Understanding Museum Exhibits using Vision-Language Reasoning Supplementary Material

Ada-Astrid Balauca[1,*,✉] Sanjana Garai[1,3,*] Stefan Balauca[1] Rasesh Udayakumar Shetty[3]
Naitik Agrawal[3] Dhwanil Subhashbhai Shah[3] Yuqian Fu[1] Xi Wang[1,2]
Kristina Toutanova[1,4] Danda Pani Paudel[1] Luc Van Gool[1]

[1]INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria [2]ETH Zürich, Switzerland
[3]Indian Institute of Technology, Varanasi (IIT BHU) [4]Google DeepMind

A. Index

Table 1. The index showing the additional information, technical details and results.

B. Data

This section provides comprehensive details about the dataset used in the task. It includes information on the raw dataset, an example of an instance, and the data format. Additionally, it outlines a category-wise list of questions, data splits, and a detailed category analysis, offering insights into the structure and distribution of the data.

B.1. Data format

All this curated information was stored in the form of json files in a dictionary format. With the object_id being the key and the information in the respective value.

B.2. Example of instance

A detailed example from the dataset, showcasing the structure of an individual data point to clarify how the data is organized and used in the task is looking as follows:

Question	["Who is the artist of the object?", "What materials is the object made of?"]
Answer	[["Leonardo Da Vinci"], ["wood","iron"]]
Image	["object1_1", "object1_2", "object1_3"]

B.3. Data splits

This section details the dataset splits, including multiple training datasets designed to analyze the impact of varying data sizes. It also covers the validation split and multiple testing splits, enabling more efficient evaluation and comparison by reducing time requirements. The 42M train set is the original training set that we were able to collect, still due to time and other resources constraints we choose to fine-tune up to the 20M instances dataset.

B.4. Dataset details

We provide an overview of the dataset origin Tab. 3, including its composition, sources, and initial structure before processing. It highlights the foundational data used to create the final dataset for the task. We also show the amount of objects, images and attributes available from each museum, highlighting the attributes used for fine-tuning (Trainable attributes). The raw dataset will also be made publicly available along the curated dataset and it will also include the attributes not used for fine-tuning (Non-trainable attributes).

*Equal Contribution ✉ astrid.mokanu@gmail.com

Dataset	Objects	Q-A pairs
1mn_train	1M	3M
10mn_train	10M	31M
20mn_train	20M	61M
42mn_train	42M	123M
val	2M	4M
test	6M	18M
tiny_test	10K	30K
small_test	100K	3M
multilingual	15M	45M

Table 2. Description the dataset splits, including multiple training sets, a validation set, and several test sets. The splits are designed to facilitate analysis of performance under different training scenarios and streamline evaluation across various testing conditions.

The links to the curated and raw datasets can be found here: MUSEUM-65 .

B.5. Datasest Curation Process

The dataset curation was done in 5 major steps:
1. Museum selection:

- The dataset comprises 65 million data points, with 95% sourced from three major cultural aggregators: Digital Public Library of America - DPLA (24M), Europeana (20M), and the Smithsonian Institution (3.5M).
- These aggregators provide access to extensive digitized collections from major museums across Europe and America and offer structured data through platform-specific APIs.
- DPLA and Smithsonian provide metadata in English, whereas Europeana includes metadata in English as well as several European languages such as French, Spanish, and German.
- To ensure broader diversity in geography, culture, variety, and language, we curated the remaining 5% of the dataset from 12 additional major museums spanning multiple continents. These additional museums were selected based on their global prominence and the richness of their collections, with data acquired through a custom scraping pipeline.
- Depending on the museum's web infrastructure, data was collected either using official APIs or through HTML parsing tools such as BeautifulSoup.
- In cases where museums provided multiple images per object, we collected the URLs of all available views to preserve multi-angle visual representations.
- All this information was stored in the form of json files in a dictionary format. With the object id being the key and the information in the respective value.

2. Data Cleaning: Each museum's data was curated by a single domain expert to ensure consistency. Curation involved minimal edits: removing redundant attributes (inventory numbers, bibliographic info); extraneous symbols and numbers. Given high quality of museum data, focus was on consistent formatting rather than content rewriting.
3. Attribute-Value Structuring: While some museums provided such structured data, others required parsing complete strings, with experts identifying logical separators and attribute boundaries through example-driven consensus.
4. Question Crafting:

- To structure the attribute–value data for the visual question answering (VQA) task, we aligned our approach with natural human curiosity—formulating questions and expecting concise answers—toward our goal of real-time deployment in interactive museum environments.
- Each museum's data was processed independently due to differing metadata formats and schema structures.
- Experts manually authored natural language questions for a total of 63 unique attributes, with the corresponding attribute value serving as the ground-truth answer.
- Human synthesis ensured that even when the same attribute appeared across different museums, the phrasing of the questions varied to maintain linguistic diversity. Example: Both the questions *"Which primary material is the object made of?"* and *"What is the material used in the object?"* are related to the attribute "material"
- After synthesis, a centralized review process was conducted to remove redundancy, normalize structure where needed, and ensure phrasing diversity.
- We computed an average intra-category question similarity of 75%, indicating a desirable balance between consistency and variation across museums and attributes.

5. Final Assembly: We download all the images from the collected image-urls. For each object, we now have a list of images and a set of question-answer pairs, omitting the answers for which the value is not known. Finally, for each museum we create 3 columns - image (having the list of images from different viewing angles), question (having the list of all questions), answer (having the list of respective answers). The answer to every question is in the form of a list as sometimes there may be multiple answers.
Quality Control: A data processing protocol covering data cleaning, consistency norms, and question design was shared with experts. Edge cases were discussed collaboratively, and the final dataset was schema-validated programatically. More details will be included in the supp. mat.

B.6. List of questions category-wise

We provide the categorization of the questions in the dataset. The questions are grouped based on their type or theme for an easier analysis during the evaluation. The Tab. 4 is showing all these questions and their categories for a better understanding of the diversity of information and the variety of asking a question included in our dataset.

B.7. Category analysis

We present in Tab. 7 the top values and their frequencies across various categories, providing insights into the most

Museum Name	#attributes	#objects	#images	Trainable attributes	Non-trainable attributes
Europeana	7	19163199	23395805	organization, subject, type, country, title, creator	description
Carnegie	6	76655	76655	creator, classification, credit, medium	nationality, date
Contemporary	3	9582	9582	artist, title	date
Harvard	9	579148	265555	technique, classification, worktypes, century, medium	division, creditline, department, period
Peabody	9	77379	77379	title, material, place of origin, artist, category, department, subjects, keywords associated, short description	NA
ArtUK	22	292358	579148	tags, artist, title, medium, worktypes	Acquisition method, Work status, Access note, Date Listing date, Installation end date, Signature/marks description, Venue, Access, Listing, Measurements, status, Unveiling date, Accession number, Installation start date, Custodian, Inscription description, Owner
Hermitage	22	12572	14135	technique, school, place, title, author, material, epoch, category	Place of creation, Date, Inventory Number, Subcollection, Acquisition date, Dimension, Place of finding, Collection, Complex., firm, Manufacture, workshop, "Book, album, seria", Information about the original, Archaeological site, Comment
SouthWales	6	27433	46380	title	exhibition history, audio, provenance, video, places
Indian	34	189838	313962	language, coin description observe, main material, main artist, inscription	Accession Number, Artist Nationality, Mint Title, Weight, Manufacturing Technique, Script, Historical Note, Detailed Description, Medium, Provenance, Museum Name, Patron Dynasty, Coin Description Reverse, Dimensions, Find Place, Origin Place, Tribe, School, Gallery Name, Title2, Number of Illustrations, Brief Description, Subject, Scribe, Culture, Artist Life Date, Number of folios, Country
DPLA	6	22984790	22984790	language, publisher, collection title, title, place of origin, subject	NA
Colbase	14	22196	22196	category, genre, material, artist, holder	Period/Century, Country/Origin, Donor, Quantity, Inscriptions, Excavation site, Cultural property designation, Size, Collection reference no.,
Tepapa	6	187595	251361	collection, title, type, additionalType	Caption, CreditLine
Penn	12	191831	556092	culture, culture area, continent, materials, technique, credit line, place	Description, length, width, height, depth
Smithsonian	4	3277593	3277593	name, sex, place of origin, taxonomy	NA
Ariadne	4	665289	665289	title, nativesubject, place	description

Table 3. The list of museums and aggregators. We display the number of attributes each museum have, the number of objects that they provided and the number of images available for them. We also present the attributes that helped the creation of the questions used during training and testing (Trainable attributes) as well as the attributes not used for questions but that we make available in the raw dataset (Non-trainable attributes).

prominent features and trends within the dataset.

B.8. Multilingual Dataset

Our multilingual dataset comprises 15 million datapoints, featuring objects described in 37 European and Asian languages. Below is the distribution of exhibits across these languages:

B.9. Bias in Dataset

Large-scale models have a profound impact on society, both positive and negative, particularly in applications involving multi-modal tasks. Their performance heavily depends on the datasets they are trained on, and research shows that biases affect certain user groups unfairly or reinforcing discriminatory patterns. Many large-scale

Category	Question
Subject	what are the subjects that the object depicts? what are the subjects that are depicted by the object? which category does this object belong to? what is the subject of this image? what tags can the object be associated with? under what category does this object fall? what are the keywords associated with objects? what is the category of the object? what category does this object fall into? what are the subjects of object ?
Creator	who is the publisher of this object? who is the holder of the object? who is the creator of the object? who has created this object? who is the author of the text? who is author of the object? to whom is this object credited to? who is the artist of the object? who created this art?
Title	what is the title of the object? what is the name of the object? what is the title of this object? what is the name of the costume? what is a suitable title for the object? what is the denomination of the coin? what can be the title of the object? what is the title of the object
Material	which primary material is the object made of? what material is the object made of? what materials is the object made of? which secondary material is the object made of? what is the medium used to create this object? which tertiary material is the object made of? what are the materials that this object is made up of? what is the medium of the object?
Type	which type of object is this? which type of object is it? what is the genre of this object? what type of work is that of the object? what is the additionaltype of the object? what is the type of the object?
Place of Origin	what is the place of origin of the object? what is the place of origin of this object? which country does this object belong to? which continent does this object belong to? what place could this object be from?
Collection	from which collection has this object been taken? what is the collection of the object? what department does this object fall into? what school does object belongs to?
Technique	what technique is used to make the object? what is the technique that has been used to make this object?
Culture	which area does the culture depicted by this object belong to? which culture does this object belong to?
Language	which language is the text in the object? what is the language of the text?
Others	what is the object about? which period does this object belong to? which style do the costumes belong to? what is inscribed on the art piece? what is the obverse of the coin? which organization does this object belong to?

Table 4. The questions generated from the attributes available for the exhibits grouped by categories.

Language	%	Language	%	Other Languages (%)	
German	16.62	Italian	2.57	Lithuanian	
Norwegian	11.89	Polish	2.56	Romanian	
Dutch	11.79	Estonian	2.41	Croatian	
Spanish	8.80	Czech	1.54	Portuguese	21.99
French	7.96	Finnish	1.16	Bulgarian	
Swedish	5.65	Catalan	1.12	Greek	
Danish	3.10	Hungarian	0.84	and more	

Table 5. **Multi-lingual dataset distribution**

models and their training datasets remain inaccessible, with most only available through a restricted input-output interface. While open-source initiatives attempt to replicate model architectures, the lack of publicly available datasets makes it challenging to thoroughly investigate and address potential biases. While bias-free datasets are unattainable [2], we ensure our dataset is bias-aware and take active steps toward inclusivity.

Selection Bias: Our primary data sources, international aggregators, naturally emphasize European and American objects, leading to a **selection bias**, further amplified by the lack of digitization in smaller museums. However, our dataset includes 5M+ objects from other continents, and we are collaborating with local museums to diversify underrepresented cultures.

Temporal bias: Given the aggregators' extensive curation, our collection spans a vast historical timeline, from ancient artifacts to modern art, covering statues, paintings, vessels, fossils, corals, war depictions, weapons, manuscripts, textiles, coins, ceramics, scientific instruments, and more.

Language bias: To mitigate language bias, we include 15M samples across 37 languages beyond English as part of our multilingual dataset with ongoing expansions.

Framing Bias: We also acknowledge framing bias, as models are trained on front-view images as per standard digitization practices, yet multi-angle experiments confirm model robustness to different image perspectives.

Bias due to model architecture: The CLIP model itself introduces biases that are challenging to fully assess, as its training data is not publicly available. Since the vision encoders of both BLIP and LLaVA models rely on CLIP embeddings, they inherit these biases as well. By providing an open large-scale image-text dataset, we enable greater transparency and facilitate the auditing of contrastive image-text models like CLIP.

Handling bias in dataset: To facilitate a thorough investigation of dataset biases, we will release MUSEUM-65 along with tools designed for large-scale data exploration using precomputed image embeddings [1]. These tools will allow researchers to retrieve images based on text or image queries, enabling a systematic study of how biases manifest within the dataset. By examining patterns in object representation, cultural distribution, and framing biases, researchers can gain insights into potential disparities and their implications. In addition to aiding bias detection, these tools will support the development of automated methods for dataset curation, helping mitigate safety concerns associated with large-scale data.

To ensure accessibility, we will publicly release the source code and essential routines, allowing users to build their own versions of these tools for customized dataset exploration. Furthermore, we plan to introduce a public portal where researchers can report undetected biases and model behaviors based on their findings, contributing to ongoing improvements in dataset transparency, fairness, and inclusivity. Through these efforts, we aim to establish MUSEUM-65 as a robust real-world large-scale dataset that not only supports cultural heritage research but also advances responsible data curation practices.

For applications requiring a minimally biased dataset, debiasing techniques such as model-agnostic training or specialized model architectures can also be applied as needed [3, 7, 8] .

B.10. Safety and Ethical considerations

Our dataset provides a foundation for training multimodal models that can enhance cultural accessibility, support educational tools, and enable virtual heritage exploration while promoting multilingual data and fostering cross-cultural appreciation by making global artifacts easily comparable. Upon randomly inspecting images and text from our dataset, we found that museums, as reputable institutions, carefully curate their collections to address potential controversies such as historical disputes, religious issues, privacy concerns, and racial biases. This curation also extends to inappropriate content tagging, ensuring the safety and quality of the dataset. This curation also extends to inappropriate content tagging, ensuring dataset safety and quality. While bias in the place of origin remains an ethical consideration, we aim to mitigate it through collaborations and diversification, with the hope that broader digitization efforts by museums will further enhance dataset diversity

In the current form, we consider this dataset a research artefact and strongly advocate **academic use only** and advise careful investigation of downstream model biases.

This dataset serves as a foundation rather than a final solution for building more balanced and carefully curated datasets for model training. We believe that this process should be open and transparent, involving the broader research community to ensure responsible data development. By introducing MUSEUM-65, a large-scale dataset with diverse image-text pairs and annotations, we provide a resource that can aid in identifying biases, refining data selection, and creating safer, more representative subsets for various applications. We encourage researchers to contribute to this ongoing effort, fostering collaboration toward more ethical and inclusive dataset curation.

C. Experimental details

This section includes detailed information on the parameters used for fine-tuning, LLaVA and BLIP. It also covers a proposed ablation for LLaVA, experimental configurations, and tuning strategies applied during fine-tuning, providing insights into the optimization process and training evolution of these models. The code can be found: MUSEUM-65

C.1. Implementation details

BLIP The BLIP model we used is *blip_vqa*. During fine-tuning we follow the same protocol as [4], having a learning rate 2e-5, a cosine annealing learning rate and the AdamW optimizer [6], with weight decay 0.05. We used a batch size of 4x8x16= 512.

LLaVA During fine-tuning we follow the same protocol as [5], having learning rate 1e-3 and a cosine annealing learning rate schedule with a warmup ratio 0.03 and the AdamW optimizer [6], with weight decay 0.1. We used LORA for fine-tuning as [5]. We used a batch size of 4x8x16= 512.

C.2. Evaluation metrics

We compute several scores to evaluate our fine-tuned methods for a more diverse assessment. We use both uni-gram and n-gram methods, and choose metrics that are intuitive and well known.

Setup. To ensure accurate and consistent metric calculations, we pre-process the answers by removing special characters, retaining only alphanumeric content before computing the metrics. The overall metric is an average of individual metric scores for each question.

Precision. Given the model's prediction and a list of valid answers for a question, the precision is the fraction of words from the model's prediction that appear in at least one of the valid answers. We consider Complete Precision as the percentage of questions for which the precision is 1.0 (the answer completely matches the ground truth) and Partial Precision as the percentage of questions with precision > 0.0 (the answer partially matches the ground truth).

Recall. For each valid answer, the recall is the fraction of words from the answer that are included in model's prediction. For each question, the recall is averaged among all valid answers. Again, we consider Complete Recall as the percentage of questions for which the recall is 1.0 and Partial Recall as the percentage of questions with recall > 0.0.

BLEU scores. We compute the BLEU score to address matching word pairs accurately. The BLEU score is the fraction of word n-grams from the model's prediction that appear in at least one of the valid answers, modified by a brevity penalty that penalizes short responses that only match a few words. We translate the score to give values between 0 and 100. We compute individual scores for BLEU 1-gram and BLEU 2-gram (referred as BLEU1 and BLEU2) and we average the scores among all the instances.

C.3. Finetuning

Why BLIP and LLaVA? BLIP excels at aligning images with descriptive text, generating accurate captions which contribute to its question answering capabilities, making it a good first choice for VQA. However, BLIP relies on a relatively small pre-trained text encoder/decoder (BERT-base with 110 million parameters), which may limit its depth of understanding, especially for more complex or nuanced instructions and queries. Therefore we also chose the LLaVA model, which uses Llama2 7B, an instruction-tuned LLM which is a much more powerful pre-trained language model that understands instructions better than BLIP.

epoch	1	2	3	4	5
LLaVA mQ	57.3	59.51	60.75	60.77	60.77
LLaVA 1Q	54.7	55.76	56.73	57.61	58.08

Table 6. Comparison of two LLaVA fine-tuning methods: LLaVA-1Q, which uses one random question per image per epoch, and LLaVA-mQ, which utilizes all available questions per image each epoch. LLaVA-mQ achieves better results and faster convergence.

LLaVA ablation During fine-tuning we wanted to observe the impact of using all the questions available for an image and we observed an improvement during evaluation for that model. As it was very time consuming (each epoch being 3 times longer), and as LLaVA already being time expensive, we continued the rest of the experiments with the version that chooses one random question for each image in every epoch. (LLaVA 1Q). See Tab. 6.

C.4. Training Evolution

We present the performance of BLIP across different epochs, highlighting its progression during training. It compares the outcomes of various BLIP and LLaVA fine-tuning approaches, see Tab. 8. We also show a comparison between BLIP1mn and BLIP20mn when having the same amounts of steps, meaning BLIP1mn is trained for 20 epochs while BLIP20mn is trained for 1 epoch (BLIP1mn-20ep and BLIP20mn-1ep), see Fig. 1. We observe that BLIP20mn-1ep is having better results than BLIP1mn-20ep highlighting that the amount of data matters.

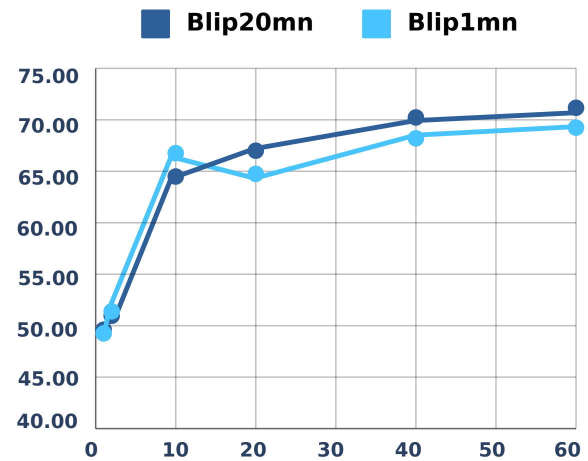

Figure 1. Comparison between BLIP20mn-1ep ad BLIP1mn-20ep across multiple epochs during fine-tuning, maintaining the same number of steps. We observe BLIP20mn-1ep having better results than BLIP1mn-20ep.

D. Additional Results

This section includes supplementary findings, expanding the primary results presented in the main study, more detailed evaluations of the experiments and graphics comparing the performance of multiple models, for a deeper understanding of their strengths and weaknesses. It also provides insights into the questions created specifically for these analyses.

D.1. Benchmark Comparison

In order to show the superior utility of our dataset compared to other existing in literature, we also fine-tuned BLIP on MUZE (MUZE-BLIP). Our model (BLIP-20m-5e) outperforms MUZE-BLIP significantly in all settings (see Tab. 9). Thanks to the large scale nature of our dataset, model trained on it achieves the best results on MUZE in a zero-shot manner.

D.2. General VQA

Following the General Visual Question Answering (VQA) settings, we present a comprehensive table comparing all BLIP and LLaVA models fine-tuned on our dataset evaluated across multiple metrics, see Tab. 10. We observe that in general, the fine-tuned models have much better results than the original models. The results show that LLaVA achieves the best performance among the models.

D.3. Category-wise VQA

For category-wise Visual Question Answering (VQA), we present the results of multiple BLIP and LLaVA models compared with each other across categories such as *subject, title, creator, material* and more (see Fig. 2). The results demonstrate improved performance of the fine-tuned models in each category. Moreover, the LLaVA fine-tuned models are having better results than BLIP ones on *subject, title, creator, collection, language* and *type*.

D.4. MultiAngle VQA

Following the MultiAngle VQA setting, we presents the table comparing multiple models on both original images and images from different viewpoints with extended metrics, helping to evaluate model performance across varying perspectives, offering deeper insights into their robustness. See Tab. 11

Subject	#instances	Types	#instances	Material	#instances	Place	#instances	Creator	#instances
united states	2096485	ship	96423	oil	189599	united kingdom	8991283	Robert John Welch	4128
university	845965	model	86022	canvas	141786	texas	1578768	British school	3428
american	716196	vessel	72338	paper	86420	california	424987	William Alfred green	3352
maps	641289	medal	47672	wood	46831	massachusetts	350232	Joseph Hardman	2323
school	561988	water transport	46829	stone	46084	new york	254190	John Everett	1798
church	419043	uniform	37773	bronze	31201	washington	253248	Henry Moore	1086
river	337474	artifact	21469	glass	26390	los angeles	248918	Godfrey Kneller	876
city	293439	accessory	18304	fiber	14692	carolina	177376	Alfred James Munnings	731
family	252846	documentary	15968	acrylic	9528	michigan	65104	Joshua Reynolds	676
company	213910	component	4463	steel	5572	milwaukee	54816	Peter Lely	629

Table 7. Detailed list of the most common values across different categories, *subject, types, material, place, creator* (left), along with the number of instances that correspond to them (right).

model \ epoch	1	2	3	4	5
LLaVA1mn-5ep	54.7	55.76	56.73	57.61	58.08
BLIP1mn-5ep	49.24	51.2	56.34	55.54	56.67
BLIP10mn-5ep	64.05	66.67	68.49	69.02	69.23
BLIP20mn-5ep	67.03	69	70.23	71.17	71.51

Table 8. Comparison of multiple models over 5 epochs, highlighting their performance progression. The results show that LLaVA achieves significantly better outcomes much earlier in training compared to other models.

Model	Test dataset	Partial Prec.	Complete Prec.	Partial Recall	Complete Recall	BLEU1
MUZE-BLIP	MUZE	67.84	51.19	67.84	51.19	27.97
Ours-BLIP	MUZE	**79.77**	**66.69**	**79.77**	**66.69**	**40.58**
MUZE-BLIP	Ours	36.38	25.93	36.38	25.93	18.7
Ours-BLIP	Ours	**71.51**	**60.58**	**71.51**	**33.95**	**48.9**

Table 9. BLIP on **MUZE vs. MUSEUM-65** datasets.

D.5. Visually Unanswerable Questions VQA

We created 510 Q&A pairs for this task, featuring 5 painters and 10 continents. The dataset includes 5 images per painter and 5 images per country, ensuring a diverse and balanced representation of artists and geographic regions. Each image is paired with 5-8 questions depending on the available information for their subject (painter, country). In Tab. 12 we show the countries and artists used during the experiment and in Tab. 14 we present the questions associated with them. As many exhibits where coming from Europe, we included Europe among the countries and designed special questions for it.

D.6. MultiLanguage VQA

Following the MultiLanguage VQA setting, we present an extended evaluation of model performance on French and German languages. This analysis provides insights into how well the models handle VQA tasks across different linguistic contexts. See Tab. 13.

E. Limitations and society impact

One limitation of the dataset is that it contains an **unequal representation of objects** from different cultures or regions, which may introduce bias in training models. This imbalance could lead to under-representation of certain cultural artifacts, affecting the model's ability to generalize well across diverse cultural contexts. Additionally, the **variability in the quality and depth of information** provided by different museums further complicates the dataset. Some museums may offer detailed descriptions for their objects, while others provide minimal or inconsistent metadata, which could impact the performance of image-text pairing models when dealing with incomplete or sparse information.

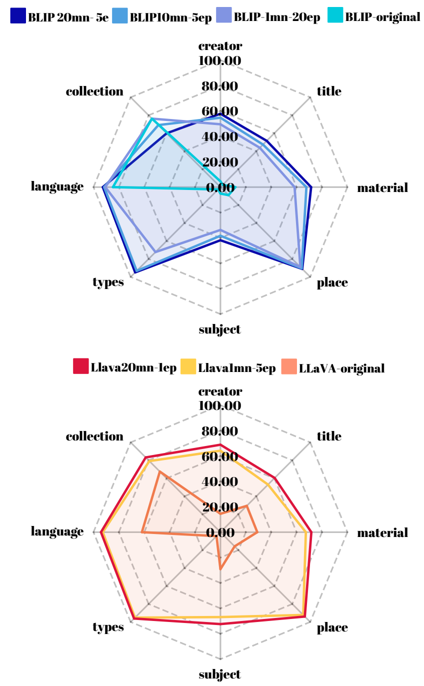

Figure 2. **VQA category-wise results.** On left be compared all BLIP models and in right all LLaVA models. The **fine-tuned models do better on all categories**. The original ones only perform well for *language* and *collection*, as they have easier, common knowledge answers (for *collection*, the results are also influenced by the reduced number of instances that have questions about this). **LLaVA20mn obtained the best results** among all models, showing significant improvement for *subjects, collection, creator* and *title*, surpassing fine-tuned BLIP.

	partial prec.	complete prec.	partial recall	complete recall	BLEU1	BLEU2	BLEU3	BLEU4
BLIP	9.1	4.65	9.1	6.27	5.76	0.13	0	0
BLIP1mn-5ep	56.67	43.5	56.67	21.82	34.57	14.01	3.56	2.45
BLIP1mn-20ep	64.75	53.65	64.74	29.6	43.01	22.37	5.27	3.67
BLIP1mn-60ep	69.24	56.97	69.24	31.48	46.08	24.51	6.32	4.37
BLIP10mn-5ep	69.23	58.18	69.23	32.8	47.16	25.85	6.34	4.38
BLIP20mn-1ep	67	55.89	67	31.18	45.02	23.91	5.5	3.84
BLIP20mn-5ep	71.51	60.58	71.51	33.95	48.9	27.22	7.27	5.13
LLaVA	23	3.97	23	4.07	5.03	0.28	0.1	0.04
LLaVA1mn-1ep	73.12	56.28	73.18	55.1	50.12	30.74	10.36	6.98
LLaVA1mn-5ep	**76.27**	**60.04**	**76.31**	**59.14**	**53.45**	**33.5**	**12.56**	**8.64**
LLaVA20mn-1ep	**81.25**	**63.96**	**81.26**	**63.21**	**57.06**	**36.38**	**14.84**	**10.38**

Table 10. **General VQA results**. Comparison of all the fine-tuned models and their no fine-tune version on precision and recall. We observe the models fine-tuned with 20mn dataset are obtaining the best results, while **LLaVA20mn-1ep is the best**, having 80% of the object with partial precision and 64% with complete precision. Also the **LLaVA models seem to have much better results for recall than the BLIP ones**, being similar with the precision results, showing that the prediction of LLaVA models are more often containing or contained in the ground truth.

	partial prec.	complete prec.	partial recall	complete recall	BLEU1	BLEU2	BLEU3	BLEU4
LLaVA20mn-1ep	58.09	46.09	58.12	41.04	42.14	7.19	2.08	0.52
changed angle	56.14	44.89	56.15	40.01	41.02	6.97	1.97	0.49
LLaVA no finetune	24.35	0.09	24.35	11.25	1.61	0.01	0	0
changed angle	23.56	0.02	23.56	10.85	1.54	0.02	0	0
BLIP20mn-5ep	52.78	42.51	52.78	35.29	38.31	8.01	1.48	0.24
changed angle	51.75	41.87	51.75	34.59	37.62	7.84	1.48	0.26
BLIP no finetune	13.82	9.7	13.82	5.22	6.52	0.02	0.01	0
changed angle	12.86	8.71	12.86	4.72	5.92	0.01	0	0

Table 11. **MultiAngle results**. Comparing fine-tuned LLaVA20mn-1ep and BLIP20mn-5ep along with the no fine-tune models. We observe the alternative angle images results remain close to the original images results across all metrics for all the models which shows **stability in regard to changing the angle**, even if the difference between the images is visible.

Countries	Artists
Germany	Abdourahmane Sakaly
France	George Victor Du Noyer
USA	Leo Swan
Netherlands	Shakespeare William
Italy	Robert John Welch
Ireland	
Denmark	
Belgium	
United Kingdom	
Europe	

Table 12. The lists of the countries and the artists used for the Visually Unaswerable Questions VQA experiment.

	partial prec.	complete prec.	partial recall	complete recall
LLaVA20mn-1ep	10.04	0.8	10.02	0.17
LLaVA nofinetune	30.11	0.27	30.59	0.56
BLIP20mn-5ep	2.37	0.27	2.41	0
BLIP nofinetune	1.40	0.44	1.43	0

Table 13. **MultiLanguage results**. (French and German). We observe that LLaVA models have better results than BLIP ones, still LLaVA20mn-1ep is **slightly forgetting the ability to answer in other languages**, due to its fine-tuning in English. However, on complete precision and BLEU2 the results of LLaVA20mn-1ep are sligthly better than for the no fine-tune version.

F. Examples

In Fig. 3 we show examples of prediction (P) for the best model finetuned with our dataset, LLaVA20mn-5ep, for the proposed tasks.

Figure 3. Examples of LLaVA20mn-5ep results for the proposed tasks. The question is denoted with (Q), the answer wit (A) and the prediction with (P).

Painters	Countries	Europe
What is the period the artist lived in?	Which continent is the country of origin of this object located in?	Which oceans border the continent of origin of this object?
What is the nationality of the artist?	Who are the neighbors of the country of origin of this object?	What are the major languages spoken in the continent of origin of this object?
What is the name of the spouse of the artist?	When did the country of origin of this object get independence or get established?	What is the largest country by area in the continent of origin of this object?
Who was the mentor of the artist?	Which part did the country of origin of this object support during World War 2?	What is the smallest country in the continent of origin of this object?
Who was influenced by the artist?	What is the main religion of the country of origin of this object?	What are some major rivers in the continent of origin of this object?
What is the capital of the country the artist was born in?	What is the form of government in the country of origin of this object?	What is the dominant climate of the continent of origin of this object?
What was the political regime when the artist lived?	Who is the president of the country of origin of this country?	What are the main religions in the continent of origin of this object?
Who was the king/president in the period the artist lived?	What is the capital of the country of origin of this object?	What are some of the major economic sectors of the continent of origin of this object?

Table 14. The questions used for the Visually Unanswerable Questions VQA task. These questions are derived from the dataset information starting from the painters or the country of origin for some images. We also added questions related to the continent due to the big number of objects located in Europe, that usually do not have precise location of origin.

References

[1] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 4

[2] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. 4

[3] Jingliang Gu and Zhixin Li. Beyond language bias: Overcoming multimodal shortcut and distribution biases for robust visual question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3767–3771, 2024. 5

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[7] Ninglin Ouyang, Qingbao Huang, Pijian Li, Yi Cai, Bin Liu, Ho-fung Leung, and Qing Li. Suppressing biased samples for robust vqa. *IEEE Transactions on Multimedia*, 24:3405–3415, 2021. 5

[8] Desen Yuan. Language bias in visual question answering: A survey and taxonomy. *arXiv preprint arXiv:2111.08531*, 2021. 5