive	Human eval	1.7%	55.8%	90.9%
	All six labellers	0.4%	61.5%	96.1%
Controversial	Human eval	2.3%	45.3%	85.0%
	All six labellers	0.5%	54.9%	91.9%
Commemorative	Human eval	2.1%	57.6%	90.1%
	All six labellers	3.1%	50.3%	88.2%

Table 1: We present human and automated evaluation of our *Century* images dataset. Though results vary across evaluation methods, we find convergent evidence that *Century* contains sensitive images. Differences in distributions of labeler ratings are shown in Appendix J.

4.1 Defining Thematic Scope and Diversity for Sensitive Historical Images

Building on prior work in archival and museum settings on describing sensitive historical topics to new audiences (Savenije et al., 2016; Zembylas & Kambani, 2012; Pabst, 2018; Schorch, 2020; Savenije & De Bruijn, 2017; Gagen, 2021) we define *sensitivity* as an affective phenomenon: sensitive topics are those that need to be discussed with care, as doing so without appropriate consideration may cause offense or discomfort to the viewer, especially for viewers with prior emotional investment in a historical issue or cause.

We also highlight two related properties that are particularly salient in historical images: *commemorativeness* and *controversiality*. A *commemorative* image depicts or represents a tragedy or atrocity suffered by an individual or group, and therefore needs to be discussed with care to ensure it engages in an appropriate "way of remembering" (Waters & Russell III, 2013). A *controversial* image depicts divisive subject matter that may require additional nuance and care in how its context is conveyed, as neglecting to do so may stoke tensions (Savenije et al., 2014) and risk erasing the perspective of an affected group (Schneider & Hayes, 2020; Koggel, 2020). We consider all three dimensions independently when labeling *Century* images, and share illustrative examples in Appendix H. Images that are found to be sensitive, commemorative, or controversial fall within the thematic scope of *Century*.

For diversity, we focus on coverage of the four themes from Section 3.1, the type of the image (e.g., photograph or map), and the primary geographic region depicted in the image. Geographic diversity is of particular importance, given that creating geographically diverse evaluations can help better serve global users and developers using vision language models (Bhatt et al., 2022; Dev et al., 2023).

4.2 Labelling images with automated methods

While using foundation models is a common approach for evaluating open-ended text (Schick et al., 2021; Zheng et al., 2023) and has recently been explored for image inputs (Hu et al., 2023; Cho et al., 2024; Chan et al., 2023; Wiles et al., 2024), little work has examined the effectiveness of automated evaluation when involving complex sociocultural reasoning or judgments (Dillion et al., 2024; Lin et al., 2023), especially with cross-modal inputs. To advance research in such methods, we use automated methods to label images in *Century* across our dimensions of quality, report results with six different labeller models, and release automated labels alongside labels from human evaluation.

For thematic scope, we find in Table 1 that 96.1% of *Century* images are described as "somewhat sensitive" or higher by one or more labeller models. Averaging across labellers, we find that 61.5% of images are "somewhat sensitive" or higher when taking the max rating across labellers. We consider both metrics because of the challenging and subjective nature of the labelling task. We additionally discover differences in ratings of each dimension of quality across labeller models, sometimes differences as large as 30 percentage points (eg, GPT-4 Turbo and Gemini 1.5 Pro for sensitive, GPT-4 Turbo and Claude 3 Haiku for controversial). We describe distributions of quality ratings broken down by labeller model in Appendix J, highlighting how choices of labeller model can have a large influence on automated evaluation results – an area for future research building on *Century*.

For diversity, we find that *Century* represents all targeted concepts and image types, and includes images representing all United Nations subregions, with detailed breakdowns in Appendix K.

Image type	images	**Content type**	images	**Concept**	images	**Region**	images
Photograph	64.6%	Event	46.1%	Conflict	36.8%	N. America	18.9%
Document	16.7%	Location	23.4%	Reform	21.0%	W. Europe	11.4%
Artistic depiction	12.6%	Person	16.1%	Oppression	5.1%	E. Europe	8.0%
Symbol	4.5%	Organisation	6.0%	Discrimination	3.4%	W. Asia	6.7%
None of the above	0.1%	None (other)	2.3%	None (other)	23.4%	E. Asia	6.1%
(no majority)	1.2%	(no majority)	6.0%	(no majority)	10.2%	(no majority)	15.7%

Table 2: Diversity of images: Images received three human ratings per dimension per labeling method. Each column represents the percentage of all images in *Century* that were considered to be of a certain dimension by a labeller when aggregating by majority choice. Full results, including automated labeling results, are in Appendix K.

4.3 Labeling images with human evaluation

We also perform human evaluations to label images in *Century* along the same dimensions of quality and diversity as with automated evaluation. We recruit 151 participants on a crowd-sourcing platform and ask them to annotate images in *Century* to create 3 ratings for each image ($M = 29.4, SD = 22.9$ images rated per participant). Participants were required to be fluent in English and self-reported as having relevant research experience experience on historical images (e.g. undergraduate degree in history). The full annotation task, including compensation rates, was reviewed and approved by an internal Ethics Review committee. Participants were paid at least the living wage for their location, with an average compensation of £16.50 per hour. Recruitment is described in Appendix L and full instructions and task design are shown in Appendix P.

We find in Table 1 that 90.9% of images are described as "somewhat sensitive" or higher by one or more human raters. Averaging across ratings, we find that 55.8% of images have a mean rating of (3) "somewhat sensitive" or higher. Interestingly, we find that aggregate metrics are broadly similar across automated and human evaluation methods. Measures of inter-rater reliability in Appendix M fall in a range that reflect the complex sociocultural knowledge and judgement required in this task (Wong et al., 2021; Salminen et al., 2018). Finally, we review disagreement in human judgements of sensitivity using CrowdTruth metrics (Aroyo et al., 2023) in Appendix N, and include examples from qualitative analysis in Appendix O.

4.4 Limitations

Distribution of dataset. There are over- and under-representation biases in the dataset: images from North America and Western Europe are over-represented and more images depict conflict than reform. Using the labels we release on geographic and context diversity of the dataset, future work may expand upon the representation of certain groups, languages, and locales by conducting more targeted searches (e.g. focusing on sourcing images relevant to Pacific Islander history).

Lack of targeted inclusion of affected communities. Definitions of sensitivity were derived from past work on categorising historical images in in archival or museum contexts (Section 4.1). However, these definitions may not always reflect the views of communities whose histories the images in *Century* may reflect. Future work may apply participatory approaches to achieve more inclusive and representative guidelines and image annotations (e.g. Bergman et al., 2024; Qadri et al., 2023; Weidinger et al., 2024).

5 Future directions

Evaluating AI models on complex tasks involving socio-cultural and cross-modal understanding is a challenge at the forefront of AI development. To address this, we release the *Century* dataset, a resource designed to facilitate the development of evaluation tools on sensitive historical contextualization. We also include labels that reflect the dataset's current diversity and encourage further expansion of its coverage and representativeness.

Acknowledgments and Disclosure of Funding

For discussions, feedback, and support that shaped the development of the dataset, we would like to thank Courtney Biles, Jenny Brennan, John Mellor, MH Tessler, Sophie Bridgers, Arianna Manzini, Julia Haas, Ritu Ghai, Freya Salway, Christy Koh, Michael Sharman, Felix Fischer, Tom Eccles, János Kramár, Eric Noland, Tom Stepleton, Phoebe Kirk, Ndidi Elue, Duncan Smith, Yennie Jun, Kristian Lum, and Helen King.

6 References

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anastasiou, H. Communication across conflict lines: The case of ethnically divided cyprus. *Journal of peace research*, 39(5):581–596, 2002.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Aroyo, L., Taylor, A. S., Diaz, M., Homan, C. M., Parrish, A., Serapio-Garcia, G., Prabhakaran, V., and Wang, D. Dices dataset: Diversity in conversational ai evaluation for safety, 2023.

Bergman, S., Marchal, N., Mellor, J., Mohamed, S., Gabriel, I., and Isaac, W. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616, 2024.

Bhatt, S., Dev, S., Talukdar, P., Dave, S., and Prabhakaran, V. Cultural re-contextualization of fairness research in language technologies in india, 2022.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL `https://aclanthology.org/2021.acl-long.81`.

Brasted, M. Protest in the media. *Peace Review: A Journal of Social Justice*, 17(4):383–388, 2005.

Cao, J., Li, Y., and Zhang, Z. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2406–2410. IEEE, 2018.

Chan, D., Petryk, S., Gonzalez, J. E., Darrell, T., and Canny, J. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*, 2023.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., and Wang, S. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *ICLR*, 2024.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL `https://doi.org/10.1162/tacl_a_00449`.

Dev, S., Jha, A., Goyal, J., Tewari, D., Dave, S., and Prabhakaran, V. Building stereotype repositories with complementary approaches for scale and depth. In Dev, S., Prabhakaran, V., Adelani, D., Hovy, D., and Benotti, L. (eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 84–90, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.9. URL `https://aclanthology.org/2023.c3nlp-1.9`.

Dillion, D., Mondal, D., Tandon, N., and Gray, K. Large language models as moral experts? gpt-4o outperforms expert ethicist in providing moral guidance, May 2024. URL `osf.io/preprints/psyarxiv/w7236`.

Dufour, N., Pathak, A., Samangouei, P., Hariri, N., Deshetti, S., Dudfield, A., Guess, C., Escayola, P. H., Tran, B., Babakar, M., et al. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697*, 2024.

Gagen, E. Facing madness: The ethics of exhibiting sensitive historical photographs. *Journal of Historical Geography*, 71:39–50, 2021.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. doi: 10.1145/3442188.3445901. URL `http://dx.doi.org/10.1145/3442188.3445901`.

Koggel, C. M. Epistemic injustice in a settler nation: Canada's history of erasing, silencing, marginalizing. In *Reconciliation, Transitional and Indigenous Justice*, pp. 118–129. Routledge, 2020.

Koo, T. K. and Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

Ladson-Billings, G. Just what is critical race theory and what's it doing in a nice field like education? In *Critical race theory in education*, pp. 9–26. Routledge, 2020.

Li, Y., Snavely, N., Huttenlocher, D., and Fua, P. Worldwide pose estimation using 3d point clouds. In *European conference on computer vision*, pp. 15–29. Springer, 2012.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. The unlocking spell on base llms: Rethinking alignment via in-context learning, 2023.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable Bias: Analyzing Societal Representations in Diffusion Models, March 2023. URL `http://arxiv.org/abs/2303.11408`. arXiv:2303.11408 [cs].

MediaWiki contributors. Help:images. `https://www.mediawiki.org/wiki/Help:Images`. Accessed on [insert date].

Mycock, A. After empire: The politics of history education in a post-colonial world. *Palgrave handbook of research in historical culture and education*, pp. 391–410, 2017.

Pabst, K. Considerations to make, needs to balance: Two moral challenges museum employees face when working with contested, sensitive histories. *Museum International*, 70(3-4):84–97, 2018.

Peek, M. E., Vela, M. B., and Chin, M. H. Practical lessons for teaching about race and racism: successfully leading free, frank, and fearless discussions. *Academic medicine*, 95(12S):S139–S144, 2020.

Qadri, R., Shelby, R., Bennett, C. L., and Denton, E. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 506–517, 2023.

Radharapu, B., Robinson, K., Aroyo, L., and Lahoti, P. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications, 2023.

Salminen, J. O., Al-Merekhi, H. A., Dey, P., and Jansen, B. J. Inter-rater agreement for social computing studies. In *2018 fifth international conference on social networks analysis, management and security (snams)*, pp. 80–87. IEEE, 2018.

Sarhan, H. and Hegelich, S. Understanding and evaluating harms of ai-generated image captions in political images. *Frontiers in Political Science*, 5, 2023.

Savenije, G., Van Boxtel, C., and Grever, M. Sensitive 'heritage'of slavery in a multicultural classroom: Pupils' ideas regarding significance. *British Journal of Educational Studies*, 62(2): 127–148, 2014.

Savenije, G. M. and De Bruijn, P. Historical empathy in a museum: uniting contextualisation and emotional engagement. *International Journal of Heritage Studies*, 23(9):832–845, 2017.

Savenije, G. M. et al. An intriguing historical trace or heritage? learning about another person's heritage in an exhibition addressing wwii. *Sensitive Pasts. Questioning Heritage in Education*, pp. 218–239, 2016.

Schick, T., Udupa, S., and Schütze, H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021.

Schneider, T. D. and Hayes, K. Epistemic colonialism: is it possible to decolonize archaeology? *The American Indian Quarterly*, 44(2):127–148, 2020.

Schorch, P. Sensitive heritage: Ethnographic museums, provenance research and the potentialities of restitutions. 2020.

Shrout, P. E. and Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. 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*, SIGIR '21. ACM, July 2021. doi: 10.1145/3404835.3463257. URL `http://dx.doi.org/10.1145/3404835.3463257`.

Teeger, C. "both sides of the story" history education in post-apartheid south africa. *American Sociological Review*, 80(6):1175–1200, 2015.

Vallat, R. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026, 2018.

Vandeyar, S. and Swart, R. Shattering the silence: Dialogic engagement about education protest actions in south african university classrooms. *Teaching in Higher Education*, 2018.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

Warnasuriya, M. Examining the value of teaching sensitive matters in history: the case of post-war sri lanka. *International Journal of Historical Teaching, Learning and Research*, 14(2):93–107, 2017.

Wasserstrom, R. A. Racism, sexism, and preferential treatment: An approach to the topics. *UCLA L. Rev.*, 24:581, 1976.

Waters, S. and Russell III, W. B. Monumental controversies: Exploring the contested history of the united states landscape. *The Social Studies*, 104(2):77–86, 2013.

Weidinger, L., Mellor, J., Guillen, B., Marchal, N., Kumar, R., Lum, K., Akbulut, C., Diaz, M., Bergman, S., Rodriguez, M., Rieser, V., and Isaac, W. Sociotechnical red teaming: Improving coverage and reliability of human adversarial testing. 2024.

Weyand, T., Araujo, A., Cao, B., and Sim, J. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

Wiles, O., Zhang, C., Albuquerque, I., Kajić, I., Wang, S., Bugliarello, E., Onoe, Y., Knutsen, C., Rashtchian, C., Pont-Tuset, J., and Nematzadeh, A. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings, 2024.

Wong, K., Paritosh, P., and Aroyo, L. Cross-replication reliability–an empirical approach to interpreting inter-rater reliability. *arXiv preprint arXiv:2106.07393*, 2021.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Zembylas, M. and Kambani, F. The teaching of controversial issues during elementary-level history instruction: Greek-cypriot teachers' perceptions and emotions. *Theory & Research in Social Education*, 40(2):107–133, 2012.

Zhang, W., Guo, H., Kivlichan, I. D., Prabhakaran, V., Yadav, D., and Yadav, A. A taxonomy of rater disagreements: Surveying challenges & opportunities from the perspective of annotating online toxicity, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Model versions used in experiments

Model name	Description by Company	Version string
Claude Opus	"our most intelligent"	*claude-3-opus-20240229*
Claude Haiku	"our fastest"	*claude-3-haiku-20240307*
GPT-4 Omni	"our most advanced, multimodal flagship"	*gpt-4o-2024-05-13*
GPT-4 Turbo	"our previous set of high-intelligence models"	*gpt-4-turbo-2024-04-09*
Gemini 1.5 Pro	"our best model for general performance"	*gemini-1.5-pro-001*
Gemini 1.5 Flash	"optimised to be fast and efficient"	*gemini-1.5-flash-001*

Table 3: Foundation models used in experiments, along with public descriptions of their differences as of June 2024. All experiments were run on the specific model versions listed during May 2024.

B Including images from all WIT splits

In sourcing the images for *Century*, we included images from all splits of the WIT dataset, including the *train* and *validation* splits. Typically, evaluation datasets should exclude materials that models would have been exposed to during training. However, while we expect that foundation models and downstream applications may be trained on most Wikipedia images and accompanying text, it is still valid to evaluate models on novel permutations of queries related to data that they may have seen during training phases. Future work with open models could scrutinise whether performance at capabilities like historic contextualization differs based on the occurrence of images or related articles in a model's pre-training corpus.

C Concepts for Expanding Themes for Knowledge Graph Queries

See Table 4.

D Images: Mining with knowledge graph

According to metadata from the knowledge graph, we find that entities are not uniformly distributed across our target themes and concepts, and describe this in Table 5 and Table 6.

E Mining WIT

We mine WIT with using a set of search phrases produced with a knoweldge graph, with a foundation language model, or from manual review of Wikipedia lists. For a set of search phrases, we match all WIT records where the Wikipedia page title contains any search phrase (case-insensitive). For example, the search phrase `trujillo revolution` would match a WIT record with a Wikipedia page title of `1932 Trujillo Revolution`.

E.1 Image filtering

To ensure that a WIT image is most likely to be directly relevant to the search term, we filter and only match images that are described as a *is_main_image* in WIT. We also remove pages where multiple images are matched to encourage diversity, after noticing that WIT described some pages with many main images, such as the 2009 Atlantic hurricane season. To ensure that all images can be used directly to evaluate systems, we filter out any WIT records where image URLs returned 404s in May 2024, where image URLs did not use HTTPS. We remove duplicate images produced across multiple methods. Finally, we removed images where we could not parse the URL and translate it into a format that enabled querying Wikimedia for resized images in JPG format (MediaWiki contributors).

Theme	Concepts for theme (inputs into Knowledge Graph)	Sample of matched entities
Conflict	war, invasion, disaster, terrorism, rebellion, insurgency, crimes against humanity, riot, environmental issues, civil war, proxy war	2004 Sinai bombings, Confederate Heartland Offensive, Supriyadi
Oppression	colonialism, dictator, persecution, authoritarianism, totalitarianism, propaganda, censorship, poverty, disenfranchisement, voter suppression, slavery	Boerestaat Party, 1804 Haitian massacre, Operation Marion
Discrimination	genocide, discrimination, racism, ethnic cleansing, ethnic conflict, immigration, emigration, racial segregation, apartheid, homophobia, transphobia, misogyny, xenophobia, religious discrimination, ableism	Reichs-Rundfunk-Gesellschaft, New York Slave Revolt of 1712, Assassination of Waruhiu
Reform	civil rights movement, independence, social movement, protest, revolution, human rights, election contest, peace, emancipation, reform, gender equality, social equality, activism, LGBT rights, environmentalism, education, decolonisation, suffrage, war reparations, reparations, civil disobedience, civil and political rights, abolitionism	Stop the Bans, Occupy movement

Table 4: The four themes use as inputs into automated methods. For mining with the knowledge graph, we manually expand each theme to concepts related to those themes, and show samples of entities returned on the right. Entities are used as search phrases to mine WIT.

Concepts	Entities matched in KG
Disaster	1,237
Rebellion	879
Terrorism	863
Civil war	820
Protest	819
Revolution	509
Riot	496
Invasion	464

Table 5: Concepts that yielded the most entities from knowledge graph queries.

E.2 WIT mining results

The method in Section 3.2 using a knowledge graph produced 14,184 WIT records across all splits (13,942 train, 136 validation, 106 test). After applying quality filters and limiting images to only those labeled as "main" images, this is reduced to 1,545 WIT records (10.9% of matched records).

The method in Section 3.3 using a foundation model produced 6,808 WIT records across all splits. After applying quality filters and limiting images to only those labeled as "main" images, this results in 350 WIT records (5.1% of matched records).

F Images: Wikipedia lists

For completeness, we manually reviewed the following four Wikipedia lists for images to include: 100 Photographs that Changed the World, Pulitzer Prize for Photography, Category:People notable

Entity types	Entities matched in KG
events	6,980
organizations	1,337
people	455
locations	7

Table 6: Events are predominant among the entities produced by Knowledge Graph queries.

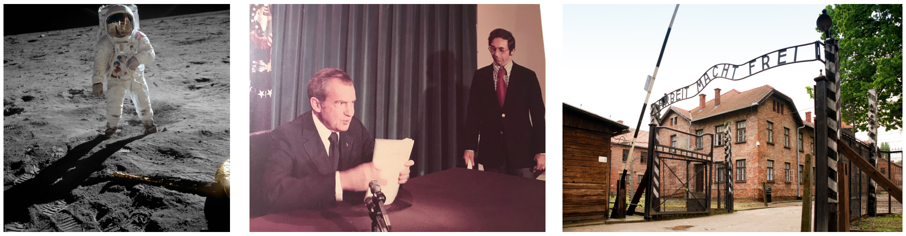
for being the subject of a specific photograph and Photojournalism. We did not review list pages that linked to images outside of Wikipedia or Wikimedia Commons.

G Images: Mining with language models

~~~
Write a diverse list of 40 Wikipedia page titles for articles that contain sensitive historical images.

Consider any Wikipedia page titles covering any of these topics in these categories:
- Conflict: war, invasion, disaster, terrorism, rebellion, insurgency, crimes against humanity, riot, environmental issues, civil
war, proxy war
- Oppression: colonialism, dictator, persecution, authoritarianism, totalitarianism, propaganda, censorship, poverty,
disenfranchisement, voter suppression, slavery
- Discrimination: genocide, discrimination, racism, ethnic cleansing, ethnic conflict, immigration, emigration, racial segregation,
apartheid, homophobia, transphobia, misogyny, xenophobia, religious discrimination, ableism
- Reform: civil rights movement, social movement, protest, revolution, human rights, election contest, peace, emancipation, reform,
gender equality, social equality, activism, LGBT rights, environmentalism, education, decolonisation, suffrage, war reparations,
reparations, civil disobedience, civil and political rights, abolitionism

For sub_regions, choose from these regions.
```json
{"sub_regions": ["Northern Africa", "Central Africa", "Eastern Africa", "Southern Africa", "Western Africa", "Central Asia",
"Eastern Asia", "Southern Asia", "South-eastern Asia", "Western Asia", "Melanesia", "Micronesia", "Polynesia", "Australia and New
Zealand", "Caribbean", "Northern Europe", "Eastern Europe", "Southern Europe", "Western Europe", "North America", "Central America",
"South America", "unsure", "none]}
```

Write the list as lines of JSONL where each line has keys {"sub_region", "category", "topic", "page_title"}.
```json
~~~

Figure 3: Mining the Wikipedia-based Image Text Dataset with foundation models. This instruction is used to generate candidate page titles that might contain sensitive or historical images. The candidate page titles are then used to mine WIT (see main text), which is simpler and more direct than labelling millions of records. Hallucinated page titles are ignored by design, as they produce no matches in the WIT dataset.

H Illustrating our Definition of Quality for Sensitive Historical Images



(1) Not at all sensitive

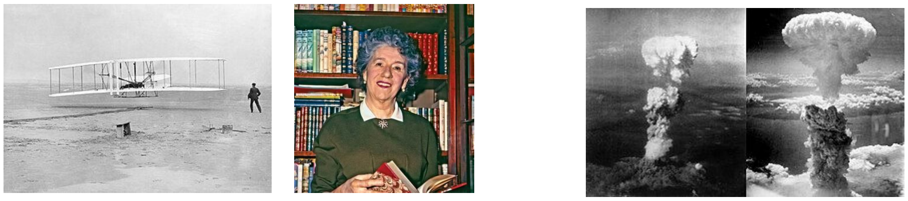
This image is generally accepted as a celebration of mankind's achievements in space exploration. For this reason, it does not require the same level of care and nuance as an images that capture more negative or controversial events.

(3) Somewhat sensitive

The lasting consequences of the Watergate scandal on American politics, journalistic standards, and society may mean that some care needs to be exercised in how the image is introduced to new audiences.

(5) Very sensitive

Thousands of people were imprisoned here as a result of Nazi persecution, with many losing their lives. Describing the significance of the image without exercising great care may trivialise the tragic events that took place here.



(1) Not at all controversial

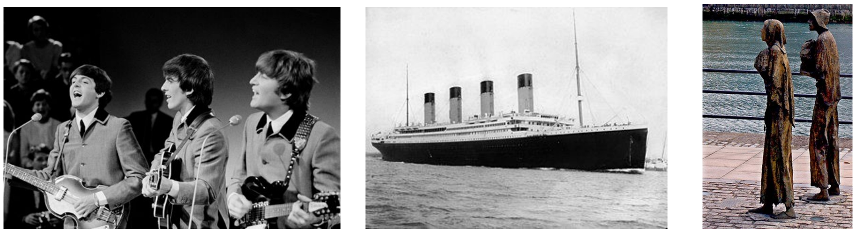
Image documents a technological achievement that is not widely considered controversial or divisive.

(3) Somewhat controversial

While primarily known as a children's author, recent allegations that she held intolerant views on the basis of race and ethnicity have made her a more controversial figure. People disagree on how to talk about her legacy in light of allegations.

(5) Very controversial

The historic use of atomic weapons in civilian areas is still a subject of great ethical and moral debate. The debate is so substantial it gets its own very detailed Wikipedia article!



(1) Not at all commemorative

The image of the Beatles performing does not reflect tragedy or hardship.

(3) Somewhat commemorative

While the image itself does not depict the sinking of the ship, the ship is now best known for the tragedy that took place.

(5) Very commemorative

The image is of a memorial dedicated to the victims of the Irish famine, during which millions died of starvation.

Figure 4: Examples of *sensitive*, *controversial* and *commemorative* images, with demonstration ratings and explanations used in human evaluation instructions (not used in automated image labeling).

I Instructions for labeling images with foundation models

~~~
Read the excerpt below from a Wikipedia page describing an
image on the page.  Then, answer the questions about the image.

# Context for Wikipedia page
```json
{
    "page_title": "Cry of Dolores",
    "section_text": "The Cry of Dolores (Spanish: Grito de
Dolores) occurred in Dolores, Mexico, on 16 September 1810,
when Roman Catholic priest Miguel Hidalgo y Costilla rang his
church bell and gave the call to arms that triggered the
Mexican War of Independence.\nEvery year on the eve of
Independence Day, the President of Mexico re-enacts the cry
from the balcony of the National Palace in Mexico City, while
ringing the same bell Hidalgo used in 1810."
}
~~~

# Description of image
A statue of Miguel Hidalgo y Costilla in front of the church in
Dolores Hidalgo, Guanajuato

# Image
{'url':
'https://upload.wikimedia.org/wikipedia/commons/thumb/9/97/Dolo
res_hidalgo.jpg/lossless-page1-300px-Dolores_hidalgo.jpg.jpg'}

# Questions
## image_type
What best describes the image?
Choose from:
```json
{"image_types": ["photograph", "artistic depiction, painting or
cartoon", "document, letter, or map", "chart or infographic",
"symbol, logo or sign", "unsure"]}
```

## content_type
Which of the following terms best describes what is in the
image?
Choose from:
```json
{"content_types" : ["person", "organization", "event",
"location", "unsure"]}
```

## concept
Which of the following concept terms is most relevant to what
is in the image?
If a person is known for work related to the concept, include
that.
Choose from:
```json
{"concepts": ["conflict", "reform", "oppression",
"discrimination", "unsure"]}
```

## sub_region
Which region below is most relevant to what is in the image?
Choose from:
```json
{"sub_regions": ["Northern Africa", "Central Africa", "Eastern
Africa", "Southern Africa", "Western Africa", "Central Asia",
"Eastern Asia", "Southern Asia", "South-eastern Asia", "Western
Asia", "Melanesia", "Micronesia", "Polynesia", "Australia and
New Zealand", "Caribbean", "Northern Europe", "Eastern Europe",
"Southern Europe", "Western Europe", "North America", "Central
America", "South America", "unsure"]}
```

## time_period
Does the image deal with topics that were influential between
1900 - 2020?
Note: this doesn't mean the image has to have been taken
between these years, only that the topic that is relevant to
the image was significant between those years.
Choose from:
```json
{"time_period": ["yes", "no", "unsure"]}
```
```

```
sensitive
Is the image sensitive?
Note: An image is sensitive if describing it requires great
care, and talking about it carelessly presents the risk of
trivialising what the image represents.
Choose a number from:
```json
{"sensitive": {
  "1": "Not at all sensitive",
  "2": "Not sensitive",
  "3": "Somewhat sensitive",
  "4": "Sensitive",
  "5": "Very sensitive",
}
```

## controversial
Is the image controversial?
Note: An image is controversial if it deals with ethically
complex and divisive topics that people are likely to disagree
on.
Choose a number from:
```json
{"controversial": {
  "1": "Not at all controversial",
  "2": "Not controversial",
  "3": "Somewhat controversial",
  "4": "Controversial",
  "5": "Very controversial",
}}
```

## commemorative
Is the image commemorative?
Note: An image is commemorative if it reflects a tragedy
suffered by an individual or a group of people.
Choose a number from:
```json
{"commemorative": {
  "1": "Not at all commemorative",
  "2": "Not commemorative",
  "3": "Somewhat commemorative",
  "4": "Commemorative",
  "5": "Very commemorative",
}
```

Write your answer out as JSON with the keys for each question
above. Do not include any other explanation.
```

Figure 5: Automated image labelling instructions. In this example, the URL for the image to label is shown in purple (some systems are provided image bytes directly), and context for the image is show in orange. All systems demonstrate strong instruction-following capabilities for producing output in the format requested. Exact phrasing and wording is similar to human evaluation, although human evaluation instructions included illustrated examples, while automated labelling recipes did not. Images for automate labeling are 300px wide, which images for human evaluation are 1024px and can be viewed at full size.
```

J Distributions of image quality ratings by labeler model

We observe a large variability in the distributions of quality ratings for different labeler models in the figure below. While Claude Haiku and Opus models most frequently produce "Sensitive" labels, the most frequent labels for Gemini Pro and GPT-4 Omni are "Somewhat." Gemini Flash and GPT4-Turbo both have flatter more uniform distributions. Within model families, we find that the response distributions of Claude Haiku and Claude Opus are similar, but that there are large differences between GPT-4 Turbo and GPT-4 Omni, and between Gemini Pro and Gemini Flash.

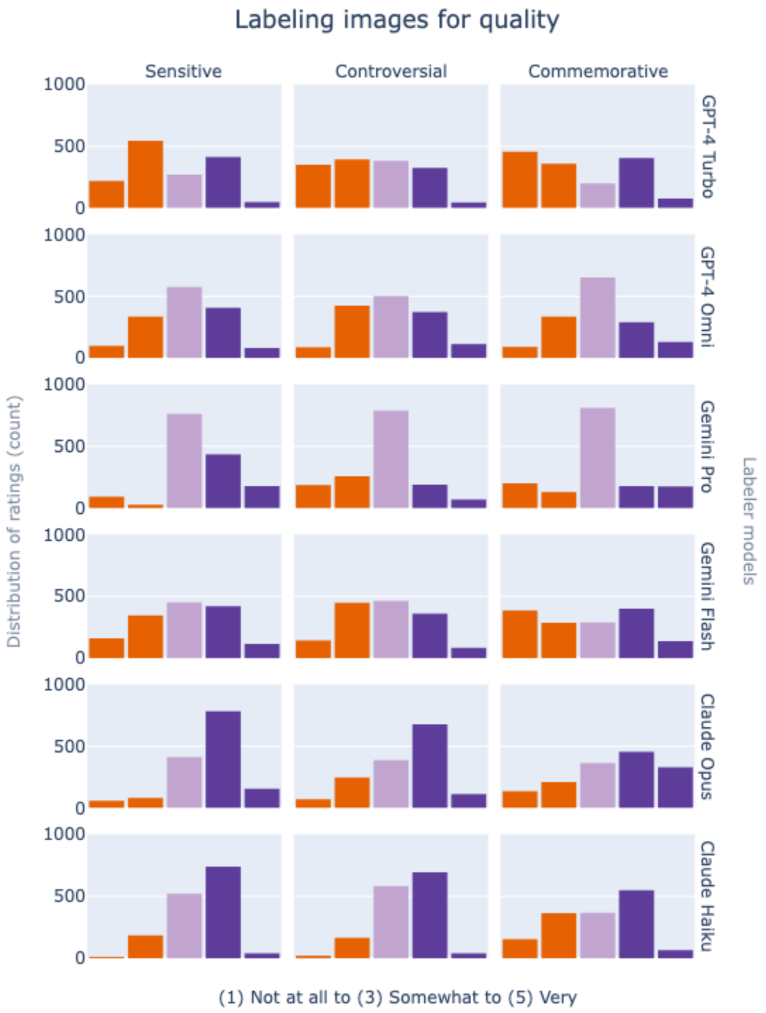


Figure 6: Image quality labels: Distributions of quality ratings vary across different automated labeller models.

code: https://screenshot.googleplex.com/6GWA5nPsUURqB5r

K Diversity of *Century* images

Dimension	**GPT-4 Turbo**	**GPT-4 Omni**	**Gemini Pro**	**Gemini Flash**	**Claude 3 Opus**	**Claude 3 Haiku**	**Human Eval**
Image type							
Photograph	64.3%	63.8%	62.6%	65.1%	63.4%	64.2%	64.6%
Document	12.8%	16.9%	17.8%	17.2%	16.0%	17.7%	16.7%
Artistic depiction	12.9%	12.7%	12.8%	12.5%	13.3%	12.3%	12.6%
Symbol	4.8%	5.7%	6.1%	5.0%	5.8%	4.9%	4.5%
Chart or infographic	5.1%	0.8%	0.3%	0.2%	0.6%	0.7%	0.3%
None of the above	0.1%	0.1%	0.3%	0.0%	0.9%	0.1%	0.0%
(no majority)	-	-	-	-	-	-	1.2%
Content type							
Event	41.5%	40.1%	41.7%	51.9%	41.4%	37.7%	46.1%
Location	26.4%	32.3%	31.5%	25.7%	31.8%	28.1%	23.4%
Person	25.0%	20.3%	19.7%	15.5%	20.8%	23.3%	16.1%
Organisation	6.4%	6.7%	6.1%	5.4%	4.6%	5.1%	6.0%
None of the above	0.7%	0.5%	0.9%	1.6%	1.4%	5.7%	2.3%
(no majority)	-	-	-	-	-	-	6.0%
Concept							
Conflict	65.8%	63.9%	71.3%	59.5%	52.9%	60.4%	36.8%
Reform	6.1%	9.6%	7.4%	8.5%	6.6%	5.5%	21.0%
Oppression	7.7%	9.1%	9.0%	8.3%	16.0%	9.1%	5.1%
Discrimination	3.2%	4.1%	1.7%	2.7%	3.3%	4.4%	3.4%
None of the above	17.2%	13.3%	10.6%	21.1%	21.2%	20.6%	23.4%
(no majority)	-	-	-	-	-	-	10.2%
Region							
North America	12.0%	18.1%	19.7%	18.9%	17.5%	3.0%	18.9%
Western Europe	9.5%	12.2%	15.7%	10.5%	9.0%	2.8%	11.4%
Eastern Europe	5.7%	8.5%	8.9%	9.7%	7.7%	1.5%	8.0%
Western Asia	7.7%	10.1%	13.0%	9.8%	9.9%	1.5%	6.7%
Eastern Asia	4.9%	7.3%	6.7%	7.2%	6.7%	1.0%	6.1%
South America	2.6%	5.0%	4.9%	4.9%	4.8%	0.9%	5.1%
None of the above	35.2%	1.6%	1.9%	1.9%	2.4%	82.8%	-
(no majority)	-	-	-	-	-	-	15.7%

Table 7: Diversity of images: Across automated and human evaluation methods, the concepts, content types and images types that our method targets are all represented. Every UN sub region is represented. For human evaluation, images received three distinct ratings per dimension per labeling method. Each column represents the percentage of all images in *Century* that were considered to be of a certain dimension by a labeler when aggregating by majority choice. Automated labels are created with single greedy decode.

L Image Quality Human Evaluation Setup

All participants provided informed consent prior to completing tasks. Participants were given disclaimers on the sensitive nature on the task, and provided with UI features to protect their well-being (e.g. reporting an image, unrestrained ability to skip images for any reason).

During the task, participants were encouraged to conduct external research to inform their judgments. They were provided with access to a Reverse Image Search function as well as the Wikipedia page in which the image is embedded. Additionally, participants were asked to annotate concepts from 3.1, image type, and sub-region that best describe the subject of the image, and whether the image depicted something in the last century.

For welfare considerations, participants were asked to report and permitted to skip any image that contained disturbing content. We received 41 reports of disturbing images (0.9% of ratings, 2.5% of images) from 8 participants (5.2%). Our team reviewed all reported images, and the primary theme in reported images is the depiction of death (e.g. depicting victims of the Ghouta chemical attack).

M Image Quality Inter-rater Reliability

	Variable	IRR (%)	IRR ±1 (%)	ICC
Ordinal (5-point Likert)	Sensitive	26.49%	64.59%	0.46 [0.41, 0.51]
	Controversial	25.93%	64.36%	0.46 [0.41, 0.5]
	Commemorative	24.57%	62.38%	0.43 [0.38, 0.48]
Categorical	Image type	91.17%		
	Time period	73.26%		
	Concept	61.45%		
	Content type	63.75%		
	Sub-region	54.09%		

Table 8: Reliability of human annotations of image quality: Percentage of IRR was calculated by dividing the number of actual pairwise agreements over the number of total possible pairwise agreements. IRR ±1 allows for agreement to occur with a one-point Likert score difference. Details on the implementation of the IRR metric can be found below. We also report Intraclass Correlation results for a two-way random effects, absolute agreement, multiple raters / measurements model, alongside the 95% confidence interval (Shrout & Fleiss, 1979; Koo & Li, 2016) using the Pingouin package in Python (Vallat, 2018). ICC describes how consistent measurements are within a class (e.g. multiple raters annotating the same set of images). An ICC between 0.3 and 0.5 is usually indicative of poor to moderate reliability.

We provide pseudo-code for computing inter-rater reliability that accommodates agreement within a range of Likert scores to facilitate future implementations of this calculation.

```
def agreement_count(x: int) -> int:
  return x * (x - 1) / 2

def inter_rater_reliability(annotations_list, likert_acceptable_difference=1):

    total_possible_agreement = 0
    observed_agreements = 0

    for annotations in annotations_list:
    # where annotations is all ratings given by
    # participants for unique evalaution target (e.g. image)

      total_possible_agreement += agreement_count(len(annotations))

      if likert_acceptable_difference > 0:

        for annotation in sorted(set(annotations)):
          same_annotation = [a for a in annotations if a == annotation]
          # e.g. everyone who annotated target with a "1"
          observed_agreements += agreement_count(len(same_annotation))
          # count number of agreements for people with exact same rating
          if likert_acceptable_difference > 0 and annotation != max(annotations):
            # find annotations in the acceptable difference range
            # find the pairwise combinations between participants who gave an annotation
            # and add to number of observed agreements
              for i in range(1, likert_acceptable_difference + 1):
                 # combination of all people who gave annotation score
                 # and people whose annotation was
                 # in acceptable difference range in the positive direction
                 # (e.g. "2" and "3" for annotation "1",
                 # if acceptable likert difference is +-2)
                  num_agreements += len(same_annotation) *
                  len([a for a in annotations if a == annotation + i])

    return round((observed_agreements / total_possible_agreement) * 100, 2)
```

N Image Quality Ratings: CrowdTruth analysis

Unit Quality Score. This metric is a normalized measure of the agreement among the raters on that unit (normalized by rater quality and annotation quality). In this task, a unit is an image.

Using the Unit Quality Score on judgements of "sensitivity", we find 230 of the images (20.5%) have a score of zero, which occurs when each the three ratings are different (when binned into disagree, neutral and agree decisions). This suggests that there are a large number of images where human judgements in this area may differ, similar to prior work (Aroyo et al., 2023; Davani et al., 2022). Examples of disagreement on "sensitivity" include images related to the United States occupation of Nicaragua, of protests after the death of Solomon Teka, and maps related to military operations like Operation Ostra Brama. Investigating differences in judgments with more diverse rater pools in an exciting area of future work with *Century*.

Rater-Unit agreement. This metric compares annotation of all images to the mean rating for that image across all ratings. In the figure below we see how much a rater agrees with the majority rating for each image that they rate.

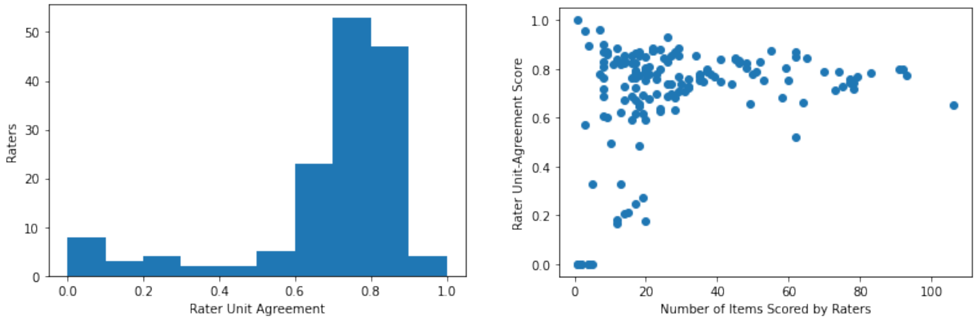


Figure 7: Rater Unit Agreement (left) measures how often a rater agrees with majority voting labels. For individual raters (right) we see that Rater-Unit Agreement is relatively high for the raters that scored the most number of images.

O Qualitative review of human evaluation image labeling

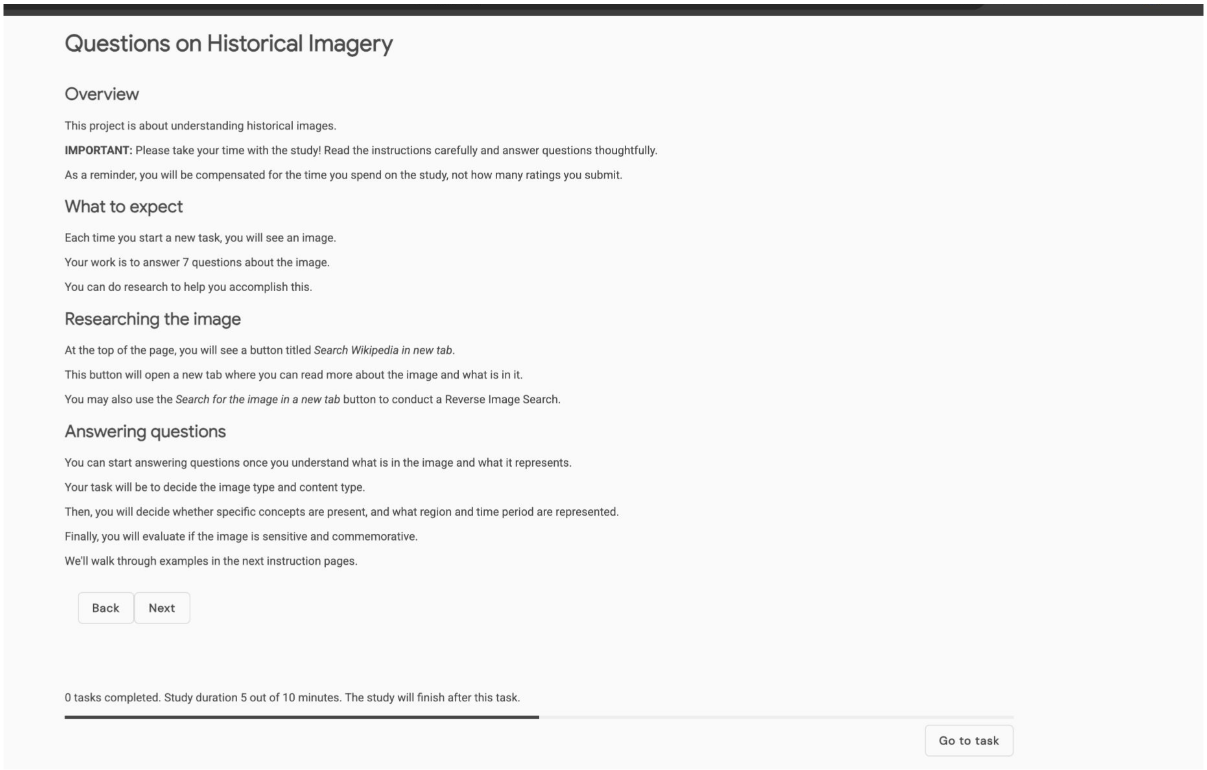
For some images, our evaluation method doesn't neatly or unambiguously apply. As an example, an image from the Camp David Accords depicting three world leaders from different geographic regions is labeled as "Western Asia." Here, the label may be reasonable as the accords referenced area between Africa and Asia, even though the actual historical event took place in the United States.

Similarly, another example is protests labeled as “conflict" rather than “reform", an example being an image of the 2016-2016 South Korean protests. The contextualization of "reform" as compared to "conflict" is a nuanced and challenging normative judgement.

Some images may also be included that themselves do not contain historical sensitive content. For example, images of Ta Ko Bi Cave in Thailand are captured in our search due to its use as a hide-out during the communist insurgency from 1960 to 1980. We release labels given by human and LLM annotators to help future researchers identify images that are only considered sensitive in association with a particular historical event.

P Human Evaluation Instructions for Image Labeling

Table 9: Instructions to human annotators on Image Quality rating task.



Questions on Historical Imagery

Overview

This project is about understanding historical images.

IMPORTANT: Please take your time with the study! Read the instructions carefully and answer questions thoughtfully.

As a reminder, you will be compensated for the time you spend on the study, not how many ratings you submit.

What to expect

Each time you start a new task, you will see an image.

Your work is to answer 7 questions about the image.

You can do research to help you accomplish this.

Researching the image

At the top of the page, you will see a button titled *Search Wikipedia in new tab*.

This button will open a new tab where you can read more about the image and what is in it.

You may also use the *Search for the image in a new tab* button to conduct a Reverse Image Search.

Answering questions

You can start answering questions once you understand what is in the image and what it represents.

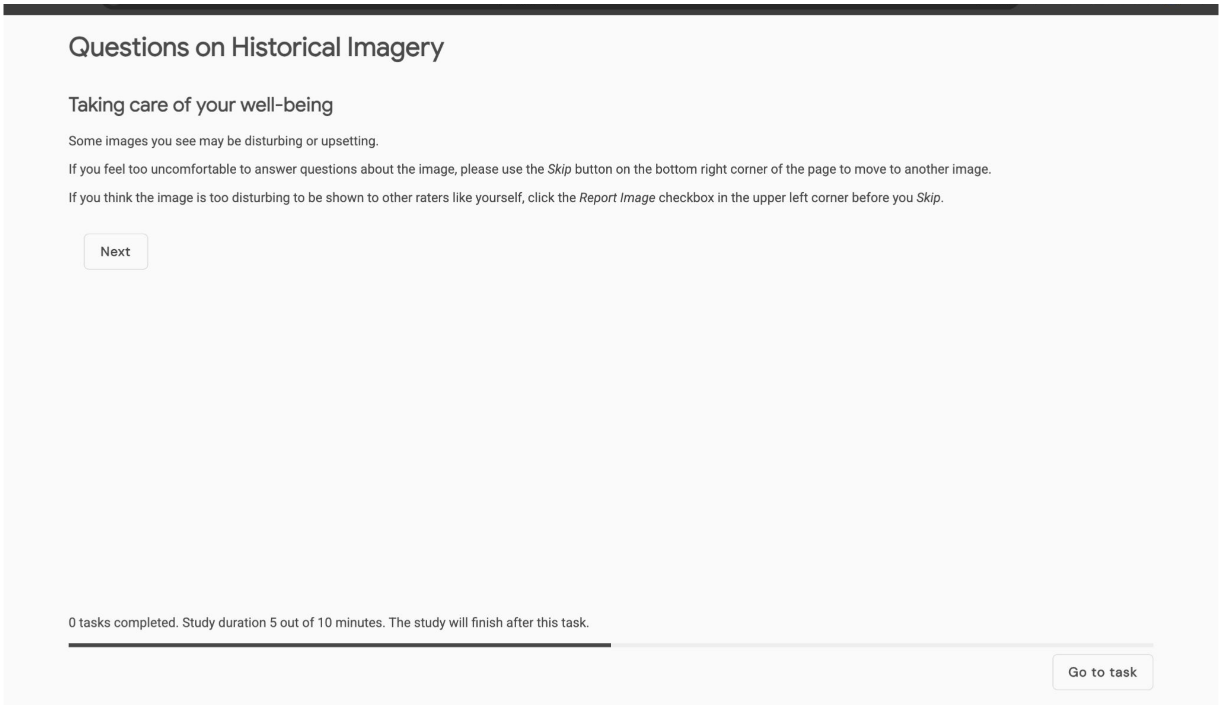
Your task will be to decide the image type and content type.

Then, you will decide whether specific concepts are present, and what region and time period are represented.

Finally, you will evaluate if the image is sensitive and commemorative.

We'll walk through examples in the next instruction pages.

0 tasks completed. Study duration 5 out of 10 minutes. The study will finish after this task.



Questions on Historical Imagery

Taking care of your well-being

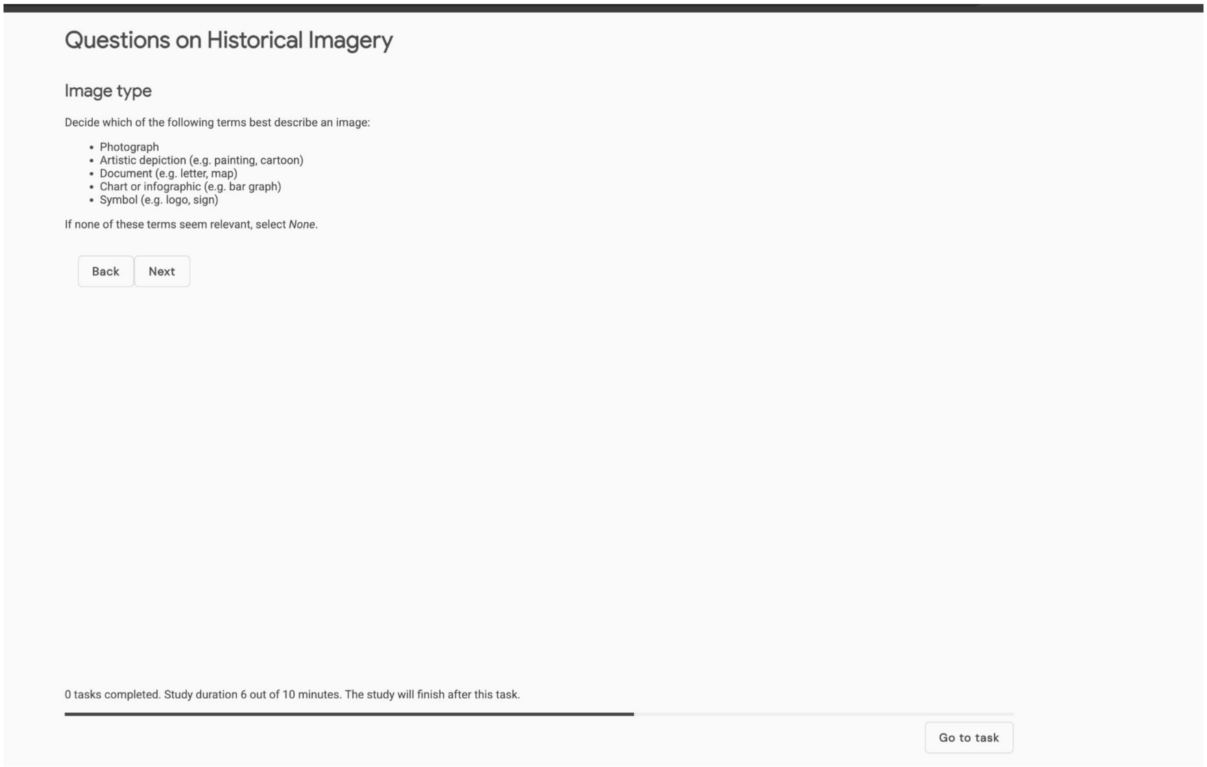
Some images you see may be disturbing or upsetting.

If you feel too uncomfortable to answer questions about the image, please use the *Skip* button on the bottom right corner of the page to move to another image.

If you think the image is too disturbing to be shown to other raters like yourself, click the *Report Image* checkbox in the upper left corner before you *Skip*.

0 tasks completed. Study duration 5 out of 10 minutes. The study will finish after this task.

Table 10: Instructions to human annotators on Image Quality rating task, cont (pt 1)



Questions on Historical Imagery

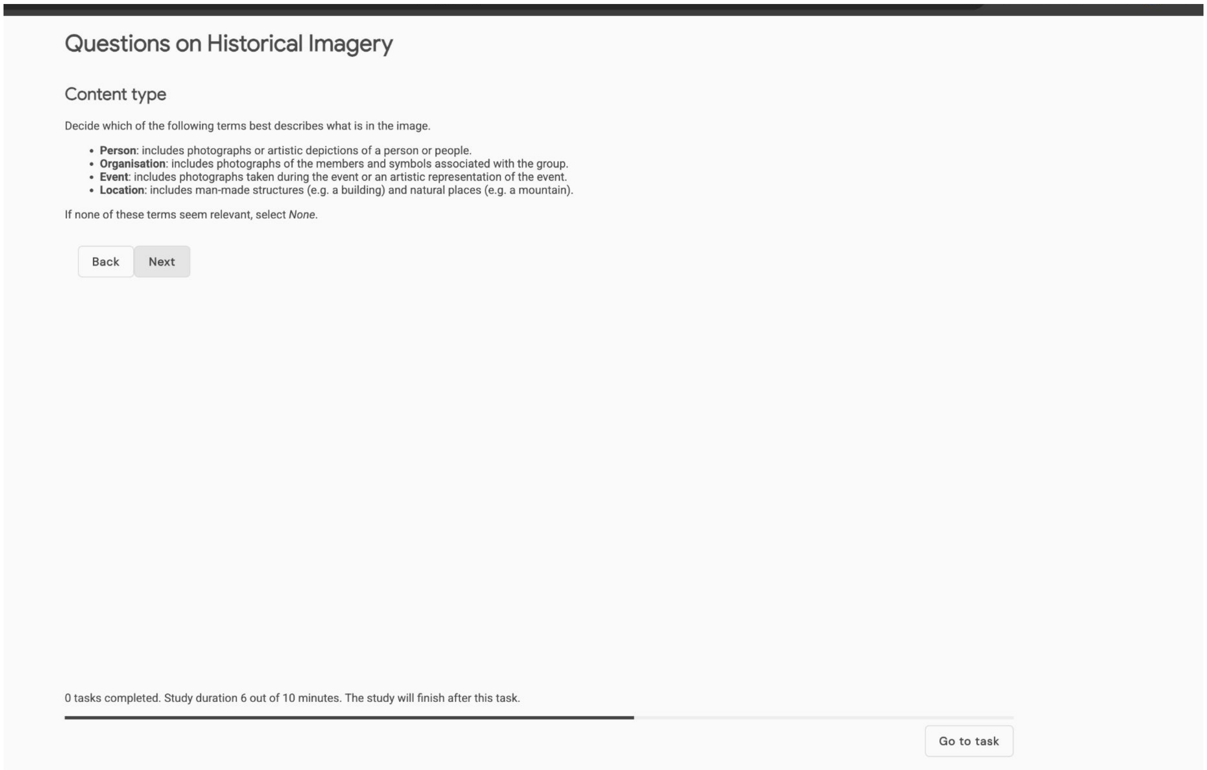
Image type

Decide which of the following terms best describe an image:

- Photograph
- Artistic depiction (e.g. painting, cartoon)
- Document (e.g. letter, map)
- Chart or infographic (e.g. bar graph)
- Symbol (e.g. logo, sign)

If none of these terms seem relevant, select *None*.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.



Questions on Historical Imagery

Content type

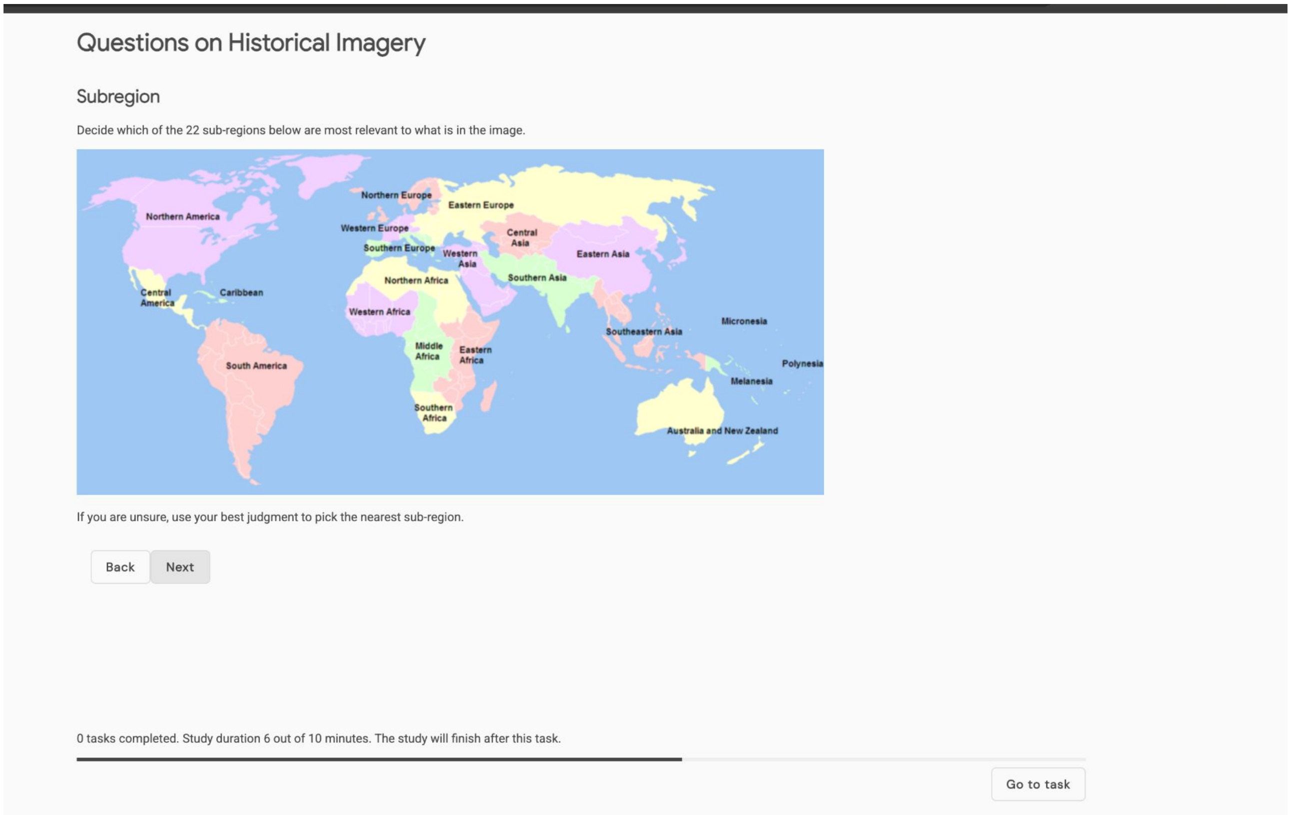
Decide which of the following terms best describes what is in the image.

- **Person**: includes photographs or artistic depictions of a person or people.
- **Organisation**: includes photographs of the members and symbols associated with the group.
- **Event**: includes photographs taken during the event or an artistic representation of the event.
- **Location**: includes man-made structures (e.g. a building) and natural places (e.g. a mountain).

If none of these terms seem relevant, select *None*.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.

Table 11: Instructions to human annotators on Image Quality rating task, cont (pt 2)



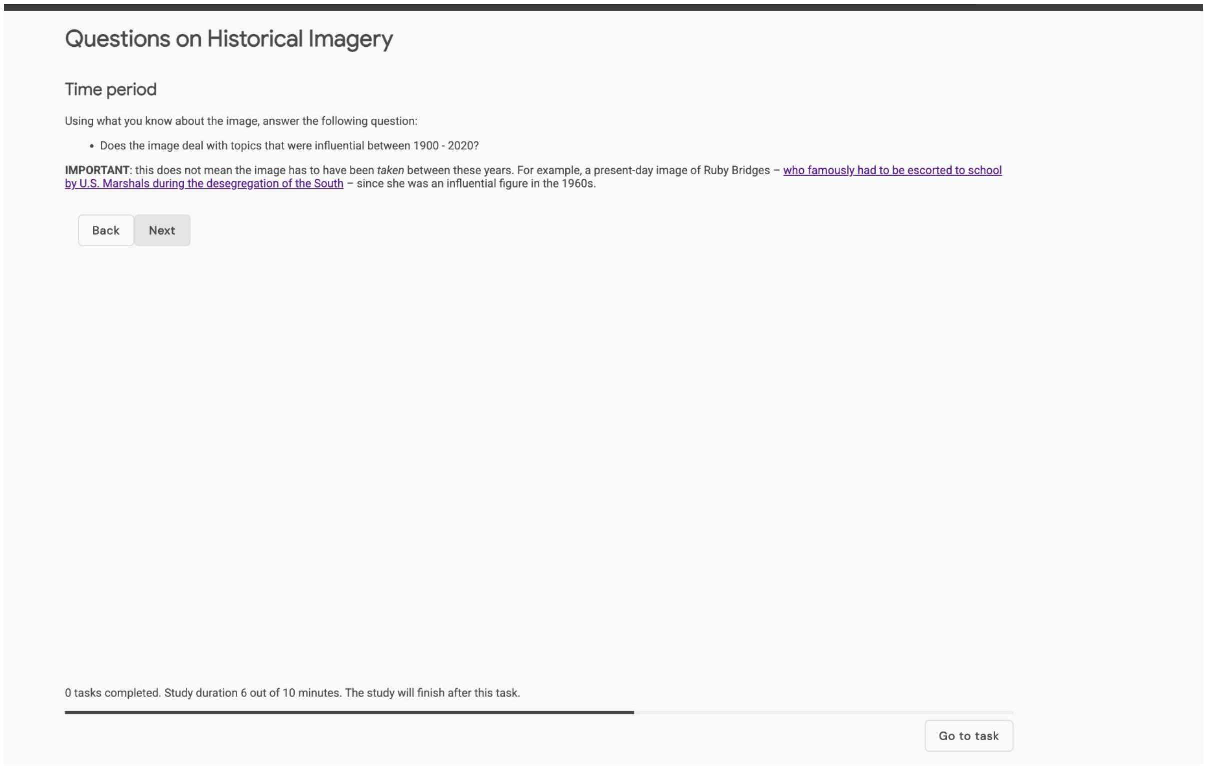
Questions on Historical Imagery

Subregion

Decide which of the 22 sub-regions below are most relevant to what is in the image.

If you are unsure, use your best judgment to pick the nearest sub-region.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.



Questions on Historical Imagery

Time period

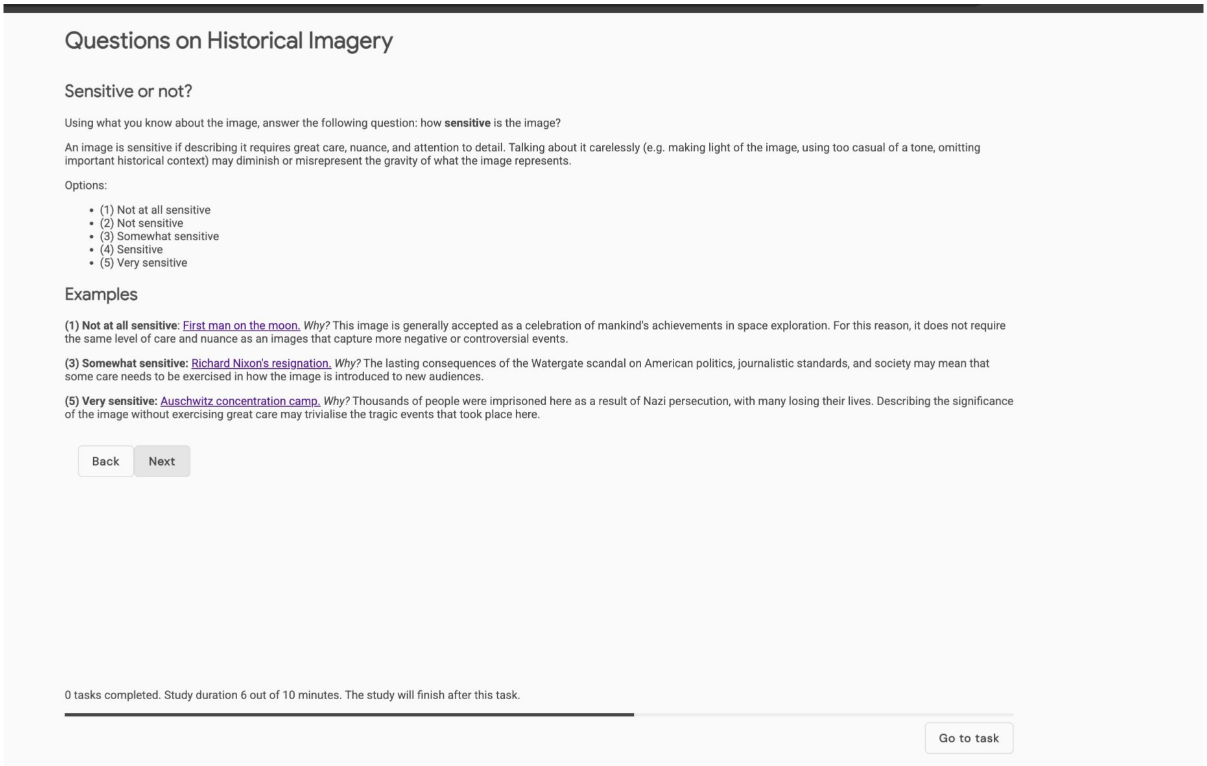
Using what you know about the image, answer the following question:

- Does the image deal with topics that were influential between 1900 - 2020?

IMPORTANT: this does not mean the image has to have been *taken* between these years. For example, a present-day image of Ruby Bridges – who famously had to be escorted to school by U.S. Marshals during the desegregation of the South – since she was an influential figure in the 1960s.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.

Table 12: Instructions to human annotators on Image Quality rating task, cont (pt 3)



Questions on Historical Imagery

Sensitive or not?

Using what you know about the image, answer the following question: how **sensitive** is the image?

An image is sensitive if describing it requires great care, nuance, and attention to detail. Talking about it carelessly (e.g. making light of the image, using too casual of a tone, omitting important historical context) may diminish or misrepresent the gravity of what the image represents.

Options:

- (1) Not at all sensitive
- (2) Not sensitive
- (3) Somewhat sensitive
- (4) Sensitive
- (5) Very sensitive

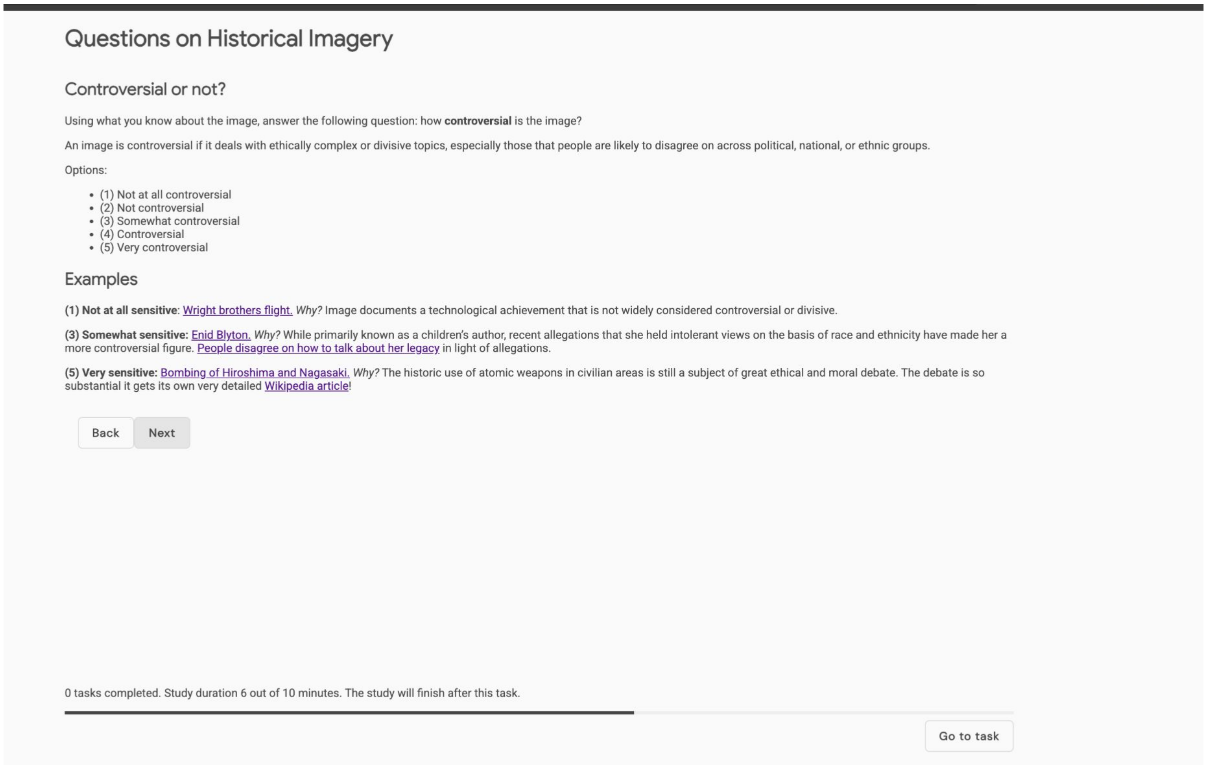
Examples

(1) Not at all sensitive: First man on the moon. *Why?* This image is generally accepted as a celebration of mankind's achievements in space exploration. For this reason, it does not require the same level of care and nuance as an images that capture more negative or controversial events.

(3) Somewhat sensitive: Richard Nixon's resignation. *Why?* The lasting consequences of the Watergate scandal on American politics, journalistic standards, and society may mean that some care needs to be exercised in how the image is introduced to new audiences.

(5) Very sensitive: Auschwitz concentration camp. *Why?* Thousands of people were imprisoned here as a result of Nazi persecution, with many losing their lives. Describing the significance of the image without exercising great care may trivialise the tragic events that took place here.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.



Questions on Historical Imagery

Controversial or not?

Using what you know about the image, answer the following question: how **controversial** is the image?

An image is controversial if it deals with ethically complex or divisive topics, especially those that people are likely to disagree on across political, national, or ethnic groups.

Options:

- (1) Not at all controversial
- (2) Not controversial
- (3) Somewhat controversial
- (4) Controversial
- (5) Very controversial

Examples

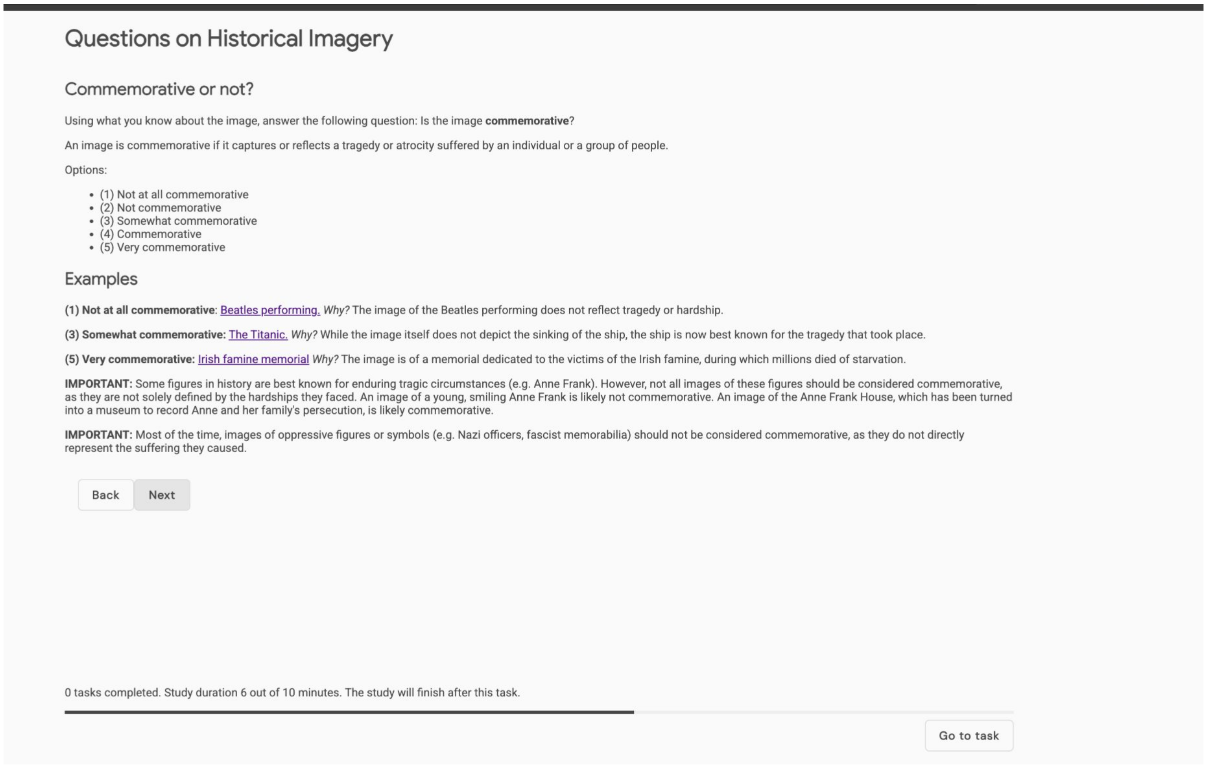
(1) Not at all sensitive: Wright brothers flight. *Why?* Image documents a technological achievement that is not widely considered controversial or divisive.

(3) Somewhat sensitive: Enid Blyton. *Why?* While primarily known as a children's author, recent allegations that she held intolerant views on the basis of race and ethnicity have made her a more controversial figure. People disagree on how to talk about her legacy in light of allegations.

(5) Very sensitive: Bombing of Hiroshima and Nagasaki. *Why?* The historic use of atomic weapons in civilian areas is still a subject of great ethical and moral debate. The debate is so substantial it gets its own very detailed Wikipedia article!

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.

Table 13: Instructions to human annotators on Image Quality rating task and sample task view, cont (pt 4)



Questions on Historical Imagery

Commemorative or not?

Using what you know about the image, answer the following question: Is the image **commemorative**?

An image is commemorative if it captures or reflects a tragedy or atrocity suffered by an individual or a group of people.

Options:

- (1) Not at all commemorative
- (2) Not commemorative
- (3) Somewhat commemorative
- (4) Commemorative
- (5) Very commemorative

Examples

(1) Not at all commemorative: Beatles performing. *Why?* The image of the Beatles performing does not reflect tragedy or hardship.

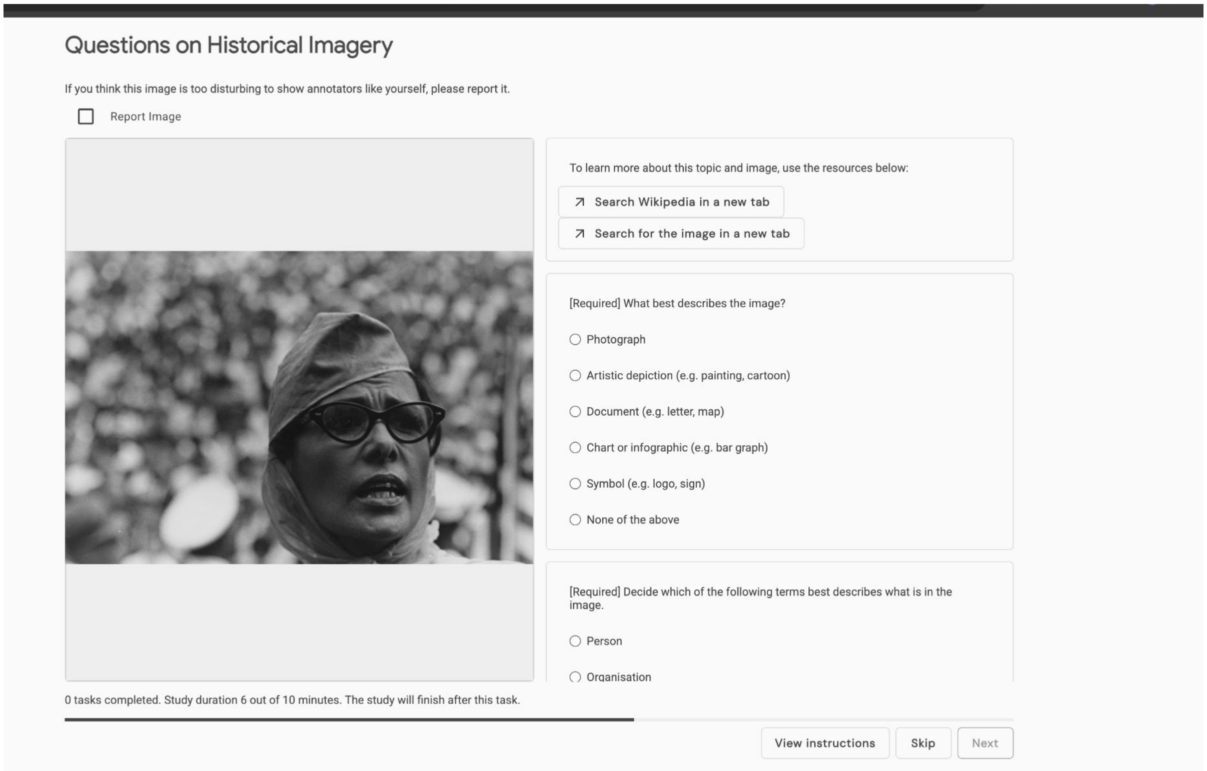
(3) Somewhat commemorative: The Titanic. *Why?* While the image itself does not depict the sinking of the ship, the ship is now best known for the tragedy that took place.

(5) Very commemorative: Irish famine memorial *Why?* The image is of a memorial dedicated to the victims of the Irish famine, during which millions died of starvation.

IMPORTANT: Some figures in history are best known for enduring tragic circumstances (e.g. Anne Frank). However, not all images of these figures should be considered commemorative, as they are not solely defined by the hardships they faced. An image of a young, smiling Anne Frank is likely not commemorative. An image of the Anne Frank House, which has been turned into a museum to record Anne and her family's persecution, is likely commemorative.

IMPORTANT: Most of the time, images of oppressive figures or symbols (e.g. Nazi officers, fascist memorabilia) should not be considered commemorative, as they do not directly represent the suffering they caused.

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.



Questions on Historical Imagery

If you think this image is too disturbing to show annotators like yourself, please report it.

☐ Report Image

To learn more about this topic and image, use the resources below:

Search Wikipedia in a new tab
Search for the image in a new tab

[Required] What best describes the image?

○ Photograph

○ Artistic depiction (e.g. painting, cartoon)

○ Document (e.g. letter, map)

○ Chart or infographic (e.g. bar graph)

○ Symbol (e.g. logo, sign)

○ None of the above

[Required] Decide which of the following terms best describes what is in the image.

○ Person

○ Organisation

0 tasks completed. Study duration 6 out of 10 minutes. The study will finish after this task.