Heritage Weaver: Classifying, Searching, and Linking Museum Data with Multimodal AI

Kaspar Beelen[1], and Natasha Kitcher[2]

[1] School of Advanced Study, University of London, London, United Kingdom
[2] The National Archives, London, United Kingdom

Abstract

Heritage Weaver investigates the use of multimodal AI to link and explore museum data across collections. Through a series of experiments, from zero-shot learning and information retrieval to record linking, we demonstrate the value of fine-tuning multimodal models on digital heritage. The paper elaborates on various evaluation strategies, leveraging existing metadata or using expert annotations, to measure improvements in the model's "understanding" of often complex and messy historical materials.

Keywords: multimodal, digital heritage, data linking

1 Introduction

Increasingly, museums and other GLAM institutions are digitising their holdings, producing digital metadata and images (or transcriptions) of their objects. These efforts enable novel ways of exploring and searching GLAM collections. But while the amount of digital information has undoubtedly grown, the datafication of heritage has remained all too often locked within institutional silos, creating a landscape that is patchy and disconnected. How can we break out of these silos, build meaningful bridges between data islands, and explore connections between collections? These questions were central to the *Towards a National Collection* (TaNC) initiative, funded by *UK Research and Innovation*'s *Arts and Humanities Research Council*, in which this research took place as part of the *Congruence Engine* (CE) Project.[1] TaNC aimed to break "down the barriers that exist between the UK's outstanding cultural heritage"[2] with CE focusing specifically on heritage related to the industrial past.

Previous attempts at linking museum data often framed the challenge as a linked-open-data (LOD) problem, which required fitting (heterogeneous) data into preconceived ontologies and/or vocabularies to encode relations between collection records [4].[3] While LOD has its merits for harmonising data, Heritage Weaver pursues a different strategy. We developed methods that harness specialised multimodal AI to navigate and connect digital heritage. More specifically, this paper presents experiments that evaluate the impact of model fine-tuning on zero-shot classification, searching, and linking museum data *across collections*. These methods enable a flexible, bottom-up approach to exploring connections across institutional silos, using both image and text to forge links across otherwise isolated information containers.

Recently, multiple papers in computational humanities investigated the application of multimodality to heritage collections [1; 2; 14]. However, researchers often rely on "off-the-shelf"

Kaspar Beelen, and Natasha Kitcher. "Heritage Weaver: Classifying, Searching, and Linking Museum Data with Multimodal AI." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 261–277. https://doi.org/10.63744/txxUt12DJeT7.

[1] `https://www.sciencemuseumgroup.org.uk/projects/the-congruence-engine`
[2] `https://www.nationalcollection.org.uk/`
[3] See for example: `https://www.sciencemuseumgroup.org.uk/projects/heritage-connector`

models without adapting them to the particularities of the historical data. This paper, however, investigates what we can gain from adapting multimodal models to digital heritage, especially in the context of exploring GLAM data across collections. In short, this paper probes the following questions:

- Does model fine-tuning improve the classification, searching, and linking of museum data taken from different institutions?
- How can we evaluate this?

While not pretending to provide definite answers, in this article we will share practical recommendations that help both scholars and GLAM professionals employ multimodal AI in their research data or collections. In the following sections, we firstly explain the content of our data and the models. Secondly, we outline the design of our experiments. Finally, we will discuss the results from both a quantitative and qualitative perspective.[4]

2 Research Context

While originally focused on text, the digital and computational humanities have undergone subsequent visual [16] and multimodal turns [15]. The recent arrival of CLIP [11] and other vision-language models has led to an increased emphasis on multimodal (as opposed to monomodal) approaches. In [15] Smits and Wevers, the authors demonstrate the value of zero-shot classification in the context of heritage data. They demonstrate how CLIP provides a meaningful helping hand when labeling data, which they apply to magic lanterns and children's literature. Moreover, they point to the (historical) biases present in these models, which arise from the composition of the training data [3]. One problem with CLIP models, when used off-the-shelf, is that training data may not be aligned with research materials and applications. We therefore explore how fine-tuning multimodal models on museum collections might make them more reliable tools for exploring and processing digital heritage. Instead of using off-the-shelf models or relying on very large but closed models, we demonstrate that researchers and curators can gain a lot from adapting small or mid-sized models to their collections. Our results reproduce (on a smaller scale, but applied to digital heritage) the recent findings by [12]. They benchmark popular foundation models (including *gpt-4o*, *gemini 1.5*, *claude 3.5 sonnet*, and more) on various tasks, demonstrating that these large multimodal models are not competitive for specialised tasks, but perform respectably when looking for a more generic solution. While large models are dominating AI research, their findings suggest there is still value in fine-tuning smaller, open-source models, especially for tasks such as image classification. Especially for computational humanities and GLAM, where researchers and curators often rely on AI to analyse, navigate and explore specific (historical) datasets, we need specialist instead of generalist models.

In the context of Computational Humanities Research, a few recent papers explored the potential of multimodality cultural heritage. In [14], Smits and Kestemont demonstrate how CLIP achieve respectable accuracy when applied to nineteenth-century magic lanterns in a zero-shot setting. However, it was still outperformed by a vision model specifically fine-tuned for the task. Another important source for multimodal analysis are historical maps, which combine visual and textual features. [7] highlight the potential of multimodal search for historical maps, using text and visual inputs (or both). They also introduce a dataset for fine-tuning multimodal models containing more than 10,000 map-caption pairs. In "Reading maps at a distance", [10] explores the combined analysis of text labels and visual features on Ordnance Survey maps, to understand different types of "railspace" in nineteenth-century Britain. Other types of cultural heritage collections have also

[4] `Forcodeandupdatessee:https://github.com/congruence-engine/heritage-weaver`

been analyzed from a multimodal perspective. For example, [8] investigated the viability of zero-shot learning and search outside of the English language. They applied CLIP and SigLIP to a collection of historical photos derived from Ajapaik, a crowd-sourced photo archive that comprises images related to Estonia or its neighboring countries. A paper by [13] looked closely at opportunities of multimodality for book collections. They presented a proof-of-concept image-search tool to explore the pre-1900 collections of the National Library of Norway. They found that, for image retrieval and classification, SigLIP performed slightly better than CLIP or ViT. Lastly, [2] proposed using advanced multimodal large language models (LLMs) to build an open-ended, interpretable interface for exploring visual collections. Their approach demonstrated how such models can power innovative clustering and recommendation mechanisms while overcoming typical limitations associated with techniques that rely solely on visual embeddings.

This research also took inspiration from recent waves of "historical" language models (LM). Using the collection of British Library books, [5] released a variety of historical BERT models fine-tuned on specific temporal segments of the collection. MacBERTh can serve as another example, introduced in [9] as an LM pre-trained on historical English (1450-1950). While researchers have adapted and "historicised" LMs using transfer learning [6],[5] this paradigm seems less prevalent in multimodal computational humanities. By demonstrating the benefits of historical adaptation of multimodal AI, this research hopes to change this.

3 Data

In the Heritage Weaver project, we investigated two collections: the Science Museum Group (SMG) and National Museums Scotland (NMS). Our data selection was guided by the Congruence Engine's emphasis on industrial heritage, and we focused on records related to three main themes of the project: communications, energy, and textiles. We obtained all records related to these themes, based on the existing taxonomy terms in the catalogues, selected by domain experts in collaboration with the collection curators. In total, we collected 21,871 records: 20,889 from SMG and 982 from NMS. Metadata and images were supplied by the relevant partner institutions. After retrieval, we converted the data to a minimal, uniform format. We described each object with the following metadata fields:

- **record_id:** the original object identifier taken from the museum catalogue
- **name:** the name mentioned in the catalogue ("object_name" in the NMS metadata, and value for the "name" key in SMG).
- **img_url:** a link to where the image is hosted online.
- **description:** a free-text description of the object. In the original metadata, this is the "description" column in NMS. For the SMG collection, we concatenated the values under the "title" and "description" fields.

We stored this information in a vector database,[6] together with an embedding of text (name and description) and the image. The database forms the infrastructural backbone for our experiments below.

4 Models

Our collection contained images of historical and industrial artefacts, whose shape and functionality may be hard to decode for both the contemporary viewer and for current AI models. Models

[5] or sometimes pre-training from scratch

[6] Using the chromadb API in Python: `https://www.trychroma.com/`

such as CLIP often exhibit recency biases, failing to properly process or "understand" the content of heritage materials [15]. Because of the specificity of our data, we set out to fine-tune open-source multimodal models on these images and texts. Below, we assess if such adaptation might be beneficial to museums, especially in facilitating the work of curators and researchers who wish to classify, search and connect records across collections.

For all of our experiments, we used SigLIP (a variation on CLIP) as our base model [17].[7] CLIP (Contrastive Language–Image Pretraining) [11] learns joint text-image embeddings using a contrastive loss between paired image and text data. Sigmoid Language–Image Pretraining (SigLIP) replaces CLIP's softmax-based contrastive loss with a sigmoid loss, improving performance by enabling better handling of large-scale unnormalised similarities and reducing the need for negative samples.

Before training our models, we selected a random subset of the data for evaluation. Our training set contained 19,545 records, and the testing set 2172 records.[8] Each item in the training set consisted of an image and a text pair from the same record. The text was either the name or the description of the image. If the description was longer than the 64 tokens allowed by the SigLIP tokeniser, we split it into different segments of a maximum of 64 tokens. In this case, we would associate the image with each segment, increasing the number of training examples. After formatting the image-text pairs this way, our total data increased to 75,378 items, which we split into a training and evaluation set.

While we refer to the process of adaptation as "fine-tuning," the more accurate term is "continued pre-training," since we are not changing the task, but rather adapting the base model to new image-text data. We trained the model for five iterations using Binary Cross-Entropy with Sigmoid activation as a loss function.[9] We left all other hyperparameters untouched for this round of experiments, but we hope to investigate training options in more detail in later research. We saved two checkpoints: *ft-best*, the checkpoint that obtained the lowest loss scores on the evaluation data, and *ft-last*, the last, i.e. fifth checkpoint. The original SigLIP model is referred to as *base*. Training took around 3 hours on a single L4 GPU.

We included *ft-last* in our analysis to understand the risks (and impact) of overfitting. When we want AI tools to act as "museum specialists", does training longer harm or improve the performance on downstream tasks?

To better contextualise the value of fine-tuning smaller models, we compared some results to *gpt-4o* and *gpt-4o-mini*.[10] When presenting this work on earlier occasions, we were often asked why we focused on SigLIP, and did not "simply" use ChatGPT (or similarly powerful closed models with multimodal capabilities).

Firstly, our investigation aimed to give pragmatic guidance to museums (and other institutions) on how AI can assist them with preserving and curating large heritage collections. This implies that we needed to take into account the resources and skills available to these organisations, as well as legal constraints. Given that museums are unlikely to possess large compute clusters or budgets to integrate commercial LLMs, we investigated whether improving cheaper and smaller models is a more fruitful avenue. These models can be fine-tuned and deployed with minimal cost and empower institutions to retain control of their data and their tools.[11] When relying on external

[7] See: `https://huggingface.co/google/siglip-base-patch16-224`

[8] Classification and retrieval scores reports below were based on records in the test set. Given that a record can contain multiple images and text fragments, we split the data based on record identifiers to ensure none of the test examples seeped into the training set.

[9] BCEWithLogitsLoss in PyTorch: `https://docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`

[10] See classification experiments

[11] Something which is not always guaranteed when partnering up with commercial companies relying on selling services built on closed models.

APIs, LLMs are convenient to implement and excellent for prototyping. However, when working with large collections and building tools for production, they saddle users with high costs.[12] Secondly, Heritage Weaver aimed to create tools that work well for specific collections and historical data. Do we need models that contain billions of parameters to improve the exploration of such heritage collections? Thirdly, for reasons of privacy, legal and environmental concerns, handing over collections to large corporations might carry significant risks that we were not willing to take.

5 Experiment Design

5.1 Background

This paper presents multiple experiments to evaluate the potential benefits of using fine-tuned multimodal AI to explore museum data and digital heritage across collections. In Heritage Weaver, we initially focused on the representations or embeddings produced by multimodal models, and established to what extent these can serve as instruments for finding meaningful links (i.e. "weaving connections") within and across collections. Using representations extracted from SigLIP, we can compute similarity based on textual and visual data. But how and when do these similarities translate to something meaningful? What models and modalities provide more reliable embeddings and a better understanding of connections in our heritage materials?

Figure 1 visualises part of the problem. It compares all records in SMG (y-axis) to NMS (x-axis) based on their embeddings. Each point contains a cosine similarity, and highly similar record pairs are marked by black dots,[13] the rest by white dots. We repeated this operation for different modalities: text-to-text (left), image-to-image (centre) and image-to-text (right). What becomes apparent is the divergence between these plots: which records are perceived as similar by the model depends on the modality through which they are compared. Moreover, it also depends on the model. Figure 1 highlights *differences* between *ft_best* and *base* for each modality. Black dots indicate pairs on which the models disagree in their similarity assessment. We observe that different regions light up; in this case, the models seem to disagree, especially in their cross-modal comparison of records (image-to-text plot at the right).

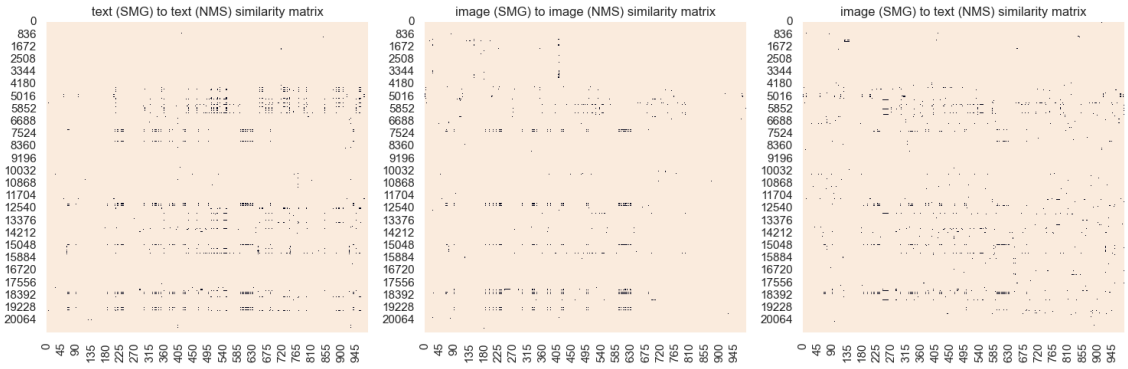

Figure 1: Visualising similar items across collections for different modalities (using *ft_best* model) Black dots indicate highly similar records.

To establish when multimodal AI "makes sense" we designed a set of targeted experiments. All experiments comprise a comparison between model outputs and a "ground truth" derived from catalogue metadata or expert annotations. In this section we explain the design of our experiments

[12] For more details, please consult a recent report written by Kaspar Beelen, published by Jisc, on "Small Language Models for libraries and computational humanities." `https://repository.jisc.ac.uk/10293/1/small-language-models-for-libraries-and-computational-humanities-18-sept-2025.pdf`

[13] These record pairs have a similarity scores that range in the 97.5th percentile.

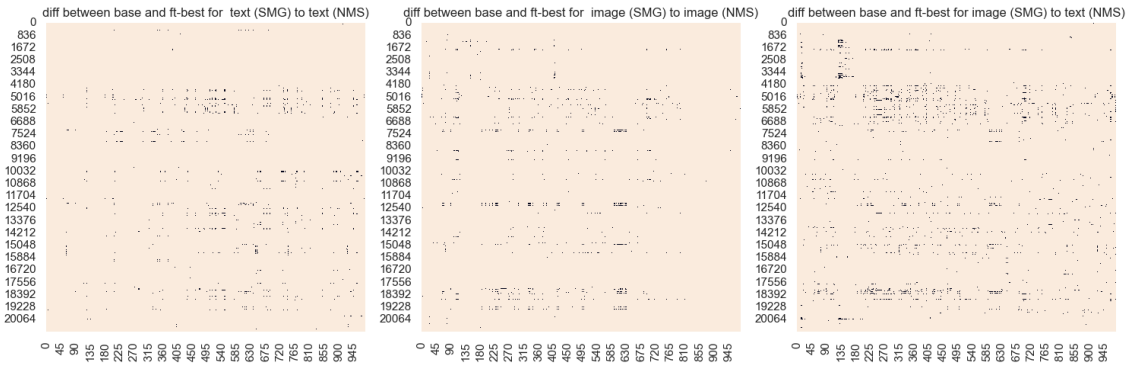

Figure 2: Visualising how models differ in their similarity assessment (*ft_best* vs. *base*). Black dots indicate disagreement.

on classification, retrieval and linking. The evaluation of these experiments is described in the next section.

5.2 Classification

We designed an automated experiment to evaluate the zero-shot classification abilities of the original and fine-tuned models. The goal of this experiment was to understand how well different models captured what the image represented by assigning meaningful labels. Zero-shot classification provides an efficient method for exploring connections between collections based on flexible label sets designed by the curator or the researcher. For it to work, however, the model needed to successful recognise what is in a given image. We used the original name of an object record (or the name's root noun[14]) as the ground truth and randomly add *n* other labels as candidates. We presented the model with a set of labels and asked it to guess the original name (i.e. the name derived from the original record). In some cases, the automatically generated ground truth may be wrong or uninformative. But as emphasised earlier, we are primarily interested in *relative* improvements to the base model.[15]

This experiment simulates, to some extent, whether a model "understands" an image. By repeatedly asking it to select the correct label from a set of candidates, we can measure if fine-tuning improves the overall alignment between visual and textual concepts. We should again emphasise that images used for evaluation were not part of the training data. Table 2 shows a sample of the data used for zero-shot classification using nouns as labels. It shows the original names and extracted root nouns. The candidate labels include the original noun and randomly selected nouns from other records. The information in *candidate_nouns* is presented to the model for classification. In the experiment report below, we apply zero-shot classification to names and nouns, and increase the difficulty of the task by expanding the number of candidates from five to ten. We also apply these tasks to subsets of the data, indicated by the *data* column.

5.3 Search

The second automated experiment simulated the influence of model fine-tuning on search and information retrieval, testing performance within and across modalities. For each record r in the evaluation data, we used the image r_{img} as a query, and retrieved n most similar entries in the vector

[14] We use SpaCy to determine the root noun, selected the token with dependency tag "ROOT" and part-of-speech tag "NOUN" or "PROPN"

[15] Also, the likelihood that one of the randomly selected labels is actually a better candidate than the one distilled from record metadata remains small.

name	noun	candidate_nouns
razor blade sharpener	sharpener	[sharpener, block, switch, unit, button]
specimen	specimen	[specimen, uniselector, telephone, lamp, gener...
hand printing block (trademark)	block	[block, apparatus, machine, mat, bit]
Dynamo, Gramme	dynamo	[dynamo, specimen, cooker, lamp, pirns]
pay check	check	[check, horseshoe, cable, representation, cup]

Table 1: Example of data used for evaluating the zero-shot capabilities of original and fine-tuned models. In this case, we test if the model can guess the root noun present in the record name.

database, based on either their image (image-to-image retrieval) or text (image-to-text retrieval) embeddings.

To evaluate if the retrieved candidates were relevant to the query, we embed the name of the query r_{name} (which is *not* included in the search) and those of the retrieved records. We computed the cosine similarity between r_{name} and $e_{i_{name}}$ for each retrieved item e_i.[16] Figure 3 shows an example of this setup: the query is an image with the name "battery". We searched the database for similar images and computed the similarity between the vector representations of the names (i.e. sim("battery", "accumulator") and sim("battery", "lamp battery")) to score the relevance of the retrieved candidates. By comparing the names, we have some way of automatically assessing if the retrieved images are relevant to the query. Because Heritage Weaver is specifically concerned with building bridges across collections, we repeated this experiment, but this time we focused on *cross-collection search*: i.e. if the image is produced by SMG, we assessed the extent to which we could retrieve relevant records from NMS (and vice versa). The results for collection-agnostic (top) and cross-collection (bottom) search are reported in Table 5.

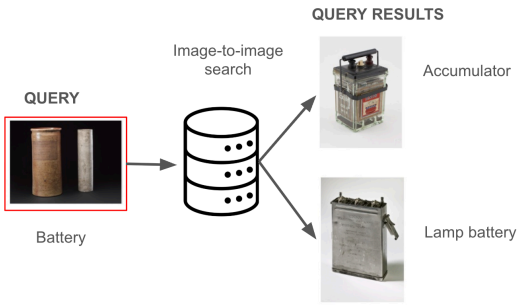

Figure 3: Example of image-to-image search.

To gauge improvements in multimodal search, we experimented with text-to-image search, in which object descriptions serve as queries for image retrieval. Put differently: Can we find relevant images based solely on textual descriptions? Also, here, we used the names of the objects to evaluate the quality of the found records. For cross-modal search, we also evaluate if the original record associated with the description was retrieved.

5.4 Linking

The automated experiments elucidate how model fine-tuning contributes to downstream tasks such as classification and information retrieval. To assess if GLAM professionals might benefit from tai-

[16] For embedding names, we used the all-MiniLM-L6-v2 model with the Sentence Transformers library.

lored multimodal models for bridging information silos, we set up an annotation experiment that mimics linking and exploring heritage data across collections. We evaluated these experiments both quantitatively and qualitatively. We recruited historians and curators to annotate, including curators from National Museum Scotland, researchers from the Congruence Engine project, a museum professional from the Discovery Museum in Newcastle, and a historian of Science and Technology. The goal was to have a diverse group of researchers and museum professionals, with different levels of expertise, familiarity with the collection, but united by a common understanding of industrial heritage objects.

The annotation session took place online. We used a LabelStudio[17] instance hosted on HuggingFace Spaces[18] as the annotation interface. We introduced the project, tasks and explained the labels, to make sure everyone was equally informed. Each annotator was provided a different subset for each task. Afterwards, the annotators were asked to share their thoughts and opinions. We used these as a form of *qualitative* evaluation, but also let these observations inform decisions for *quantitative* evaluation.

Figure 4: Example of link annotation.

To evaluate the impact of model fine-tuning, we followed a specific sampling strategy. We sampled record pairs where each item is from a different collection (i.e. each pair links SMG to NMS). Figure 4 shows an example of a record pair that annotators were asked to label. Because of the wealth of possible combinations, randomly sampling record pairs would yield mostly unrelated items. Therefore, we focused on object pairs whose embeddings were highly similar.[19] Then, we zoomed in on occurrences where models diverged in their assessments, e.g. record pairs that a fine-tuned model recognises as similar while an off-the-shelf SigLIP model did not (or vice versa). Table 2 provides an overview of the procedure used for sampling record pairs. The first row covers images where all models agreed on their image similarity (i.e. the similarity was above a set threshold for all models). The texts *might* be similar, but this was not a criterion for selection (therefore marked by 0/1). The second row points to pairs of images which the *base* model considers similar, but the fine-tuned model disagrees.

Annotators were tasked with labelling these sampled record pairs. We should stress that we did

[17] `https://labelstud.io/`
[18] `https://huggingface.co/spaces`
[19] In the 99th percentile bracket

image	text	base	ft_best	ft_last
1	0/1	1	1	1
1	0/1	1	0	0
1	0/1	0	1	1
1	0/1	0	1	0
0/1	1	1	1	1
0/1	1	1	0	0
0/1	1	0	1	1
0/1	1	0	1	0

Table 2: Overview of sampling strategies for record pairs.

not disclose to the annotators the sampling method, meaning that they didn't know which model or modality produced the examples they were labelling. They, therefore, couldn't be biased by our research question. Each annotator was presented with 100 pairs to annotate, choosing to assign each pair as either "same object", "similar object", "same category", and "unrelated".

- **Same object**: A rare annotation used for when the object is identical. For example, two 700 series telephones, two gas lamps, or two Jacquard looms.
- **Similar object**: Objects which are highly similar but not the same, for example, two telephones, two light bulbs, two looms.
- **Same category**: Denotes when two objects have similar uses, perhaps they would be assigned to the same category within a CMS, but they are different 'things'. This is probably the loosest and most subjective form of linkage.
- **Unrelated**: These objects are not the same, similar, or used for similar purposes. They have no meaningful connection.

We did not compute inter-annotator agreement, as our main goal was *not* to create a gold standard, but to establish which aspects (i.e. modalities and models) yielded results useful to these particular experts.[20]

6 Results

6.1 Classification

To evaluate the zero-shot classification capabilities, we measured if the SigLIP models can accurately guess the correct label for an image from a set of candidates. The target or correct label is the name associated with the image in the metadata, to which we randomly added other names from the catalogue. We varied the experiments slightly to better understand the robustness of our findings:

- we changed the number of candidates
- we used the original name of the object or the root noun (both as target label and other candidates)

[20] Also, we gave annotators different pairs to label, maximizing the number of different observation, but making computing agreement impossible.

data	model	target	n_cand	accuracy	precision	f1
all	base	noun	5	0.721	0.829	0.751
all	ft-last	noun	5	0.876	0.910	0.884
all	ft-best	noun	5	**0.892**	**0.920**	**0.898**
all	base	noun	10	0.625	0.764	0.658
all	ft-last	noun	10	0.805	**0.849**	0.816
all	ft-best	noun	10	**0.814**	0.848	**0.821**
all	base	name	5	0.793	0.854	0.809
all	ft-last	name	5	0.935	**0.955**	0.941
all	ft-best	name	5	**0.939**	0.953	**0.941**
all	base	name	10	0.716	0.794	0.735
all	ft-last	name	10	0.902	**0.934**	0.911
all	ft-best	name	10	**0.907**	0.927	**0.912**
filter	base	noun	5	0.683	0.721	0.679
filter	ft-last	noun	5	0.849	0.871	0.852
filter	ft-best	noun	5	**0.850**	**0.872**	**0.853**
filter	base	name	5	0.779	0.781	0.769
filter	ft-last	name	5	0.927	0.938	0.928
filter	ft-best	name	5	**0.932**	**0.943**	**0.934**
sample	base	noun	10	0.635	0.707	0.656
sample	ft-last	noun	10	0.805	**0.844**	0.811
sample	ft-best	noun	10	**0.825**	0.836	**0.825**
sample	chatgpt-4o-mini	noun	10	0.660	0.729	0.672
sample	chatgpt-4o	noun	10	0.770	0.801	0.771

Table 3: Zero-shot classification results.

In Table 5.4 the *target* column indicates whether we classified names or nouns; *n_candidates* points to the number of candidates (either 5 or 10). Other columns show the overall accuracy and the weighted precision and f1 scores for each experiment. We first applied zero-shot classification to the complete test set, which is indicated by the *all* value in the *data* column.

Across all experiments, the fine-tuned models substantially improved upon the off-the-shelf base model. Overall, we observed an increase in accuracy scores between 15% to 20%. When expanding the candidates, these numbers slightly drop, as the task gets more difficult, but the gap between the original and fine-tuned models persists. For example, SigLIP *base* obtains an accuracy of 0.721 for zero-shot classification with five candidates. When asked to pick the right label from a set of ten, the accuracy declines to 0.625. The best fine-tuned models achieve a score of 0.892, respectively 0.814. Names seem somewhat easier to classify compared to the more abstract and generic nouns: the scores climb and the fine-tuned models outperform SigLIP *base* by a margin of more than 10%.

To understand divergence between original and fine-tuned models, we calculated the difference in accuracy for each object type (in this case, the head noun in the object name).[21] We compared the fine-tuned models to SigLIP *base* disaggregating the scores by noun; scores above zero indi-

[21] These numbers are based on the first experiment, zero-shot classification for all the data with five different labels

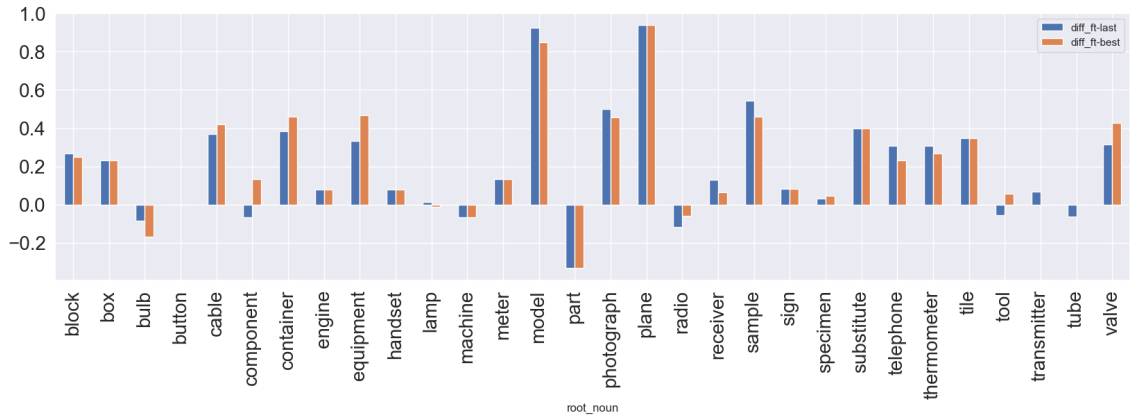

Figure 5: Difference in accuracy compared to *base* SigLIP model.

cate better performance of *ft-last* or *ft-best* for this noun, negative numbers point to objects where the fine-tuning might harm performance. The results for the most common objects are reported in 3. Inspecting these results more closely shows that part of the performance difference stems from the naming conventions in museum catalogues. For example, fine-tuned models excel for the categories “specimen”, “model”, or “sample”. In this sense, one could argue that adapting models does not necessarily improve their ability to understand what is represented in these (historical) images, but tailors them to curatorial practices and conventions. However, we reran these experiments, leaving out the records where the name does not describe the object but characterises the type of representation (“model”, “specimen”, “sample”, etc.). These experiments are marked by *filter* in the *data* column. While zooming in on the “objective” names changes the accuracy somewhat, it does not alter the core of our findings: the gaps in performance persist and the adapted models score better than the base models, with margins of around 15% or more. Also in this scenario, *ft-best* leaves all others behind.

Lastly, we compared the performance of our historical, adapted models to the popular and powerful ChatGPT. To keep the costs down, we restricted this experiment to a subsample of 200 records. We instructed ChatGPT as follows:

- System Prompt: You are a helpful assistant that classifies images into specific categories.
- User Prompt: Classify this image into one of these labels: {{labels}} Respond only with the label.

Lastly, we ensured that all responses could be mapped to at least one of the classes, with the only exception when ChatGPT replied that “none of the labels applied”. This was considered to be an incorrect answer, or failure. Interestingly, even an immensely large model as *gpt-4o*—which admittedly hasn’t been trained on our museum data, at least as far as we know—performs considerably worse than our small (but fine-tuned) models. While *gpt-4o* does better than SigLIP *base*, its accuracy is still 5% below *ft-best*. Moreover, ChatGPT took around 5 minutes to process 200 records, while SigLIP completed the task in about 1 minute on a MacBook Pro *without* GPU acceleration. It suffices to say that, based on these findings, museums and other heritage institutions, better prioritise small and customizable AI instead of relying on expensive, commercial LLMs.

6.2 Search

The search experiments simulate various content retrieval scenarios: how does fine-tuning change the ranking of records based on their intra-modal (i.e. image-to-image) or cross-model (text-to-

model	n	relevance	sd.	n	relevance	sd.	n	relevance	sd.
base	3	1.959	0.928	10	4.678	3.621	20	8.019	7.379
ft-last	3	1.998	0.932	10	4.868	3.723	20	8.510	7.694
ft-best	3	**2.005**	0.925	10	**4.882**	3.729	20	**8.510**	7.682
base	3	0.024	0.191	10	0.084	0.581	20	0.138	0.921
ft-last	3	0.033	0.225	10	0.109	0.696	20	0.180	1.116
ft-best	3	**0.035**	0.220	10	**0.111**	0.659	20	**0.189**	1.116

Table 4: Average similarity (for image-to-image search) between names of the query record and retrieved records. The first set of results is collection agnostic; the second set shows results for cross-collection search.

model	n	experiment	relevance	found	experiment	relevance	found
base	10	text-image	2.631	0.179	image-text	2.234	**0.122**
ft-last	10	text-image	**4.639**	**0.307**	image-text	**3.876**	0.109
ft-best	10	text-image	4.225	0.288	image-text	3.327	0.054

Table 5: Average similarity between the name of the query record and retrieved records for image-to-text and text-to-image search.

image) similarity? Table 5 reports scores for image-to-image search. The first set of results is based on collection agnostic search (we search the whole database), the second set shows performance for cross-collection document retrieval (i.e. we use records from SMG to find information in NMS and vice versa). Query records were not part of the training data. Therefore, for each image in the test set, we evaluated the relevance of the retrieved records by comparing the average similarity between the name of the query image and the names of the retrieved images. If the similarity between the names was above 0.95 we coded this pair as one[22], otherwise as zero. The scores in Table 5 should be read as follows: a *relevance* score of 2.005 for n=3, indicates that, on average, two out of three retrieved images have record names whose cosine similarity is above 0.95 (to the query image name). the *sd.* column shows the standard deviation of these relevance scores across all searches.

The findings in Table 5 align with those of zero-shot classification. Regardless of the number of retrieved records, the fine-tuned models return more relevant images. However, we should note that the difference can remain small, i.e. looking only at the top three, the gap between *base* and *ft-best* is only 0.046. This difference widens to almost 0.5 records when we increase n to twenty. When searching across collections, the numbers drop, but this was to be expected: there simply might not be enough relevant entries in the other collection. However, in all cases, we observe a slight improvement attributable to fine-tuning.

Table 4 tests the cross-modal search capabilities: we use a description to search for images (or vice versa, use images to search descriptions). Similar to 5, the *relevance* column compares the name of the query record to the retrieved items. The *found* columns indicate to what extent we could find the original record across all searches, e.g. whether we could fetch the correct image based on its descriptions. The values in this column range between zero (never) and one (always).

[22] This value is somewhat arbitrary, from the previous experiments with sentence embeddings we found that scores around or above 0.95 likely indicate similarity in meaning.

The former indicates that we could not retrieve the original records in any of the searches; the latter that all cross-modal queries managed to return the original record (i.e. the original record was found among the top n retrieved items).

The scores for text-to-image search are generally higher than image-to-text search. Interestingly, the performance gap between base and fine-tuned models tends to widen when looking at the *relevance* scores. In these cases, the model trained for more epochs achieved better results. However, the numbers reported in the *found* column deviate somewhat from the patterns observed in previous experiments. *ft-last* managed to retrieve the original image based on the description in around 30% of the searches (meaning that the original record appeared in the top ten results). This drops to around 11% when searching the other way. In this case, we have evidence that fine-tuning might hurt performance, as both *ft-last* and *ft-best* fare worse than *base*. At the moment, we can only speculate about reasons, but this will be part of future research.

6.3 Linking

So far, we have relied on metadata as proxies of ground truth. However, to establish how experts read and interpret the results returned by our models, we designed an annotation experiment. The following experiment established where experts might actually experience benefits from using fine-tuned multimodal AI for connecting records across museum collections. With this linking experiment, we gauged how similarity, measured in terms of model representations (or embeddings), translates to meaningful connections between historical objects (as perceived by experts). As discussed above, each of the sampled pairs connected objects between SMG and NMS, and scored high with respect to their cosine similarity.

6.3.1 Quantitative Evaluation

Overall, we collected around 551 annotations, for which Figure 5 shows the distribution. The majority category (46%) is “unrelated”, which points to the fact that even high vector similarity does not equate to a meaningful connection. However, in most cases (54%), these pairs do exhibit a valuable relationship between objects, though the strength of this relationship varies, from “same object” (1%) over “similar object” (30%), to “same category” (23%).

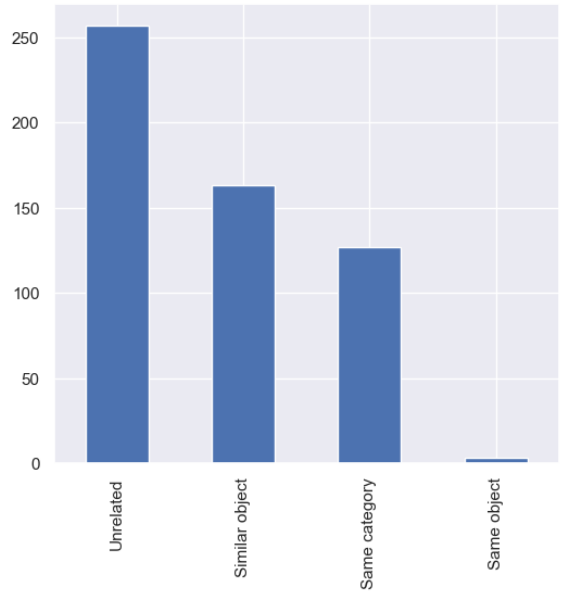

Figure 6: Label distribution of link annotations

	coef	std err	z	P>|z|	[0.025	0.975]
base	-0.1960	0.148	-1.329	0.184	-0.485	0.093
ft-last	**0.3692**	0.122	3.024	**0.002**	0.130	0.608
base	-0.2392	0.148	-1.613	0.107	-0.530	0.051
ft-best	**0.4543**	0.124	3.654	**0.000**	0.211	0.698

Table 6: Logistic regression that predict the effect of the model selection on link annotation.

	coef	std err	z	P>|z|	[0.025	0.975]
img_base	-0.1189	0.207	-0.576	0.565	-0.524	0.286
img_ft-best	**0.3919**	0.204	1.917	**0.055**	-0.009	0.792
txt_base	-0.1473	0.209	-0.706	0.480	-0.556	0.262
txt_ft-best	0.1870	0.223	0.837	0.403	-0.251	0.625
img_base	0.0006	0.197	0.003	0.998	-0.386	0.388
img_ft-last	**0.3630**	0.184	1.974	**0.048**	0.003	0.723
txt_base	-0.1583	0.212	-0.745	0.456	-0.575	0.258
txt_ft-last	0.0606	0.202	0.301	0.764	-0.334	0.456

Table 7: Logistic Regression that estimates the effect of modality and model on link annotation.

The principal question, however, was to establish if fine-tuning models provided *better* candidates for record linking across collections. We analysed the annotations as a function of the models and modalities (i.e. as a function of the sampling strategy). We binarised the dependent variable as 0 ("unrelated") or 1 (for the other categories). Our independent variables are the models, *base*, *ft-last* and *ft-best*. Each observation was then dummy-coded: e.g. [1,0] would indicate that a record pair was similar for *base* but not the *ft-last*. We then performed a logistic regression and reported the coefficients with their standard errors. We tested the performance of each of the fine-tuned models against the base model in separate regressions (*base* vs *ft-last* and *base* vs *ft-best*). In both regressions (see Table 6), the fine-tuned models appear to contribute to the likelihood of a positive link. We observe the largest and most significant effect for *ft-best* with a z-score of 3.654 and a p-value smaller than 0.001.

To get a more precise estimate of how different modalities contribute to linking, we broke down the predictors to *modality_model* variables, e.g. *image_base* indicates that image embeddings for the *base* SigLIP model obtained a high cosine similarity; *text_ft-best* in turn means that the text embeddings for the *ft-best* model were very similar, etc. We then reran the regression using the expanded set of independent variables. While the effects remained weak and only marginally significant, we observed that improvements in linking came primarily through visual modality. Again, this finding is consistent across fine-tuned models, with the largest effect of 0.3919 observed for *ft-best* (p-value > 0.1), followed by *img_ft-last* (with a coefficient of 0.3630 and p-value > 0.05). These results not only underline the value of fine-tuning models on heritage data, but they also point towards the importance of multimodality for exploring and linking museum collections. The improvements of fine-tuning seem to be located at the visual level, rather than the textual.

6.3.2 Qualitative Evaluation: Experts' Reflections on Linking

The preceding evaluation scrutinised the annotations from a statistical angle. However, we were also interested in how the experts experienced the annotation session; how they approached the tasks and what made it meaningful for them. After the session, the participants were invited to note down their reflections, which we collected and analysed to inform both the quantitative evaluation and the future development of multimodal tools linking records across collections.

Curators who knew their objects were keen to have more specific ways to link objects, i.e. more refined categories. Curators less familiar with the data reflected that this level of specificity would be overwhelming to those who lacked relevant specialist knowledge. This curator, alongside the other participants who were all researchers and academics, felt that more categories would be confusing as they lacked the necessary domain expertise. They felt they were able to infer with reasonable confidence what linked and did not link, but could not go into more detail. This informed our decision to binarise the labels for evaluating the link annotations. All participants reflected they were more inclined to look at the object images than the text when categorising the object pairs, which may partly explain why image similarity proved a good predictor linking record pairs (see Table 6). Participants noted that resemblances picked up by the models only contribute meaningfully to a curator's or historian's work when a human partakes in the pipeline. While helpful, multimodality is meaningless until a human has corroborated the results.

The "human" and "expert" aspects of the linking experiment revealed that participants perceived different priorities and pursued varying strategies. These must be taken into account when developing tools for labelling and developing record linkage for future applications in the GLAM context. Curators, who were inspecting their own collections, required more fine-grained, precise functionalities. Their interaction with objects is a more precise and objective endeavour compared to how researchers and historians explore these collections. Researchers were more likely to engage in the pursuit of exploration, and in the process made more creative associations between objects that are particular to their interests. For example, they'd link a series of telephones because they appear to have similar components, which enables a historian to investigate material culture in relation to the history of technology.

7 Conclusion

This paper argued for the value of "specialist" models when analysing digital heritage collections. We demonstrated the value of fine-tuning multimodal models on text and images derived from museum data, focusing on records pertaining to the industrial past. We have shown how adapting SigLIP improved its overall capabilities for zero-shot classification and information retrieval. When classifying images, it even outperformed much larger models such as *gpt-4o*. Fine-tuning seems especially powerful to improve cross-modal capabilities of these models (image-to-text and vice versa). Moreover, this paper explored a wide range of evaluation strategies: from using metadata as proxy-labels to expert annotations. We found that experts preferred the fine-tuned models for linking (even though they were not aware which models produced which results). Interestingly, experts relied to a large extent on visual cues when annotating the data, even though textual information was also crucial for many examples. This observation also emphasises the importance of multimodal approaches.

While LLMs are dominating public and scientific discourse, this paper argues that museums and other GLAM institutions have other options at their disposal. Fine-tuning a small or mid-sized open-source model might be a more effective and accurate strategy for making collections more accessible and navigable.

References

[1] Arnold, Taylor and Tilton, Lauren. “Automated Image Color Mapping for a Historic Photographic Collection”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 37–47.

[2] Arnold, Taylor and Tilton, Lauren. “Explainable Search and Discovery of Visual Cultural Heritage Collections with Multimodal Large Language Models”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 559–574.

[3] Beelen, Kaspar, Lawrence, Jon, Wilson, Daniel CS, and Beavan, David. “Bias and representativeness in digitized newspaper collections: Introducing the environmental scan”. In: *Digital Scholarship in the Humanities 38*, no. 1 (2023), pp. 1–22.

[4] Dutia, Kalyan and Stack, John. “Heritage connector: A machine learning framework for building linked open data from museum collections”. In: *Applied AI Letters 2*, no. 2 (2021), e23.

[5] Hosseini, Kasra, Beelen, Kaspar, Colavizza, Giovanni, and Ardanuy, Mariona Coll. “Neural language models for nineteenth-century english”. In: *arXiv preprint arXiv:2105.11321* (2021).

[6] Howard, Jeremy and Ruder, Sebastian. “Universal language model fine-tuning for text classification”. In: *arXiv preprint arXiv:1801.06146* (2018).

[7] Mahowald, Jamie and Lee, Benjamin Charles Germain. “Integrating Visual and Textual Inputs for Searching Large-Scale Map Collections with CLIP”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 528–547.

[8] Maksimova, Erika, Meimer, Mari-Anna, Piirsalu, Mari, and Järv, Priit. “Viability of Zero-shot Classification and Search of Historical Photos”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. CEUR Workshop Proceedings. 2024, pp. 1242–1258.

[9] Manjavacas, Enrique and Fonteyn, Lauren. “Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950)”. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. 2021, pp. 23–36.

[10] McDonough, Katherine, Beelen, Kaspar, Wilson, Daniel CS, and Wood, Rosie. “Reading maps at a distance: Texts on maps as new historical data”. In: *Imago Mundi 76*, no. 2 (2024), pp. 296–307.

[11] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[12] Ramachandran, Rahul, Garjani, Ali, Bachmann, Roman, Atanov, Andrei, Kar, Oğuzhan Fatih, and Zamir, Amir. “How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks”. In: *arXiv preprint arXiv:2507.01955* (2025).

[13] Roald, Marie, Birkenes, Magnus Breder, and Johnsen, Lars Gunnarsønn Bagøien. “Visual Navigation of Digital Libraries: Retrieval and Classification of Images in the National Library of Norway’s Digitised Book Collection”. In: *arXiv preprint arXiv:2410.14969* (2024).

[14] Smits, Thomas and Kestemont, Mike. “Towards Multimodal Computational Humanities. Using CLIP to Analyze Late-Nineteenth Century Magic Lantern Slides.” In: *CHR*. 2021, pp. 149–158.

[15] Smits, Thomas and Wevers, Melvin. “A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections”. In: *Digital Scholarship in the Humanities 38*, no. 3 (2023), pp. 1267–1280.

[16] Wevers, Melvin and Smits, Thomas. “The visual digital turn: Using neural networks to study historical images”. In: *Digital Scholarship in the Humanities 35*, no. 1 (2020), pp. 194–207.

[17] Zhai, Xiaohua, Mustafa, Basil, Kolesnikov, Alexander, and Beyer, Lucas. “Sigmoid loss for language image pre-training”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 11975–11986.